Second Edition

ADVANCED SURVEYING

TOTAL STATION, GPS, GIS AND REMOTE SENSING

Satheesh Gopi
R. Sathikumar
N. Madhu

Pearson

About Pearson

Pearson is the world's learning company, with presence across 70 countries worldwide. Our unique insights and world-class expertise comes from a long history of working closely with renowned teachers, authors and thought leaders, as a result of which, we have emerged as the preferred choice for millions of teachers and learners across the world.

We believe learning opens up opportunities, creates fulfilling careers and hence better lives. We hence collaborate with the best of minds to deliver you class-leading products, spread across the Higher Education and K12 spectrum.

Superior learning experience and improved outcomes are at the heart of everything we do. This product is the result of one such effort.

Your feedback plays a critical role in the evolution of our products and you can contact us - reachus@pearson.com. *We look forward to it.*

This page is intentionally left blank

ADVANCED SURVEYING

Total Station, GPS, GIS and Remote Sensing

Second Edition

Satheesh Gopi
Deputy Chief Hydrographer
and Chartered Engineer
Hydrographic Survey Wing
Kerala Port Department
Thiruvananthapuram

R. Sathikumar
Principal (Retired)
Government Engineering College
Burton Hill
Thiruvananthapuram

N. Madhu
Professor (Retired)
Civil Engineering Department
College of Engineering
Thiruvananthapuram

Pearson

Published by Pearson India Education Services Pvt. Ltd, CIN: U72200TN2005PTC057128, formerly known as TutorVista Global Pvt. Ltd, licensee of Pearson Education in South Asia.

ISBN 978-93-528-6072-2
eISBN 978-93-528-XXXX-X

Head Office: A-8 (A), 7th Floor, Knowledge Boulevard, Sector 62, Noida 201 309, Uttar Pradesh, India.
Registered Office: Module G4, Ground Floor, Elnet Software City, TS-140, Block 2 & 9, Rajiv Gandhi Salai, Taramani, Chennai 600 113, Tamil Nadu, India.
Fax: 080-30461003, Phone: 080-30461060
www.pearson.co.in, Email: companysecretary.india@pearson.com

BRIEF CONTENTS

CONTENTS

This page is intentionally left blank

PREFACE

Surveying is fundamental to all civil engineering activities, which include construction of buildings, roads, railways, bridges, dams, canals, airports and harbours. Survey data is the integral part of design and execution of all engineering projects. Time is very important in the present scenario and advanced surveying techniques employs precise electronic surveying instruments that are capable of performing most accurate data collection in short time.

There have been far rapider advancements in the area of surveying compared with other divisions of civil engineering. During the last two decades, developments in electronics and optics have led to significant advancements in surveying instruments and techniques. The conventional equipment such as compass, vernier theodolites, dumpy levels have given way to versatile modern digital instruments such as digital levels, electronic theodolites, electronic distance measuring devices, total stations, Global Positioning Systems (GPS) and LIDAR. The traditional surveying techniques was completely replaced by advanced techniques. Manual methods were completely replaced by advanced computerized data collection, digital processing and computer-aided drafting and mapping systems. Modern techniques had evolved with the introduction of satellite-based remote sensing systems and Global Navigational Satellite Systems (GNSS). The field of surveying had undergone rapid changes due to the introduction of most modern electronic instruments and widespread adoption: total station, GPS, digital levels and geographical information systems (GISs). With the introduction of the latest instruments mentioned above and advanced techniques made the preparations of survey maps is very simple and precise.

The main aims of compiling this book is to provide the knowledge of modern techniques and modern instruments such as total station, electronic distance measuring (EDM), GPS and LIDAR and the knowledge of GIS in a comprehensive manner for civil engineering students as well as practicing engineers.

In this new edition, a major organizational change has been done. The revised edition has been specially written keeping in mind the requirements of undergraduate students of engineering. Most of the subject prescribed in undergraduate curricula of civil engineering has been covered in an easy to comprehend manner. This book is also useful to students pursuing GIS. The book deals with modern instruments and techniques used in the field of geoinformatics.

Content of the book is as follows: The book has been organized into four sections. The first part discusses the Fundamentals of GIS, Applications of GIS, Map Projection and Coordinate Systems. The second part deals with the Principles of Total Station, EDM, Surveying Using Total Station and detailed descriptions of most modern surveying equipment such as digital level, micro-optic theodolites, LIDAR and terrestrial scanner. The third part deals with Photogrammetry and Aerial Surveying in detail. The fourth part covers Basics of GPS and Surveying Using GPS. Appendix A deals with the basic Geodetic Aspects required for GPS-based surveying. Appendices B–D provide step-by-step field procedure of various models of total stations.

The detailed structure of the book is as follows:

Chapter 1 presents the fundamental concepts of GIS such as GIS architecture, components of GIS and GIS work flow and categories of GIS.

Chapter 2 discusses about the GIS data models such as spatial, vector, raster and attribute data models. The chapter also discusses digital elevation model (DEM), image data and application of DEM.

Chapter 3 deals with data acquisition in GIS using analogue maps, aerial photographs, satellite imagery, ground surveys, GPS, etc. The chapter also focuses on digitizers, scanners as a source of data input, rasterization and vectorization using various equipment and methods.

Chapter 4 provides details on maps and map projection. It narrates the characteristics of maps, different map projections and construction of map projection.

Chapter 5 presents various coordinate systems in detail. It also covers geoid and reference ellipsoids, and different transformations.

Chapter 6 is devoted to spatial analysis. It presents classification of analytic of GIS, overlay function and network analysis.

Chapter 7 covers various applications of GIS including case studies.

Chapter 8 deals with the basics of total station.

Chapter 9 focuses on measuring principle of EDM, characteristics of EDM, field procedure of EDM and various types of EDMs.

Chapter 10 describes execution of field surveys using total station, which includes field procedures, overview of computerized survey data systems, total station job planning and estimating and trigonometric levelling.

Chapter 11 elaborates data collection procedures to be followed while surveying using total station.

Chapter 12 covers most modern surveying instruments such as automatic levels, digital levels, micro-optic theodolites, LIDAR, digital planimeters and laser levels.

Chapter 13 deals with aerial surveying and details about aerial photogrammetry, terrestrial photogrammetry, different photogrammetric techniques and LIDAR.

Chapter 14 presents the fundamentals of remote sensing.

Chapter 15 deals with the basics of GPS.

Chapter 16 introduces the modern methods of surveying using GPS.

Lecture PPTs are available for download from the Instructor Resource centre at www.pearsoned.co.in/satheeshgopi

Satheesh Gopi
R. Sathikumar
N. Madhu

ACKNOWLEDGEMENTS

We are extremely grateful to the editorial team of M/s. Pearson Education for the excellent effort they made to bring out the revised edition of the book. We wish to thank M/s. Toshnitek International, Chennai for providing catalogues of various modern survey equipment and step-by-step field procedures of Sokkia total station. Despite of our best efforts, it is possible that some errors may be unnoticed and we shall be grateful if these are brought to our notice.

We would like to thank all those reviewers who took out time to go through the book and gave us their valuable feedback. Their names are as follows:

K. Nirmalkumar
Professor
Kongu Engineering College, Tamil Nadu

Tapas Karmaker
Assistant Professor
Thapar University, Punjab

M. Vijai Kishore
Assistant Professor
University of Petroleum and Energy Studies, Uttarakhand

H. B. Nagaraj
Professor
B.M.S. College of Engineering, Karnataka

Bhavna Tripathi
Associate Professor and Head of Department
Vivekananda Institute of Technology, Rajasthan

Narendra Kumar
Professor
Poornima University, Rajasthan

Abhijit M. Zende
Assistant Professor
Dr. Daulatrao Aher College of Engineering, Maharashtra

Kuldeep Singh
Assistant Professor
Radharaman Engineering College, Madhya Pradesh

Reshma Raskar Phule
Assistant Professor
Sardar Patel College of Engineering, Maharashtra

N. Gladwin Gnana Asir
Associate Professor
Dr. Sivanthi Aditanar College of Engineering, Tamil Nadu

Goutam Bairagi
Professor
Jalpaiguri Government Engineering College, West Bengal

Mukul C Bora
Professor
Dibrugarh University Institute of Engineering and Technology, Assam

Deepak Kumar Mandal
Professor
University of North Bengal, West Bengal

Satheesh Gopi
R. Sathikumar
N. Madhu

ABOUT THE AUTHORS

Satheesh Gopi, the principal contributor to this book, has over 27 years' experience as a hydrographer and is currently a Deputy Chief Hydrographer in the Hydrographic Survey Wing of Kerala Port Department. He graduated in civil engineering from the College of Engineering, Thiruvananthapuram and holds master's degree in Construction Engineering and Management from Anna University, Chennai. He also holds master's degrees in Information Technology, and Disaster Management. He has worked with the Survey Department of Government of Dubai for 5 years as a Senior Hydrographer.

He is the author of the book *Global Positioning System – Principle and Applications* and *Basic Civil Engineering*. He has published research papers on global positioning system, advanced surveying, digital cartography and hydrography in various journals of national repute. Most of his research papers were presented in international conferences, such as International Federation of Surveyors (FIG), International Hydrographic Organization (IHO) and Map Middle East. He has worked as guest faculty in reputed engineering colleges and universities. He is a member of Institution of Engineers and life member of Indian Cartographic Association.

R. Sathikumar is currently the Principal, at the Rajadhani Institute of Engineering and Technology, Thiruvananthapuram. He retired from the Government service as Principal, Government College of Engineering, Burton Hill, Thiruvananthapuram. He received his post-graduate degree, in Transportation Engineering from IIT Kanpur in 1989 and also pursued his Ph.D. from IIT Roorkee in 1996.

N. Madhu is a retired Professor (Civil) from the College of Engineering, Thiruvananthapuram. He obtained his M.Tech. in Traffic and Transportation Engineering from IIT Madras in 1991.

This page is intentionally left blank

FUNDAMENTAL CONCEPTS OF GEOGRAPHIC INFORMATION SYSTEM 1

Chapter Outline

1.1 INTRODUCTION

GIS stands for *geographic information system*. An information system is a computer program that manages data. A GIS is a type of information system that deals specifically with geographic, or spatial, information. GIS is a rapidly growing technological field that incorporates graphical features with tabular data to assess real-world problems. The GIS began to develop in the 1960s, with the discovery that maps could be programmed using simple code and then stored in a computer, allowing future modifications when necessary. This was a marvellous change from the era of hand cartography when maps had to be created by hand. The earliest version of a GIS was known as computer cartography and involved simple line work to represent land features. From that evolved the concept of overlaying different mapped features on top of each other to determine patterns and causes of spatial phenomena.

The capabilities of GIS have been known from the simple beginnings of computer cartography. At the simplest level, GIS can be thought of as a high-tech equivalent of a map. Paper maps cannot be produced for quicker and more effective storage of data. An easily accessible digital format of maps in GIS enables complex analysis and modelling of data, which was previously not possible with paper maps. The reach of GIS expands into all disciplines and has been used in wide range, such as prioritizing sensitive species habitat, and determining optimal real-estate locations for new businesses.

The keyword to GIS technology is geography. This means that the data, or at least some proportion of the data, are spatial; or in other words, data in some way referred to the locations on earth. Geographic information describes the spatial or location factors of an object or area. This can simply be latitude and longitude, but in most cases more complex factors are included.

A GIS allows users to view, update, query, analyse, combine and manipulate map data. It can take information from different maps and tables and register them to a desired base. It can manage large collections of natural resources and environmental data and the complex data sets needed for urban studies. It can overlay maps to eliminate or include areas based on multiple layers of tabular criteria. It can automatically generate buffers around features such as sensitive land use types.

A GIS operates on many levels. At basic level, it is used as computer cartography for automated mapping. The real power in GIS is through using spatial and statistical methods to analyse attributes and geographic information. The end result of the analysis can be derivative, interpolated or prioritized information.

A GIS can be made up of a variety of software and hardware tools. The important factor is the level of integration of these tools, which provide a smooth operating and fully functional geographic data processing environment. GIS should be viewed as a technology, and not simply as a computer system. GIS provides facilities for data capture, data management, data manipulation and analysis, and the presentation of results in both graphic and report form. It provides a particular emphasis on preserving and utilizing inherent characteristics of spatial data. Today, GIS is a multi-billion-dollar industry employing hundreds of thousands of people worldwide. GIS is taught in schools, colleges and universities throughout the world. Professionals and domain specialists in every discipline are becoming increasingly aware of the advantages of using this technology for addressing their unique spatial problems.

GIS provides the necessary set of tools required for gaining the appropriate insights into the places and processes that drive our environment, society and economy. It provides the means of finding hidden patterns and trends that would have remained undetected using conventional means of looking at data, by combining data from a variety of different sources such as corporate, commercial and government databases, images from cameras and satellites, and data derived from any type of sensor or human observation, and presenting it in the form of a map. GISs are not restricted to the conventional view of geography, that is, that of people and places on the earth's surface. Hidden geographies lie everywhere, and a GIS is the perfect tool to take on voyages of discovery. GIS will help to pave the way for the success of the expedition by providing the means of visualizing and exploring these uncharted territories.

Computer revolution and the advent of cheap desktop computers made an explosion in the use and availability of GIS. GISs are no longer the sole domain of scientists funded by generous research grants, but are available to everyone from a small shop owner to a major corporation at a highly competitive price, allowing virtually anyone to become a player in the global marketplace through the insights fuelled by the geo-revolution. Our perception is built upon the ability to distinguish boundaries and outlines, and effectively store them in our brains as mental images. Thus, maps with their clear and concise definition of area and space provide ample food for thought, a crystal clear concept provided by reams of data in a spreadsheet.

Dr John Snow's well-known cholera map is one of the earliest known examples of GIS. In September 1854, during an outbreak of cholera in a section of the city of London, Dr John Snow, a local physician, decided to test his hypothesis that cholera is a waterborne disease and that the outbreak was a result of contaminated water supplies—a view contrary to the medical beliefs of the time—by drawing a map showing where the victims lived and where the local water pumps were

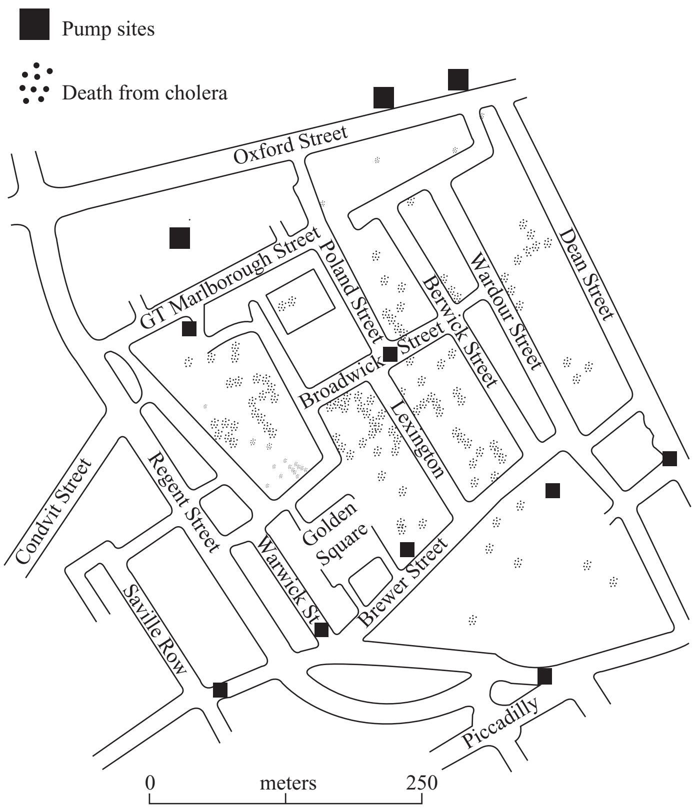


FIGURE 1.1 The first GIS used for identifying an epidemic disease

located (see Fig. 1.1). The map clearly showed a clustering of cholera cases around one of the pumps, and when that pump was shut down, the outbreak stopped. Using the simple means of pen and paper, Dr Snow was able to prove a cause and effect relationship between a contaminated supply of water and instances of disease by using a map to visualize the instances of disease and the location of water pumps. This mapping in epidemiology is cited as the first example to highlight the application of geographic analysis.

GISs, backed by the power of modern computers, allow us to benefit from the visual power of maps and to transcend the limitations of the past. Paper maps, for all their beauty and communicative power, are limited; they are static representations of the world, and the result of a great amount of effort. GIS maps, on the other hand, are dynamic; users can pan, zoom, turn layers on or off, change the scale and their point of perspective a thousand times without any effort at all. The data in a GIS are dynamic and constantly change as new data become available in the system, and through the enormous power of modern computers, a GIS can turn out thousands of maps representing the data from millions of records at a minimum amount of time.

The introduction of topological techniques permitted the data to be connected in a relational sense, in addition to their spatial connections. Thus, it became possible not only to determine where a point (e.g., a hydrant) or a line (e.g., a road) or an area (e.g., a stadium, park) was located, but also to analyse those features with respect to their relation to other special features, connectivity (network analysis) and direction of vectors. The present generation GIS is distinguished from geo-processing of the past by the use of computer automation to integrate geographic data processing tools in a friendly and comprehensive environment.

The advent of sophisticated computer techniques have proliferated the multidisciplinary application of geo-processing methodologies, and provided data integration capabilities that were logistically impossible before. The ability to incorporate spatial data, manage and analyse it, and answer spatial questions are the distinctive characteristics of GIS.

GIS is now becoming an independent discipline in the name of geomatics, geoinformatics or geospatial information science and is used in many departments of government and university.

1.2 VARIOUS DEFINITIONS OF GIS

The GIS is a rapidly growing technology and has been defined in various ways. The following definition of GIS gives a concise description of the term/system: a GIS is a computerized, integrated system used to compile, store, manipulate and output mapped spatial data.

Different definitions of Geographic Information System given by various organizations are as follows:

- Geographic information system, commonly referred to as a GIS, is an integrated set of hardware and software tools used for the manipulation and management of digital spatial (geographic) and related attribute data.
- Geographic information system (GIS) is a computer-based tool for mapping and analysing things that exist and events that happen on earth. GIS technology integrates common database operations such as query and statistical analysis with the unique visualization and geographic analysis benefits offered by maps.
- GIS is an integrated system of computer hardware, software and trained personnel linking topographic, demographic, utility, facility, image and other resource data that are geographically referenced.
- Geographic information system (GIS) is a computer-based information system that enables capturing, modelling, manipulation, retrieval, analysis and presentation of geographically referenced data.

GIS has become an essential part of day-to-day life. People might have not realized about its applications on everyday life. If someone uses an Internet mapping program to find directions, they are actually using a GIS. A new supermarket chain on the corner was probably located using GIS to determine the most effective place to meet customer demand. The system-based definition of GIS may be called ‘functional’, because it is based on the functions that a GIS performs. As each of these functions is performed on the geographic data associated with the GIS, the definition may also be called ‘data-centred’.

From the above definitions, the following observations can be made:

1. GIS database use geo-references as a primary mean to store and access information.
2. GIS integrates technology.
3. GIS can be viewed as a process rather than software/hardware.
4. GIS helps to make better and quick decisions.

1.3 ORDINARY MAPPING TO GISs

Topographic features called as entities have a long history of being portrayed on scaled maps and plans. These maps or plans provided an inventory of scaled or general features that were found in a given geographic area. With the emergence of large databases, which were collected primarily for mapping, attention was given to a new technique for analysing and querying the computer-stored data, which later became the GIS. GIS data can be assembled from existing databases, digitized or scanned from the existing maps and plans, or collected using conventional surveying techniques or by using the global positioning system (GPS).

A GIS is a system which uses geo-referenced data to answer questions. A computer-assisted cartographic system is a set of graphic elements for map display and printing, and is not a GIS. A computer-aided-drafting (CAD) system is a set of graphic elements for engineering and architectural design in which some have GIS modules as add-ons.

GIS in an integrated multidisciplinary science consisting of the following traditional disciplines such as geography, cartography, statistics, remote sensing, computer science, photogrammetry, mathematics, operational research, surveying, civil engineering, geodesy and urban planning. Table 1.1 summarizes how the above disciplines make up GIS with respect to the functions. GIS has many alternative names used over the years with respect to the range of applications and emphasis, as listed below:

1. Land Information System (LIS)
2. AM/FM—Automated Mapping and Facilities Management

TABLE 1.1 Relations of traditional disciplines with GIS

	Functions of GIS								
Discipline	Data Acquisition	Mapping	Pre-Processing	Data Structure	Database	Spatial Analysis	Modeling	Display	Application
Geography		×				×			×
Cartography	×	×						×	×
Remote sensing	×	×				×		×	×
Photogrammetry	×	×						×	×
Surveying	×	×							
Geodesy		×							
Statistics			×		×	×			
Operational research						×	×		
Computer science			×	×	×	×	×		
Mathematics				×		×	×		
Civil engineering	×	×				×	×		
Urban planning	×					×	×		×

3. Environmental Information System (EIS)
4. Resources Information System
5. Planning Information System
6. Spatial Data Handling System

The ability to store data on feature-unique layers permits the simple production of special feature maps. In a GIS, spatial entities have two characteristics, namely, the location and attributes. Coordinates, street address, etc. can give the location and attributes and describe some characteristics of the feature being analysed. GIS has now blossomed into a huge and diverse field of activity. Most activities can be identified as being in one of the two broad fields:

1. Geographic feature-specific activities such as mapping, engineering, environment, resources and agriculture.
2. Cultural or social activities such as market research, census, demographics and socio-economic studies.

The switch from hard copy maps to computerized GIS has provided many benefits. This system can help to do the following:

1. Store and easily update large amount of data.
2. Sort and store spatial features called entities into thematic layers. Data are stored in layers so that complex spatial data can be manipulated and analysed efficiently by layer, rather than trying to deal with the entire database at the same time.
3. Zoom into sections of the displayed data to generate additional graphics, which may be hidden at default scales.
4. Query items of interest to obtain tables of attribute information that may have been tagged to specific points of interest.
5. Analyse both entities and their attribute data using sophisticated computer programs.
6. Prepare maps showing selected thematic layers of interest. One can update maps quickly as new data are assembled.
7. Import stored data (both spatial and non-spatial) electronically from different agencies, and thus save the costs of collecting data.
8. Use the stored data to prepare maps at different scales, for a wide range of purposes.
9. Build and augment a database by combining digital data from all the data-gathering techniques, that is, from surveying, remote sensing, map digitization, scanning and from the Internet.
10. Create new maps by modelling or re-interpreting existing data.

1.4 COMPARISON OF GIS WITH CAD AND OTHER SYSTEMS

Both GIS and CAD are computer based, but CAD is often associated with high-precision engineering and surveying applications, and GIS is more often associated with the lower levels of precision usually required for mapping and planning.

CAD is similar to mapping because it is essentially an inventory of entered data and computed data that can provide answers to the questions: What is it? Where it is located? But CAD has no

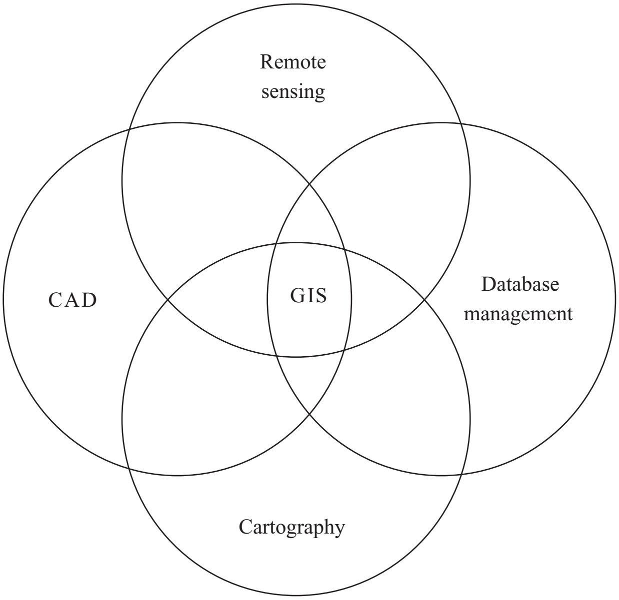


FIGURE 1.2 Relationship of GIS to other main systems

analytical tools to perform spatial analysis. On the other hand, in addition to answering the where and what questions, GIS can model data and provide answers to the spatial and other questions like what occurred? what if ? and what patterns exist? Topology gives GIS the ability to determine spatial relations, such as adjacency and connectivity between physical features or entities.

For establishing a definition of GIS, it is important to outline the relationship of GIS with CAD, computer cartography, database management systems (DBMS) and remote sensing information systems. The relationship of GIS to the other systems is given in Fig. 1.2. A GIS can be termed as a subset of the four listed technologies as in the figure. A true GIS can be distinguished from other systems by the fact that it can be used to conduct special searches and overlays, which actually generate new information. Typical subsets of geographic feature-specific GIS include LIS, AM/FM and geographic information system for transportation (GIS-T).

1.4.1 Land Information System

LIS is a subsidiary of the GIS. LIS is related entirely to land data. It includes information such as size, shape, location, legal description, topography, flood plains, water resources, easements and zoning requirements for each piece of land in the system. The legal descriptions are those bound with the local administration departments. A GIS includes data for a much wider range than that of LIS. The GIS includes not only the data mentioned for LIS, but also includes such things as soil types, depth of ground water, census data, school bus routes, tax maps, fire, police jurisdiction and so on. With an LIS, a user can determine the ownership of property, tax assessments, mortgages, utilities, boundaries, improvements made to the land and other information needed for land appraisal, land acquisition or for other various uses of the land.

1.4.2 Automated Mapping and Facility Management

AM/FM is an important field of GIS activity with macro applications in the management of municipal utilities. Engineers and surveyors can map inventory roadways, pipelines, cables and other municipal infrastructure utilities using AM/FM software programs. These programs also record all relevant characteristics of the utility. For example, in the case of a sewage line the following data have to be recorded: the pipe type, length of run between manholes, inverts and pipe slopes, date of installation and record of maintenance. The program can be designed to issue work orders for a scheduled maintenance automatically and to prepare plans and profiles that show the locations of all services. AM/FM applications manage the physical plant and services of large office and industrial complexes. All services, including electrical, heating, air-conditioning, elevators, fire protection, communications, as well as building design and layout are stored on three-dimensional layer-specific sections of the computer storage.

1.4.3 GIS-T

GIS-T is a subset of GIS. GIS-T refers to the principles and applications of applied GIS to solve transportation problems. GIS can effectively be employed to find, analyse and solve transportation problems. Now GIS-T has been evolved as a specific branch in GIS, to address transportation issues, and now GIS-T has become one of the major applications of GIS. Highway data are usually modelled in vector fashion by digitizing the centre line as lines of arcs and nodes. Lines or arcs join two coordinated points, and nodes are points of intersection with other highway centre lines or other linear features. Highway reference methods are based on the presence or absence of specific attributes to be analysed or displayed, that is, attributes are stored in separate databases categorized by their attribute characteristics.

1.5 GIS ARCHITECTURE (GIS SUBSYSTEMS)

GIS can be understood as a group of subsystems within the framework of a main system. Accordingly, GIS has the following four functional subsystems (see Fig. 1.4):

1. Data input (data input from maps, aerial photos, satellite imageries and from other sources)
2. Data storage and retrieval (data storage retrieval and query)
3. Data manipulation and analysis (data transformation, analysis and modelling)
4. Data output and display (data reporting such as maps, reports and plans)

1.5.1 Data Input

A data input subsystem allows the user to capture, collect and transform spatial and thematic data into digital form. The data inputs are usually derived from a combination of hardcopy maps, aerial photographs, remotely sensed images, reports and survey documents.

There are five types of data entry systems commonly used in a GIS:

1. Keyboard entry (manually entering the data at a computer terminal)
2. Coordinate geometry (entering coordinates of surveyed data using a keyboard)
3. Manual digitizing (widely used method for entering spatial data from maps)

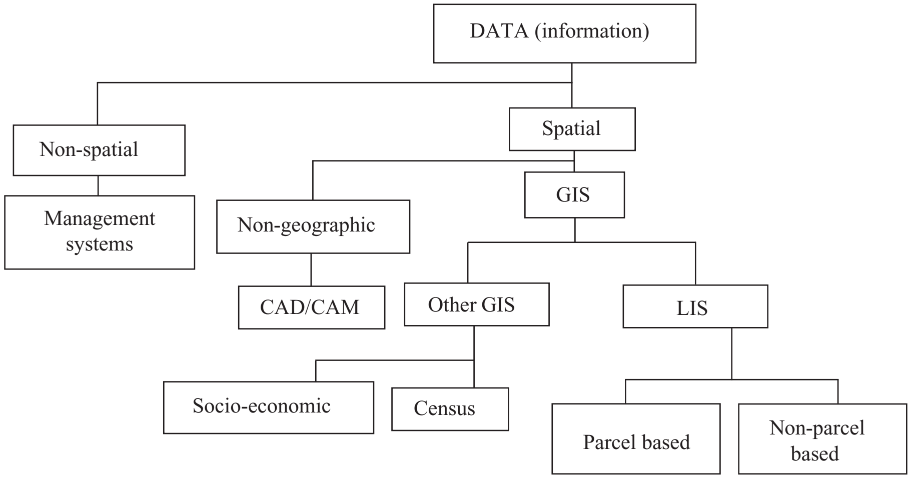


FIGURE 1.3 Flow diagram showing the relation between spatial and non-spatial information with various systems

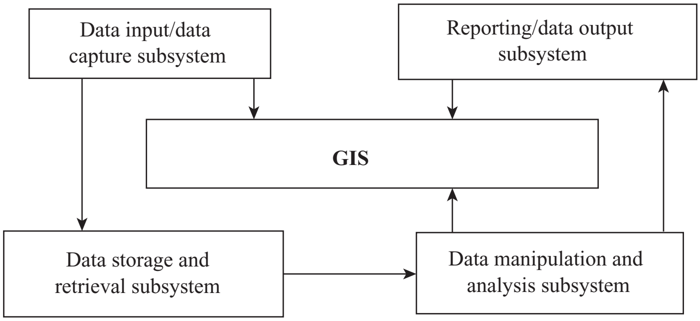


FIGURE 1.4 GIS architecture/GIS subsystems

4. Scanning
5. Input of existing digital files.

1.5.2 Data Storage and Retrieval

The data storage and retrieval subsystem organizes the data, spatial and attribute, in a form which permits it to be quickly retrieved by the user for analysis, and permits rapid and accurate updates to be made to the database. This component usually involves the use of a database management system (DBMS) for maintaining attribute data. Spatial data are usually encoded and maintained in a proprietary file format.

1.5.3 Data Manipulation and Analysis

The data manipulation and analysis subsystem allows the user to define and execute spatial and attribute procedures to generate derived information. This subsystem is commonly thought of as the heart of a GIS, and usually distinguishes it from other DBMS and CAD systems.

1.5.4 Data Output

Output is the procedure by which information from the GIS is presented in a form suitable to the user. The data output subsystem allows the user to generate graphic displays, normally maps and tabular reports representing derived information. Data output can be in three forms: hardcopy, softcopy and the electronic form (binary form). Hardcopy outputs are permanent means of display. The information is printed on paper, mylar, photographic film or other similar material. Softcopy output is in a format which can be viewed on a computer monitor.

Output in electronic formats consists of computer-compatible files (in binary forms, etc.). Softcopy outputs are used to allow operator interaction and to preview data before the final output. A softcopy output can be changed interactively, but the view is restricted by the size of the monitor. A hardcopy output takes a longer time to produce and requires more expensive equipment such as colour plotters, but it is a permanent record.

The GIS flowchart is shown in Fig. 1.5. GIS may be further described by listing its typical components as below:

1. The computer, along with the GIS software, is the heart of GIS. Typically, GIS works on three main platforms such as Windows7/8/10, Linux and Unix operating systems.
2. Data storage devices such as computer hard drives, optical disks and external memory cards.
3. Data collection devices can be divided into components such as field surveying and point capturing using GPSs, remote sensing, digitization of existing maps and plans, and digital data transfer via Internet.
4. GIS software is designed to download, edit, sort and analyse the data. For example, database software, relations database software, geometric and drawing software, softcopy photogrammetry and satellite imagery analysis software.
5. GIS software is designed to process and output the data in the form of graphics, maps and plans.
6. Hardware components include surveying equipment (GPS, Total Station), remote sensing equipment, digitizers, scanners, interactive graphic terminals, plotters and printers.

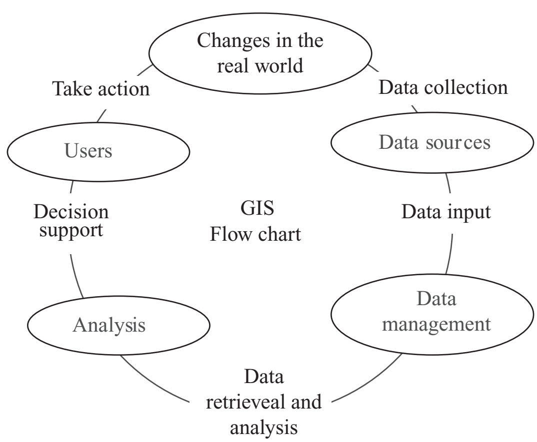


FIGURE 1.5 The GIS flowchart

1.6 COMPONENTS OF A GIS

An operational GIS also has a series of components that combine to make the system work. GIS have five important components, computer hardware, sets of application software modules, required data, people who manage the system and develop plans, and a well-designed implementation method. These components are critical for GIS.

A working GIS integrates the following five key components:

1. Hardware
2. Software
3. Data
4. People
5. Methods

1.6.1 Hardware

Hardware is the computer system on which a GIS operates. A GIS relies on a computer for storage and processing of data. The size of the computing system will depend on the type and nature of the GIS. A small-scale GIS will need only a small personal computer to run on, while a large, enterprise-wide system will need larger computers and a host of client machines to support multiple users.

In addition to computers, a variety of other devices can be used to capture and feed data into the system. Scanners and digitizing tables are used to scan existing paper maps, charts and drawings into the system. GPS receivers are used to create map features in the field and transmit the current location of moving vehicles. Today, GIS software runs on a wide range of hardware types from centralized computer servers to desktop computers used in stand-alone or networked configurations.

1.6.2 Software

GIS software provides the functions and tools needed to store, analyse and display geographic information. The core of any GIS system lies in the GIS software itself, providing the functionality to store, manage, link, query and analyse geographic data. In addition to the core, various other software components can be added to GIS software to provide access to additional sources of data and forms of functionality. Imaging systems used to analyse satellite imagery, DBMS used to store additional sets of data, and CAD systems can all be integrated into a GIS solution to provide the data and full functionality needs of the GIS.

1.6.3 Data

The most important component of a GIS is the data. Geographic data and related tabular data can be collected in-house, compiled to custom specifications and requirements, or purchased from a commercial data provider. A GIS can integrate spatial data with other existing data resources often stored in a corporate DBMS. The integration of spatial data (often proprietary to the GIS software), and tabular data stored in a DBMS is the key functionality afforded by a GIS.

Data for a GIS comes in two forms—geographic or spatial data, and attribute of spatial data. Spatial data are data that contain an explicit geographic location in the form of a set of coordinates. Attribute data are descriptive sets of data that contain various information relevant to a particular location, for example, depth, height and sales figures; and can be linked to a particular location by means of an identifier, for example, address and zip code.

Sources of spatial data include paper maps, charts and drawings scanned or digitized into the system, digital files imported from CAD or other graphic systems, coordinate data recorded using a GPS receiver, and data captured from satellite imagery or aerial photography.

Sources of attribute data include databases, workflow, messaging and any other form of computer system that stores data sets that can be linked to the GIS by means of a common identifier. Data streams generated by sensors or data loggers of any kind can be stored in the GIS and that data can be linked to it by means of an identifier. Data from satellite imagery derived through image analysis techniques that can be linked to a location or set of locations.

1.6.4 People

The main objective of a GIS is to support its users with the appropriate data and decision support tools. Thus, careful consideration of particular needs of the users must be given at the design stages of the system, so that each group of users will be given access to the data and functionality of the system in the most appropriate way. A system must be highly accessible and usable, otherwise it may not to be used effectively, or may not be used at all.

Some users need access to the most advanced functionality features a GIS offers, and are willing to invest the time needed to learn how to effectively use these tools. Other users only need specific answers to their questions. In such cases, they have no need for advanced functionality and there is no point in investing the funds required to train them. Hence for maximum utilization of GIS by the users, it is necessary to invest in workstations and training, allowing them to directly interact with the GIS. A careful attention to the needs of each user group will lead to higher utilization rates, and, therefore, a greater return result on the investment in GIS technology.

GIS technology is of limited value without the people who manage the system and develop plans for applying it to real world problems. GIS users range from technical specialists, who design and maintain the system, to those who use it to help them perform their everyday work. The identification of GIS specialists versus end users is often critical to the proper implementation of GIS technology.

1.6.5 Methods

A successful GIS operates according to a well-designed implementation plan and business rules, which are the models and operating practices unique to each organization.

As in all organizations dealing with sophisticated technology, new tools can only be used effectively if they are properly integrated into the entire business strategy and operation. To do this properly, it requires not only the necessary investments in hardware and software, but also in the hiring of personnel to use the new technology in the proper organizational context. Failure to implement the GIS without regard for a proper organizational commitment will result in an unsuccessful system. It is simply not sufficient for an organization to purchase a computer with some GIS software, hire some enthusiastic individuals and expect instant success.

1.7 THE FOUR MS

There are four key activities in a GIS. These are measurement, mapping, monitoring and modelling (see Fig. 1.6). The scientists, engineers, resource managers and urban planners observe and manage

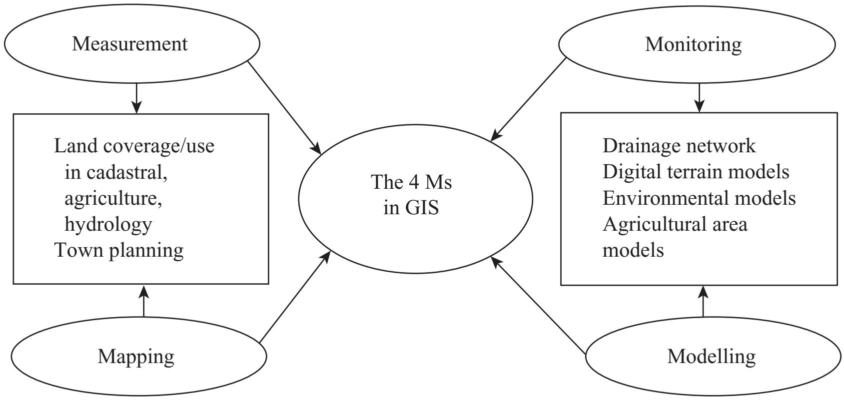


FIGURE 1.6 Schematic representation of the four Ms in a GIS

these four key parameters, and develop maps which portray characteristics of the earth. The above groups are monitoring the changes in our surroundings in space and time. They model alternatives of action and process operation in the environment.

1.8 GIS WORK FLOW

The five essential elements that define a GIS are data acquisition, data preprocessing, data management, data manipulation and analysis, and product generation (output). The above elements are in a continuous flow process in a GIS as in Fig. 1.7. Data acquisition is the process of identifying and collecting the data required for a given application. It includes locating and acquiring of existing data such as aerial maps, ground maps, images from photogrammetry and images gathered using remote sensing. Data preprocessing operations, which are concerned with storage and retrieval of data, are associated with receiving data into the system by manual digitising, scanning, keyboard entry of attributes and online retrieval from other database.

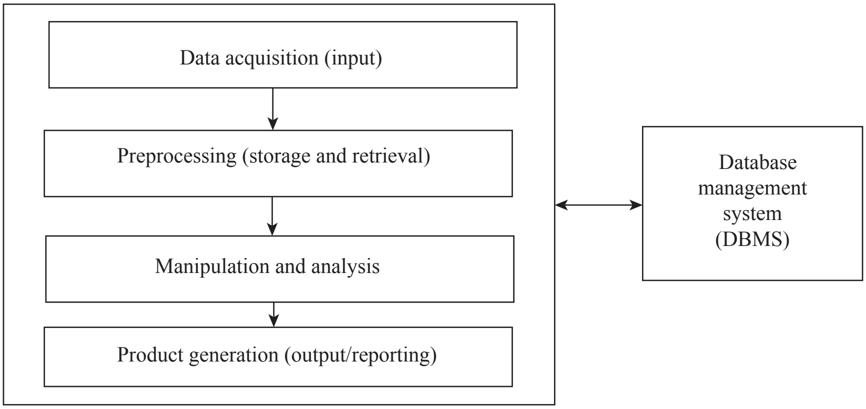


FIGURE 1.7 GIS workflow diagram

Database management governs the creation of an access to the database itself. It provides an easy and consistent method for data entry, updating data, deletion of data and retrieval of data. A modern DBMS is usually used for the creation of a GIS database. Storage and retrieval mechanisms include the control of physical storage of the data in memory and its retrieval for the needs of the other components. The data manipulation and analysis includes:

1. Classification and aggregation
2. Geometric operations such as rotation, translation, scaling, rectification and registration
3. Controlled determination
4. Data structure conversion
5. Spatial operations of connectivity and neighbourhood operations
6. Measurement of distance and direction
7. Statistical analysis as descriptive statistics regression, correlation and cross-tabulation
8. Modelling

These elements represent the whole spectrum of techniques available for the transformation of the digital model by mathematical means. Product generation is the phase where final outputs from the GIS are created. These output products include statistical reports, maps and graphics of various kinds.

1.9 FUNDAMENTAL OPERATIONS OF GIS

The fundamental operations of a GIS include:

1. The overlay operation, which involves the combination of two or more maps, according to the Boolean conditions and results in the delineation of new boundaries.
2. Re-classification operations that transform the attribute information associated with single map coverage.
3. Distance and connectivity measurements, which include both a simple measure of inter-point distance and more completed operations such as the construction of zones of increasing transport cost away from specified locations.
4. Neighbourhood characterization, which involves the values to a location, both summary and mean measure of a variable, and includes smoothing and enhancement filters. Sequences of such manipulation operations are known as cartographic modelling.

1.10 LEVELS OF USE OF A GIS

There are three levels at which a GIS can be used. They are data management, analysis and prediction.

1. *Data management:* This is the lower level of application, which is used to input and store data, retrieve that data through spatial and conditional queries, and to display the result. For the data management type of application, the GIS is merely used as an inventory system with the

purpose of storing and displaying information about spatial features. These features include width, number of lanes and traffic count for a particular highway.

2. *Analysis:* The second level of a GIS application is the analysis. Here the user uses the spatial analysis capability of the system. Examples: determining the shortest path between two locations, grouping of areas of land into larger ones depending on certain criteria and so on.
3. *Prediction:* The highest application level of GIS falls into the prediction category (what if?). It is here that the data management and analysis capabilities of a GIS are combined into a modelling operation. Examples: modelling operation for the prediction of the effect of traffic on a particular area at a particular time, predicting the effect of flooding and predicting a particular effect of an earthquake.

1.11 OBJECTIVE OF GIS

The main objective of a GIS is to reduce the time for the expensive activities of handling, recording, researching, handling of huge amount of data and generating and establishing its relation to the geographical position or land features. The rapidly decreasing costs of computer hardware and software make the GIS handle more tasks, thus making it reach to a wider audience. In addition, the cost of data storage devices in a computer has decreased drastically in recent years, and the computer memory has increased from kilobytes to gigabytes.

The primary goal of a GIS is to take a raw data and transform it by overlap and by various analytical calculations into new information that can help to make a sudden decision. A GIS is not built for one or two specific applications, but it is a common problem-solving tool. A GIS provides an organization with the capability of applying geographic analyst methods to designated geographic areas to solve various problems.

The following are the reasons why a GIS is needed:

1. Geospatial data are poorly maintained.
2. Maps and statistics are out of date.
3. Data and information are inaccurate.
4. There is no data retrieval service.
5. There is no data sharing.

Once a GIS is implemented, the following benefits are expected:

1. Geospatial data are better maintained in a standard format.
2. Revision and updating are easier.
3. Geospatial data and information are easier to search, analyze and represent.
4. More value added products are obtained.
5. Geospatial data can be shared and exchanged freely.
6. Productivity of the staff is improved with more efficiency.
7. Time and money are saved.
8. Better decisions can be made within a short time.

1.12 THE THEORETICAL FRAMEWORK OF A GIS

Raw data are collected from the real world and that data, with its geographic positions, are fed into the GIS software for data manipulation and required data analysis, and the output is gathered after analysis through that application software. This is the theoretical framework of a GIS. Figure 1.8 gives the transformation-based view of a GIS operation.

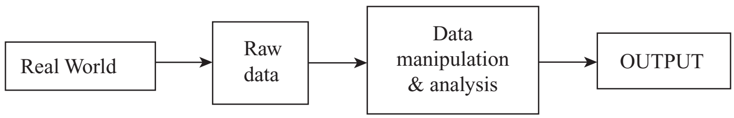


FIGURE 1.8 The transformation-based view of a GIS operation

1.13 ACCURACY IN A GIS

Engineers and surveyors always give more emphasis to the accuracy of a GIS. For other people like planners, rescue squad, fireman and security personnel are having less care about the accuracy. To develop an accurate GIS, the required geo-references have to be professionally surveyed and located, and then carefully described in the computer form. Unfortunately, the usual GISs are not done completely in this fashion, and various shortcuts are taken. The usual practice is that the data on existing maps that comes under the area of interest are converted to digital form and adjusted to the coordinate system being used. Such values will often be short on accuracy. When a map is digitized for a particular GIS application, the accuracy of the map which is digitized will define the accuracy of that GIS application.

But when control stations are set up at various points of land that have been previously located in that area and surveyed with a good degree of accuracy and then those values are taken for calibrating the digitized maps, then that GIS will achieve a good level of accuracy. The different parcels of land will have to be shrunk, stretched, warped or bent to make them fit with control point values.

1.14 DATA EXPLORATION

Data exploration is a data centred query and analysis. Data exploration can be a GIS operation by itself or a predecessor to formal data analysis. Data query allows the user to explore the general trends in the data, to take a close look at data subsets, and to focus on relationships between data sets. The purpose of data exploration is to understand the data better to formulate research questions and hypothesis. An important component of effective data exploration consists of interactive and dynamically linked visual tools including maps, graphs and tables. By using a map or a table, a GIS user can perform data query. Following a query, a GIS user can view both the spatial and attribute components of the data subset.

1.15 THEMATIC LAYERING

A thematic map shows a particular theme connected within a specific geographic area. Thematic map focuses on different themes such as physical, social, political, economical, sociological and agricultural

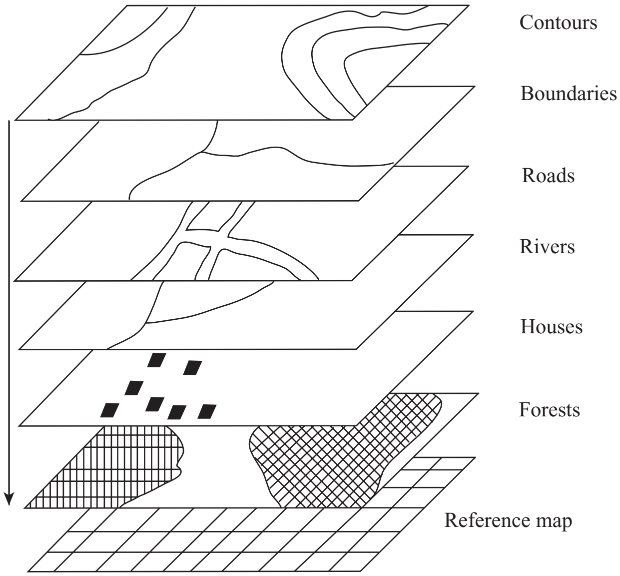


FIGURE 1.9 Thematic layering of a few of the typical items needed for a planimetric GIS map

as aspects of a geographical area. The real world entities are so complex that they should be classified into object classes with some similarity through thematic data modelling in a spatial database.

Attributes are often termed thematic data or non-spatial data, which are linked with spatial data or geometric data. An attribute has a defined characteristic of entity in the real world.

With a GIS, the relevant information for a particular area can be called on the computer screen. The system can further be used to call up maps with the needed data displayed graphically. Furthermore, all or parts of the data from various maps can be overlayed, enabling users to examine different sets of data simultaneously. Figure.1.9 shows a few of the typical items that might be digitally stored for a particular piece of land. The items are really not layered within the computer as shown in the figure, but this type of figure is commonly used to list the types of data that are stored. The term thematic layering is used in relation to a GIS, where the word thematic means relating to or consisting of a theme or themes.

1.16 LEVELS OF MEASUREMENT IN GIS

Spatial features or entities contain information about how they occupy space, how important they are and what they are. The non-spatial information or attributes help us to describe the object we observe in space. In GIS analysis, the character of attribute data themselves can influence the utility of data sets. In GIS there is a well-established measurement framework for all forms of data, including geographic data. Most natural phenomena or characteristics are best represented using fields or levels, particularly those that are continuous. There are different types of levels or fields. These levels of geographic measurement are well illustrated in Fig. 1.10. The four commonly used levels of measurements are nominal scale, ordinal scale, interval scale and ratio scale.

Nominal scales are variables, which are defined by name, without a specific order. Examples of nominal scale are land use, commercial areas and residential areas. Ordinal values are variables in

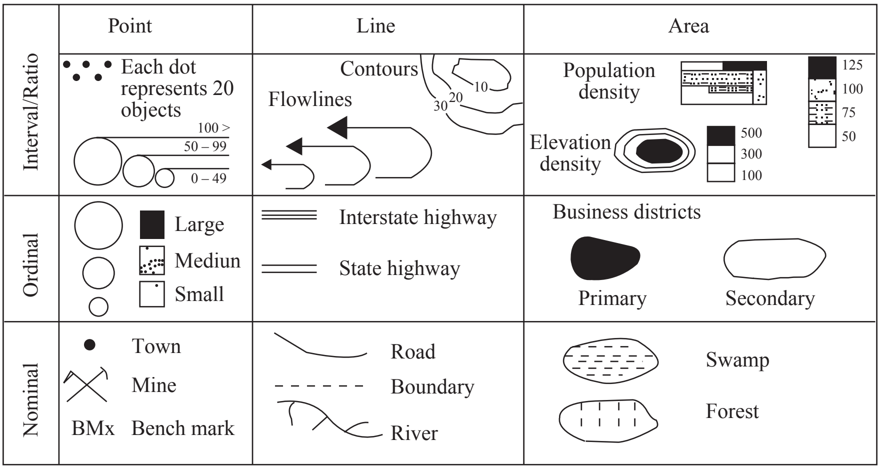


FIGURE 1.10 Levels of measurement in a GIS

discrete classes with an inherent order. Examples for ordinal variables are roads and highways. In an interval scale of measurement, numbers are assigned to the items measured. Examples for interval scale are pressure, temperature and wind speed. Here the data measured can be compared as in the case of an ordinal scale of measurement. The ratio variables have the same characteristics as interval variables, but they have a natural zero or a starting point, such as rainfall in a month. Thus:

Nominal: different qualitative values which cannot be ordered (land cover)
Ordinal: ordered data but not directly comparable (landslide risk categories)
Interval/ratio: usual measurements taking real values (elevation)

1.17 CATEGORIES OF GIS

In the case of geo-referenced entities in a GIS, if the entity is represented by the x- and y-coordinate, it is said to be a two-dimensional GIS. When the two-dimensional x, y geometry integrates, the third dimension, namely z-coordinate (the height parameter) , it is called a 2.5-dimensional GIS. 2.5 dimensional GIS is used to create a digital terrain model (DTM). In addition, when the two-dimensional x, y geometry integrates any other predefined simple equation for an attribute like soil type, as in the equation $z = f(x, y)$, it is called a 2.5-dimentional GIS.

In a three-dimensional GIS, for the same pair of x, y coordinates, a number of spatial locations (points) with different z-coordinates could exist, as the z-coordinates are calculated independently from the x, y pair. The four-dimensional GIS, or the temporal GIS, is the fourth category of GIS, in which the fourth dimension represents time.

1.18 TOPOLOGY

Mathematically, topology can be defined as the study of geometrical properties and spatial relations unaffected by the continuous change of shape or size of figures. Topology in GIS is generally defined as the spatial relationships between adjacent or neighbouring features. It is the term used to describe the geomatic characteristics of objects, which do not change under transformations such as stretching and bending, and are independent of any coordinate systems. The topological character of an object is also independent of the scale of measurement. Topology relates to spatial data and consists of three elements, such as connectivity, containment and adjacency. Adjacency and containment describe the geometric relationship that exists between area features. Areas sharing a common boundary can be defined with adjacency. Containment is an extension of the adjacency theme, and describes area features contained within another area feature such as an island in a lake. Connectivity is the geometric property used to define the linkage between line features, such as roads connected to form the road network.

REVIEW QUESTIONS

1. What is a geographic information system? Why is data in a GIS called geo referenced?
2. How did Dr John Snow help the city of London in 1854? Explain.
3. Define thematic layering as it applies to GIS.
4. Mention the objectives of a GIS
5. What does the term spatial mean?
6. Explain the concept of four Ms in a GIS.
7. Explain the various levels of the use of a GIS.
8. Write the four fundamental operations of a GIS.
9. What are the components of a GIS? Explain.
10. Write a short note on data exploration.
11. Explain the workflow of a GIS with the help of a workflow diagram.
12. Describe the architecture of a GIS.
13. Discuss about the theoretical framework of a GIS.
14. Write a short note on the categories of a GIS.
15. What are the levels of measurements in a GIS?

2 GIS DATA MODELS

Chapter Outline

2.1 INTRODUCTION

A model is a simplified representation of a phenomenon or a system. The data model represents a set of guidelines to convert the real world called entity to the digitally and logically represented spatial objects consisting of the attributes and geometry. The attributes are managed by a thematic structure, whereas the geometry is represented by a geometric-topological structure.

Geographic information system (GIS) stores information about the world as a collection of thematic layers that can be linked together by geography. This simple but extremely powerful and versatile concept has been proven invaluable for solving many real-world problems from tracking delivery vehicles, to recording details of planning applications and to modelling the global atmospheric circulation. The thematic layer approach allows us to organize the complexity of the real world into a simple representation to help facilitate our understanding of natural relationships.

Conceptually, a GIS can be visualized as a stacked set of map layers, where each layer is aligned or registered to all other layers. Typically, each layer will contain a unique geographic theme or data type. These themes might include, for example, topography, soils, land use, cadastral (land ownership) information or infrastructure such as roads, pipelines, power lines or sewer networks. This image of GIS is shown in Fig. 2.1. By sharing mutual geography, all layers in the GIS can be combined or overlaid in any user-specified combination. In some cases, data type may define the GIS that the system is designed to handle. For example, the term ‘Land Information System’ or ‘LIS’ is often applied to a type of GIS used by districts, cities and municipalities to manage land parcel information.

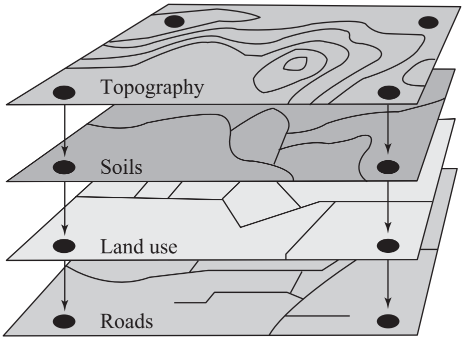


FIGURE 2.1 A conceptual model of a GIS visualizes a set of map layers or themes, all overlaid together to a common map base of a particular geographic area. Each layer typically contains one type of data

2.2 GIS DATA TYPES

The basic data types in a GIS reflect traditional data found on a map. Accordingly, GIS technology uses two basic types of data. They are spatial data and attribute data.

2.2.1 Spatial Data

Spatial data describes the absolute and relative location of geographic features. Spatial features may be discrete or continuous. A GIS can represent spatial data, which has a physical dimension on earth. The various components of geographic data can be reduced in the form of point, arcs or polygons for effective processing. In the case of a GIS, the data may be available in tabular form, a geographical map, a digital map or a remotely sensed map.

Discrete features are those that do not exist between observations, form separate entities and are individually distinguishable. Wells, roads, buildings and land use types are examples of discrete features. Continuous features exist specially between observations. Precipitation of rain and elevation are examples of continuous features.

Spatial features in the real world come in four types—point, line, area and surface (see Fig. 2.2). These are used both in GIS and analogue map systems. GIS uses two basic data models to represent the spatial feature, the vector and raster. Spatial data are stored in graphic files and managed by a file management system.

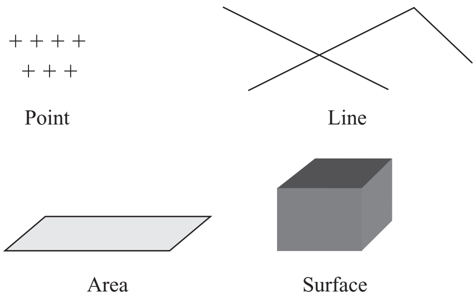


FIGURE 2.2 Different types of spatial features

2.2.2 Attribute Data

Attribute data describes the characteristics of the spatial features. It holds the characteristics of the spatial features, and the descriptive information about the geographic features. These characteristics can be quantitative and qualitative in nature. Attribute data are non-spatial data associated with time and area entities. Attribute data are often referred to as tabular data. The GIS attributes are represented using colours, textures, and linear or graphic symbols—for instance, school/college locations are designated using special symbols and contour lines with brown colour. The actual value of the attribute that is measured or sampled and stored in the database is called attribute value. In addition, each spatial entity may have more than one attribute associated with it, for example, a point representing a hotel can have a number of rooms, different standards of accommodation, different parking space and so on. The amount of attribute data to be attached to a spatial feature varies significantly, depending on the feature type and the application.

The coordinate location of a wildlife sanctuary would be spatial data, and the characteristics of that wildlife sanctuary, for example, cover group, dominant species, special and rare group of species, and nature of vegetations and climatic conditions, would be attribute data. Other data types, in particular, image and multimedia data, are becoming more prevalent with changing technology. Depending on the specific content of the data, image data may be considered either spatial (e.g., photographs, animation and movies) or attribute (e.g., sound, descriptions and narrations), Attributes may be classified into two types—primary attributes and secondary attributes. Socio-economic characteristics and physical properties of objects are some of the examples of primary attributes; and flow of information levels, districts, capitals and names of constituencies are examples of secondary attributes. Many commercial GIS packages store attribute data separate from spatial data in a split data system, known as the georelational model. Spatial data are stored in graphic files and managed by a file management system, but attribute data are stored in a relational database. A relational database is a collection of tables, which can be connected to each other by attributes whose values can uniquely identify a record in a table. Instead of storing spatial data and attribute data in a split data system, the object-oriented data model stores both data in a single database. A field called geometry is used to store spatial data. The object-oriented data model system eliminates the complexity of coordinates and synchronizes the spatial data and attribute data, and brings GIS closer to other information systems.

2.3 SPATIAL DATA MODELS

Traditionally, spatial data has been stored and presented in the form of a map. Three basic types of spatial data models have evolved for storing geographic data digitally. They are vector, raster and image.

Figure 2.3 reflects the two primary spatial data encoding techniques. They are vector and raster. Image data uses techniques very similar to raster data, however, typically it lacks the internal format required for analysis and modelling of the data. Images reflect pictures or photographs of the landscape.

2.4 VECTOR DATA MODEL

All spatial data models are approaches for storing the spatial location of geographic features in a database. Vector storage implies the use of vectors (directional lines) to represent a geographic feature. Vector data are characterized by the use of sequential points or vertices to define a linear segment. Each vertex consists of an x-coordinate and a y-coordinate.

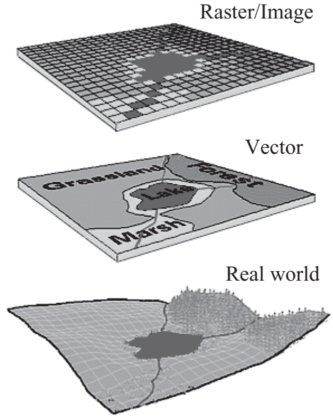


FIGURE 2.3 Two primary spatial data encoding techniques

Vector lines are often referred to as arcs and consist of a string of vertices terminated by a node. A node is defined as a vertex that starts or ends an arc segment. One coordinate pair, and a vertex define point features. Polygonal features are defined by a set of closed coordinate pairs. In vector representation, the storage of the vertices for each feature is important, as well as the connectivity between features, for example; the sharing of common vertices where features connect. The geometry of a point is given by two-dimensional (2D) coordinates (x, y), whereas line, string and area are given by a series of point coordinates, as shown in Fig. 2.4(a). The topology, however, defines additional structure as shown in Fig. 2.4(b).

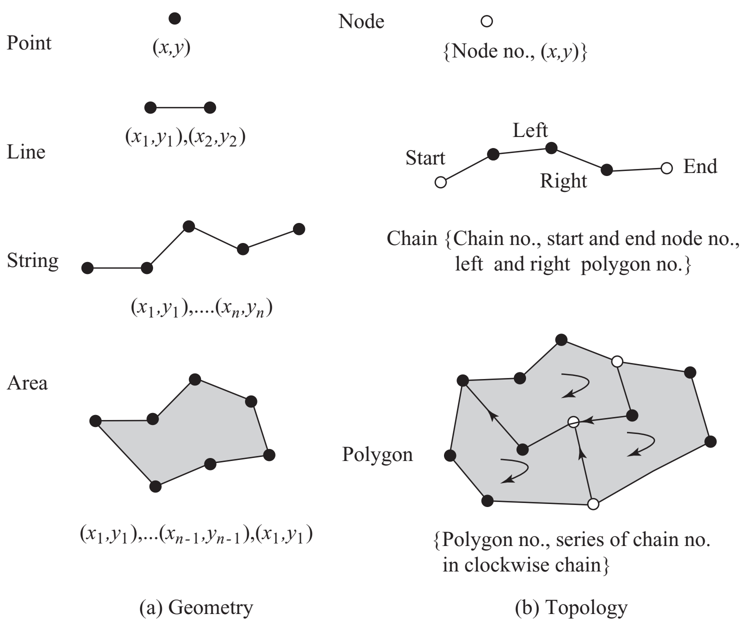


FIGURE 2.4 (a) Geometry and (b) topology of vector data

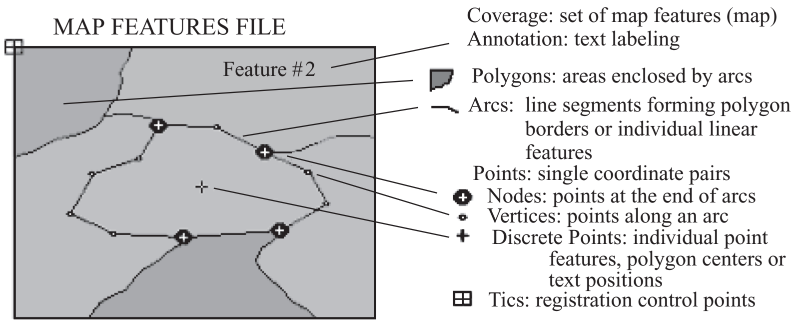


FIGURE 2.5 GIS vector data structure

To understand topology, the following terms should be clearly understood:

1. **Node:** An intersect of more than two lines or strings, or start and end points of a string with node number.
2. **Chain:** A line or a string with chain number, start and end node number, left and right neighbour polygons.
3. **Polygon:** An area with polygon number, series of chains that form the area in clockwise order (minus sign is assigned in case of anti-clockwise order).

Several different vector data models exist; however, only two are commonly used in GIS data storage. The most popular method of retaining spatial relationships among features is to explicitly record adjacency information in what is known as the topologic data model. Topology is a mathematical concept that has its basis in the principles of feature adjacency and connectivity.

The topologic data structure is often referred to as an intelligent data structure, because spatial relationships between geographic features are easily derived when using them. Primarily for this reason, the topologic model is the dominant vector data structure currently used in GIS technology. Many of the complex data analysis functions cannot effectively be undertaken without a topologic vector data structure. Topology is reviewed in greater detail later in the book.

The secondary vector data structure, which is common among GIS software, is the computer-aided drafting (CAD) data structure. This structure consists of listing elements, not features, defined by strings of vertices, to define geographic features such as points, lines or areas, as shown in Fig. 2.5. There is considerable redundancy with this data model, because the boundary segment between two polygons can be stored twice, once for each feature. The CAD structure emerged from the development of computer graphics systems without specific considerations for processing geographic features. Accordingly, as features such as polygons are self-contained and independent, questions about the adjacency of features can be difficult to answer. The CAD vector model lacks the definition of spatial relationships between features that is defined by the topologic data model.

2.5 RASTER DATA MODEL

Raster data models incorporate the use of a grid cell data structure, where the geographic area is divided into cells identified by row and column. This data structure is commonly called raster. While the term raster implies a regularly spaced grid, other tessellated data structures do exist in grid-based GIS systems. In particular, the quad tree data structure has found some acceptance as an alternative raster data model.

The size of cells in a tessellated data structure is selected on the basis of the data accuracy and the resolution needed by the user. There is no explicit coding of geographic coordinates required, as that is implicit in the layout of the cells. A raster data structure is in fact a matrix where any coordinate can be quickly calculated if the origin point and the size of the grid cells is known. As grid cells can be handled as 2D arrays in computer encoding, many analytical operations are easy to program. This makes tessellated data structures a popular choice for many GIS software. Topology is not a relevant concept with tessellated structures, as adjacency and connectivity are implicit in the location of a particular cell in the data matrix.

The geometry of raster data is given by point, line and area objects as follows (see Fig. 2.6):

a. **Point objects:** A point is given by point ID, coordinates (i, j) and the attributes.

b. **Line object:** A line is given by line ID, series of coordinates forming the line, and the attributes.

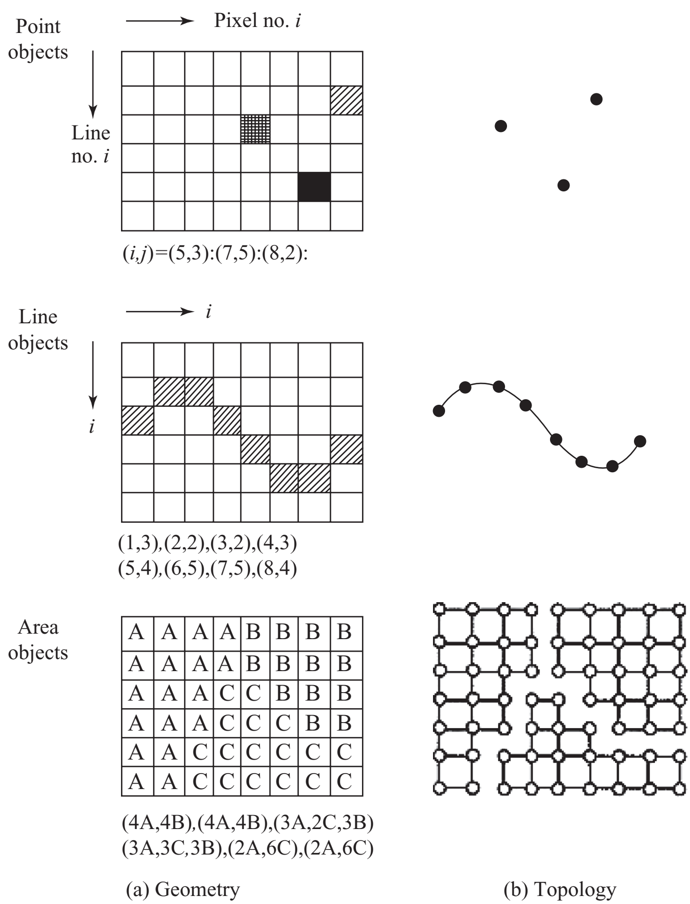


FIGURE 2.6 (a) Geometry and (b) topology of raster data

c. **Area objects:** An area segment is given by area ID, a group of coordinates forming the area and the attributes. Area objects in raster model are typically given by 'Run Length' that rearranges the raster into the sequence of length (or number of pixels) of each class as shown in Fig. 2.6.

The topology of a raster model is rather simple as compared with the vector model. The topology of line objects is given by a sequence of pixels forming the line segments.

The topology of an area object is usually given by 'Run Length' structure as follows:

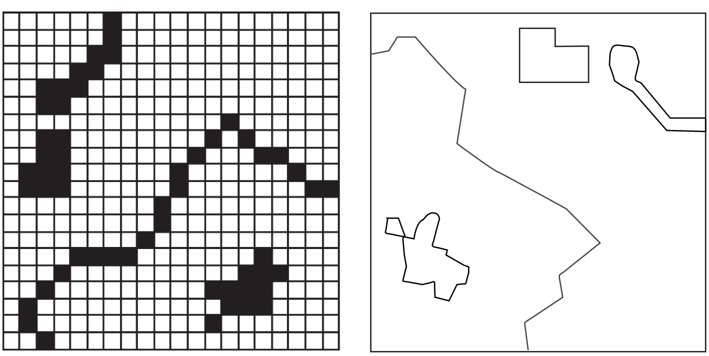
- Start line no. (start pixel no., number of pixels)
- Second line no. (start pixel no., number of pixels)

Several tessellated data structures exist; however, only two are commonly used in geographic information systems. The most popular cell structure is the regularly spaced matrix or raster structure. This data structure involves a division of spatial data into regularly spaced cells. Each cell is of the same shape and size. Squares are most commonly used.

As regularly spaced shapes rarely distinguish geographic data, cells must be classified with respect to the most common attribute for the cell. The problem of determining the proper resolution for a particular data layer can be a concern. If one coarse a cell size, then the data provide result as a generalized overlay. If one selects too fine a cell size, then too many cells may be created resulting in large data volumes, slower processing times and a more cumbersome data set. In addition, one can imply accuracy greater than that of the original data capture process, and this may result in some erroneous results during analysis.

Besides, as most data are captured in a vector format, for example, digitizing, data must be converted to the raster data structure. This is called vector-raster conversion. Most GIS software allows the user to define the raster grid (cell) size for vector–raster conversion. It is imperative that the original scale, for example, accuracy of the data, be known before conversion. The accuracy of the data, often referred to as the resolution, should determine the cell size of the output raster map during conversion (Fig. 2.7).

Most raster-based GIS software requires that the raster cell should contain only a single discrete value. Accordingly, a data layer, for example, forest inventory stands, may be broken down into a series of raster maps, each representing an attribute type, for example, a species map, a height map and a density map (Fig. 2.8). These are often referred to as attribute maps. This is in contrast to most conventional vector data models that maintain data as multiple attribute maps, for example,



(a) (b)

FIGURE 2.7 The difference between (a) raster and (b) vector representations of space

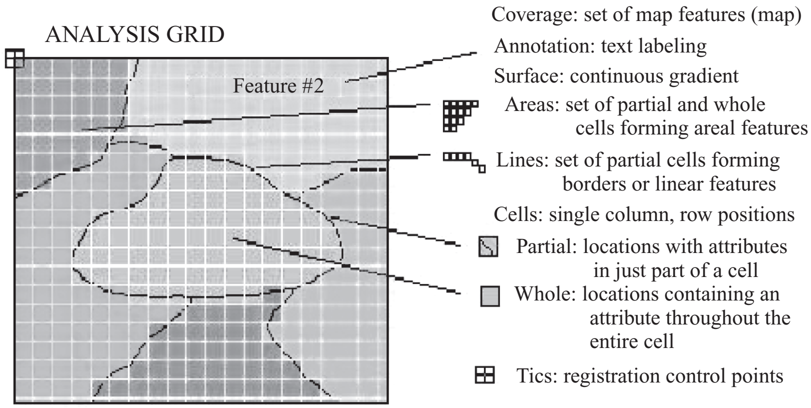


FIGURE 2.8 GIS raster map structure

forest inventory polygons linked to a database table containing all attributes as columns. This basic distinction of raster data storage provides the foundation for quantitative analysis techniques. This is often referred to as raster or map algebra. The use of raster data structures allows for sophisticated mathematical modelling processes, while vector-based systems are often constrained by the capabilities and language of a relational database management system (DBMS).

This difference is the major distinguishing factor between vector- and raster-based GIS software. It is also important to understand that the selection of a particular data structure can provide advantages during the analysis stage. For example, the vector data model does not handle continuous data, for example, elevation, very well, whereas the raster data model is more ideally suited for this type of analysis. Accordingly, the raster structure does not handle linear data analysis, for example, shortest path, very well, whereas vector systems do. It is important for the user to understand that there are certain advantages and disadvantages to each data model.

The selection of a particular data model, vector or raster, is dependent on the source and type of data, as well as the intended use of the data. Certain analytical procedures require raster data, while others are better suited to vector data.

2.6 IMAGE DATA

Image data is most often used to represent graphic or pictorial data. The term image inherently reflects a graphic representation, and in the GIS world, it differs significantly from raster data. Most often, image data is used to store remotely sensed imagery, for example, satellite scenes, orthophotos or ancillary graphics such as photographs and scanned plan documents. Image data is typically used in GIS systems as background display data (if the image has been rectified and geo-referenced), or as a graphic attribute.

Remote sensing software makes use of image data for image classification and processing. Typically, this data must be converted into a raster format (and perhaps vector) to be used analytically with the GIS. Image data is typically stored in a variety of the industry standard proprietary formats. These often reflect the most popular image processing systems. Other graphic image formats, such as TIFF, GIF and PCX, are used to store ancillary image data. Most GIS software will read such formats and allow you to display this data.

FIGURE 2.9 A remote sensed image

Image data is most often used for remotely sensed imagery, such as satellite imagery or digital orthophotos. Figure 2.9 shows a remote sensed image.

2.7 VECTOR AND RASTER—ADVANTAGES AND DISADVANTAGES

There are several advantages and disadvantages for using either the vector or raster data model to store spatial data. These are summarized below.

Advantages of Vector Data

1. Data can be represented in its original resolution and form without generalization.
2. Graphic output is usually more aesthetically pleasing (traditional cartographic representation).
3. As most data (e.g., hard copy maps) is in the vector form, no data conversion is required.
4. Accurate geographic location of data is maintained.
5. It allows for efficient encoding of topology, and as a result more efficient operations that require topological information, for example, proximity, network analysis.

Disadvantages of Vector Data

1. The location of each vertex needs to be stored explicitly.
2. For effective analysis, vector data must be converted into a topological structure. This requires intensive processing and extensive data cleaning. In addition, topology is static, and any updating or editing of the vector data requires re-building of the topology.
3. Algorithms for manipulative and analysis functions are complex and may be required for intensive processing. Often, this inherently limits the functionality for large data sets, for example, a large number of features.
4. Continuous data, such as elevation data, is not effectively represented in vector form. Usually, substantial data generalization or interpolation is required for these data layers.
5. Spatial analysis and filtering within polygons is impossible.

Advantages of Raster Data

1. The geographic location of each cell is implied by its position in the cell matrix. Accordingly, other than an origin point (e.g., bottom left corner), no geographic coordinates are stored.
2. Due to the nature of the data storage technique, data analysis is usually easy to program and quick to perform.
3. The inherent nature of raster maps, for example, one attribute maps, is ideally suited for mathematical modelling and quantitative analysis.
4. Discrete data, for example, forestry stands, is accommodated equally well as continuous data, for example, elevation data, and facilitates the integrating of the two data types.
5. Grid-cell systems are very compatible with raster-based output devices, for example, electrostatic plotters and graphic terminals.

Disadvantages of Raster Data

1. The cell size determines the resolution at which the data is represented.
2. It is especially difficult to adequately represent linear features depending on the cell resolution. Accordingly, network linkages are difficult to establish.
3. Processing of associated attribute data may be cumbersome if a large amount of data exists. Raster maps inherently reflect only one attribute or characteristic for an area.
4. As most input data is in vector form, data must undergo vector-to-raster conversion. Besides increased processing requirements, this may introduce data integrity concerns due to generalization and choice of inappropriate cell size.
5. Most output maps from grid-cell systems do not conform to high-quality cartographic needs.

It is often difficult to compare or rate GIS software that uses different data models. Some personal computer (PC) packages use vector structures for data input, editing and display, but convert to raster structures for any analysis. Other more comprehensive GIS offerings provide both integrated raster and vector analysis techniques. They allow users to select the data structure appropriate for the analysis requirements. Integrated raster and vector processing capabilities are most desirable, and provide the greatest flexibility for data manipulation and analysis.

2.8 ATTRIBUTE DATA MODELS

The real world entities are so complex that they should be classified into object classes with some similarity through thematic data modelling in a spatial database. The objects in a spatial database are defined as representations of real world entities with associated attributes. Generally, geospatial data have three major components, such as position, attributes and time. Attributes are often termed thematic data or non-spatial data that are linked with spatial data or geometric data. An attribute has a defined characteristic of entity in the real world.

One method of defining attribute data is by the data types allowed in a GIS package. The data type used in a GIS includes character string, integers, real numbers, dates and time interval as in the case of computer programming. Here each field in an attribute table is defined with a data type, which applies to the domain of the field. The second method for defining attribute data is defining the attribute data by a measurement scale. By a measurement scale, an attribute can be categorized as *nominal, ordinal, interval* and *ratio. Nominal* data describe different kinds of different categories of data such as land use types or soil types. *Ordinal* data differentiate data by a ranking relationship. For example, cities are classified as big cities, medium cities and small cities according to the population. *Interval* data describes intervals between values such as temperature and pressure readings. For example, the temperature reading of 60°C is hotter than a temperature reading of 40°C. *Ratio* data are the same as interval data, except that ratio data are based on meaningful, or absolute zero value. Population density is an example for ratio data, where a density of 0 is an absolute zero. The above four measurement scales can be grouped into two higher-level categories such as *categorical,* which includes nominal and ordinal scales, and *numerical,* which includes interval and ratio scales.

A separate data model is used to store and maintain attribute data for GIS. These data models may exist internally within the GIS software, or may be reflected in external commercial database management software. A variety of different data models exist for the storage and management of attribute data. The most common are:

1. Tabular
2. Hierarchical
3. Network
4. Relational
5. Object-oriented

The tabular model is the manner in which most early GIS software packages store their attribute data. The next three models are those most commonly implemented in DBMS. The object oriented is newer but rapidly gaining popularity for some applications. A brief review of each model is provided.

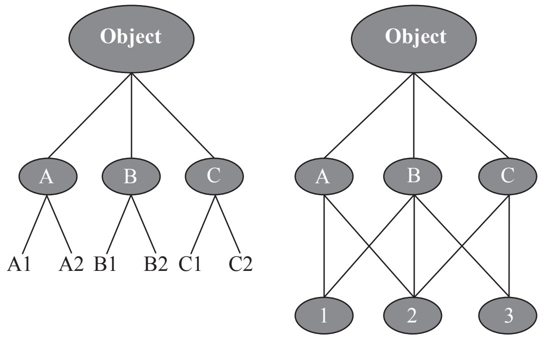
2.8.1 Tabular Model

The simple tabular model stores attribute data as sequential data files with fixed formats (or comma delimited for ASCII data), for the location of attribute values in a predefined record structure. This type of a data model is outdated in the GIS arena. It lacks any method of checking data integrity, as well as being inefficient with respect to data storage, for example, limited indexing capability for attributes or records.

2.8.2 Hierarchical Model

The hierarchical database organizes data in a tree structure. Data are structured downward in a hierarchy of tables. Any level in the hierarchy can have unlimited children, but any child can have only one parent. Hierarchical DBMS have not gained any noticeable acceptance for use within GIS. They are oriented for data sets that are very stable, where primary relationships among the data change infrequently or never at all. Additionally, the limitation on the number of parents that an element may have is not always conducive to actual geographic phenomena.

Several records or files are hierarchically related with each other. For example, an organization has several departments, each of which has attributes such as name of director, number of staffs and annual products. Each department has several divisions with attributes of name of manager, number of staffs, annual products and so on. Then each division has several sections with attributes such as name of head, number of staff and number of PCs. An hierarchical model is of a tree structure (see Fig. 2.10(a)). A set of links connects all record types in a tree structure. The advantages of an hierarchical model are high speed of access to large datasets and ease of updating. However, the



(a) Hierachical model (b) Network model

FIGURE 2.10 Concept of hierarchical and network model

disadvantage is that linkages are only possible vertically, but not horizontally or diagonally, which means there is no relation between different trees at the same level, unless they share the same parent.

2.8.3 Network Model

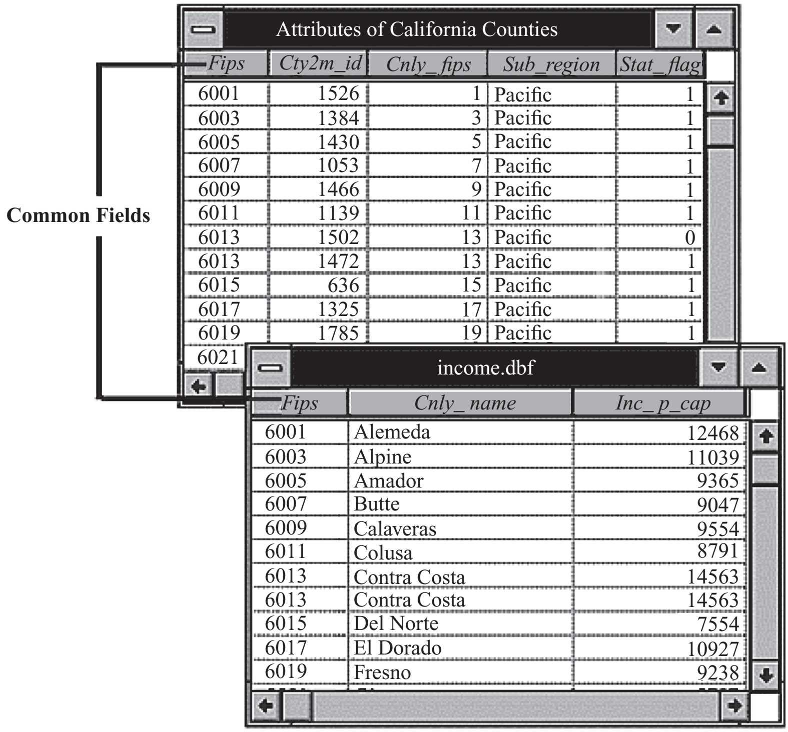
The network database organizes data in a network. Network DBMS have not found much more acceptance in GIS than the hierarchical DBMS. They have the same flexibility limitations as hierarchical databases; however, the more powerful structure for representing data relationships allows a more realistic modelling of geographic phenomena. However, network databases tend to become overly complex easily. In this regard, it is easy to lose control and understanding of the relationships between elements (see Fig. 2.10(b)).

2.8.4 Relational Model

The relational database organizes data in tables. Each table is identified by a unique table name, and is organized by rows and columns. Each column within a table also has a unique name. Columns store the values for a specific attribute such as cover group and tree height. Rows represent one record in the table. In a GIS, each row is usually linked to a separate spatial feature such as a forestry stand. Accordingly, each row would be composed of several columns, each column containing a specific value for that geographic feature. Table 2.1 presents a sample table for forest inventory features. This table has four rows and five columns. The forest stand number would be the label for the spatial feature, as well as the primary key for the database table.

TABLE 2.1 Table showing forest inventory features

Unique Stand Number	Dominant Cover Group	Average Tree Height	Stand Site Index	Stand Age
001	DEC	3	G	100
002	DEC-CON	4	M	80
003	DEC-CON	4	M	60
004	CON	4	G	120



Attributes of California Counties

Fips	Cty2m_id	Cnly_fips	Sub_region	Stat_flag
6001	1526	1	Pacific	1
6003	1384	3	Pacific	1
6005	1430	5	Pacific	1
6007	1053	7	Pacific	1
6009	1466	9	Pacific	1
6011	1139	11	Pacific	1
6013	1502	13	Pacific	0
6013	1472	13	Pacific	1
6015	636	15	Pacific	1
6017	1325	17	Pacific	1
6019	1785	19	Pacific	1
6021				

income.dbf

Fips	Cnly_name	Inc_p_cap
6001	Alemeda	12468
6003	Alpine	11039
6005	Amador	9365
6007	Butte	9047
6009	Calaveras	9554
6011	Colusa	8791
6013	Contra Costa	14563
6013	Contra Costa	14563
6015	Del Norte	7554
6017	El Dorado	10927
6019	Fresno	9238

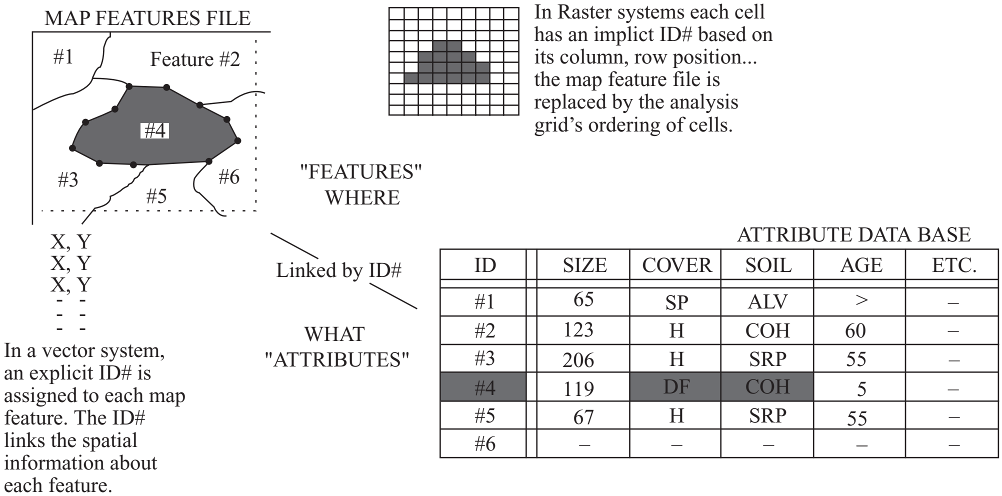
FIGURE 2.11 DBMS model

This serves as the linkage between the spatial definition and the attribute data for the feature.

Data are often stored in several tables. Tables can be joined or referenced to each other by common columns (relational fields). Usually, the common column is an identification number for a selected geographic feature, for example, a forestry stand polygon number. This identification number acts as the primary key for the table. The ability to join tables through use of a common column is the essence of the relational model. Such relational models are usually ad hoc in nature and form the basis for querying in a relational GIS product. Unlike the other previously discussed database types, relationships are implicit in the character of the data, as opposed to explicit characteristics of the database set up. The relational database model is the most widely accepted for managing the attributes of geographic data.

There are many different designs of DBMS, but in a GIS, the relational design has been the most useful (Fig. 2.11). In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together. This surprisingly simple design has been very widely used, primarily because of its flexibility and very wide deployment in applications in GIS.

In fact, most GIS software provides an internal relational data model, as well as support for commercial off-the-shelf (COTS) relational DBMS. COTS DBMS are referred to as external DBMS. This approach supports both users with small data sets, where an internal data model is sufficient, and customers with larger data sets, who use a DBMS for other corporate data storage requirements. With an external DBMS, the GIS software can simply connect to the database, and the user can make use of the inherent capabilities of the DBMS. External DBMS tend to have much more extensive querying and data integrity capabilities than the GIS internal relational model. The emergence and use of the



ID	SIZE	COVER	SOIL	AGE	ETC.
#1	65	SP	ALV	>	–
#2	123	H	COH	60	–
#3	206	H	SRP	55	–
#4	119	DF	COH	5	–
#5	67	H	SRP	55	–
#6	–	–	–	–	–

FIGURE 2.12 Basic linkage between vector data and attributes mandated in a relational database file

external DBMS is a trend that has resulted in the proliferation of GIS technology into more traditional data processing environments.

The relational DBMS is attractive because of:

1. Simplicity in organization and data modelling
2. Flexibility—data can be manipulated in an ad hoc manner by joining tables
3. Efficiency of storage—with a proper design of data tables, redundant data can be minimized
4. The non-procedural nature—queries on a relational database do not need to take into account the internal organization of the data

The relational DBMS has emerged as the dominant commercial data management tool in GIS implementation and application. Figure 2.12 illustrates the basic linkage between a vector spatial data (topologic model) and attributes mandated in a relational database file.

2.8.5 Object-Oriented Model

The object-oriented database model manages data through *objects*. An object is a collection of data elements and operations that together are considered a single entity. The object-oriented database is a relatively new model. This approach has the attraction that querying is very natural, as features can be bundled together with attributes at the database administrator's discretion. To date, only a few GIS packages are promoting the use of this attribute data model. However, initial impressions indicate that this approach may hold many operational benefits with respect to geographic data processing. Fulfilment of this promise with a commercial GIS product remains to be seen.

2.9 DIGITAL ELEVATION MODEL

When modern aerial photogrammetry and satellite remote sensing started to provide continuous surface information by means of optical cameras, radar or laser beams and the derivation of terrain elevation was made possible by stereoscopy and interferometry, topology gained a whole new meaning in spatial studies. The ability to perceive and analyse the physical, biological, chemical and cultural character of the earth's surface has since been greatly expanded. The elevation information in GIS can be represented in a digital format. This format is usually called digital elevation models (DEMs). Thus, a DEM is a computerized representation of the earth's relief. Different format exist; among the most usual are triangulated irregular networks (TINs), regular grids, contour lines and scattered data points.

A DEM is usually described either by a wire frame model, or an image matrix in which the value of each pixel is associated with a specific topographic height. Digital elevation models, in combination with other spatial data, are an important database for topography-related analyses or three-dimensional (3D) video animations. Different geo-referenced 3D products can be derived and complemented by a coordinate system, and presented in a 2D map projection or as a 3D perspective view.

2.9.1 The Availability of DEM Data

The scientific community and the commercial market are increasingly aware of the importance of DEMs in their applications. However, there are certain limitations in their employment. On the one hand, there is only limited access to existing digital topographic data due to high costs and military secrecy. On the other hand, free available global data sets with a horizontal resolution of 1 km and a vertical one of 100 m do not meet the requirements of detailed spatial and geographical analysis. In addition, when using the mosaic of existing DEMs, which are derived by different resolutions, data, frequencies and polarizations, to arrive at global representations of the earth's relief, the results are inhomogeneous, inconsistent and, consequently, incomparable.

The currently available digital elevation models exhibit discontinuities with respect to coverage, resolution, accuracy and reference datum. They are also inhomogeneous because they are derived through different acquisition methods. The currently best, most comprehensive, global data set comes with a horizontal resolution of one kilometre and an elevation accuracy of 100 m.

2.10 DIGITAL ELEVATION MODELS AND GEOGRAPHIC INFORMATION SYSTEMS

DEMs can be used together with other spatial data, image data in GISs, for instance. A GIS is an information system designed to acquire, store, process and display data referenced by spatial or geographical coordinates. In a sense, a GIS may be thought of as a higher-order map, being both a database system with specific capabilities for spatially referenced data, as well as a set of operations for processing and analysing the data.

The DEM provides a basic spatial reference system to the GIS spatial data set. Images or vector information can automatically be draped over and integrated with the DEM for more advanced analysis.

2.11 APPLICATIONS OF DEM

Geographical information technology and digital image processing has become a rapidly expanding field in recent years, with particular significance in the treatment of geo- and image information for scientific, commercial and operational applications. For most applications in these three domains, DEMs are an important, integral part. In the following sections, the various applications and areas of application of DEMs are described in more detail. These have been considered as a selection of representative activities in the domains of:

1. Scientific applications
2. Commercial applications
3. Industrial applications
4. Operational applications
5. Military applications

DEMs play a significant role in the improvement of analysis results, product development and decision-making. Thus, DEMs are an asset in a variety of both commercial and public business and management fields within telecommunications, navigation, energy, disaster management, transportation, weather forecast, remote sensing, geodesy, land cover classification, civil engineering and many more. The wide range of different applications in which DEMs will be useful reflect the overall importance of the availability of global, consistent, high-quality digital elevation models.

2.11.1 Scientific Applications

The exact information about the earth's surface is of fundamental importance in all geosciences. The topography exaggerates control over a range of earth surface processes (evaporation, water flow, mass movement, forest fires) important for the energy exchange between the physical climate system in the atmosphere and the biogeochemical cycles at the earth surface. Ecology investigates the dependencies between all life forms and their environment such as soil, water, climate and landscape. Hydrology needs knowledge about the relief to model the movement of water, glaciers and ice. Geomorphology describes the relief recognizing form-building processes. Climatology investigates fluxes of temperature, moisture, air particles all influenced by topography.

The relationship between the topography and the shape of the land surface with a variety of variables of geo-processes like evaporation, runoff and soil moisture are influencing the weather forecast and climate modelling on a local and global scale.

Another area of application is a global land cover classification. Precise mapping and classification of the earth's surface at a global scale is the most important prerequisite for large-scale modelling of geo-processes. In numerous studies, it was demonstrated that radar images are suitable for documentation and classification of natural vegetation and agricultural areas. In remote sensing, DEMs are used together with GIS to correct images or retrieve thematic information with respect to sensor geometry and local relief to produce geo-coded products. Thus, for the synergic use of different sensor systems (and GIS), digital elevation models are a prerequisite for geo-coding satellite images and correcting terrain effects in radar scenes. The scientific community employs DEMs for research on:

1. Climate impact studies
2. Water and wildlife management
3. Geological and hydrological modelling
4. Geographic information technology
5. Geomorphology and landscape analysis
6. Mapping purposes
7. Educational programmes

2.11.2 Commercial Applications

Commercial applications are more marked and business-oriented applications related to sale and distribution of DEM and DEM products. From this point of view, two main market sectors are interested in digital terrain models. One sector employs basic DEM products where data are preprocessed and geo-coded, but have no application associated with them. The other sector encompasses value-adding services, which couple DEMs with a specific application. The following overview for both sectors addresses questions like: Is there a market for digital elevation models? Do the users know the available products? Are there any competitors on the market?

Commercial providers offer DEMs or DEM-associated products, which are of interests for issues in:

1. Telecommunication
2. Air traffic routing and navigation
3. Planning and construction
4. Geological exploration
5. Hydrological and meteorological services
6. Geo-coding of remote sensing
7. Market of multimedia applications and computer games

2.11.3 Industrial Applications

For industrial applications, digital elevation models are used for the development of market-oriented product technology, improved services and to increase the economic outcome of industrial production. Such applications are found within different fields such as telecom, telematics, avionics, mining, mineral exploration, tourism and engineering.

2.11.4 Operational Applications

Geo-information is a substantial part of the need of modern society for communication and information technologies. These data are increasingly used throughout all levels in administration, management and planning. Geo-information is the basis for regional planning and its availability prerequisite for decision-making processes to locate infrastructure development and investment. Operational applications include applications where DEM is used to improve management and planning of natural resources and within areas of regional planning, environmental protection, hazard reduction, military and other security- relevant applications, insurance issues, health services, agriculture, forestry and soil conservation. Operational applications are mostly related to state and governmental services and management operations. DEMs may ultimately replace printed maps as the standard means of portraying landforms. Finally, operational users need DEMs for:

1. Generating and updating geo-information for governmental issues
2. Administering assistance in areas inflicted by disasters
3. Airline operation safety
4. Security relevant activities (risk and hazard)

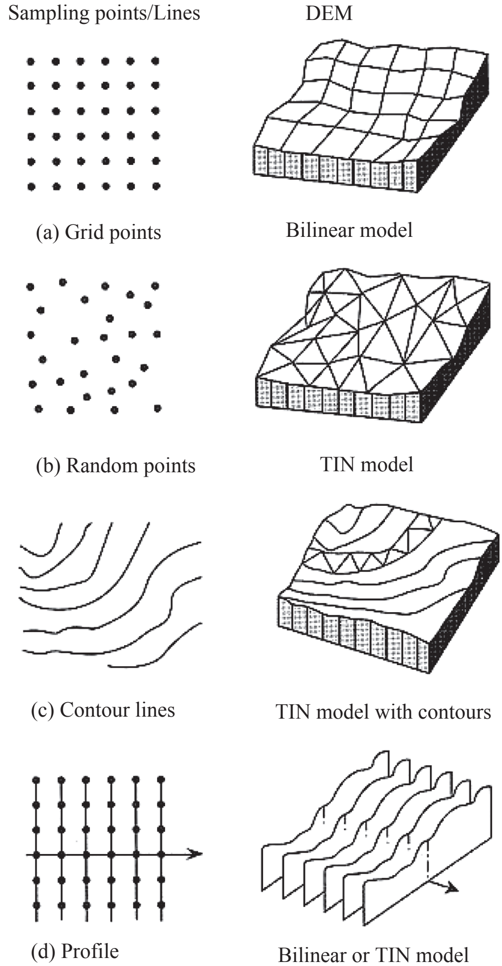


FIGURE 2.13 Different types of digital elevation models

2.12 DATA STRUCTURE FOR CONTINUOUS SURFACE MODEL

In GIS, continuous surface such as terrain surface, population density meteorological observation such as rainfall, temperature and pressure have to be modelled. For this sampling, points are observed at a discrete interval, and a surface model to present the 3D shape (such as $z = f(x, y)$) should be built to allow the interpolation of values at arbitrary points of interest. The following four types of sampling point structure are modelled into DEM for creating a surface model to present a 3D shape.

1. **Grid at regular intervals:** For girding at regular intervals, a bilinear surface with four points or a bicubic surface with sixteen points is commonly used.
2. **Random points:** In the case of random points, a TIN is commonly used.

3. **Contour lines:** In a contour line sampling point structure, interpolation based on proportional distance between adjacent contours is used. TIN is also used in this method.
4. **Profile:** Profiles are observed perpendicular to an alignment of a curve in highways. In case the alignment is a straight line, grid points will be interpolated, and if the alignment is a curve, TIN will be generated.

REVIEW QUESTIONS

1. What is a data model? Enumerate different types of GIS data.
2. Write short notes on
   i. Spatial data
   ii. Attribute data
   iii. Conceptual model of GIS
   iv. Spatial data models
   v. Vector data models
   vi. Raster data models
3. What is an image data and how is it gathered?
4. Enumerate the advantages and disadvantages of vector data.
5. Enumerate the advantages and disadvantages of raster data.
6. What is attribute data and how it is defined in GIS?
7. List out the different models used for the storage and management of attribute data. Explain each model with examples.
8. What is the peculiarity of DBMS in GIS?
9. What is a relational DBMS? Enumerate its advantages.
10. What is an object-oriented model?
11. Define DEM and TIN.
12. How is DEM represented?
13. Quote the different applications of DEM?
14. How is a continuous surface modelled in GIS?

DATA ACQUISITION 3

Chapter Outline

3.1 DATA ACQUISITION IN GEOGRAPHIC INFORMATION SYSTEM

About two-thirds of the total cost of implementing a geographic information system (GIS) involves building the GIS database. Therefore, a good GIS programme will build the GIS database at the lowest possible cost. Apart from cost, many other factors, such as data accuracy, have a significant role in the usefulness of the GIS. As data acquisition or data input of geospatial data in digital format is more expensive (about 80% of the total GIS project cost), and procedures are time consuming in a GIS, the data sources for data acquisitions should be carefully selected for specific purposes.

For GIS, data sources such as analog maps, aerial photographs, satellite imagery, Global Positioning System (GPS) surveying, digitizers, scanners and reports and publications are widely used.

3.2 ANALOG MAPS

Topographic maps with contours and other terrain features, and thematic maps with respect to defined object classes are digitized manually by digitizers, or semiautomatically by scanners. Problems of analog maps are lack of availability, inconsistency in map production time, outdated maps and inaccuracy.

3.3 AERIAL PHOTOGRAPHS

An aerial photograph is a snapshot of the earth at a particular instant of time. It contains a mass of data and it is necessary to carry out some form of interpretation to make effective use of the information photographed. Aerial photographs are useful for monitoring changes, as repeated photographs of the same area are relatively cheap. Aerial photographs may be used in GIS as a background for other data, to give those data spatial context and to aid interpretation. Hence, the user can abstract information on land use, vegetation type, moisture or heat levels or other aspects of the landscape from the photograph. Interpretation of a sequence of photographs allows in ascertaining events such as major floods, which cause changes to the landscape. The six characteristics of aerial photographs, which make them of immense value as a data source for GIS, are:

1. Low cost compared with remote sensed satellite images
2. Wide availability
3. Three-dimensional (3D) perspective
4. Time-freezing abilities
5. High spectral and spatial resolution
6. Wide area views

Aerial photographs can also be used to obtain data that is not available from other secondary sources, such as the location and extent of new housing estates and the extent of forest fires.

Aerial photographs do not provide spatially referenced data. Spatial referencing has to be added to features on the image by reference to other sources such as paper maps. Several different types of aerial photographs are available. They are:

1. Black and white photographs
2. Colour photographs
3. Colour infrared photos: Colour infrared film is often called *false-colour* film. Objects that are normally red appear green, green objects (except vegetation) appear blue, and infrared objects, appear red, which normally do not appear in the case of ordinary photographs.

The primary use of colour infrared photography is vegetation studies. This is because healthy green vegetation is a very strong reflector of infrared radiation, and appears bright red on colour infrared photographs. In aerial photography, vertical and oblique aerial photographs are available; of which the vertical photographs are taken directly below from an aeroplane, and oblique ones are taken at an angle. Oblique photographs generally cover larger areas and are cheaper than vertical photographs. Vertical photographs are widely used for GIS applications.

In aerial photography, factors such as the scale of the photograph and the factors affecting the interpretation of the data should be considered. The date of the photograph may be important to ensure that the data taken from it are contemporaneous with the rest of the data in the GIS. Aerial photographs represent a versatile, relatively inexpensive and detailed data source for many GIS applications. At a larger scale, photographs can be used to provide data on drainage or vegetation conditions within individual fields or parcels that could not be obtained from conventional photographic maps. Analytical or digital photogrammetry is rather expensive, but is the best method for updating.

3.4 SATELLITE IMAGERY

Satellite images are collected by sensors on board a satellite and then relayed to earth as a series of electronic signals, which are processed by computers to produce an image. These data can be processed in a variety of ways, each giving a different digital version of the image. Satellite imagery can also be used as a raster backdrop to vector GIS data. Satellite images have supported numerous GIS applications, including environmental impact analysis, site evaluation for large facilities, highway planning, the development and monitoring of environmental baselines, emergency disaster response, agriculture and forestry. Satellite images are useful for urban planning and management, where they are used to detect areas of change, monitor traffic conditions, measure water levels in reservoirs, building heights and many other applications.

There are large numbers of satellites orbiting the earth continuously, collecting data and returning them to ground stations all over the world. Some satellites are geostationary (e.g., Meteosat) and others orbit the earth to provide full coverage over a period of a few days (e.g., Landsat, SPOT). Landsat offers repeat coverage of any area on a 16-day cycle. Sensors on board these satellites detect radiation from the earth for different parts of the electromagnetic spectrum. The multispectral scanner (MSS) on board Landsat simultaneously detects radiation in four different wave bands—near infrared, red, green and blue. After processing, the images can be used to detect features more readily apparent to the naked eye, such as subtle changes in moisture content across a field, sediment dispersal in a lake or heat escaping from roofs in urban areas.

Scanned images are stored as a collection of pixels, which have a value representing the amount of radiation received by the sensor from that portion of the earth's surface. The size of the pixel gives a measure of the resolution of the image. The smaller the pixels, the higher is the resolution. (The Landsat thematic mapper collects data for pixels of size 30 m 30 m).

In GIS, satellite images are helpful for agriculture and forestry, providing information for crop and forest identification and inventory, growth and health monitoring and even measuring height of trees. Satellite images can be used to create image maps, which combine the raster satellite image with vector line work and text that delineate special features such as boundaries, roads and transmission lines. The four important aspects of satellite imagery are spatial resolution, spectral resolution, temporal resolution and extent. Spatial resolution refers to the smallest spatial element that can be sensed or resolved by the satellite. These elements are commonly referred to as pixels. Imagery from satellites has a resolution of 10–80 m. Modern, new generation satellites provide a resolution of 1–3 m. The high resolution in the order of 1-m satellite imagery is useful for GIS applications such as acquisition of more detailed aerial photography over more intensive areas.

Digital sensors provided in the satellites divide the electromagnetic spectrum into various bands. The visible portion of the electromagnetic spectrum is the portion humans are capable of seeing. But the photographic film is capable of capturing the visible and near-infrared portions of this spectrum. Satellite data from the near and mid-infrared portions of the electromagnetic spectrum allows GIS users to distinguish things such as vegetation types.

Temporal resolution defines when and how often an image of a particular area is captured. The Landsat satellite can view the same area once every 16 days, and SPOT satellite; on the other hand, revisits the same area every 3 days. The most modern satellite systems can provide images of an area once in 2–3 days, thereby increasing temporal resolution of satellite imagery. Extent is the amount of area captured by one satellite image. The coverage of image extends over a thousand to three million square kilometres.

For GIS, remotely sensed data offers many advantages. The images are always available in digital form, providing an easy transfer of the data to a computer system. But some processing is usually necessary to ensure integration with other data. Processing may be necessary to reduce data volumes, adjustment of resolution, change of pixel shape or to alter the projection of the data. The other advantage is the opportunity to process the image or the use of different wavebands for the collection of data to highlight features of interest (e.g., water and vegetation). The repeated coverage of the earth is a further advantage, allowing the monitoring of change at regular intervals. In adition, a small scale of images provides data useful for regional studies, and applications have included mapping remote areas, geological surveys and land use monitoring. Advantages of remotely sensed data for GIS applications in the area of natural resource management are:

1. Low cost relative to other data sources
2. Gather information about inaccessible areas
3. Accuracy
4. Completeness of data
5. Uniform standards across an area of interest
6. Saves time and efforts as information about large areas can be gathered quickly
7. Remote sensing data are extremely useful for different disciplines such as geology, fisheries, agriculture and land use

Satellite images or data are available for land use classification, digital elevation model (DEM), updating highway networks. But the image map scale would be around 1:50,000–1:100,000. High-resolution satellite image with ground resolution of 1–3 m will produce 1:25,000 topo-maps in the near future.

3.5 GROUND SURVEY

There are several methods of collecting raw data in the field for direct input into the GIS. These are most often used when the required data do not exist in any other readily available format such as a map or satellite image. Traditional manual surveying techniques using chains, plane tables, levels and theodolites are example of direct field measurement. For the above methods, the data are collected and are recorded manually. Modern digital instruments have replaced these manual methods, and data collected are stored in digital formats that are ready for direct input into GIS. Topographic surveying using Total Station (a surveying equipment) together with GPS will gather more details with much more accuracy. It is very accurate, but too expensive to cover wide areas.

3.6 GLOBAL POSITIONING SYSTEM

A relatively new technique for field data collection, which has found particular favour with GIS users, is the use of the satellite navigation system of GPS, which comprises a series of 24 satellites (owned and operated by US Department of Defense), which orbit the earth. These satellites send location information back to earth. Commercially, available receivers can capture that data from a minimum of three satellites at any one time, to give the GPS receiver operator a coordinate location. These are portable backpack or hand-held devices that use signals from GPS satellites to work out the exact location of the user on the earth's surface in terms of x, y, z coordinates using trigonometry. Position fixes are obtained quickly and accurately, literally at the push of a button. The accuracy obtainable from civilian GPS receivers ranges from 10 cm to 30 cm depending on type of GPS receivers used, method of observation and time of observation. Originally designed for real time navigation purposes, most GPS receivers will store collected coordinates and associated attribute information in their internal memory, so they can be downloaded directly into a GIS database. The ability to walk or drive around collecting coordinate information at sample points in this manner has obvious appeal for those involved in field data collection for GIS projects.

3.7 REPORTS AND PUBLICATIONS

Social economic data are usually listed in the reports of statistics and census with respect to administration units. Census and survey are a collection of related information. They may be spatial in character if each item in a collection has a spatial reference, which allows its location on the surface of the earth to be identified. Examples are population census, employment data, agricultural census data and marketing data.

Population census data normally have some element of spatial referencing. Knowing how many people are there in the given country may be useful in its own right, but details of where the population are will be of additional interest. Most population census use an hierarchical series of spatial units to publish data. This is the area over which one census enumerator delivers and collects census forms. Enumeration districts aggregate into wards, and wards into local government districts. Census data are only one example of the myriad of spatial data, which are collected by the use of survey techniques.

3.8 DIGITIZERS (FOR VECTOR DATA INPUT)

Digitizing is the process of tracing maps into computer format. Most of the vector GIS data are collected by this method. Digitization is a simplification process that converts all spatial data to either a point (e.g., a well), a line (e.g., a stream), a polygon formed by a closed, complex line (e.g., a lake) or a grid cell. Digitization reduces all spatial entities to these simple forms, because they are easy to store in the computer. A GIS database cannot readily recognize features or entities as human map users do. For example, we cannot enter the entity ‘lake’ into a GIS. Rather, we enter the spatial data coordinates for the lake’s shoreline as a polygon. Later, the attributes of the lake will be entered into the GIS database and will be associated with the polygon.

When using a digitizing table, the paper map is carefully taped down on the table’s surface. A grid of fine wires is embedded in this surface. This grid senses the position of the cross-hair on a hand-held cursor. When the cursor button is depressed, the system records a point at that location in the GIS database. The operator also identifies the type of feature being digitized, or its attributes. In this way, the map features can be traced into the system. The process more commonly used today is to first digitize the entire paper map using a scanner, such as a drum scanner.

Tablet digitizers with a free cursor connected with a personal computer are the most common devices for digitizing spatial features with the planimetric coordinates from analog maps. The analog map is placed on the surface of the digitizing tablet as shown in Figs 3.1 and 3.2. The size of the digitizer usually ranges from A3 to A0 size.

For the digitization of map features, the user completes the compilation phase by relating all spatial features to their respective attributes, and by cleaning up and correcting errors introduced as a result of the data conversion process. The end result of compilation is a set of digital files, each accurately representing all of the spatial and attribute data of interest contained on the original map

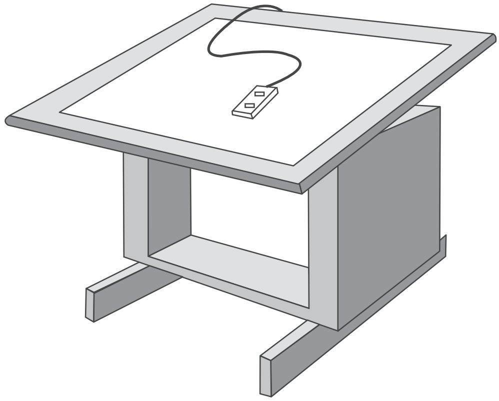


FIGURE 3.1 A tablet digitizer can convert paper maps into numerical digits that can be stored in the computer. The digitizing simplifies map data into sets of points, lines and cells which are stored as digital map files

FIGURE 3.2 A tablet digitizer

manuscripts. These digital files contain geographic coordinates for spatial objects (points, lines, polygons and cells) that represent mapped features. Although we conceptualize the GIS as a set of registered map layers, the GIS actually stores these data at a much more primitive level.

3.8.1 The Map Digitizing Operation

The following steps describe the digitization of a map using a digitizer:

1. A map is affixed to a digitizing table
2. Control points or ticks at four corners of this map sheet are digitized by the digitizer and input to PC, together with the map coordinates of the four corners
3. Map contents are digitized according to the map layers and map code system in either point mode or stream mode at a short time interval

4. Editing of errors such as small gaps at line junctions, overshoots and duplicates are made for a clean data set without errors
5. Conversion from digitizer coordinates to map coordinates is done to store in a spatial database

3.8.2 Major Problems of Map Digitization

The following are the major problems in the digitization of a map:

1. The map will stretch or shrink day by day, which makes the newly digitized points slightly off from the previous points
2. The map itself has errors
3. Discrepancies across neighbouring map sheets will produce non-connectivity

3.8.3 Advantages of Digitized Storage

There are many reasons for using computers in cartography, such as:

1. They make existing maps more quickly and cheaply
2. They make maps for specific user needs
3. The map production is possible without skilled staff
4. It facilitates map-making and updating
5. It permits interaction between statistical analyses and mapping
6. The printed map is not the database
7. It allows experimentation with different graphical representations of the same data
8. It creates maps that are difficult to do by hand, for example, 3D maps
9. The automation fosters a review of the map-making process
10. The selection and generalization procedures are explicitly defined and consistently executed

Whether a GIS is raster-based or vector-based, spatial data from map manuscripts must be compiled and stored in a simplified form that the computer can recognize. Each of these models handles spatial and attribute data in such a way as to allow the digital processing capability of a computer to be applied to managing and storing data. Computers cannot store maps in their original form; they must be converted via a digitizing process. The digitizing and storage process are often a time-consuming and expensive part of building a GIS; the cost of initially digitizing all data may comprise a majority of the total cost of system implementation. Despite the cost, digital maps stored on a computer are in a much more dynamic form, and can be easily and rapidly processed. Digital data can also be exchanged between GIS systems, copies can easily be made and distributed to multiple users, and data can be shared using telecommunication options available for modern computer users. Thus, the end result of converting data into GIS-compatible format is a more dynamic, flexible data environment.

3.9 SCANNERS (FOR RASTER DATA INPUT)

Scanners are used to convert analog maps or photographs to digital image data in raster format. A light sensor in the scanner encodes the map as a large array of dots, much like a facsimile machine,

FIGURE 3.3 A drum scanner

which scans a letter. High-resolution scanners can capture data at resolutions as fine as 2500 dots per inch (dpi), but maps and drawings are typically scanned at 100–400 dpi. The image of the map is then processed and displayed on the computer screen. The raster map is then registered to the coordinate system of the GIS, and the map features manually traced as vectors. This is called heads-up digitizing. Vectorizing software is available to help speed the process of converting raster map features to vector format. This software automatically recognizes most of the features on the raster image, such as line work and symbols, and traces vectors over them.

Digital image data are usually integer-based with a one byte grey scale (256 grey tones from 0 to 255) for a black and white image, and a set of three grey scales of red (R), green (G) and blue (B) for a colour image.

The following four types of scanners are commonly used in GIS and remote sensing:

a. **Mechanical Scanner:** A mechanical scanner is also called a drum scanner (Fig. 3.3), because a map or an image placed on a drum is digitized mechanically with rotation of the drum and shift of the sensor. It is accurate but slow.

b. **Video Camera:** A video camera with cathode ray tube (CRT) is often used to digitize a small part of a map. Even though this type of digitizing is not very accurate, it is cheap.

c. **CCD Camera:** CCD is the acronym for charged coupling device and is the key element in most of the digital cameras. A CCD camera (called digital still camera), instead of a video camera, is also convenient to acquire digital image data. It is more stable and accurate than a video camera.

d. **CCD Scanner:** A flatbed-type or roll-feed-type scanner (Fig. 3.4) with linear CCD is now commonly used to digitize analog maps in raster format, either in mono-tone or colour mode. It is accurate but expensive.

FIGURE 3.4 A flatbed scanner

3.10 DIGITAL MAPPING BY AERIAL PHOTOGRAMMETRY

Although aerial photogrammetry is rather expensive and slow in air flight, as well as subsequent photogrammetric plotting and editing, it is still very important to input accurate and up-to-date spatial information. Aerial photogrammetry needs a series of the procedures, including aerial photography, stereo-plotting, editing and output. There are two types of aerial photogrammetry—analytical and digital.

a. **Analytical Photogrammetry:** Although computer systems are used for aerial triangulation, measuring map data, editing and output with a pen plotter, a stereo pair of analog films are set up in a stereo plotter, and the operator will manually read terrain features through a stereo photogrammetric plotter called analytical plotter.
b. **Digital Photogrammetry:** In digital photogrammetry, aerial films are converted into digital image data. DEM is automatically generated with stereo matching, using a digital photogrammetric workstation. Digital orthophoto and 3D bird's eye view using DEM will also be automatically created as by- products. Even though it is very expensive, this method is adopted for automated mapping. There is a need for further research for identifying the patterns of houses, roads, structures and other terrain features automatically, which is called image understanding.

3.11 REMOTE SENSING WITH SATELLITE IMAGERY

Satellite remote sensing is a modern technology to obtain digital image data of the terrain surface in the electromagnetic region of visible, infrared and microwave. Multi-spectral bands including visible, near-infrared and/or thermal infrared are most commonly used for the production of land use maps, soil maps, geological maps, agricultural maps and forest maps at the scale of 1:50,000–250,000. A lot of earth observation satellites, for example, Landsat, SPOT, ERS-1, JERS-1, IRS and Radarsat are available.

Synthetic Aperture Radar (SAR) is now becoming a new technology in remote sensing, because SAR can penetrate through clouds, which enables cloud-free imagery in all weather conditions. Very high-resolution satellite imagery with ground resolution of 1–3 m have become available since 1998. Such high-resolution satellite images are expected to identify individual houses in urban areas. The high-resolution satellite images are highly expected to apply to urban GIS.

3.12 RASTERIZATION

Conversion between raster and vector data is very useful in practical applications of GIS. Rasterization refers to conversion from vector to raster data. The raster format is more convenient to produce colour coded polygon maps, such as a colour coded land use map, whereas map digitizing in the vector format is easier to trace only the boundary. Rasterization is also useful to integrate GIS with remote sensing, because remote sensing images are in the raster format.

A simple algorithm for calculation of the trapezoid area can be applied to convert a vectorized polygon to a rasterized polygon with grid cells as shown in Fig. 3.5. If vertical lines are dropped to the x-axis from two adjacent vertices, a trapezoid will be formed as shown in Fig. 3.5.

The area of the trapezoid is given by *Ai* (*xi* 1 *xi*) (*yi yi* 1)/2

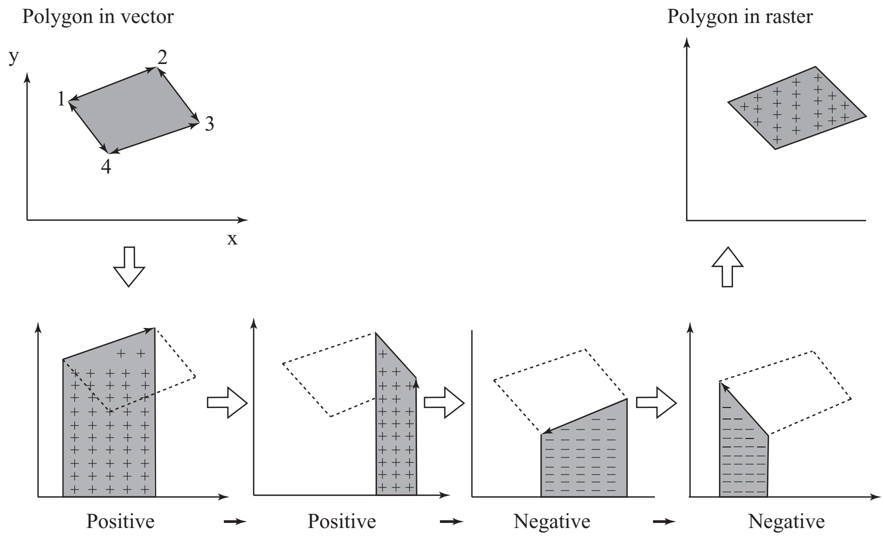


FIGURE 3.5 Conversion from vector to raster

The sum of all trapezoids will give the area of the original polygon. Using this algorithm, the grid cells in the polygon are easily identified.

3.13 VECTORIZATION

Vectorization refers to the conversion from raster to vector data, which is often called raster–vector conversion. Vectorization is not very easy as compared with rasterization, because the vector format needs a topological structure, for example, direction of line or chain, boundaries and nodes of polygons, order of series of chains that form a polygon, left and right polygons, ID of a chain and so on.

A simple algorithm of vectorization is explained in Fig. 3.6, in which the original image in raster format is converted to vector data through thinning and chain coding. This algorithm is useful to convert raster image to vector data with the coordinates, but it is not sufficient because the algorithm

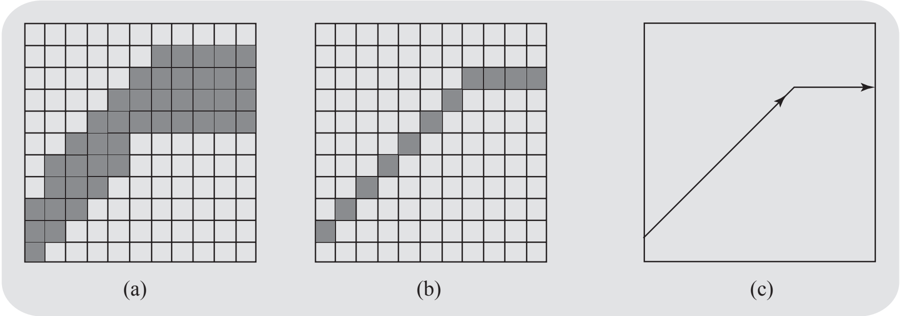


FIGURE 3.6 Vectorisation algorithm (a) Original data, (b) thin ring with chain coding and (c) vector data

will not build a topological structure. Raster vector conversion with automatic building of topology is possible if a 2 2 window is continuously moved along the boundary from a node.

Figure 3.6 shows schematically the raster vector conversion by which left and right polygons are identified. In order to automate raster vector conversion as much as possible, a clean image without noises or unnecessary marks should be scanned in the beginning.

3.14 ADVANCED TECHNOLOGIES FOR PRIMARY DATA ACQUISITION

Several advanced technologies have become available for primary data acquisition of geospatial data as well as DEM. The following advanced technologies will be useful for future GIS.

a. **Electronic Plane Surveying System:** An integrated system of total station with automated tracking function, kinematic GPS and a pen computer will replace the conventional plane surveying. Direct data acquisition in digital form, at the field site, will be very useful for large scale GIS data set, for example, in the application to cadastral surveys, utility facilities and urban structures.
b. **Mobile Mapping System:** Different sensors such as GPS, inertia navigation system (INS), more than two digital cameras and voice recorder are fixed on a vehicle to map objects in close range. For example, the centre line of highways, utility lines and railways are mapped with the help of GPS mounted on a vehicle, and it is used to determine the trajectory of the moving vehicle.
c. **Laser Scanner:** Airborne laser scanner, together with GPS and INS, will directly measure the terrain relief or DEM, with the height accuracy of 10 cm up to the altitude of 1,000 m.
d. **SAR Interferometry:** SAR interferometry is a new technology to produce DEM automatically by special interferometric processing of a pair of SAR images. Airborne and spaceborne SAR interferometry are now available if the interferometric condition meets the standard.

3.15 DIGITAL MAPPING BY AERIAL PHOTOGRAMMETRY

Although aerial photogrammetry is rather expensive and slow in air flight as well as subsequent photogrammetric plotting and editing, it is still very important to input accurate and up-to-date spatial information. Aerial photogrammetry needs a series of the procedures including aerial photography, stereo-plotting, editing and output. There are two types of aerial photogrammetry.

a. **Analytical Photogrammetry:** Although computer systems are used for aerial triangulation, measuring map data, editing and output with a pen plotter, a stereo pair of analog films are set up in a stereo plotter, and the operator will manually read terrain features through a stereo photogrammetric plotter called an analytical plotter.
b. **Digital Photogrammetry:** In digital photogrammetry, aerial films are converted into digital image data. DEM is automatically generated with stereo matching using a digital photogrammetric workstation. Digital orthophoto and 3D bird's eye view using DEM can also be automatically created as by-products. It is still very expensive but only a method for automated mapping. There is a need for further research for identifying the patterns of houses, roads, structures and other terrain features automatically, called image understanding.

3.16 DIGITAL DATA ACQUISITION

In any GIS project, the first action in the data collection phase is to check for the existence of digital data for the geographic region of interest. Before doing so, we should know the map scale, map projection system, datum, minimum mapping unit (MMU) that one wishes to recognize as a distinct entity, units of resolution and the map features one is interested in, such as roads, hydrography, topography and vegetation.

3.17 DATA PROCESSING

A GIS is of no use unless one has digital data that are in the proper format and structure. Generally, a considerable amount of data processing must be done before any spatial analyses can be performed. This processing may involve:

1. Media conversion—analog to digital
2. Geographic data conversion—vector to raster
3. Registration and coordinate system transformation
4. Tiling and edge matching

3.17.1 Media Conversion

Paper or Mylar map sheets (analog) must be converted to a digital form using either or both of the methods listed below. In either case, this process is very time consuming, labour intensive and error prone. If more than one person is doing this work, standards must be established so that the results are consistent and repeatable. As a rule-of-thumb, 80% of a GIS project's funds are consumed by this activity.

1. Manual encoding (vector or raster)
2. Automated encoding (scanner or digitizer)

Manual Encoding

This endeavour involves placing a plastic grid over a map sheet and coding a thematic value for each cell if one is interested in the raster data structure. For vector data, one could read (x, y) coordinates from the map sheet and enter them into an ASCII file, or use a GPS receiver to create a point data file.

Automated Encoding

A scanner or a digitizer can perform this process. Scanners continue to decrease in cost and possess more capabilities with time, but are generally not used because extensive editing is required to remove undesirable artefacts from the output. The other alternative is a digitizing tablet, which comes in different sizes, resolutions and accuracies. This piece of equipment utilizes an embedded grid of wires that form a Cartesian coordinate system. The resolution of such grids is from 0.01 to 0.002.

A 16-button puck is used to manually trace the various elements located on the map manuscript taped to the tablet surface. Usually, vector data files are produced. Before elements can be traced, the map must be registered. This process uses a set of four or more control points (ticks) to create a coordinate system into which the traced map elements will be placed. These control points may be known as earth coordinates or an arbitrary relative coordinate system. In any event, the locations of these control points must be identifiable on the map manuscript. One should realize that paper maps are unstable and subject to swelling and shrinkage, depending on the relative humidity and temperature changes. Note that discrepancies will exist across map borders due to different photo interpretations. There are four digitizing modes—point, line, string and stream mode. In each case, points are generated and it is a matter of how many and whether or not there is connectivity (joined by lines or arcs). A line only has nodes or endpoints, whereas a string has the same to/from nodes or starting/ending points, in addition to one or more intermediate points called key points or vertices. The number of points (density) and their placement are controlled by the user in the first three modes. The fourth mode (stream) causes points to be generated based on either a distance or time criterion, and should generally be avoided.

3.17.2 Geographic Data Conversion

Vector-to-raster conversion leads to loss of local detail (generalization), moves point data and could result in black holes (pixel dropout). It is generally not reversible. Raster-to-vector conversion involves the use of cell centroids and smoothing algorithms. Cell boundaries become vectors, which must be weeded (removal of excess points).

3.17.3 Registration/Coordinate Transformation

Through a series of control points, map manuscript features are made coincident using either a geographic/absolute or relative (digitizer) coordinate system. Once the digitized map elements are cleaned, the coordinate system used to build a digital map can be transformed into any of the 103 earth-based coordinate systems.

3.17.4 Tiling and Edge Matching

Maps having large geographic extents may need to be further divided into smaller units called tiles, so that the spatial data contained within their boundaries can be managed more efficiently. Often the geographic area of interest spreads over multiple map sheets and requires the matching of features at the edge. This operation changes features and their topology. In addition, a multi-segmented neat line may need to be dissolved. This type of editing takes considerable time and effort.

3.18 DIGITIZING ISSUES

Considerable planning must occur before any digitizing takes place. The user must address the following points:

1. What is the smallest area on the ground that is to be recognized as a distinct entity (MMU—minimum mapping unit)?
2. Which map projection system should be used?
3. What is the unit of horizontal and vertical resolution?
4. What is the map scale?
5. What map features are to be placed on what levels?

6. What map levels require linked databases?
7. What is the structure of those databases?
8. What line and text characteristics will be used for what map features?
9. How will individual map sheets be matched?

3.19 FUNCTIONS OF GIS

Another productive way to study GIS is through the original definition: a GIS is a computerized, integrated system used to compile, store, manipulate and output mapped data. This section examines each of these functions.

3.19.1 Compilation

Data compilation involves assembling all of the spatial and attribute data that are to be stored in a computerized format within the GIS. Map data with common projections, scales, and coordinate systems must be pulled together to establish the centralized GIS database. Data must also be examined for compatibility in terms of content and time of data collection. Ultimately, the data will be stored in a GIS according to the specific format requirements set by both the user and the chosen GIS software/hardware environment.

When all of the common data requirements are set by the GIS user, a base map has been established. In a GIS, a base map is a set of standard requirements for data. It provides accurate standards for geographic control, and also defines a model or template that is used to shape all data into a compatible form.

Once the data are assembled and base map parameters are set, the user must translate the map and attribute data into computer-compatible form. This conversion process, referred to as conversion or digitizing, converts paper maps into numerical digits that can be stored in the computer. Digitizing can be performed using various techniques.

Scanning is one technique. Another technique is line digitizing, which uses a tablet and a tracing stylus. Digitizing simplifies map data into sets of points, lines or cells that can be stored in the GIS computer. Each GIS software package will impose a specific form and design on the way that these sets of points, lines and cells are stored as digital map files.

3.19.2 Storage

Once the data have been digitally compiled, digital map files in the GIS are stored on magnetic or other (e.g., optical) digital media. Again, different GIS software packages will employ different storage formats. In most cases, however, data storage will be based on a generic data model that is used to convert map data into a digital form. The two most common types of data models are *raster* and *vector*. Both types are used to simplify the data shown on a map into a more basic form that can be easily and efficiently stored in the computer.

3.19.3 Manipulation

Once data are stored in a GIS, many retrieval, analysis and output options are available to users. These functions are often available in the form of toolkits. A toolkit is a set of generic functions that a GIS user can employ to manipulate and analyse geographic data. Toolkits provide processing functions such as data retrieval, measuring area and perimeter, overlaying maps, performing map algebra, and reclassifying map data. A GIS usually includes a basic set of computer programs or 'tools'. The functions provided by the toolkit vary with the software package. Figures 3.9 and 3.10 provide an overview of various tool functions.

3.19.4 Output

The final functional task of a GIS is to generate output; usually a map. GIS-generated maps are compiled from the many data sets contained in the digital GIS and match the exact user specifications. Map output may employ several colour and symbol schemes, and will be sized and scaled to meet user needs. These output products resemble hand-drafted maps and fulfil essentially the same purposes. However, it is incorrect to refer to GIS simply as a mapping system. Although GIS is able to generate high-quality map output, its ability to perform analysis and management sets it apart from the more limited computer-mapping packages.

Another form of output from a GIS is tabular or report information. Data summarized according to user-defined classes or within user-defined areas can readily be generated in a textual format. This output may also be routed to another computer application such as a statistical analysis package or a graphing package for subsequent analysis and display.

The digital data is often overlooked as a type of GIS output. Data files can be readily shared between users or systems. Because the data is in a digital format, it can easily be copied, transmitted via cable or phone line, or distributed on media such as diskettes, computer-compatible tapes or optical media such as compact disks. This greatly facilitates data sharing, and provides increased access to data and information across the entire user spectrum.

3.20 SPATIAL DATA RELATIONSHIPS

The nature of spatial data relationships is important to understand within the context of GIS. In particular, the relationship between geographic features is a complex problem, which we are far from understanding in its entirety. This is of concern, because the primary role of GIS is the manipulation and analysis of large quantities of spatial data. To date, the accepted theoretical solution is to topologically structure the spatial data.

A topologic data model best reflects the geography of the real world, and provides an effective mathematical foundation for encoding spatial relationships, providing a data model for manipulating and analysing vector-based data.

Most GIS software segregate spatial and attribute data into separate data management systems. Most frequently, the topological or raster structure is used to store the spatial data, while the relational database structure is used to store the attribute data. Data from both structures are linked together for use through unique identification numbers, for example, feature labels and database management system (DBMS) primary keys. An internal number assigned by the GIS software usually maintains this coupling of spatial features with an attribute record. A label is required so that the user can load the appropriate attribute record for a given geographic feature. Most often, the GIS software automatically creates a single attribute record once a clean topological structure is properly generated. This attribute record normally contains the internal number for the feature, the user's label identifier, the area of the feature, and the perimeter of the feature. Linear features have the length of the feature defined instead of the area.

3.21 TOPOLOGIC DATA

The topologic model is often confusing to initial users of GIS. Topology is a mathematical approach that allows one to structure data based on the principles of feature adjacency and feature connectivity. It is in fact the mathematical method used to define spatial relationships. Without a topologic data structure in a vector-based GIS, most data manipulation and analysis functions would not be practical or feasible.

The most common topological data structure is the arc/node data model. This model contains two basic entities, the arc and the node. The arc is a series of points, joined by straight-line segments, which start and end at a node. The node is an intersection point where two or more arcs meet. Nodes also occur at the end of a dangling arc, for example, an arc that does not connect to another arc such as a dead end street. Isolated nodes, not connected to arcs, represent point features. A polygon feature is composed of a closed chain of arcs.

In GIS software, the topological definition is commonly stored in a proprietary format. However, most software offerings record the topological definition in three tables. These tables are analogous to relational tables. The three tables represent the different types of features, for example, point, line and area. A fourth table containing the coordinates is also used. The node table stores information about the node and the arcs that are connected to it. The arc table contains topological information about the arcs. This includes the start and end node, and the polygon to the left or right. The polygon table defines the arcs that make up each polygon. While arc, node and polygon terminology is used by most GIS vendors, some also introduce terms such as edges and faces to define arcs and polygons. This is merely the use of different words to define topological definitions. Do not be confused by this.

As most input data does not exist in a topological data structure, topology must be built with the GIS software. Depending on the data set, this can be a central processing unit (CPU) intensive and time-consuming procedure. This building process involves the creation of the topological tables and the definition of the arc, node and polygon entities. To properly define the topology there are specific requirements with respect to graphic elements, for example, no duplicate lines and no gaps in arcs that define polygon features. These requirements are reviewed in the Data Editing section of the book.

The topological model is used because it effectively models the relationship of spatial entities. Accordingly, it is well suited for operations such as contiguity and connectivity analyses. Contiguity involves the evaluation of feature adjacency (e.g., features that touch one another) and proximity (e.g., features that are near one another). The primary advantage of the topological model is that spatial analysis can be done without using the coordinate data. Many operations can be done largely, if not entirely, by using the topological definition alone. This is a significant advantage over the CAD or spaghetti vector data structure that requires the derivation of spatial relationships from the coordinate data before analysis can be undertaken.

The major disadvantage of the topological data model is its static nature. It can be a time-consuming process to properly define the topology, depending on the size and complexity of the data set. For example, 2,000 forest stand polygons will require considerably longer time to build the topology of 2,000 municipal group boundaries. This is due to the inherent complexity of the features, for example, groups tend to be rectangular while forest stands are often long and sinuous. This can be a consideration when evaluating the topological building capabilities of GIS software. The static nature of the topological model also implies that every time some editing has occurred, for example, forest stand boundaries are changed to reflect harvesting or burns, the topology must be rebuilt. The integrity of the topological structure and the DBMS tables containing the attribute data can be a concern here.

This is often referred to as referential integrity. While topology is the mechanism to ensure integrity with spatial data, referential integrity is the concept of ensuring integrity for both linked topological data and attribute data.

3.22 COMPARISON OF ANALOG MAP VERSUS DIGITAL MAP

An analog map is a printed drawing on a piece of paper or a scanned image of a map. It is static. A digital map is a data set stored in a computer in digital form (not as a picture). It is not static, and the flexibility of digital maps is vastly greater than paper maps. Inherent in this concept is the point that data on which the map is based is available to examine or question. Analysis capabilities are much greater with digital maps, and reporting outputs are available faster and in more formats.

REVIEW QUESTIONS

1. How is data acquisition done in GIS?
2. Mention the six characteristics of aerial photographs, which make them a good data source for GIS.
3. What are the advantages of satellite imagery?
4. How are GIS data gathered using GPS?
5. What is a digitizer? How is a map digitized using a digitizer?
6. Write the steps followed in the digitization of a map using a digitizer
7. What are the major problems in map digitization?
8. List the advantages of digitized data storage.
9. What is a scanner? What are the different types of scanners used?
10. Describe briefly:
    - i. Rasterization
    - ii. Vectorization

MAPS AND MAP PROJECTIONS 4

Chapter Outline

4.1 INTRODUCTION

Geodesy is the science concerned with the study of the shape and size of the earth in the geometrical sense, and the study of certain physical phenomena, such as gravity, seeking explanation of fine irregularities in the earth's shape. The subject is intimately linked with surveying and mapping. A major part of the evidence about the shape and size of the earth is based upon surveys. The earth is a nearly spherical planet upon which the surface of irregularities created by land and sea, highlands and lowlands, mountains and valleys are superimposed. These topographical irregularities represent little more than a roughening of the surface. As the radius of the earth is approximately equal to 6371 km, and the major relief features do not rise more than 9 km above or fall more than 11 km below sea level, they are relatively less than, say, the indentation on the surface of a golf ball. For example, if the earth is drawn to scale as a circle of radius 6 cm, the variation in line thickness of the circumference which would show the entire height range from Mount Everest to the Mariana Trench at the same scale, will be less than 0.2 mm (i.e., less than the thickness of the line drawn with a microtip pen).

A globe represents a small model of the earth. It represents all the aspects of the earth such as its spherical shape, sizes of the countries and continents, direction of one object from the other and distances. For various analyses, the whole earth or parts of it have to be represented on a plane surface, such as a flat sheet of paper or cardboard. But before attempting to represent the earth on a plane surface, how much ground distance in kilometres or metres are decided, and shown by a centimetre on the plane surface. That is, a relation between the distance on the ground and the distance on the plane surface has to be established. This relation between the distance of two points on the ground and the distance of the same two points on a plane surface is called a scale. If a ground distance of 4 km is represented by 1 cm on a plane surface, the scale of the plane surface is 1 cm to 4 km or 1cm to 4,00,000 cm or simply 1:4,00,000.

A map is defined as a representation to scale of the features of the whole earth or a part of it on a plane surface. The size of a map is very small as compared with the size of the earth it represents. Therefore, it represents only important features of the earth. Natural and man-made features, and district, state and international boundaries are depicted on maps by symbols called conventional signs. Distribution of population, crops, etc. are shown on maps with special symbols such as dots, circles, colours and shading. A key explaining all the symbols used on a map will be appended to it. Parallels of latitude and meridians of longitude are also drawn on a map. These lines are drawn on all maps, except those that show distributions. In the case of distribution maps, the lines of latitude and those of longitude are marked just on the marginal lines bounding the maps. The title of the map and the scale on which it is drawn are also given to it.

A plan and a chart are also maps. A plan is a large-scale map on which every object of the ground is drawn to scale. For example, on a topographical map, the conventional signs representing roads, railways, wells and temples are not drawn to scale, whereas on a plan, they are drawn to scale. Maps used for air and marine navigation are called charts. A map showing weather conditions is called a weather chart. An air navigation chart, also called an aeronautical chart, delineates landing-grounds, hills, peaks and high towers, which may prove dangerous to aeroplanes. The sea navigation charts (e.g., the British admiralty chart, the Indian Admiralty Chart) depict coastlines, depths of the seas, lighthouses and ports. A weather chart shows weather conditions such as atmospheric pressure, wind direction and precipitation.

A map is a selective, symbolized, generalized and planimetric picture of spatial distribution of the earth's surface on a definite scale with annotations. This definition of a map leads us to some fundamental characteristics of maps, which are:

1. Every map should have a reference system.
2. Every map is made on a certain projection.
3. Every map is drawn to a particular and definite scale.
4. A map has to be selective in showing the details.
5. In each map, a certain feature or information is emphasized.
6. Every map is generalized.
7. Every map uses symbols.
8. Maps are lettered, titled and labelled.

4.2 TYPES OF MAPS

Maps are made for a variety of users. Geographers, military personnel, economists, planners, civil engineers, architects, air and marine navigators and a host of other users use maps. The earth's features are numerous, and hence it is not possible to represent all of them on one map. They are drawn on various scales to suit various requirements. Thus, many maps have only selected information on it. Attempts have been made to classify maps either on the basis of scale or on the basis of their usage. The theme of the map is the particular aspect of the world that the map attempts to show, such as roads, borders, vegetation or statistical data. Maps can be mainly classified into four categories as discussed below.

4.2.1 Cadastral Maps

Cadastral maps (planimetric maps) are used to demarcate the boundaries of fields and buildings, and for registering the ownership of landed properties.

Cadastral maps have the location of property-ownership lines. They are prepared and compiled by government agencies, and are used for revenue and tax purposes. Village maps or city plan maps can be put into this category. These maps show shorelines and boundaries, plus certain basic geographic features such as rivers, roads, administrative zones and other information (Fig. 4.1). These maps do not attempt to show the relief features in measurable form.

4.2.2 Topographic Maps

Topographic maps show the shape and elevation of terrain (Fig. 4.2). Topographic maps are used for designing gardens and parks, building roads and pipes, planning hiking routes, Examples for topographic maps are flood control maps and engineering maps. They are prepared on a large scale to show the general surface features in detail. The distinctive characteristic of a topographic map is that the shape of the earth's surface is shown by contour lines. Contours are imaginary lines that join

FIGURE 4.1 A cadastral map

FIGURE 4.2 A topographic map

points of equal elevation on the surface of the land above or below a reference surface such as mean sea level. Contours make it possible to measure the height of mountains, depths of the ocean bottom and steepness of slopes. A topographic map shows more than contours.

These maps represent features such as streets, buildings, streams and woods using symbols. These symbols are constantly refined to better relate to the features they represent, improve the appearance or readability of the map, or to reduce production cost. Consequently, within the same series, maps may have slightly different symbols for the same feature. Examples of symbols that have changed include built-up areas, roads, intermittent drainage and some type styles.

4.2.3 Thematic Maps

A thematic map shows information about a special topic, which is superimposed on a base map (Fig. 4.3). Types of thematic maps include geologic, forestry, soil, land use and historical. The information on these maps can be presented in a variety of ways. The following are some commonly used techniques.

Single Symbol All the features in a map are displayed with the same colours and symbols. This is useful when it is needed to show where a theme's feature is located.

Graduated Colour Features are displayed with the same symbol type, but the colours represent the progression of values for a data attribute specified.

Graduated Symbol Features are displayed with the same colours and symbols, representing a progression of values. This is the best way to symbolize data that expresses size or magnitude. The graduated symbol is available only for point represented and line represented data (Fig. 4.4).

Unique Value Each unique value in a theme is represented with a unique symbol. This is the most effective method for displaying categorical data such as landuse type (Fig. 4.5).

Dot Density The features of a polygon theme are displayed with a number of dots corresponding to a value. This method is good for showing how particular things are distributed throughout an area.

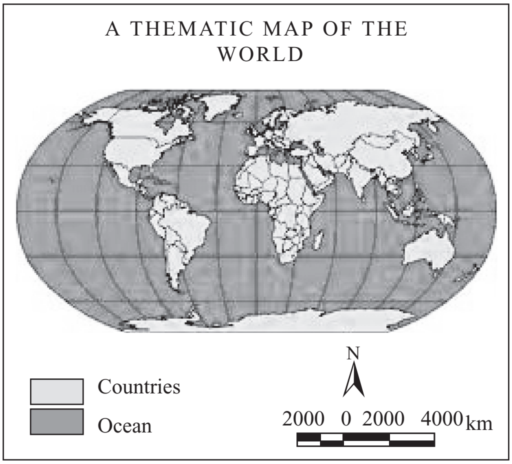


FIGURE 4.3 A thematic map of the world

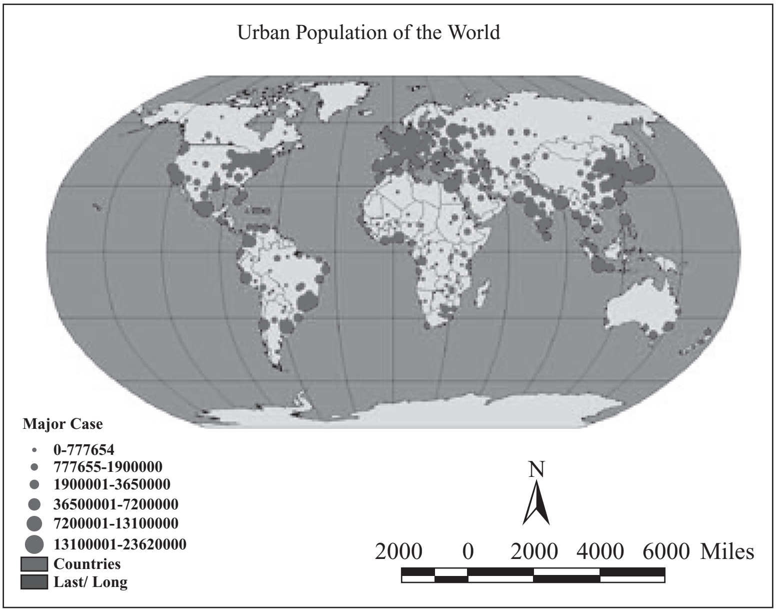


FIGURE 4.4 A thematic map with graduated symbol

FIGURE 4.5 A thematic map with unique value

For instance, a dot map depicting population will most likely have the strongest concentration of dots along rivers and near coastlines (Fig. 4.6).

Chart The features are displayed with a pie chart or column chart. The components of the chart correspond to a data variable, and the size of each part in a chart is deter mined by the value of each data attribute. This is a good method for displaying the values of many attributes such as population composition.

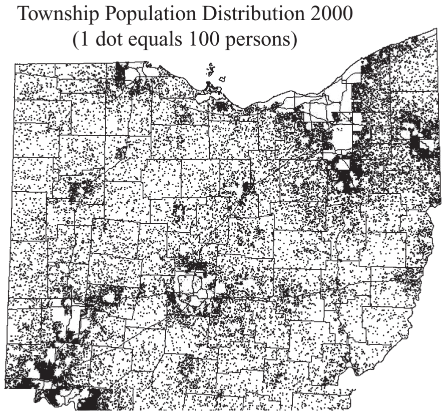


FIGURE 4.6 A thematic map with dot density

4.2.4 Remotely Sensed Images

Remote sensed images are geographic information gathered by means of a sensor (Fig. 4.7). The common remote sensing images include aerial photographs, radar images, and satellite images. These images are important sources for producing digital maps, and are useful for monitoring environmental changes and human activities. Digital maps in the form of raster images are produced as digital copies of the printed maps at various scales such as 1:50,000 or 1:2,50,000. The raster image is produced as a by-product from the computer-assisted cartographic system. These map images are most useful in various applications such as navigation systems, decision-support systems and value-added applications using the map as a reference.

Most types of maps have a number of things in common. In a map, the typographical information includes titles, legends, and names. This information is crucial in understanding the maps because it tells what the map is showing, and helps to understand what information one can get from the map. Every map should also include a map scale and orientation.

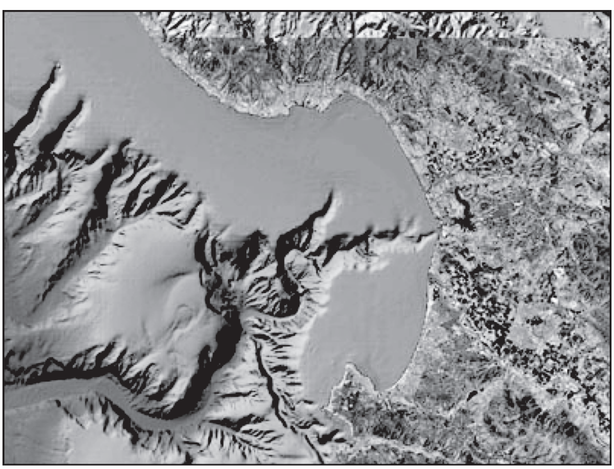


FIGURE 4.7 Remote sensed images

4.3 SCALE OF A MAP

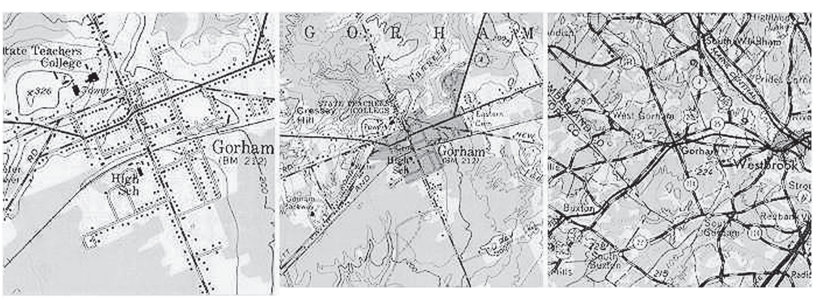
A map scale provides the map viewer a ratio between the size of the features on the map and the size of the features on the ground. The orientation of a map shows the way that the map is aligned relative to the earth's surface. Maps are always oriented with the north (normally true north) at the top. One should know the differences between Northing, Southing, Easting and Westing, and be able to locate these orientations on a map. On the basis of scale, maps are categorized as small-scale and largescale maps. A small-scale map shows a large area, and has very little detail. On the other hand, a large-scale map shows a small area, and includes a lot of detail.

The relation between the map distance and the ground distance is called a scale, and is defined as the proportion between a distance on the map and the corresponding distance on the ground. To make a map useful, cartographers establish and indicate a consistent relationship between size on the map and size in real life. This relationship is the map scale. It is also an expression of how much area represented has been reduced on the map. Map scale is important for understanding maps both in paper and computer form, so it will help one to understand the types and uses of scales.

The scale is usually located in the legend box of a map, which explains the symbols and provides other important information about the map. For instance, a map may use one cm to represent what is actually 10,000 cm on earth. Cartographers express scale as a mathematical ratio or fraction. A scale of 1:10,000 would be read as one in ten thousand. It is not easy to visualize a value such as 10,000 cm, so maps usually not only show the scale ratio, but also convert the ratio to units of measurement. For example, a 1:10,000 map can also be expressed as 1 cm = 100 m.

Map scale is the basis of drawing maps (Fig. 4.8). It enables us to represent the features of the ground on a map of convenient size. The distances between various objects on the map correspond to those on the ground. The words 'large', 'small' and 'medium' are frequently placed before scales to qualify them. Thus, a scale can be termed as a large scale, medium scale or a small scale. Every country has its own way of categorizing the scales, and no definite system is followed for differentiating one category from the other. With the adoption of the metric system of measurements, the 'Survey of India' has started publishing maps on the scales of 1:25,000, 1:50,000, 1:2,50,000, 1:10,00,000 etc. It may be helpful to categorize the scale as below:

Large scale = 1:50,000 and larger
Medium scale = 1:50,000 to 1:2,50,000
Small scale = smaller than 1:2,50,000



Scale 1:24000 Scale 1:62500 Scale 1:250,000

FIGURE 4.8 Examples of map scales

4.4 REPRESENTING THE SCALE OF A MAP

The scale of a map can be represented by using a bar scale, a representative fraction, or a text scale.

Bar Scale A bar scale shows a graphic representation, where you can actually measure the distance on the map, and then compare it to the bar scale (Fig. 4.9).

Representative Fraction A representative fraction gives the map viewer a number scale represented by a fraction (Fig. 4.10). An example of a representative fraction would be 1:24,000.

Textual Scale A text scale actually describes the scale in words. For example, 1,00,000 cm equals one kilometre, and a text scale on the map for a representative fraction of 1:1,00,000 would look like, 1 cm on the map equals to 1 km on the ground, or 1 cm equals 1 km.

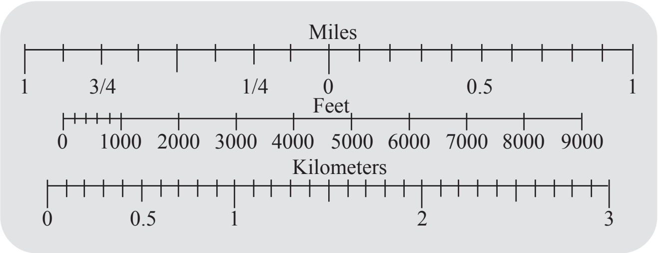


FIGURE 4.9 A bar scale

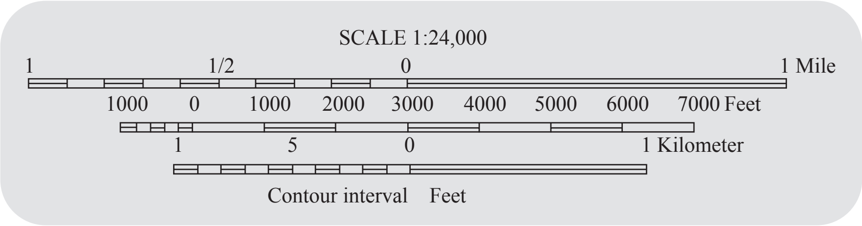


FIGURE 4.10 Representative fraction

4.5 MAP SYMBOLS

Maps depict a lot of information in a small amount of space. They use symbols and the symbols allow a map to show lots of information in a small space. By using symbols, maps can have more content and less clutter. A map should have enough symbols, so that the users can observe the interrelationships among the features. The type of symbols to include in a map depends on the purpose and audience of the map.

As symbols, maps use a variety of types of points, lines, and areas:

1. A point, for a feature that has a specific location, but is too small to show as an area. It often looks like a simplified version of the feature, as when the symbol for airport is a small aeroplane. In some schemes, colour is important.
2. A line, for a feature that has length, but is too narrow to show as an area, typically roads and rivers. A line may be thick or thin, solid or broken, in various colours.

3. An area, for a feature that is too large to show as a point or a line. An area may have different patterns or colours.

Map scale influences what types of symbols are used. On a state map, a park may be just a point. But a local map may show specific aspects of the park as areas, such as lakes, playing fields, and parking. To find out what a map symbol stands for, one has to look at the legend. The legend is a list of symbols used on a map, indicated by a sample symbol with an explanation showing what feature each represents. Each symbol should appear in the legend exactly as it looks on the map. In addition to symbols, a map may also have labels, so that the users know what specific feature is depicted. For instance, a label may give the name of a river or road.

4.6 USES OF MAPS

A map defines the following:

Position (Location) A map gives the location or position of places or features. The positions are usually given by the coordinates of the place, either in the Cartesian coordinates (x, y) in meters, or in geographical coordinates (latitude and longitude) in degrees, minutes and seconds.

Spatial Relationships A map gives the spatial relationship between features. For example, which district is the neighbour of another district? On which side of the road is the river? Is there a dam on the farm? Where is the nearest railway station?

Distance, Direction, Area One can determine a lot of information from a map such as distances, directions and areas. In determining distances and areas, the scale of the map has to be taken into consideration. Directions are based on true north, but when using a magnetic compass, it must be remembered that the compass needle points to the magnetic north, which is different from true north. The difference between magnetic north and true north is called the magnetic declination. If using the magnetic north, it should be specified in the map. If possible, the annual variation in the magnetic declination of the area of interest also have to be specified.

Different Types of Maps A map is a representation of the real world on a limited size of paper, and is restricted in what can be shown. The cartographer decides what to show and what to leave out. The cartographer is guided by the main purpose of the map, such as a road map, a topographical map or a thematic map. A road map emphasizes the roads and towns giving more importance to the roads, while a topographic map, also called a general map, shows as much of the landscape, elevations, roads, towns as possible. A thematic map is designed to depict a specific theme such as the population of various magisterial districts, the agricultural area in different districts, or annual rainfall.

4.7 CHARACTERISTICS OF MAPS

The characteristics of maps are:

1. All maps are concerned with two primary elements—location and attributes.
2. All maps are reductions of the reality—scale.
3. All maps are transformations of space—map projections and coordinate systems.

4. All maps are abstractions of reality—generalizations and its components.
5. All maps use signs and symbolism—cartographic symbolization.
6. All maps are static and dependent upon the stability of the data.

4.8 MAP PROJECTION

Map projection is the method by which the curved surface of the earth, or a part of it, is represented on a flat surface on a certain scale by making parallels and meridians. In other words, transformation geographical coordinates to a plane grid coordinate system is referred to as map projection.

Globes and maps of the world generally show lines of latitude and longitude, also known as parallels and meridians, that cross each other on the surface of the earth. This is called a graticule. Thus map projection may be defined as the preparation of the graticule on a flat surface. Mathematical formulae are used to construct a graticule on a map, corresponding to the intersecting graticule and meridian on the earth.

Due to the spherical shape of the earth and the plane surface on which this shape has to be represented, cartographers have to devise complex, graphical, geometrical and mathematical methods of transforming the earth, and the resulting transformations are collectively known as map projections. The actual meaning of the word projection is to project an image on something. In a map projection, the network of latitudes and longitudes of a globe are projected on a plane surface. Map projection is defined as the transformation of the spherical network of latitudes and longitudes on a plane surface, irrespective of the method of transformation. A map projection is a way of representing the three-dimensional surface of the earth on a flat piece of paper.

A is a method of showing the earth's surface on a map with least distortions or changes of shape, area, distance, and direction of feature. Globes represent the shape, area, distance and direction of

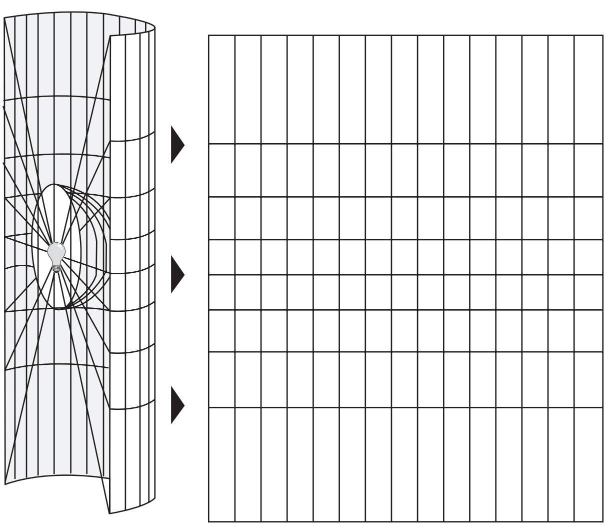


FIGURE 4.11 The graticule of a geographic coordinate system is projected onto a cylindrical projection surface

features on the surface of the earth. Map projections are designed to show shape, area, distance or direction on a flat surface, but necessarily have to distort some of these features when correctly representing others. This is called distortion.

4.9 AN IDEAL MAP PROJECTION

An ideal map projection represents the meridians and the parallels in the same way on a globe. The global network of meridians and the parallels has the following characteristics:

1. The equator divides the globe into two halves, the northern hemisphere and the southern hemisphere.
2. The equator is the only great circle line of latitude. All other lines of latitude are shorter than the equator and are not great circles.
3. Each meridian is one half a great circle in length. It is the shortest line between the two poles.
4. All meridians converge at the north and south polar points.
5. The parallels and the meridians intersect at right angles.

Due to the above characteristics, a globe possesses the following properties:

1. It represents the features of the earth's surface in their true shape. It therefore has the property of conformality.
2. All the features represented on it maintain their proportional sizes; it therefore has the property of equivalence or equal area.
3. Geodetic lines on the earth, which give the shortest distance between any two points, appear as shortest lines on the globe. In other words, the distances are correctly maintained.
4. The longitudes and latitudes are so arranged that it is convenient to locate a point.

In other words, the network of graticules has the property of simplicity.

4.10 PROJECTION CHARACTERISTICS

An issue with all projections is that they have to distort or change in some way the representation of shape, area, direction, or distance of land features or graticule in one way or another, in order to flatten the globe to a piece of paper.

Some projections are good at keeping the true shape of land features on a flat map, but they make the representation of distance on a map longer or shorter than it is in real life, or the direction of features may not be correct, or the representation of the area of a country or ocean, may not be representative of the actual area of that feature in real life.

4.10.1 Features of Various Projections

Shape Map projections that represent the true or correct shape of the earth's features are called conformal projections. Usually, these map projections can only show small areas of the earth's surface at one time.

Area Other projections, called 'equal-area' projections, are drawn so that they illustrate the same representation of the area of a feature. This would be important if a map user was, for example, measuring the amount of old growth forest that still grew in a particular country, or how much land in a country was used for growing crops. An example of an equal-area map is called a sinusoidal.

Distance Equidistant maps, or maps that keep the correct representation of actual distance on the earth are very important to travellers. An example of an equidistant map projection is a 'polar (meaning at the north or south pole) azimuthal equidistant' projection.

Direction Another feature of map projection is direction. The shortest distance between two points is a straight line, except on a globe. A straight line on a globe is actually curved due to the spherical shape of the globe, and is called a Great Circle. On all map projections, except Mercator projections, lines of constant direction, also known as rhumb lines or loxodromes, are curved. Navigators, as one might imagine, find it much easier to just draw a straight line with their ruler between two points. The Mercator projection, which is a cylindrical projection, is the only projection where rhumb lines or loxodromes are illustrated as a straight line. As Mercator projections preserve straight-line direction, they have to give up representing accurate area and shape of globe features at the north and south pole.

Map projections are very important when more than one data source is used. Different types of projections are used for specific areas of the earth to minimize the distortions for that part of the globe. Map projections are very efficient for managing spatial data for small areas, because only a known amount of distortion will be introduced through the map projection transformation.

4.11 THE STANDARD PARALLEL AND STANDARD MERIDIAN

A plane may be tangent at any point on a globe. A polar aspect refers to tangency at the pole, an equatorial aspect at the equator and an oblique aspect anywhere between the equator and the pole. The *standard line* in a map projection refers to the line of tangency between the projection surface and the reference globe. The standard line is called a standard parallel, if it follows a parallel, and a standards meridian if it follows a meridian. There is no projection distortion along the standard line, because it has the same scale as that of the reference globe. The scale factor is defined as the ratio of the local scale to the scale of the reference globe or the principal scale. A scale factor of one means that the standard line has the same scale as the scale of the reference globe.

4.12 DIFFERENT MAP PROJECTIONS

There are many types of map projections. There are different criteria considered for the classification, they are map projections according to the developable surface, map projections according to the method of deviation (source of light) and the map projection according to the global properties.

4.12.1 Map Projection According to the Development of Surface

As maps are flat, some of the simplest projections are made onto geometric shapes that can be flattened without stretching their surfaces. These are called developable surfaces. Some common examples are cones, cylinders, and planes. A map projection systematically projects locations from the surface of a spheroid to representative positions on a flat surface using mathematical algorithms. The first step in projecting from one surface to another is creating one or more points of contact. Each contact is called a point (or line) of tangency. A planar projection is tangential to the globe at one point. Tangential cones and cylinders touch the globe along a line. If the projection surface intersects the globe instead of merely touching its surface, the resulting projection is a secant rather than a tangent case. Whether the contact is tangent or secant, the contact points or lines are significant because they define locations of zero distortion. Lines of true scale are often referred to as *standard line*. In general, distortion increases with the distance from the point of contact. Many common map projections are classified according to the projection surface used. The map projections according to the projection surface are classified as:

1. Conic projection.
2. Cylindrical projection.
3. Planar or azimuthal projection.

4.12.1.1 The Conical Projection Surface

The first type of projection according to the development of surface is the conical projection. Consider a piece of light cardboard rolled up into a cone. It is then placed on a transparent globe. Now the cone and globe will somewhat resemble an ice-cream cone. Wherever the cardboard touches the globe, the most accurate representation would occur. Wherever the cardboard is farther away and not touching the globe, much more distortion will occur.

The simplest conic projection is tangent to the globe along a line of latitude. This line is called the standard parallel. The meridians are projected onto the conical surface, meeting at the apex, or point, of the cone. Parallel lines of latitude are projected onto the cone as rings (see Fig. 4.12). The cone is then cut along any meridian to produce the final conic projection, which has straight converging lines for meridians and concentric circular arcs for parallels. The meridian opposite the cut line becomes the central meridian. In general, the further you move from the standard parallel, the more the distortion is. Thus, cutting off the top of the cone produces a more accurate projection. Conic projections are used for mid-latitude zones that have an east–west orientation.

More complex conic projections contact the global surface at two locations. These projections are called secant projections and are defined by two standard parallels (see Fig. 4.13). It is also possible to define a secant projection by one standard parallel and a scale factor. The distortion pattern for secant projections is different between the standard parallels than beyond them. Generally, a secant projection has less overall distortion than a tangent projection. In still more complex conic projections,

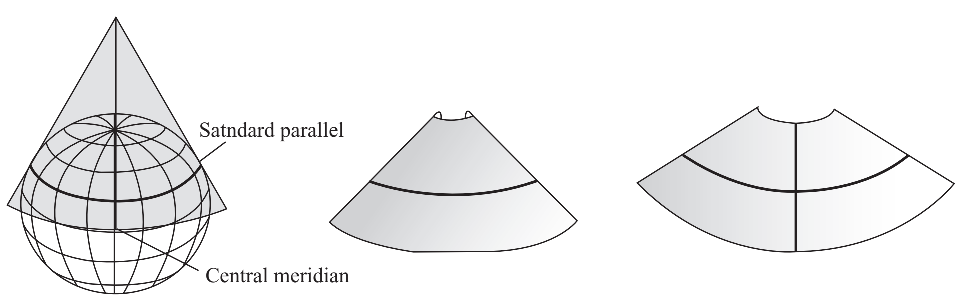


FIGURE 4.12 The conical projection and the developed portion of the plane surface in a conical projection

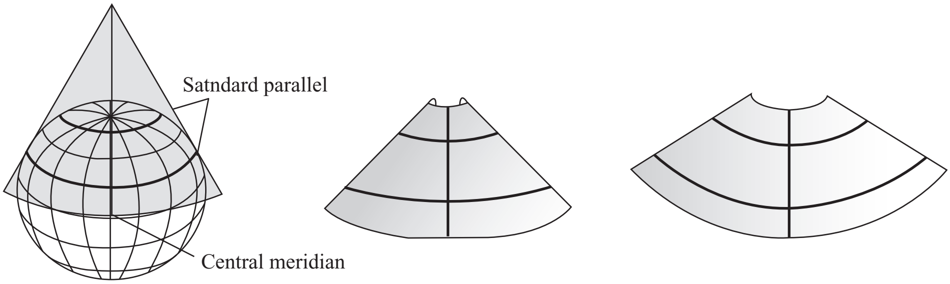


FIGURE 4.13 The secant conical projection and the developed portion of the plane surface in a conical projection

the axis of the cone does not line up with the polar axis of the globe. These types of projections are called oblique.

The representation of geographic features depends on the spacing of the parallels. When equally spaced, the projection is equidistant north–south, but neither conformal nor equal area. An example of this type of projection is the equidistant conic projection. For small areas, the overall distortion is minimal. On the Lambert Conic Conformal projection, the central parallels are spaced more closely than the parallels near the border, and small geographic shapes are maintained for both small-scale and large-scale maps.

4.12.1.2 The Cylindrical Projection Surface

The second type of projection according to the development of surface is the cylindrical projection. Imagine a transparent globe with a light source that casts the shadow of the outlines of its graticule and continents. Imagine rolling a piece of light cardboard into a cylinder around the globe and the shadow cast on to the piece of light cardboard could be captured. The resulting image, when rolled flat, would be a cylindrical projection (see Fig. 4.14). Like conic projections, cylindrical projections can also have tangent or secant cases. The Mercator projection is one of the most common cylindrical projections, and the equator is usually its line of tangency. Meridians are geometrically projected onto the cylindrical surface, and parallels are mathematically projected. This produces graticular angles of 90 degrees. The cylinder is 'cut' along any meridian to produce the final cylindrical projection. The meridians are equally spaced, while the spacing between parallel lines of latitude increases toward the poles. This projection is conformal and displays true direction along straight lines.

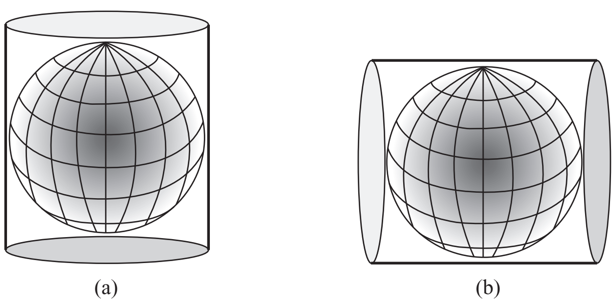


FIGURE 4.14 (a) Normal and (b) transverse cylindrical projections

On a Mercator projection, rhumb lines (lines of constant bearing) are straight lines, but most great circles are not. For more complex cylindrical projections the cylinder is rotated, thus changing the tangent or secant lines. Transverse cylindrical projections such as the transverse Mercator use a meridian as the tangential contact or lines parallel to meridians as lines of secancy.

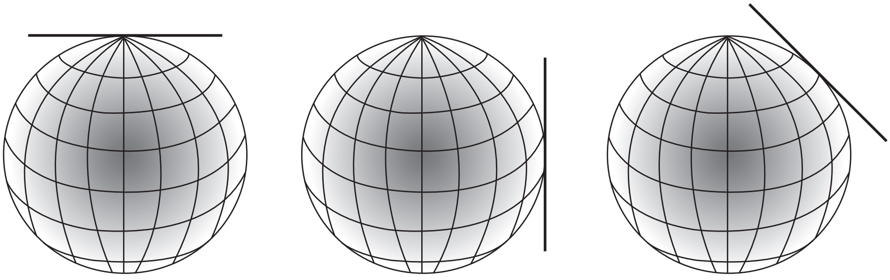
The standard lines then run north–south, along which the scale is true. Oblique cylinders are rotated around a great circle line located anywhere between the equator and the meridians. In these more complex projections, most meridians and lines of latitude are no longer straight. A Mercator projection does not show the north or south pole.

In all cylindrical projections, the line of tangency or lines of secancy have no distortion, and thus are lines of equidistance. Other geographical properties vary according to the specific projection.

4.12.1.3 The Azimuthal Projection or Plane Projection Surface

Planar projections project map data onto a flat surface touching the globe. A planar projection is also known as an azimuthal projection or a zenithal projection. This type of projection is usually tangent to the globe at one point, but may be secant also. The point of contact may be the north pole, the south pole, a point on the equator, or any point in between. This point specifies the aspect and is the focus of the projection. The focus is identified by a central longitude and central latitude. Possible aspects are polar, equatorial, and oblique. The shadow that was cast on the cardboard, the way the shadow lands on the cardboard would be called an azimuthal or plane projection (see Fig. 4.15).

Polar aspects are the simplest form. Parallels of latitude are concentric circles centred on the pole, and meridians are straight lines that intersect with their true angles of orientation at the pole. In other aspects, planar projections will have graticular angles of 90 degrees at the focus. Directions from the focus are accurate. Great circles passing through the focus are represented by straight lines; thus, the shortest distance from the centre to any other point on the map is a straight line. Patterns of area and shape distortion are circular about the focus. For this reason, azimuthal projections accommodate circular regions better than rectangular regions. Planar projections are used more often to map polar regions. Some planar projections view surface data from a specific point in space. The point of view determines how the spherical data is projected onto the flat surface. The perspective from which all locations are viewed varies between the different azimuthal projections. The perspective point may be the centre of the earth, a surface point directly opposite from the focus, or a point external to the globe, as if seen from a satellite or another planet.



(a) (b) (c)

FIGURE 4.15 Planar projections. (a) Polar, (b) equatorial and (c) oblique

4.12.2 Map Projection According to the Method of Deviation (Source of Light)

According to the method of deviation, map projections are classified as:

1. Perspective projection
2. Non-perspective projection
3. Mathematical or conventional projection

Perspective projections are those, which can be derived by projecting the image of the network of meridians and parallels of a globe on any developable surface. The lines forming the network are straightened or curved, and the spacing between the parallels and the meridians are reduced or enlarged to make a perspective projection equivalent (orthomorphic or azimuthal). Such projections are called non-perspective projections. Mathematical or conventional projections are those, which are devised by mathematical computations and formula, and have little relation to a projected image.

4.12.3 Map Projection According to the Global Properties

According to the global properties, map projections are classified as:

1. Homographic or equivalent or equal area projection
2. Orthomorphic or conformal projection
3. Azimuthal or correct bearings projection

The projections under global criteria can be developable as well as non-developable. They can be conical, cylindrical, zenithal or mathematical.

4.13 CONSTRUCTION OF MAP PROJECTION

For the construction of a map projection, one must know the scale on which the projection has to be drawn. The earth measurements have to reduce with respect to the true earth measurements on a suitable scale for the construction of map projection. The equatorial circumference of the earth is about 40,000 km and mean radius 6,371 km. The radius of the earth when converted into metres would amount to about 40,000,000 m. The length of the equator on the globe will be equal to the circumference. From the equator, we have to build the network of latitudes and longitudes. One minute of latitude is equal to about 1,848 m.

4.14 CYLINDRICAL MAP PROJECTION

In these projections, the parallels and the meridians of a globe are transferred to a cylinder, which is a developable surface. Take a cylinder, the inside diameter of which is equal to the diameter of the globe, and place that globe inside the cylinder in such a way that its equator touches the cylinder. Then bring the parallels and the meridians of the globe to the inner surface of the cylinder by using certain methods. On unrolling the cylinder, we get a flat rectangular surface. The length of the equator on the cylinder is equal to the length of the equator on the globe. Therefore, these projections are quite suitable for showing equatorial regions.

4.14.1 Cylindrical Map Projection Characteristics

The following are the characteristics of a cylindrical map projection:

1. Lines of latitude and longitude are parallel intersecting at 90 degrees.
2. Meridians are equidistant.
3. Forms a rectangular map.
4. Scale along the equator or standard parallels is true.
5. It has a simple construction.
6. It can have the properties of equidistance, conformality or equal area.

4.14.2 Types of Cylindrical Projection

There are various types of cylindrical projections, and the important ones are discussed below.

4.14.2.1 Mercator Projection

The Mercator projection is not normally used for topographic maps. It is a conformal projection. However, its description serves as a basis for understanding the transverse Mercator projection. The Mercator projection can be visualized as a spheroid projected onto a cylinder tangent to the equator and parallel to the polar axis (see Fig. 4.16). When the cylinder is opened and flattened, a distortion appears. The distortion becomes more pronounced as the distance from the equator increases.

The Mercator projection has straight meridians and parallels that intersect at right angles. The scale is true at the equator or at the two standard parallels equidistant from the equator. The projection is often used for marine navigation, because all straight lines on the map are lines of constant azimuth.

Properties of Mercator Projection

1. Parallels and meridians are straight lines.
2. The meridians intersect the parallels at right angles.
3. The distances between the parallels go on increasing towards the poles, but the distances between the meridians remain the same.
4. All the parallels are of the same length and each one of them is equal to the length of the equator.
5. The length of the equator on this projection is equal to the length of the equator on the globe.
6. The meridians are longer than the corresponding meridians on the globe.

Limitations of Mercator Projection

1. The scale along the parallels and the meridians increases rapidly towards the poles. Being a great exaggeration of the scale along the parallels and the meridians in high latitudes, the sizes of the countries on this projection are very large in the polar areas. For this reason, the polar areas cannot be shown satisfactorily on this projection.
2. Poles cannot be shown on this projection because the exaggeration in the scales along the 90 degrees where the parallel and the meridians touch them will become infinite.

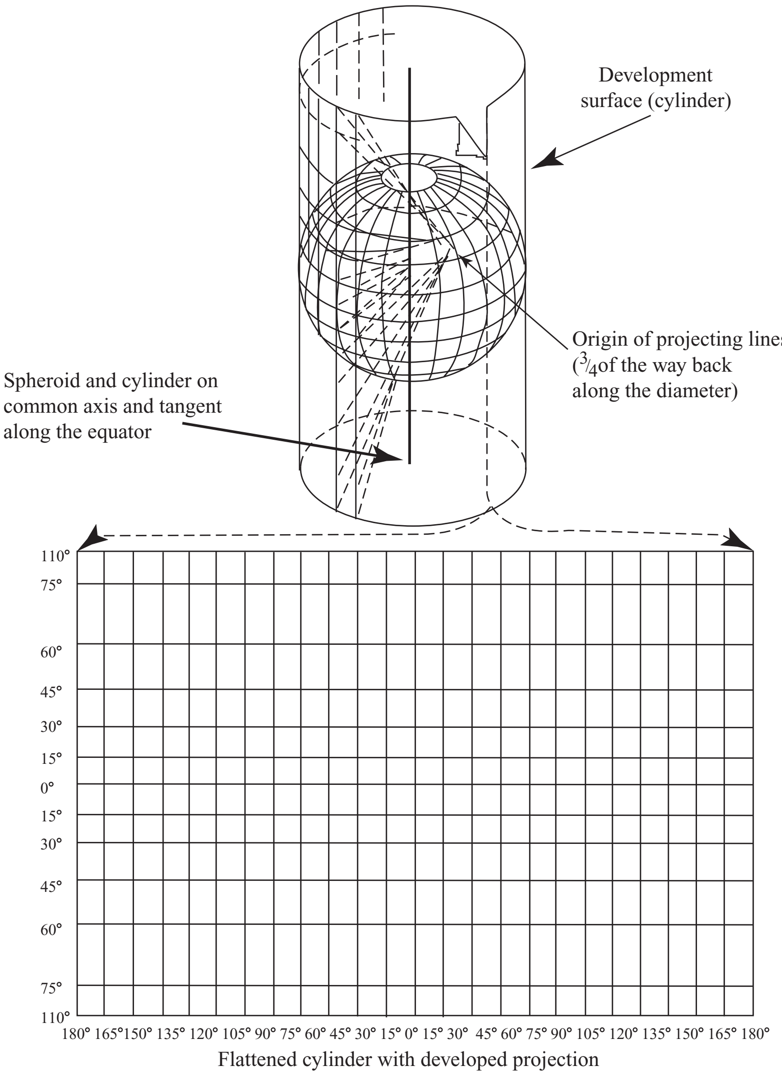


FIGURE 4.16 Construction of a cylindrical map projection

Uses of Mercator Projection

1. This projection is used for navigational purposes both on the sea and in the air.
2. Ocean currents, wind directions and pressure systems are shown on this projection, as the directions are maintained truly on this projection.
3. Maps of tropical countries are shown on this projection, when they are to be used for general purposes. The reason is that exaggeration in the size of an area is small within the tropics, and the shapes of the countries are preserved without much distortions.

FIGURE 4.17 Mercator projection

4.14.2.2 Transverse Mercator

The transverse Mercator projection is a variation of the Mercator projection and the best-known projection of the world. Transverse Mercator projections result from projecting the sphere onto a cylinder tangent to a central meridian. Instead of using the standard parallel as in the case of Mercator projection, the transverse Mercator projection uses the standard meridian. Transverse Mercator projection is formed when the Mercator projection is transversed by rotating the cylinder again until the spheroid is parallel to a second axis (the meridian), which is then open and flattened (see Fig. 4.18).

For survey purposes and to minimize distortion, the transverse Mercator projection uses 60 longitudinal zones of 6 degrees wide. Transverse Mercator maps are often to portray areas with larger north–south than east–west extent. Distortion of scale, distance, direction and area increase away from the central meridian. British National grid systems and Indian National grid systems are based on the transverse Mercator projection. The transverse Mercator projection requires parameters like scale factor at central meridian, longitude of central meridian, latitude of origin (central parallel), values of false easting and false northing.

4.14.2.3 The Universal Transverse Mercator

This projection is used to define horizontal positions worldwide by dividing the surface of the earth into 6 degree zones, each mapped by the transverse Mercator projection with a central meridian in the centre of the zone. Universal transverse Mercator (UTM) zone numbers designate 6 degree longitudinal strips extending from 80 degrees South latitude to 84 degrees North latitude. UTM zone characters designate 8-degree zones extending north and south from the equator (see Fig. 4.19). Eastings are measured from the central meridian (with a 500-km false easting to ensure positive coordinates). Northings are measured from the equator (with a 10,000-km false Northing for positions south of the equator). The scale factors increases from 0.9996 along the central meridian of the UTM zone to 1.000000 at 180,000 m to the east and west. The universal transverse Mercator grid system has the following characteristics.

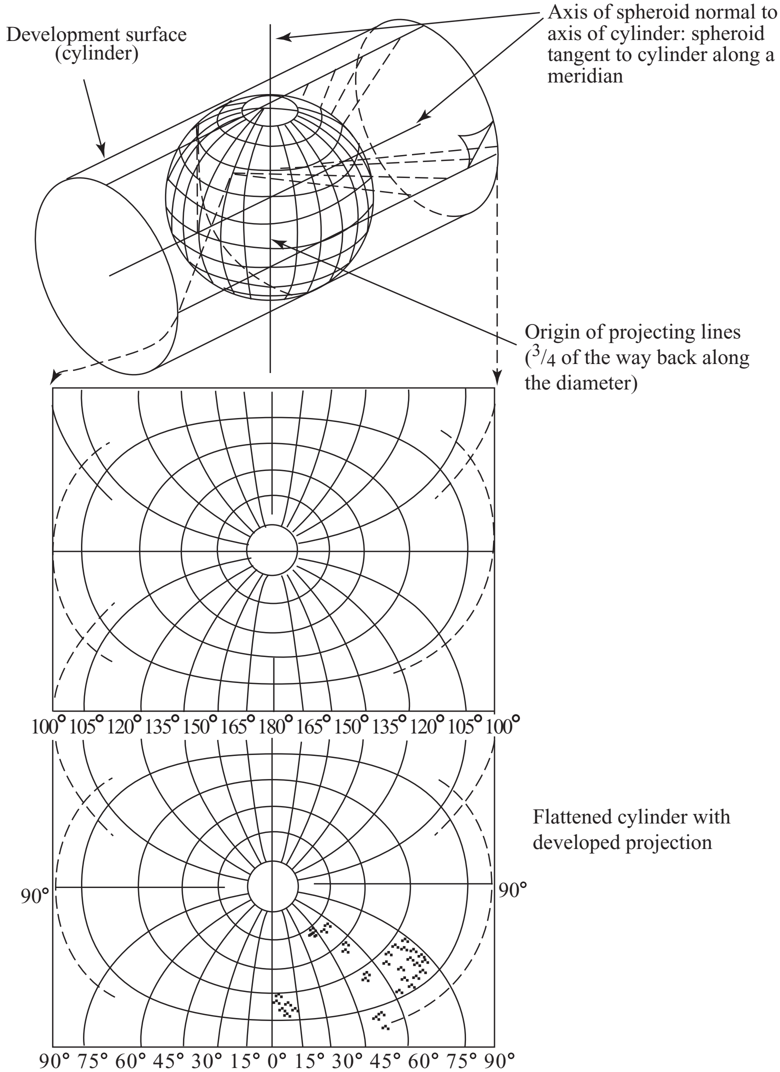


FIGURE 4.18 The transverse Mercator projection

1. **Projection:** transverse Mercator (Gauss-Kruger), in zones of 6 degrees wide.
2. **Longitude of origin:** central meridian of each zone.
3. **Latitude of origin:** zero degree, the equator.
4. **Unit:** meter.
5. **False northing:** meters (10,000,000 for southern hemisphere).
6. **False easting:** 500,000 m.
7. **Scale factor at the central meridian:** 0.9996.

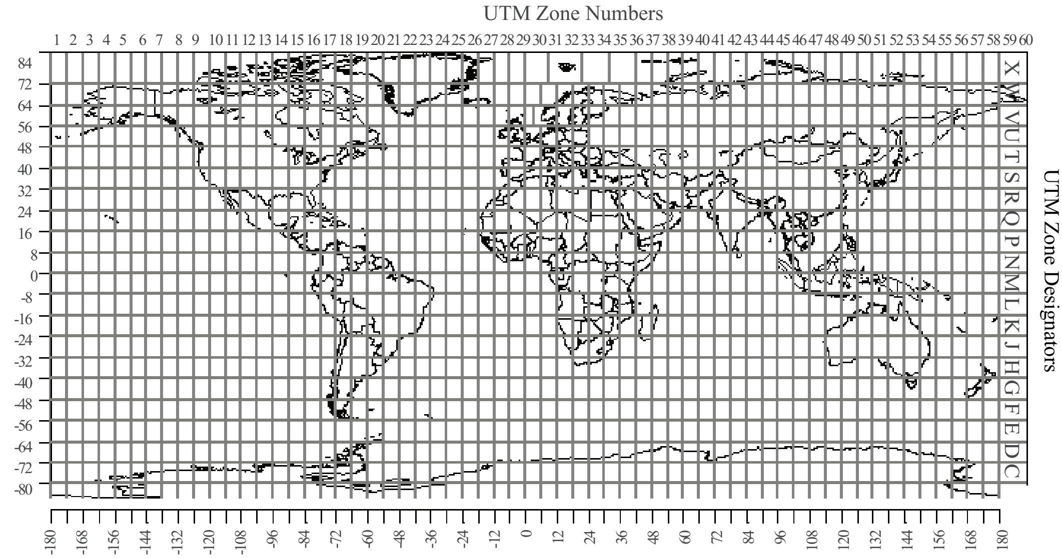


FIGURE 4.19 UTM projection

8. **Zone numbering:** starting with 1 on the zone from 180° W to 174° W and increasing eastward to 60 on the zone from 174° E to 180° E.
9. **Latitude limits of the system:** north—80° N, South—80° S.
10. **Limit of zones and overlap:** The zones are bounded by meridians whose longitudes are multiples of 6-degree West or East of Greenwich. On large-scale maps and in trig lists, an overlap of approximately 40 km on either side of the junction is given. However, this overlap is never used in giving a grid reference.

4.15 CONICAL PROJECTIONS

The parallels and the meridians of a globe are transferred to a case placed on the globe in such a way that its vertex is above one of the poles, and it touches the globe along a parallel. The parallel along which the cone touches the globe is called the standard parallel. The cone is unrolled into the flat surface.

4.15.1 Properties of Conical Projection

1. Parallels are arcs of circle which are concentric in most of the conical projections.
2. The central meridian is a straight line.
3. The distance between the meridians decreases towards the pole.
4. They can represent only one hemisphere at a time, either northern or southern.

4.15.2 Equidistant Conic Projection

Conic equidistant projections can be constructed using one or two standard parallels. As with azimuthal and cylindrical projections, the equidistant conic projections are obtained by adjusting the spacing of the parallels, so that they are equally spaced along meridians and the distance between the parallels on the map is equal to the arc length between parallels on the generating globe. Distances measured along all meridians and along the standard parallels are true to scale, but other distances are distorted.

4.15.3 Simple Conic

The simple conic projection uses one standard parallel. Any parallel of latitude can be selected as the standard parallel. Parallels are equally spaced along meridians, with the distance between parallels being equal to the arc length between parallels on the globe. Adjustment of the spacing of the parallels results in the pole being represented by a circular arc rather than by a point (see Fig. 4.20).

Properties of Simple Conic Projection

1. The parallels are concentric arcs of the circles.
2. The pole is represented by an arc.
3. The meridians are straight lines and they intersect the parallels at right angles.
4. The distance between the meridians decrease towards the pole.

Limitations of Simple Conic Projection The scale along the meridian is true. But it goes on increasing along the parallels away from the standard parallel. Therefore, away from the standard parallel, areas are exaggerated and their shapes distorted. It is suitable only for a narrow strip of land lying adjacent to the standard parallel.

Uses of Simple Conic Projection

1. Railways, roads, narrow river valleys and international boundaries running for a long distance in the east–west direction can be shown on this projection.
2. In this projection, the scale along the meridians is correct. Hence a narrow strip along a meridian is represented satisfactorily.

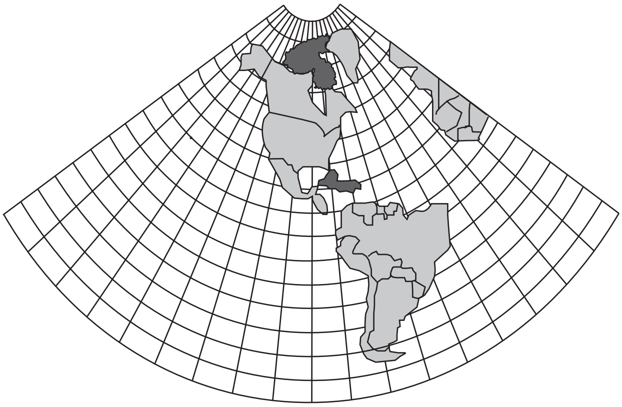


FIGURE 4.20 Simple conic projection

4.15.4 Lambert Conformal Conic Projection

The Lambert conformal conic projection is a good choice for a mid-latitude area of greater east–west than north–south extent. The Lambert conformal conic projection is analogous to the Mercator projection. Both are conformal projections, meaning that at any point, the scale is constant in all directions about the point. As a result, shapes of small areas are represented with minimal distortion, but shapes of larger areas are distorted because of changes in scale from point to point. The projection can be constructed using either one or two standard parallels. Use of two standard parallels is more common, because it gives a better distribution of distortion over the entire map (see Fig. 4.21). The Lambert conformal projection is extensively used for maps of Canada.

4.16 AZIMUTHAL PROJECTIONS

Azimuthal (or zenithal) projections are projections onto a plane that is tangent to some reference point on the globe (Fig. 4.22). If the reference point is one of the poles, the projections are polar azimuthal (zenithal) views. If the reference point lies on the equator, the projections are termed normal. For all other reference points, the projections are oblique. There are five azimuthal projections considered below—equal area, gnomonic, equidistant, stereographic and orthographic.

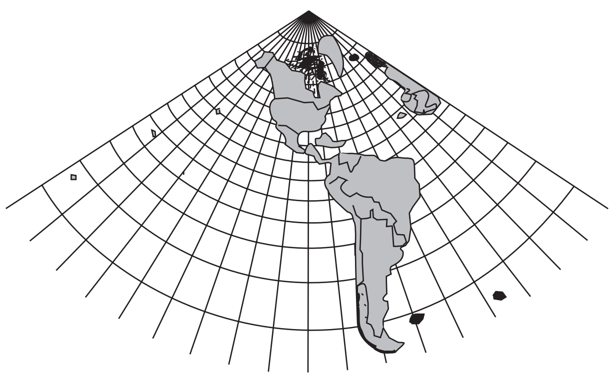


FIGURE 4.21 Lambert conformal projection

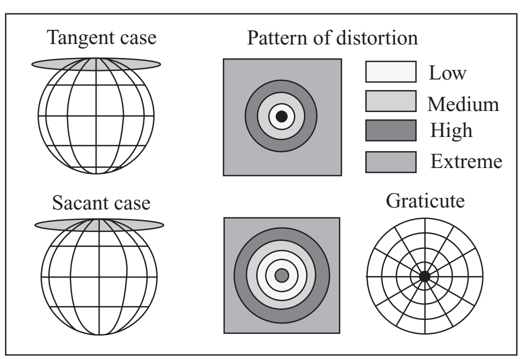


FIGURE 4.22 Azimuthal projections

In the normal (or polar) aspect, the point of tangency is either the north or the south pole, and meridians of longitude are represented as radial straight lines through the pole, while parallels of latitude appear as concentric circles. Distortion in the map increases with distance from the point of tangency. As distortion is minimal near the point of tangency, azimuthal projections are useful for representing areas having approximately equal extents in the north–south and east–west directions.

Properties of Azimuthal Projections

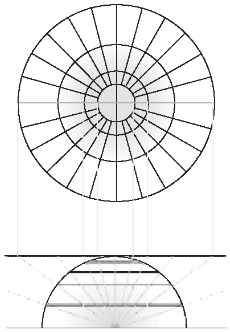
1. The pole is the centre of a projection.
2. The parallels are concentric circles.
3. The meridians on a globe are great circles.
4. The meridians are straight lines radiating from the centre of the projection.
5. The plane on which the globe is projected is placed tangentially at one of the poles and the centre of the projection coincides with the pole.
6. The outlines of maps on these projections are circular.

4.16.1 Normal Azimuthal Projections

Normal azimuthal projections share the same basic pattern of radial meridians and concentric parallels. By modifying the spacing of parallels along the meridians to preserve selected geometric properties, different projections are obtained.

4.16.2 The Gnomonic Projection

The gnomonic projection is a perspective projection with the light source located at the centre of the generating globe. Meridians appear as radial straight lines, and parallels are represented as concentric circles. The spacing of parallels and deformation of areas and angles increases rapidly with distance from the pole, which is the point of tangency. Only an area within about 60 degrees from the point of tangency can be represented on the map, because beyond this range distortion becomes extreme. The gnomonic projection has the special property in which all great circles are represented by straight lines on the map (see Fig. 4.23). The gnomonic projection is therefore useful for navigation, and is often used in conjunction with the Mercator projection for compass navigation.



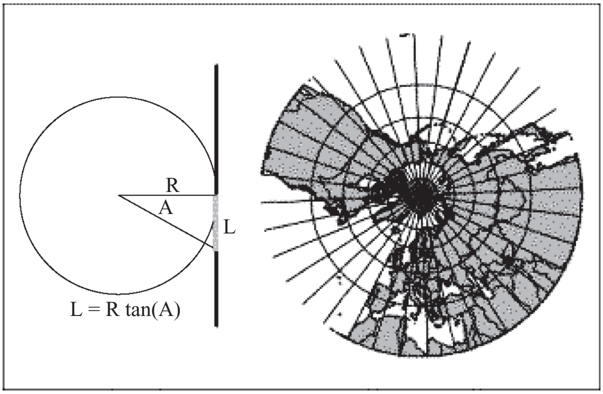


FIGURE 4.23 The gnomonic projection

Properties of Gnomonic Projection

1. The parallels are concentric circles.
2. The meridians intersect the parallels at right angles.
3. The scale along the parallels increases from the centre of the projection.
4. The spacing of parallels is not equal.

Limitations of Gnomonic Projection

1. The equator cannot be shown in this projection.
2. Shapes are generally distorted and areas greatly enlarged from the centre of the projection.

Uses of Gnomonic Projection The gnomonic projection is widely used for preparing navigation charts.

4.16.3 Stereographic Projection

The stereographic projection positions the light source at the antipode of the point of tangency. Thus, if the north pole is the point of tangency, the light source would be at the south pole. The spacing of parallels increases with distance from the pole, but not as rapidly as was the case with the gnomonic projection. As a result, deformation of areas and angles is less severe and it is possible to show an area of up to about 135 degrees from the pole on a single map. Stereographic projections are capable of showing one hemisphere. The stereographic projection is a conformal projection and is commonly used for maps of the polar region. The stereographic projection also has an important special property that, with the exception of great circles passing through the pole, circles on the globe appear as circles or circular arcs on the map. This makes the stereographic projection useful for representing radial phenomena such as shock waves from earthquakes (see Fig. 4.24).

Properties of Stereographic Projection

1. The parallels are not spaced at equal distances.
2. The scale along the parallels also increases away from the centre of the projection.
3. The areas are exaggerated on this projection, the exaggeration increases away from the centre of the projection.

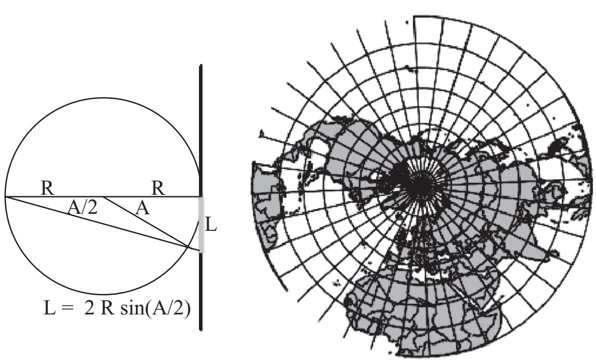


FIGURE 4.24 Stereographic projection

Limitations of Stereographic Projection Areas are enlarged away from the centre of the projection. Only a small area in the central part of the projection can be represented in a satisfactory way.

Uses of Stereographic Projection

1. The entire northern or southern hemisphere can be shown on this projection.
2. This projection is used for preparing sea navigation routes of Arctic regions.

4.16.4 Orthographic Projection

The orthographic projection assumes that the light source is an infinite distance from the point of tangency, resulting in the rays of light being parallel to each other and perpendicular to the projection surface (Fig. 4.25).

The resulting projection can show only one hemisphere. The spacing between parallels decreases towards the equator. The orthographic projection has no special properties, but it gives an approximate perspective view of the earth from outer space. It is therefore useful for visualizing spatial relationships.

Properties of Orthographic Projection

1. As the scale along the meridians decreases rapidly away from the centre, the shapes are much distorted, the distortion increasing away from the centre of the projection.
2. The parallels are concentric circles.
3. The meridians intersect the parallels at right angles.

Limitations of Orthographic Projection The shapes are much distorted near the margin of the projection. The sizes of the areas are diminished away from the centre of the projection. It is only a small area in the central part of the projection that can be represented in a satisfactory way.

Use of Orthographic Projection This projection is used to prepare charts for showing heavenly bodies.

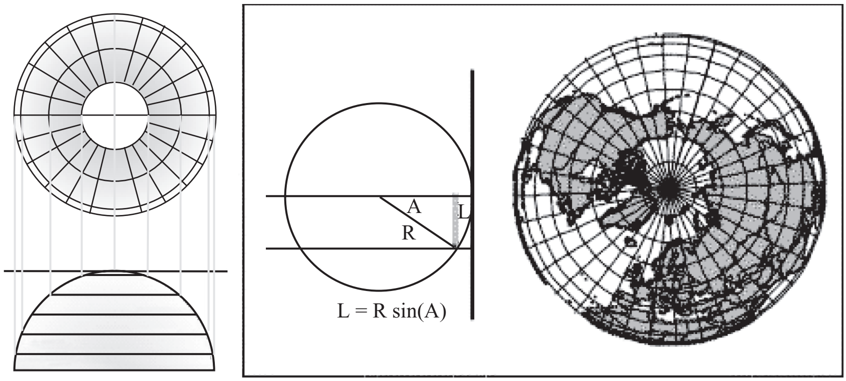


FIGURE 4.25 Orthographic projection

4.16.5 Equidistant Projection

The perspective projections correctly represent azimuths about the point of tangency, but do not correctly represent distances. An azimuthal equidistant projection can be produced by starting with the basic pattern of radial meridians and concentric parallels and modifying the spacing of the parallels so that they are equally spaced along the meridians (see Fig. 4.26). The result is an azimuthal projection that can represent the entire earth on a single map. All distances measured from the point of tangency are true to scale and angles about the point of tangency are also correctly represented. Other distance and angular relationships are distorted. As distances and angles about the central point are correct, the azimuthal equidistant projection is useful for representing routes from a single location to all other locations of interest.

Properties of Equidistant Projection

1. Parallels are concentric circles.
2. As the spacing between the parallels represents true distances, the scale along the meridians is incorrect.
3. The scale along the parallels increases away from the centre of this projection.

Limitations of Equidistant Projection Shapes are greatly distorted away from the centre of the projection. It is only a small area in the central part of the projection that can be represented in a better way. The area lying between the pole and 60-degree parallel is shown satisfactorily.

Uses of Equidistant Projection

1. This projection is commonly used for preparing maps.
2. As the scale along the meridians is correct, narrow strips running along the meridians are shown fairly correctly.

4.16.6 Lambert Equivalent Projection

On this projection, the spacing between parallels decreases with distance from the pole to compensate for the stretching that occurs along parallels of latitude (Fig. 4.27). As with the azimuthal equidistant projection, the azimuthal equivalent projection can represent the entire earth, although it is often limited to showing one hemisphere.

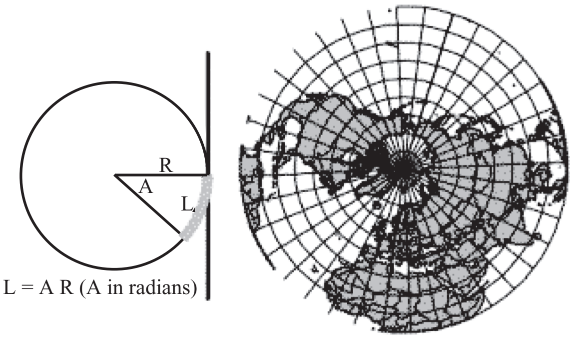


FIGURE 4.26 Equidistant projection

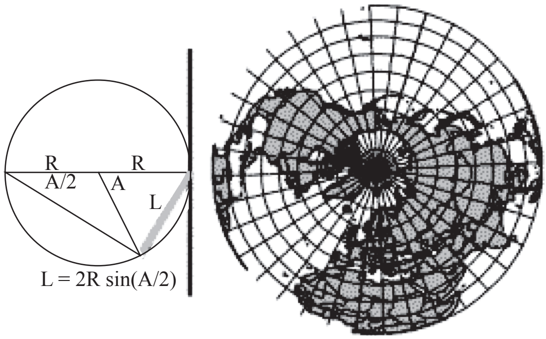


FIGURE 4.27 Lambert equidistant projection

Properties of Lambert Equivalent Projection

1. The parallels are concentric circles and meridians are straight lines radiating from the centre.
2. The meridians intersect the parallels at right angles.
3. The scale along the parallels increases away from the centre of the projection.

Limitations of Lambert Equivalent Projection

1. Shapes are distorted away from the centre of the projection. It is only the central part of the projection that can be represented satisfactorily.
2. In general, the area lying between a pole and 45-degree parallel can be shown satisfactorily on this projection.

Uses of Lambert Equivalent Projection

1. As it is an equal area projection and the shapes of the countries in the central part are preserved in a fairly satisfactory way, this projection is used for the preparation of political and distribution maps of polar regions.
2. It is also used for making general purpose maps of large areas in the northern hemisphere.

REVIEW QUESTIONS

1. Define geodesy.
2. What are the basic differences between a map and a chart? What are the fundamental characteristics of a map?
3. Explain the different classifications of a map.
4. How is a map scale suitable for a map taken? What are the different ways of representing map scales?
5. Mention the uses of a map and list its characteristics.
6. Define map projections, meridian, parallels and graticule.

7. What is an ideal map projection?
8. Explain projection characteristics. Explain the different map projections according to the projection surface and different map projections according to the method of deviation.
9. What are the characteristics of a cylindrical map projection? Explain the different types of cylindrical projection.
10. Differentiate between transverse Mercator projections and UTM projections.
11. Recall the characteristics of UTM projections.
12. List the properties and limitations of simple conic projections.
13. Write the properties and limitations of orthographic projections.

THE COORDINATE SYSTEM 5

Chapter Outline

5.1 INTRODUCTION

Coordinates are a conventional method of recording position in space. They may be used to locate position in two dimensions, such as a point on a graph. The definition of a coordinate position on the surface of a three-dimensional (3D) body such as a sphere or spheroid is more difficult. However, one should be aware of the method of describing location by means of latitude and longitude, which are geographical coordinates. In addition to providing a means of reference, coordinates can also be used as a convenient way of solving certain geometrical problems. The branch of mathematics known as coordinate geometry analyzes problems through the relationship between points as defined by their coordinates. Plane coordinate geometry is usually studied through the medium of the conic sections or the definition of the different kinds of curves formed by the surface of a cone where it is intersected by a plane. Two of the resulting sections, the ellipse and the circle, are of fundamental importance to the theory of distortions in map projections.

A reference system is used to measure horizontal and vertical distances on a planimetric map. A coordinate system is usually defined by a map projection, a spheroid of reference, a datum, one or more standard parallels, a central meridian and possible shifts in the x- and y-directions to locate x and y positions of point, line and area features.

5.2 PLANE COORDINATE SYSTEMS

There are infinite number of ways in which one point on a plane surface may be referred to another point on the same plane. Every map projection creates a unique reference system that satisfies this requirement and different map projections could theoretically be described. It is desirable to use some kind of coordinate system to describe, analyze and construct each of these projections. Any system to be used for such purposes ought to be easy to understand and simple to express algebraically. For plane representation, the choice lies between plane Cartesian coordinates and polar coordinates.

5.3 PLANE CARTESIAN COORDINATES

Most of us are familiar with the method of plotting two variables graphically, specially on ruled paper with grids. In general, any plane coordinate system that makes use of linear measurements in two directions from a pair of fixed axis can be regarded as a Cartesian system. The coordinate system comprises sets of families of lines, which intersect one another to form a network when plotted. The necessary conditions which must be fulfilled are:

1. The two families of lines should be distinct from one another.
2. Every line of one family should intersect every line of the other family at one point only.
3. No two lines of the same family should intersect one another.

Thus, a Cartesian coordinate system can comprise families of straight lines or curves, which may intersect at any angle. However, it is a distinct advantage if the special case is chosen in which both families of lines are straight and they are orthogonal or intersect at right angles.

The units into which the axes are subdivided for the purpose of linear measurement are quite arbitrary. For example, graph paper is available with millimetre ruling with various combinations of multiples and fractions of a millimetre. There is a sign convention to be observed in the use of rectangular coordinates.

This states that the X-axis is positive towards the right and the Y-axis is positive towards the top of the page. In other words, a point in the top right-hand quarter of a graph paper is defined as positive values of x and y, whereas a point in the bottom left-hand quarter has negative values for x and y.

The quarters are termed as quadrant and are numbered 1–4 in an anti-clockwise direction commencing with the top right quadrant as shown in Fig. 5.1.

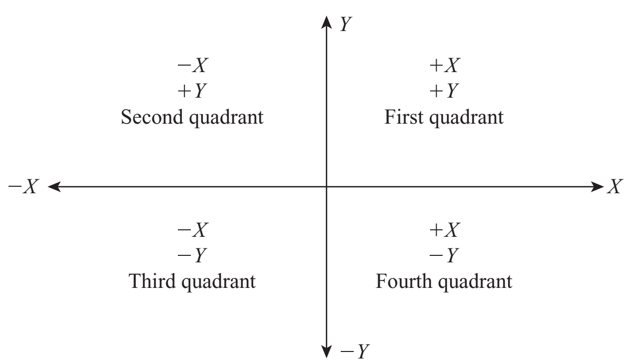


FIGURE 5.1 Plane Cartesian coordinate system

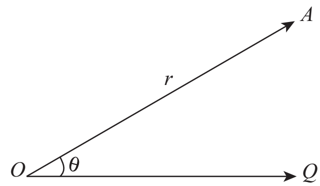


FIGURE 5.2 Plane polar coordinates

5.4 PLANE POLAR COORDINATES

Polar coordinates define positions by means of one linear measurement and one angular measurement. In Fig. 5.2, a single line *OQ*, replaces the pair of orthogonal axes passing through the origin, *O*, or the pole of the system. The position of any point *A* may be defined with reference to this pole and the polar axis or initial line, *OQ*, by means of the distance $OA = r$ and the angle θ. The line *QA* is known as the radius vector and the angle θ is the vectorial angle which the radius vector makes with the initial line. Hence, the position of *A* is defined by the coordinates *r* and θ.

The order of referring the radius vector followed by the vectorial angle is standard in all branches of mathematics. In the theoretical derivation of map projections, where θ enters directly into an equation and is not introduced as some trigono metrical function of the angle, it is necessary to this angle in absolute units, or radians. This is because both elements of the coordinate system must have the same character of length. The direction in which the vectorial angle is measured depends upon the purpose for which polar coordinates are used. Usually, the $+\theta$ is taken as the anti-clockwise angle measured from the initial line. This sign convention is used in vector algebra. On the other hand, navigators, surveyors and cartographers have become accustomed to measure a positive angle in the clockwise direction. This is because direction on the earth's surface is conventionally measured clockwise from north or clockwise from a reference object, and, therefore, we tend to measure all directions according to this rule. In most practical problems, formal recognition of the sign of an angle is unimportant, because the user can understand the convention used in the 360-degree circle measured clockwise. However, difficulties arise in automatic data processing, because computer subroutines are usually programmed according to the mathematical conventions.

5.5 CARTESIAN 3D COORDINATE SYSTEMS

There are many different coordinate systems based on a variety of geodetic datums, units, projections and reference systems in use today, which are also called rectangular coordinates. The three axes of 3D Cartesian coordinates are conventionally denoted by the *X*-, *Y*- and *Z*-axes.

There are many basic coordinate systems familiar to students of geometry and trigonometry. These systems can represent points in two-dimensional (2D) or 3D space. René Descartes (1596–1650) introduced a system of coordinates based on orthogonal (right angle) coordinates. The 2D and 3D system used in analytic geometry are often referred to as Cartesian system (see Fig. 5.3 and Fig. 5.4). Similar systems based on angles from baselines are often referred to as polar systems.

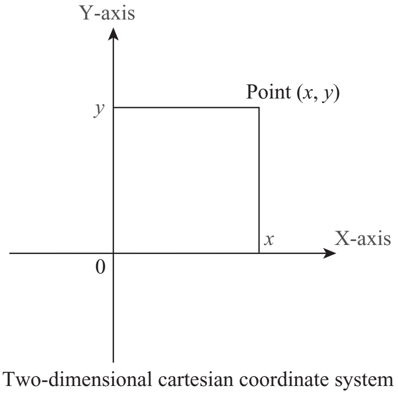


FIGURE 5.3 2D coordinate systems defined with respect to a single plane

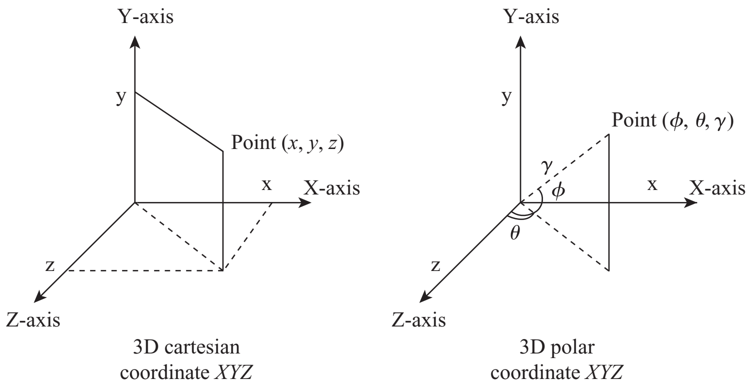


FIGURE 5.4 3D coordinate systems defined with respect to two orthogonal planes

5.6 GEOGRAPHIC COORDINATE SYSTEMS

Geographic coordinates of positions are always expressed as a latitude and longitude. The latitude and longitude are related to a particular earth figure, which may be a sphere for most atlas maps and imprecise work, but for rigorous purposes, it is a mathematical figure, which much more closely approximates the shape of the real earth. The earth's shape is assumed to be an ellipsoid of revolution (the figure defined by an ellipse rotated round its minor axis), and is sometimes termed as a spheroid. An ellipsoid appropriate for one region of the earth is not always appropriate to other regions. Recent work based on satellite data has seen the development of ellipsoids that are used worldwide, such as GRS80 (geodetic reference system 1980) and WGS84 (world geodetic system 1984).

Another property is needed to uniquely specify geographical positions. This is the position and orientation of the ellipsoid relative to the earth. The term used to describe this fitting of an ellipsoid to the earth is geodetic datum. Many geodetic datums exist throughout the world, each usually associated with the national survey of a particular country or continent. More recently, several new geodetic datums have been successively derived from steadily accumulating satellite and other data, to provide for a best worldwide fit.

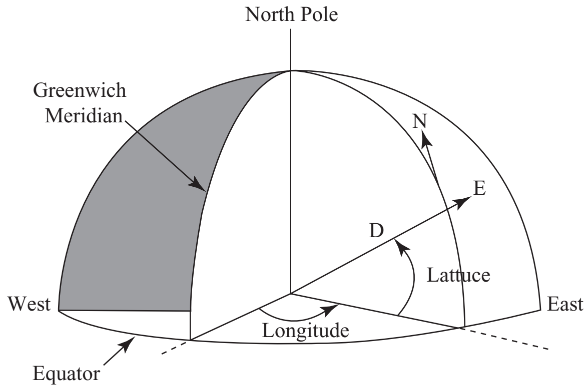


FIGURE 5.5 The latitude and longitude

Each geodetic datum is defined by a set of numerical elements, which define where the imaginary geographical graticule of lines of latitude and longitude lies on the earth's surface, that is, the direction of the poles and the rotated position of the graticule about this direction. The latitude and longitude and an azimuth at the datum point was originally established using the best means possible at that time and selecting an ellipsoid which is a best fit for the earth's surveyed surface for the country or continent in question. The geodetic datum may have taken the name of the fundamental surveyed point. In the case of the new earth centred and satellite derived geodetic datums, where the datum is effectively the centre of the earth ellipsoid, a different type of label with world or international annotations are used, for example, WGS 84 (world geodetic system 1984).

Thus, all precisely surveyed and mapped points and features on the earth's surface will be uniquely defined in position by stating their coordinates (grid or geographical), the projection and the geodetic datum, including its earth ellipsoid. Therefore, to be unambiguous and precise in the definition of position, quoting latitude, longitude and height, as well as specifying the geodetic datum, including ellipsoid, and the level to which the height coordinate values are related are essential. Geographic coordinates may be transformed to grid coordinates, or projected coordinates of epicentre, or may be transformed to other geographic coordinates on a different datum.

Latitude–longitude is a unit of measure for a spherical (3D) coordinate system. Latitude–longitude is angles measured from the earth's centre point to locations on the earth's surface (see Fig. 5.5). Lines of latitude run east–west parallel to the equator; while longitude lines run north–south and converge at the poles. The length of one degree of longitude varies depending on the latitude at which it is measured. For example, one degree of longitude at the equator is 111 kilometres in length, but the length of one degree of longitude converges to zero at the poles. Latitude–longitude is measured in degrees, minutes and seconds, and because degrees are not associated with a standard length, they cannot be used as an accurate measure of distance or area.

5.7 PROJECTED COORDINATE SYSTEMS

As all map projections attempt to represent the curved surface of the earth on a flat sheet of paper, there are inevitable distortions in relative areas, in scale, in shape or in all three. Although a great number of different map projections are in general use for atlas and sheet maps, and there are many others which have been devised to theoretically provide better representations of the earth's surface for showing particular countries or features, only a limited number has been adopted for national topographic mapping purposes. In general, those used for large- and medium-scale topographic mapping have the particular property that they preserve shape, that is, the shape of the land and sea areas, and thus preserve the topographic, geological and other features. Hence, they are also deemed most suitable for large- and medium-scale hydrocarbon exploration/production mapping. Such maps are said to be orthomorphic or conformal. The latter term is used in epicentre documentation.

Indeed, these projections are generally arranged to minimize scale distortion, such that often several elements of the same projection, usually arranged in regular bands of longitude or latitude, may be needed to achieve complete coverage of a particular country. All projections have mathematical formulas that define the relationship between the latitude and longitude, graticule on the earth and its representation on the map sheet, or the relationship between the geographic coordinates (latitude and longitude) of points and their projected coordinates (grid or rectangular coordinates) on the projection or map.

5.7.1 The Elevation

While geographic coordinate systems and projected coordinate systems in the epicentre deal with 2D locations, there is a third dimension that must be considered when giving a location. The third dimension is the height or elevation. Definition of heights or levels requires consideration of further parameters and reference levels or surfaces, and are relevant to some of the vertical coordinate systems of the epicentre.

5.8 ASTRONOMICAL COORDINATE SYSTEMS

One of the basic needs of astronomy, as well as other physical sciences, is to give reasonable descriptions for the positions of objects relative to each other. Scientifically, this is done in mathematical language or by properly assigning numbers to each position in space. These numbers are called coordinates, and the system defined by this procedure a coordinate system. The coordinate systems considered here are all based at one reference point in space with respect to which the positions are measured. The origin of the reference frame may be the location of the observer, the centre of earth, the sun or the milky way galaxy. Any location in space is then described by the radius vector or arrow between the origin and the location, namely by the distance (length of the vector) and its direction. A reference plane containing the origin is fixed, or equivalently, the axis through the origin is perpendicular to it (typically, an equatorial plane and a polar axis). Elementarily, each of these uniquely determines the other. One can assign an orientation to the polar axis from negative to positive, or south to north, and simultaneously to the equatorial plane by assigning a positive sense of rotation to the equatorial plane. These orientations are, by convention, usually combined by the right-hand rule: 'If the thumb of the right hand points to the positive (north) polar axis, the fingers show the positive direction of rotation' (and vice versa, so that a physical rotation defines a north direction).

On earth, positions are usually given by two angles, the longitude and the latitude coordinates on the surface of earth, and the elevation of the location above (or below) sea level. The circles along the earth's surface, which are parallel to the equator, are the latitude circles, where the angle at the planet's centre is constant for all points on these circles. Half circles from pole to pole, which are all perpendicular to the equatorial plane, are called meridians. Geographical longitude is measured as the angle between meridians and the meridian under consideration (or more precisely, between the half planes containing them). It is of course the same for all points of the meridian.

The earth is not exactly circular, but slightly flattened; its surface (defined by the ocean surface, or the corresponding gravitational potential) forms a specific figure, the so-called geoid, which is very similar to a slightly oblate spheroid (the reference ellipsoid). This is the reason why there are two common but different definitions of latitude on the earth as given below:

- Geocentric latitude, measured as an angle at the earth's centre, between the equatorial plane and the direction to the surface point under consideration, and
- Geographical latitude, measured on the surface between the parallel plane to the equatorial plane and the line orthogonal to the surface, the local vertical or plumb line, which may be measured by the direction of gravitational force.

5.9 GEOID AND REFERENCE ELLIPSOIDS

The earth is an oblate spheroid with the minor axis 1/300th shorter than the major axis, but the earth also has an irregular undulating surface that varies by +/– 100 m from the oblate spheroid. So the geoid is the approximation for theshape of the earth at 'sea level' that takes into account gravitational and rotational inconsistencies. The geoid is a surface along which the direction of gravity is always perpendicular (see Fig. 5.6). The latter is particularly significant because optical instruments containing levelling devices are commonly used to make geodetic measurements.

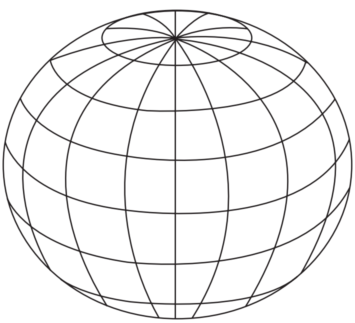


FIGURE 5.6 An ellipsoid

This irregular shape is approximated by 'ellipsoids' with a major and minor axis that fits particular parts of the globe better than others. A reference ellipsoid must have a zero point somewhere and the mean sea level must be established from the geoid. These are fixed for a particular Universal Transverse Mercator (UTM) zone, for example, the standard parallel.

5.10 CARTOGRAPHY

Cartography is a branch of science and the art of making maps and charts. The term may be taken broadly as comprising all the steps needed to produce a map such as planning, aerial photography, field surveys, photogrammetry, editing, colour separation and multicolour printing. Map-makers, however, tend to limit the use of the term to map-finishing operations, in which the master manuscript is edited and colour separation plates are prepared for lithographic printing.

5.10.1 Traditional Cartography

The traditional views of cartography focused on map making implicitly, and assumed that if one knows how to make a map, then one also knows how to use maps. This assumption was legitimate when the mapmaker was also the map user, for example, a geographer making maps of data collected through fieldwork as part of data analysis. In this context, the map user will be familiar with the data represented on the map, will be aware of the limitations of the data and the mapping techniques employed, and could therefore be expected to use the map effectively for its intended purpose. However, this context is not typical for most uses of maps.

5.10.2 Computer Cartography

Computer cartography has revolutionized traditional cartography vastly to improve map making and visualization of geographic information in a multimedia environment. Computer cartography (also known as digital mapping, automated cartography or computer-aided cartography) is the generation, storage and editing of maps using a computer.

5.11 GPS MAPPING

GPS mapping is used to locate and map features of the earth such as power lines, rivers, highways, crops, weeds, pests or soil types. It can be used to estimate the number of insects in a field, measure the area of a crop or create a map of sample locations. There are difficulties using GPS with desktop GIS applications. Most desktop applications are large and expensive. Substantial training is required. They can layout, format, analyze, view, theme and print finished maps. These features are not very useful in the field, as they slow down the user and take up memory space and slow the processor power. Desktop GIS applications are also slower. Therefore, in addition to increased software cost, they require more expensive hardware and additional training.

Most of the maps based on spatial data sources used in GIS have a projection associated with them. To undertake meaningful analysis, it is necessary to know something about the projections being used. The results of analysis will be affected in different ways by different map projections. If a GIS application requires the accurate calculation of areas, then using a projection which distorts areas is obviously not suitable. When using data at large scales the effects can be substantial.

5.12 TRANSFORMATION METHODS

The Cartesian coordinates (x, y) of a point on a map are functionally related to a position on the earth's surface expressed in geographical coordinates (φ, λ).

That is,

$$x = f_1(\varphi, \lambda)$$
$$y = f_1(\varphi, \lambda) \qquad (5.1)$$

There are three basic methods of relating (x, y) to (φ, λ) or various forms of plane coordinates used on other maps, aerial photographs or scanned imagery. These are referred to here as:

1. Analytical transformation.
2. Direct or grid-on-grid transformation.
3. Polynomial transformation.

5.12.1 Analytical Transformation

Analytical transformation is the most obvious and straightforward solution to the problem of relating Cartesian coordinates on a map to geographical coordinates on the earth's surface. This is because it approximates the methods of classical cartography, that is, locating and plotting points from their geographical coordinates. In the automated applications, the objective is to convert the (x', y') coordinates of points digitized on a source map into their geographical coordinates. These, in turn, are used to determine the (x, y) coordinates for the GIS framework or to create a new map.

The conversion from geographical coordinates into plane coordinates is the normal practice of constructing a map projection and is regarded as the forward solution. The preliminary conversions required to find the geographical coordinates from the (x', y') system of digitized coordinates is correspondingly called the inverse solution. Thus, the transformation model is:

$$(x', y') \rightarrow (\varphi, \lambda) \rightarrow (x, y)$$
$$\text{[Inverse solution] [Forward solution]} \tag{5.2}$$

Most of the standard works on map projections only provide the equations for the forward solution. This is because in the days before digital mapping became a practical possibility, only the forward equations were needed to construct a graticule, and all subsequent transfer of detail was manual. It was only in the field of topographic mapping, using the transverse Mercator and Lambert conformal conical projections in particular, that the two conversions 'geographicals to grid' and 'grid to geographicals' were likely to be employed and the projection tables provide both.

5.12.1.1 The Analytical Transformation Equations for Mercator's Projection

The relationship between the forward and inverse coordinate expressions may be exemplified by the sets of equations used to define the normal aspect of Mercator's projection, which is the basis of virtually all nautical charts. For the projection of the sphere (Eqn. 5.3), the forward solution is to be found in most of the standard works on map projections.

$$x = R \bullet \lambda$$
$$y = R \bullet ln \tan\left(\frac{\pi}{4} + \frac{\varphi}{2}\right) \tag{5.3}$$

where *ln* is the natural logarithm (to base ε), the longitude, *A*, is expressed in radians and the radius of the earth, *R*, is expressed in millimeters at the scale of the proposed map. In order to express (φ, λ) in terms of (x, y), which is the inverse problem, it is necessary to write:

$$\varphi = \frac{\varphi}{2} - 2\tan^{-1}(\varepsilon^{-y/R}) \tag{5.4}$$

$$\lambda = \frac{x}{R} + \lambda_0$$

where λ_0 is the datum meridian from which longitudes are measured. Hence, ε is the base of natural logarithms (= 2.7182818 . . .). It is written as the Greek epsilon to avoid confusion with the eccentricity of the spheroid, usually denoted by *e*.

The first complication which needs to be considered is the corresponding relationships for the projection of the spheroid, having semi-axes *a* and *b* with eccentricity *e* derived from

$$e^2 = \frac{(a^2 - b^2)}{a^2} \approx 0.0067.... \tag{5.5}$$

For the forward solution of Mercator's projection of the spheroid, Eqn. 5.3 has to be modified to the corresponding equation

$$\begin{aligned} x &= a \bullet \lambda \\ y &= a \bullet ln \tan\left(\frac{\pi}{4} + \frac{\varphi}{2}\right)\left[\left(1 - e \bullet \sin\varphi\right)/\left(1 + e \bullet \sin\varphi\right)\right]^{e/2} \end{aligned} \tag{5.6}$$

For the inverse calculation, the equation to find latitude is transcendental, needing an iterative solution.

$$\begin{aligned} \varphi &= \frac{\pi}{2} - 2\tan^{-1}(t)\left[\left(1 - e \bullet \sin\varphi\right)/\left(1 + e \bullet \sin\varphi\right)\right]^{e/2} \\ t &= \varepsilon^{-y/a} \end{aligned} \tag{5.7}$$

and the first trial solution is to find

$$\varphi = \frac{\pi}{2} - 2\tan^{-1}(t) \tag{5.8}$$

The result is inserted as φ in the right hand side of Eqn. 5.7 to calculate a new value for φ on the left hand side. The process is repeated until the results have converged to a difference between the two values for φ, which the user considers to be insignificant, and the final value for φ may be accepted. Longitude is obtained from a simple modification for the λ expression in Eqn. 5.6, namely:

$$\lambda = \frac{x}{a} + \lambda \tag{5.9}$$

5.12.1.2 Further Transformations

The number of stages in the inverse solution may have to be extended for various other reasons. Because most digitizing is done in Cartesian coordinates and it is sometimes appropriate to deal with a map projection, which is best derived in polar coordinates, it may be necessary to change plane rectangular coordinates (x', y') into plane polar coordinates (ρ, δ) before determining the geographical coordinates. Thus, the transformation model contains an additional stage, as follows:

$$(x', y') \rightarrow (\rho, \delta) \rightarrow (\varphi, \lambda) \rightarrow (x, y) \tag{5.10}$$

[Inverse solution] [Forward solution]

Similarly, a change in aspect, for example, to a transverse or oblique projection, involves yet another stage in the succession of transformations. Change in aspect is commonly effected through the system of (z, a) bearing and distance coordinates, using spherical trigonometry to convert from (φ, λ) into (z, a). Thus,

$$(x', y') \rightarrow (\rho, \delta) \rightarrow\rightarrow\rightarrow (\varphi, \lambda) \rightarrow\rightarrow\rightarrow (z, a) \rightarrow (x, y) \tag{5.11}$$

[Inverse solution] [Change in aspect] [Forward solution]

An alternative to this method of changing aspect is to use a three-dimensional Cartesian system (X, Y, Z) instead of geographical coordinates to relate positions on the spherical surface. The changes in the aspect may also be obtained by rotating the three axes through the three Eulerian angles at the center of the sphere. Hence, the transformation becomes.

$$(x', y') \rightarrow (\varphi, \lambda) \rightarrow (\mathrm{X, Y, Z}) \rightarrow (\mathrm{X', Y', Z'}) \rightarrow (z, a) \rightarrow (\varphi', \lambda') \rightarrow (x, y) \tag{5.12}$$

[Inverse solution] [Change in aspect] [Forward solution]

where (X', Y', Z') are the rotated coordinates of the point (X, Y, Z).

5.12.1.3 The Advantages and Disadvantages of the Analytical Method

The analytical methods are rigorous and independent of the size of the area to be mapped. However, it can be inconveniently slow. It seems at first sight that this is no longer a problem that modern high-speed computers have reduced these considerations to insignificance. However, the clumsiness of the analytical method becomes apparent when applied to larger data files. This is well demonstrated by Eqn. 5.11 relating to change in aspect, where each additional transformation stage may involve either the solution of a separate spherical triangle or in Eqn. 5.12, the determination of three-dimensional coordinates and rotation of them for every point on the map.

5.12.2 Direct Transformation by the Grid-on-Grid Method

The grid-on-grid method does not require inverse solution of the geographical coordinates (φ, λ) of the original map, but is based upon the relation between the rectangular coordinates of the same points on the two projections. This technique was used in traditional cartography for such purpose as re-gridding or plotting a second grid on a military topographical map, hence the name 'grid-on-grid'. This method is also important in mapping from remote sensing and is the method adopted in modern analytical plotters for use with conventional aerial photography. Practically all the methods of applying geometrical corrections to remote sensing imagery, including that derived from Landsat MSS, Landsat TM and SPOT sensors, utilize such techniques employing ground control points of known position to determine transformation parameters.

The simplest transformation model is

$$(x', y') \rightarrow (x, y) \tag{5.13}$$

The two relatively simple numerical procedures, which are commonly employed in the mapping science, are the linear transformations from one plane Cartesian coordinates system into another.

There are two major kinds of transformation—the Helmert (similarity or linear conformal) transformation; and the affine transformation.

The linear conformal or Helmert transformation is expressed in the general form

$$\begin{aligned} x &= A + Cx' + Dy' \\ y &= B - Dx' + Cy' \end{aligned} \tag{5.14}$$

The affine transformation is as follows:

$$\begin{aligned} x &= A + Cx' + Dy' \\ y &= B - Ex' + Fy' \end{aligned} \tag{5.15}$$

In these equations, the known or digitized coordinates (x', y') of a point in one system are transformed into the (x, y) coordinates of the second system, through the use of four or six coefficients A, B, C, D, E and F. In the Helmert transformation, the C and D coefficients are common to both the equations for x and y. But in the affine transformation, it is necessary to introduce separate corrections for each direction. Both transformations may be resolved into three components:

1. Transformation of the axes or change of origin corresponding to the coefficients A and B in both Eqns. 5.14 and 5.15
2. A change in scale from one grid system to the other
3. The rotation of the axes of one grid system with respect to their directions in the other (see Fig. 5.7)

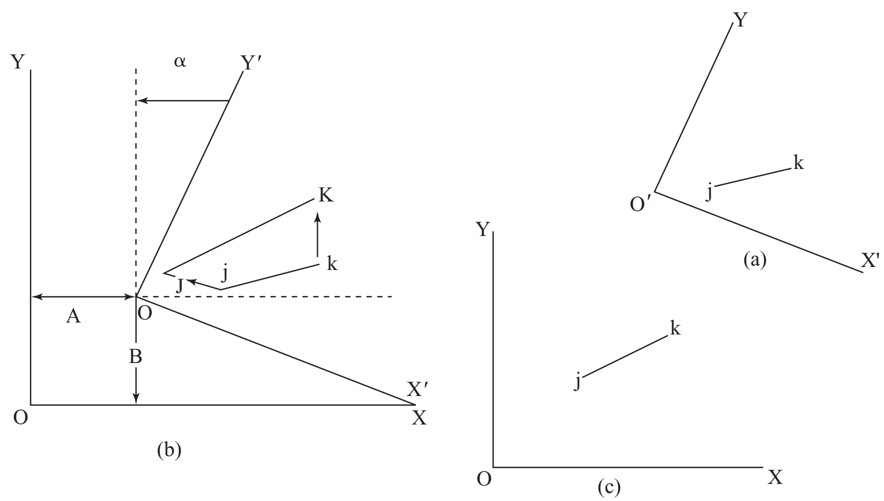


FIGURE 5.7 The geometry of Helmert transformations. (a) Initial conditions, showing two points *j*, *k* referred to Cartesian axes O'X' and O'Y', which are orthogonal. (b) The three stages in transformation superimposed upon one another. These are: First the scale change by which the line *jk* is transformed into the line *JK*. Secondly, the rotation of X' and Y' axes through the angle α about the point O' to make the axes parallel to the final OX and OY system. The third is the translation of the origin *O'* through the distance *A* and *B*, respectively, to refer *J* and *K* to the (X, Y) system. (c) The final condition indicating the position of *J* and *K* within the (X, Y) system

5.12.2.1 Helmert Transformation

For the Helmert transformation, the effect of all three displacements are combined to produce the pair of equations

$$x = (m \cdot x' \cdot \cos \alpha + m \cdot y' \cdot \sin \alpha) + A \tag{5.16}$$

$$y = (-m \cdot x' \cdot \sin \alpha + m \cdot y' \cdot \cos \alpha) + B \tag{5.17}$$

where A and B are the coefficients in Eqn. 5.14, which correspond to the shifts in the origin of the coordinates parallel with the x and y axes, the angle a is the rotation of the axes required to make these axes parallel and m is a scale factor. Thus, if two points j and k in the first system correspond to j and K in the second, the ratio of the distances jk/JK must be applied to the first system to bring it to the same scale as the second.

The complete transformation may be expressed in matrix form as:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} C & D \\ -D & C \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} A \\ B \end{pmatrix} \tag{5.18}$$

where $D = m' \sin a$ and $C = m' \cos a$. The inverse transformation is that of determining the (x', y') coordinates of points whose (x, y) coordinates are already known. Thus,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} C' & -D' \\ D' & C' \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} A \\ B \end{pmatrix} \tag{5.19}$$

where $C' = \cos \alpha/m$ and $D' = \sin \alpha/m$.

If there are only two points (x'_1, y'_1) and (x'_2, y'_2) on the first surface corresponding to (x_1, y_1) and (x_2, y_2) on the second surface whose coordinates are known or have been measured, the method of finding C and D is through Eqns. 5.20 and 5.21:

$$C = \left[\frac{(\delta x \cdot \delta y') - (\delta y \cdot \delta x')}{\delta x'^2 + \delta y'^2} \right] \tag{5.20}$$

$$D = \left[\frac{(\delta y \cdot \delta y') - (\delta x \cdot \delta x')}{\delta x'^2 + \delta y'^2} \right] \tag{5.21}$$

where $\delta x = (x_1 - x_2)$, $\delta x' = (x'_1 - x'_2)$, $\delta y = (y_1 - y_2)$ and $\delta y' = (y'_1 - y'_2)$.

If there are more than two common points, such as occurs in vector digitizing, the adjustments of aerial triangulation or fitting a remotely sensed image to many ground control points, the determination of the coefficients from only two or three of them is inadequate, because the coordinates of any of those points may contain small errors, which, in turn, introduces errors into the transformation of all other points. Therefore, all of the data available for the determination of C and D ought to be taken into consideration. This involves a solution of the coefficients by the method of least squares.

5.12.2.2 Affine Transformation

The assumption made in the Helmert transformations is that the scalar, m, is a single unique value. In other words, the ratio jk/JK is the same whatever directions of these lines. This is a reasonable assumption for some purposes, but it may not always be justifiable.

For example, in photogrammetry the location of image points on a film may be affected by deformation of the film base by stretching and shrinking, and this is not usually the same in all directions. In the extraction of positional information by digitizing a paper map, the influence of differential stretching or shrinking of the paper is even more occasional. For these applications, it is desirable to use the affine transformation or even a higher order polynomial, because this allows for different scales in the directions of the two axes, m_x and m_y.

This may also be combined with small departures of the coordinate axes from the perpendicular, as illustrated in Fig. 5.8. Here it can be seen that the (x, y) axes intersect at an angle $\gamma \neq 90°$.

5.12.3 Numerical Transformation Methods

The third method of relating Cartesian coordinates on a map to geographical coordinates on the earth's surface is to construct polynomial expressions to fit the data and use the resulting coefficients to transform the coordinates of the remaining points of the map in detail. This method is, of course, important in numerical analysis and has many different applications. In the narrower field of transforming positional data of GIS applications, this method may be used with equal efficiency for transformation from geographical into grid coordinates (Eqn. 5.22) as for making the grid-to-grid transformations (Eqn. 5.23). The required number of common points needed to determine the coefficient and the amount of computation needed vary according to the order of the degree of the polynomial. For example, a third-order polynomial, containing terms in φ and λ up to φ^3 and λ^3 requires 10 coefficients denoted a_{ij}, and 10 in b_{ij} as in Eqn. 5.22 determined by solving 10 or more equations.

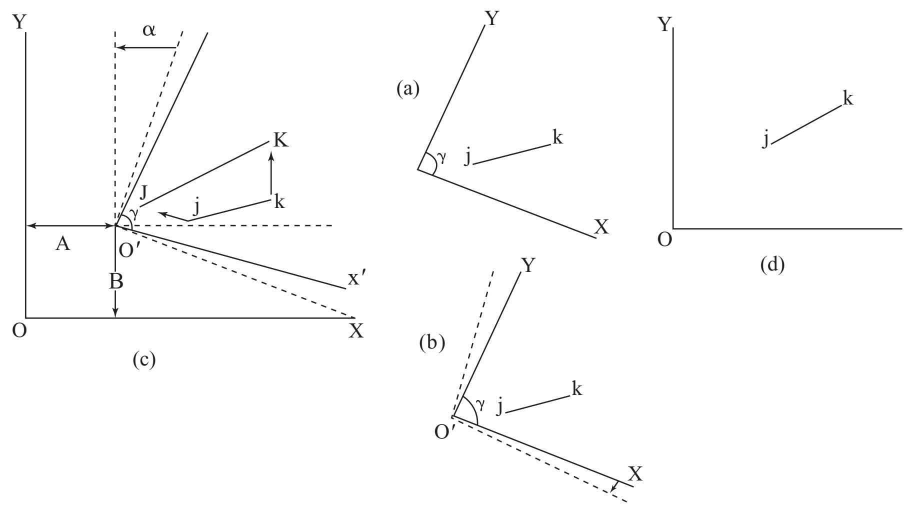


FIGURE 5.8 The geometry of affine transformation (a) Initial condition showing two points *j* and *k* referred to Cartesian axes O' X' and O'Y', which are not orthogonal, but intersect at some angle γ. Different scales in the directions of the two axes produce the same effect. (b) Creation of the orthogonal axes, shown here by the broken line. This is also equivalent to making the linear scale in the X' direction equal to that in the Y' direction. (c) All transformation stages are superimposed. The third, fourth and fifth stages comprise those in the Helmert transformation, i.e. the uniform scale change to represent *jk* by *JK* and the rotation of the axes through the angle a and finally, the translation through A and B to refer the points *J* and *K* to the OX and OY axes. (d) The final transformation illustrates this stage, where *J* and *K* are shown within the OXY system

The third-order polynomial expression relating grid to geographical coordinates may be written in the form

$$x = a_{00} + a_{10}\,\lambda a_{01}\varphi + a_{20}\lambda^2 + a_{11}\lambda\varphi + a_{02}\varphi^2 + a_{30}\lambda^3 + a_{21}\lambda^2\varphi + a_{12}\lambda\varphi^2 + a_{03}\varphi^3 \qquad (5.22)$$

$$y = b_{00} + b_{10}\,\lambda + b_{01}\varphi + b_{20}\,\lambda^2 + b_{11}\lambda\varphi + b_{02}\,\lambda^2 + b_{30}\,\lambda^3 + b_{21}\lambda^2\varphi + b_{12}\,\lambda\varphi^2 + b_{03}\varphi^3$$

Similarly, the polynomial equations used to transform from grid to grid are

$$x = c_{00} + c_{10}x' + c_{01}y' + c_{20}x'^2 + c_{11}x'y' + c_{02}\,y'^2 + c_{30}x'^3 + c_{21}x'^2 + c_{12}x'y'^2 + c_{03}y'^3$$

$$y = d_{00} + d_{10}x' + d_{01}y' + d_{20}x'^2 + d_{11}x'y' + d_{02}\,y'^2 + d_{30}x'^3 + d_{21}x'^2y' + d_{12}x'y'^2 + d_{03}y'^3 \qquad (5.23)$$

Before the advent of computers, polynomial expressions were usually left in this form because it was generally easier to compute each term individually. However, in view of what has already been said about economy of the design equations, a nested form of each equation may be obtained from a little algebraic rearrangement. For example, the expression for x in Eqn. 5.22 may also be written as:

$$x = a_{00} + \varphi(a_{01} + a_{02}\varphi) + \lambda(a_{10} + \varphi(a_{11} + a_{12}\varphi) + \lambda^2(a_{20} + a_{21}\varphi + a_{30}\,\lambda)... \qquad (5.24)$$

This example is particularly instructive. Also in addition, it has been proved that the savings, which result from using Eqn. 5.22 rather than the expression for x in Eqn. 5.22 are between 20 and 30 percent in the solution of a fifth-order polynomial. It is also proved that the conformal projections of the spheroid may be transformed more accurately by using isometric latitude ψ, in place of geodetic latitude in Eqn. 5.22.

5.12.3.1 Determination of the Polynomial Coefficients

In order to find the 20 coefficients namely a_{ij}, b_{ij} in the third-order polynomial above, it is necessary to know the plane rectangular coordinates of 10 corresponding points x_i, y_i, and φ_i, λ_i to form the linear equations from which the coefficients can be solved. The amount of data needed to determine the coefficients for a polynomial depends upon the order of the polynomial, which, in turn, depends upon the highest powers of the independent variables used in the terms. For example, the first, second, third, fourth and fifth degree polynomials require a minimum of 3, 6, 10, 15 and 21 corresponding points, respectively.

The common solution is to use even more than these minimum numbers, to obtain the required coefficients by the method of least squares. This is the condition that the sum of squares of differences between the measured and the theoretical coordinates in the new projection should be minimized. The following ($m \times n$) matrix solution is applicable for any number of coefficients, n, and common points m. It is well known in the numerical analysis that although a polynomial may be extended to include higher-powered terms in φ^4, λ^4, φ^5, λ_5, etc., the task of determining coefficients is time consuming. Hence, it is enough that increasing the degree of the polynomial from third order to fourth order barely justifies the greater accuracy obtained for any purpose other than the geodetic network.

In Eqns. 5.25 and 5.26, the individual coefficients from the column matrix on the left hand side and the control, or common point coordinates are the column matrix on the right hand side.

$$\begin{pmatrix} a_{00} \\ a_{01} \\ \cdots \\ \cdots \\ a_m \end{pmatrix} = D \bullet \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ \cdots \\ x_m \end{pmatrix} \qquad (5.25)$$

$$\begin{pmatrix} b_{00} \\ b_{01} \\ \cdots \\ \cdots \\ b_m \end{pmatrix} = D \bullet \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ \cdots \\ y_m \end{pmatrix} \tag{5.26}$$

The matrix **D** is calculated from

$$\mathbf{D} = [\mathbf{AT} \bullet \mathbf{A}]^{-1} \bullet \mathbf{AT} \tag{5.27}$$

where the $(m \times n)$ matrix **A** is formed from the geographical or grid coordinates of the corresponding points. Thus, for the third-degree polynomial requiring 10 terms per line, $n = 10$ is as follows:

$$A = \begin{pmatrix} 1 & \lambda_1 & \varphi_1 & \lambda_1^2 & \lambda_1\varphi_1 & \varphi_1^2 & \lambda_1^3 & \lambda_1^2\varphi & \lambda\,\varphi_1^2 & \varphi_1^3 \\ 1 & \lambda_2 & \varphi_2 & \lambda_2^2 & \lambda_2\varphi_2 & \varphi_2^2 & \lambda_2^3 & \lambda_2^2\varphi & \lambda\,\varphi_2^2 & \varphi_2^2 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 1 & \lambda_m & \varphi_m & \lambda_m^2 & \lambda_m\varphi_m & \varphi_m^2 & \lambda_m^2 & \lambda_m^2\varphi_m & \lambda_m\varphi_m^2 & \varphi_m^2 \end{pmatrix} \tag{5.28}$$

This method depends upon the size of the area mapped for its accuracy. This is because a polynomial transformation works well enough within a homogeneous data, but a file comprising data digitized from a paper map may not be homogenous, because different parts of it have been affected differently by paper deformation. Just as it is necessary to treat separately the panels for a map which has at some time been folded, it may be necessary to divide the whole map into blocks and transform each block separately.

5.13 FACTORS INFLUENCING THE CHOICE OF SUITABLE MAP PROJECTIONS

The principle and methods of transformation that have been described are applicable to maps of any scale. However, an application of a GIS to a large country to even a continent necessitates the choice of a projection, first to serve as the GIS framework and possible as a suitable projection for displaying the results. It is a fundamental principle of the distortion theory that the particular scale and, therefore, exaggeration of areas and angles increases from the origin of the projection towards its edges. Therefore, it is desirable to choose a projection in which either the average or the extreme distortions are small. The amount of distortion on a map depends upon the location, size and shape of the area to be mapped. Distortion is least in the representation of a small, compact country and greater in maps of the whole world. The three variables, namely, the location, size and shape usually determine the choice of origin, aspect and class of a suitable projection. These are based upon the principle that the distortion pattern, its fundamental property, remains constant within a particular projection even when the aspect of the projection is changed. Therefore, the plotted pattern of distortion isograms may be regarded as a frame, which can be used to imagine how the distortion will occur, just as an artist may compose a picture by looking at objects through a small rectangular cardboard frame, or a photographer uses the rectangular ground glass screen of the camera viewfinder.

In the pre-computer period when the methods were evolving, this was carried out using transparent overlays, which were placed singly, or in groups over a rough outline sketch map of the country or continent drawn at the same scale. By shifting the position and orientation of the overlay, it is possible to estimate many advantages to be gained from a change in origin or change in orientation of the lines of zero distortion. The actual choice of projection depends upon the comparison of the patterns of distortion isograms for different projections. When two or more overlays for different projections are superimposed, the extreme values for area scale or maximum angular distortion may be estimated from the isograms.

REVIEW QUESTIONS

1. What do you mean by a coordinate system? Why is it required for mapping purposes?
2. Write short notes on:
   (i) Plain Cartesian coordinates
   (ii) Plain polar coordinates
   (iii) Cartesian 3D coordinates
   (iv) Geographic coordinate systems
3. What is a geodetic data?
4. Explain astronomical coordinate systems.
5. Define geoid, spheroid and ellipsoid.
6. What do you mean by cartography? Explain its importance.
7. Write short notes on:
   (i) Analytical transformation
   (ii) Polynomial transformation
   (iii) Grid-on-grid transformation.

6 SPATIAL ANALYSIS

Chapter Outline

6.1 INTRODUCTION

Geographical information systems (GISs) are increasingly used for decision supports, which demands sophisticated data analysis tools to turn geospatial data into useful spatial knowledge and intelligence. The spatio-analytic capabilities of GIS use the spatial and non-spatial data in spatial database to solve various problems. The main objective of spatial data analysis is to transform and combine data from different sources into useful information, and to satisfy the objectives of decision makers. The problems in planning such as what should be the best location for a new port or the predictions such as what should be the size of basin of a dam are the typical problems that use the spatio-analytic capabilities of GIS. The solution for these kinds of problems depends on large number of parameters which are often interrelated, and their interaction should be made more precise in an application. Users of geospatial information are now able to access a wide variety of data analysis software applications. External data analysis software products can take advantage of the geo-processing tools in GIS to enhance their spatial analysis capabilities. The concept and techniques of spatial analysis are instrumental in changing the role of GIS from geospatial data management to special decision support in various organizations. Spatial analyses in GIS include basic data manipulation processes. The aim of spatial analysis was to select, combine or reformat existing geospatial data sets to generate new data suitable for answering specific spatial questions.

6.2 CLASSIFICATION OF ANALYTIC FUNCTIONS OF A GIS

The main classifications of analytic functions of a GIS are:

1. Measurement, retrieval and classification functions
2. Overlay function
3. Neighbourhood functions
4. Connectivity function

6.2.1 Measurement, Retrieval and Classification Functions

This function helps to explore data without making fundamental changes, and they are used at the beginning of data analysis. Measurement functions include computing distances between features and computation of area of two-dimensional features or volume of three-dimensional features with or without frequency of features. Spatial queries retrieve features using user-defined logical conditions. Classification means the assignment or reassignment of a thematic value and/or characteristic values to features in a data layer. Their measurement, retrieval and classification function are performed on single vector or raster data layer using associated attribute data.

Geometric measurements on spatial features include counting, distance and area size computations. Measurements on vector data are more advanced, and are more complex than measurements on raster data.

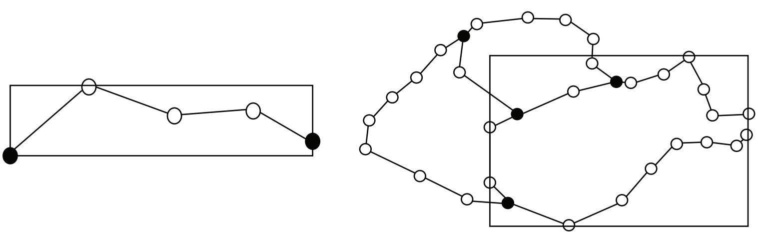
6.2.1.1 Measurements on Vector data

Basic vector data sets are point, line and polygon, and its geometric measurements are location length and area. Location length and area are geometric properties of a single isolated feature and distance requires two features to be identified. The location property of a victory feature is always stored by the GIS as a single coordinate pan for a point, or a list of pairs for polyline or polygon boundary length is associated with polylines or polygon boundary by themselves or in their function as polygon boundary. Area size is associated with polygon features and it can be easily computed, but usually stored with the polygon as an extra attribute value. This helps in quick computation of other functions that require area size values. Measuring distances between two features is another important funetion. The measuring distance function uses Pythagorean distance function. One can understand all of the above measurements by merely a look in stored data and does not require computation. Minimal bounding box applies on polylines and polygons and determines the minimum rectangle.

If both features are points say *a* and *b*, the computation in a Cartesian spatial reference system is given by the distance formula:

$$\text{Dist}\,(a,b) = \sqrt{(x_a - x_b)^2 (y_a - y_b)^2}$$

Another geometric measurement used by the GIS is the minimal bounding box computation. It applies to polylines and polygons, and determines the minimal rectangle with the sides parallel to the axes of the spatial reference system that covers the feature. It is illustrated in Fig. 6.1. If the bounding boxes of two polygons do not overlap, it means that the polygons cannot intersect with each other.



(a) (b)

FIGURE 6.1 Minimal bounding box computations

The use of area size measurement is done when one wants to sum up the area sizes of all polygons belonging to same class. The class may be the size of an area covered by sugarcane crop. If the crop classification is in a stored data layer, the computation would include, selecting the sugarcane crop areas and summing up their area sizes. Hence, geometric computations are required in the case of stored features. For example, consider the raster image data in which the distances between A and B can be determined using the geometry of ABC (see Fig. 6.3).

In a vector GIS, distances are measured using simple trigonometric/geometric properties. The vector coordinate describe points, lines or area available for calculating distances. The distance formula between two points ($X1$, $Y1$) and ($X2$, $Y2$) is SQRT $((X1-X2)^2+(Y1-Y2)^2)$). (see Fig. 6.2). The geometric formulae can be used to calculate perimeters and areas. For example, the area of the feature shown in Fig. 6.3 can be computed by calculating the area of triangle ABC using mathematical method.

Many vector-based systems measure distances along existing vector line networks, like streets, sewers and railroads. This type of distance measurement relies on topological network relationships. In addition, some vector systems automatically generate length measurements for line features entered. They store the length result in an attribute field within the layer's data file.

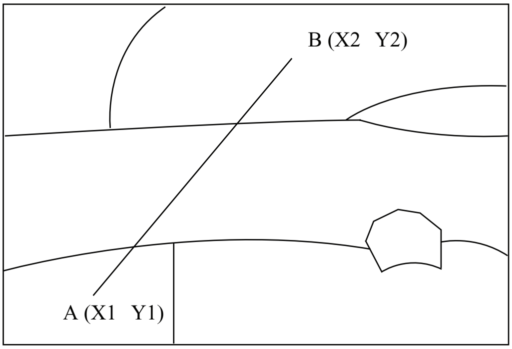


FIGURE 6.2 Area calculation of geometrical shapes

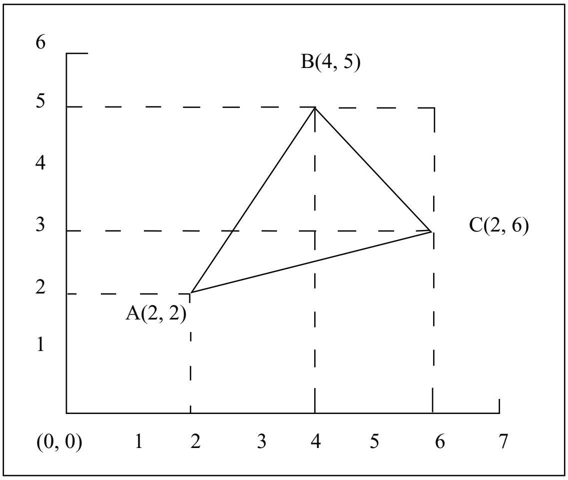


FIGURE 6.3 Distance calculation (vector)

6.2.1.2 Measurement of Raster Data

Measurement of raster data is simple due to regularity of cells and area size of cell is constant. The measurement depends on determination of the resolution of cells. The horizontal and vertical resolutions of cells are typically the same but may differ in some case. Due to the regularity of cells in the case of raster data, the measurements are simple on raster data layers. The area of a cell is constant and is determined by the cell resolution.

Location of an individual cell derives from the raster's anchor point, the cell resolution and the position of the cell in the raster (See Fig. 6.4). Anchor point is fixed by convention to be the lower left (or in some cases upper left) location of the raster. Again there are two conventions, the cell location can be its lower left corner, or the cell's mid-point.

In case of low-resolution data cell location become more critical. All geometric inhumation is stoned with the anchor point.

Area size of a selected part of the raster is calculated as the number of the cell multiplied with the cell area size. *Distance* between two raster cells is the standard distance function applied to the local locations of their respective mid-points. Length of line is calculated as the sum of distances between consecutive cells. In a raster image, the length is calculated using geometrical/trigonometric principles. Consider the raster image data in which the distances A and C can be determined (Fig. 6.5) by making use of geometry of the triangle ABC. Calculating the length AC in raster is by adding the number of grid cells together to achieve a total. One can begin by working with a straight line entry in grid format that occurs as a set of vertical or horizontal grid cells.

To obtain a perimeter measurement in a raster GIS, the number of grids cell sides that make up the boundary of a feature is multiplied by the known resolution of the raster grid (see Fig. 6.6).

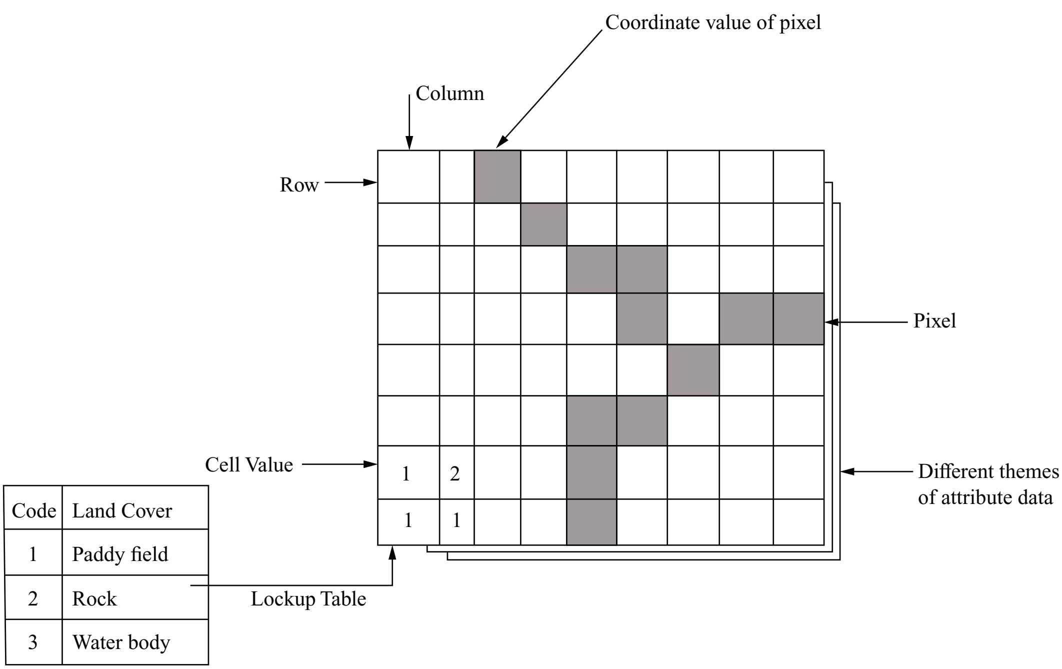
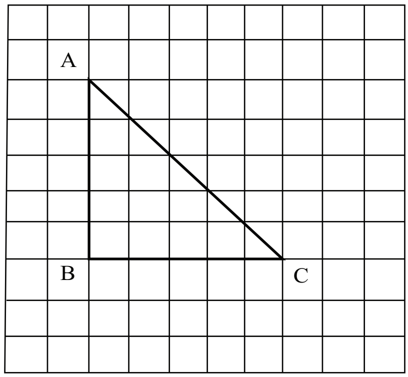


FIGURE 6.4 Measurement of raster data



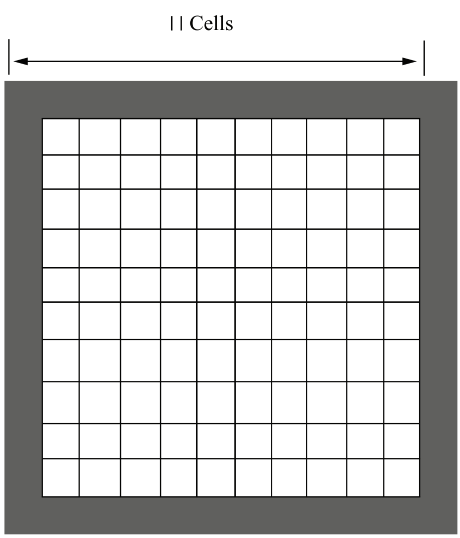
AB = 5 grid cells × 1 km/side = 5km

BC = 5 grid cell × 1 km/side = 5km

$AC = \sqrt{AB^2}+\sqrt{BC^2} = \sqrt{25 + 25} = \sqrt{50}$ km

FIGURE 6.5 Calculating the length AC by employing simple trigonometry

For area calculations, the number of cells, a feature occupies is multiplied by the known area of an individual grid cell (see Fig. 6.7).



Let each grid cell be 1km/side

Perimeter of grid cells = number of cells × 4 sides × 1km

= 11 × 4 × 1 = 44km

FIGURE 6 6 Calculation of perimeter of selected grid cells

6.2.2 Spatial Selection Queries

When exploring a spatial data set one has to select certain features to temporarily restrict the exploration. Such selections can be made on geometric/spatial grounds, or on the basis of attribute data associated with the spatial features.

a. Interactive spatial selection

In interactive spatial selection, we define the selection condition by pointing at spatial objects on the display, after having indicated the spatial data layers from which to select features. Interactive spatial

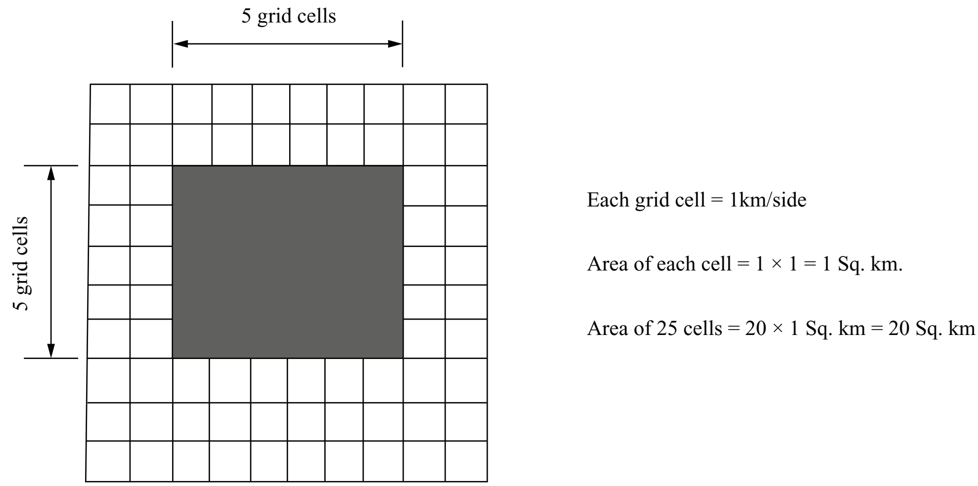


FIGURE 6.7 Area calculation

data analysis involves the use of software environments that permit the visualization, exploration and modeling of geo-referenced data. In interactive spatial selection, the interactively defined objects are called the selection objects, and they can be points, lines or polygons. The GIS than selects the features in the indicated data layers that overlap, intersect, meet or contained in with the selection objects. There become the selected objects. Spatial data in uonally associated.with its attribute data which is stored in tables through a key link. Selection of features through these links are lead to selection of records, or selection of records may lead to selection of features. One can select features by stating selection condition on the features attributes.

b. Combining attribute conditions

When multiple criteria have to be used for selection one need to carefully express all of these in a single composite condition. The tools for this come from a field of mathematical logic, known as propositional calculus.

c. Spatial selection using topological relationships

Various forms of topological relationship between spatial objects can be useful to select features as well, The steps carried out in containment, overlap, neighbourhood and at selections on the basis of distance functions are:

1. To select one or more features as the selection objects and
2. To apply the chosen spatial relationship function to determine the selected features have the relationship with the selected objects

d. Selecting features that are inside selection objects.

This type of query uses the containment relationship between spatial objects. It should be noted that, polygons can contain polygons, line or points, and lines can contain lines or points, but no other containment relationships are possible.

e. Selecting features that intersect

The intersect operators identifies features that are not disjoint, but extended to points and lines.

f. Selecting features adjacent to selection objects

Adjacency is the meet relationship. It expresses that features share boundaries, and, therefore, it applies only to line and polygon features.

g. Selecting features based on their distance

It may be necessary to use the distance function of the GIS as a tool in selecting features. Such selections can be searches within a given distance from the selection objects, at a given distance, or even beyond distance.

6.2.3 Classification

Classification is a technique of purposefully removing the detail form an input data set, for revealing important patterns of spatial distribution. In this process one produce an output data set, so that the input set can be left intact. This is done by assigning a characteristic value to each element in the input set, which is usually a collection of spatial features, either raster cells or points, lines or polygons. If the number of characteristic values is small in comparison to the size of the input set, then the input set is classified. If the pattern considered is monthly salary of the employees in a company, that is, income, then monthly salary is called the classification parameter. If income of the people of a city is considered, then one can define five different categories or classes of income such as low, below average, average, above average and high income and can provide value ranges for each category. If these five categories are mapped in a sensible colour scheme, this may reveal a good information.

6.2.3.1 User Controlled Classification

In user controlled classification one should indicate which attribute is, or which one are the classification parameters and should define the classification method. The classification method involves declaring the number of classes as well as the correspondence between the old attribute values and new classes. This is done using a classification table. Another case exists when the classification parameter is nominal or at least discrete. In this case one must also define the data format of the output, as a spatial data layer, which will contain the new classification attribute. The data type of this attribute is always categorical, that is, integer or string, no matter what is the data type of the attributes from which the classification was obtained. The table below is an example of a classification on a discrete parameter, namely, land use of a city. Different colour codes can be used in colour scheme, such as code 1, 2, 3, 4, 5, etc. represents different colours.

6.2.3.2 Automated Classification

User controlled classification require a classification table or user interaction. GIS software can also perform automatic classification, in which a user only specifies the number of classes in the output data set. The system automatically determines the class break point. Two techniques of determining break points are equal interval technique and equal frequency technique. In equal interval technique the minimum and maximum values are determined and the constant interval size for each category is calculated as (maximum value − minimum value)/n, where n is the number of classed chosen by the user. In equal frequency technique, the objective is to create categories with roughly equal numbers of features per category. The total number of features is determined first and then by the required number of categories, the number of features per category is calculated. The class break points are then determined by counting off the features in order of classification parameter value.

6.3 OVERLAY FUNCTION

Overlay is one of the most common and powerful GIS function. It investigates the spatial association of features by vertically stacking feature layers to investigate geographic patterns and determine location that meet. Specific criteria, standard overlay operators take two input data layers and assume they are geo-referenced in the same system, and overlap in study area. If either condition is not met the use of an overlay operator is meaningless. The principle of spatial overlay is to compare the characteristics of the same location in both data layers, and to produce a new characteristic for each location in the output data layer. In raster data, these comparisons are carried out between pairs of cells one from each input raster. In vector data, the same principle of comparing locations pair-wise applies, but the underlying computations rely on determining the spatial intersections of features one from each input vector layer, pair-wise.

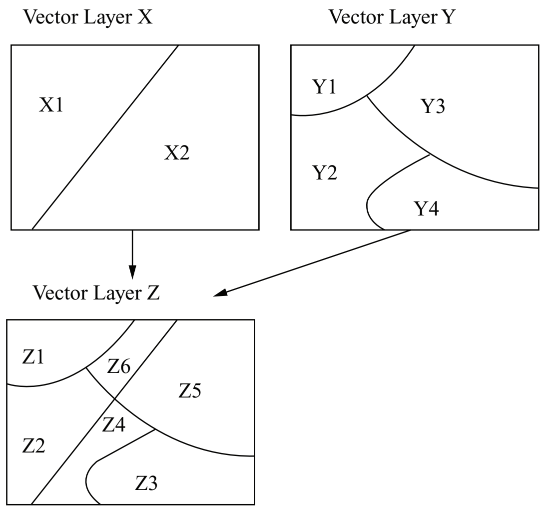
6.3.1 Vector Overlay

In the vector overlay domain, the overlaying of data layers is computationally more demanding than in the raster domain. The standard overlay operators for two layers of polygons are the polygon intersection operator. It is fundamental as many other overlay operators implemented in systems can be defined in terms of it. The result of this operator is the collection of all possible polygon intersections. The attribute table results is a join in the relational database of the two input attribute tables (See Table 6.1 and Fig. 6.8). Two polygon layers X and Y produce a new polygon layer with associated attribute table, which contains all intersections of polygons from X and Y. This operator is sometimes called as spatial joint.

Two or more polygon overlay operators as shown in Fig. 6.9. The first is known as the polygon clipping operator. It takes a polygon data and restricts its spatial extent to the generalized outer boundary obtained from all polygons in a second input layer. Besides this generalized outer boundary, no other polygon boundaries from the second layer play a role in the result. A second layer operator is

TABLE 6.1 The polygon intersect overlay operator

Code	Old Plan	New Plan
1.	Residential	Residential
2.	Industrial	Commercial
3.	Commercial	Commercial
4.	Institutional	Public
5.	Transport	Public
6.	Recreational	Public
7.	Non built-up	Non built-up
8.	Unplanned residential	Residential



X and Y Intersects		
X	Y	Z
X1	Y1	Z1
X1	Y2	Z2
X2	Y4	Z3
X2	Y2	Z4
X2	Y3	Z5
X1	Y3	Z6

FIGURE 6.8 The polygon intersect overlay operator

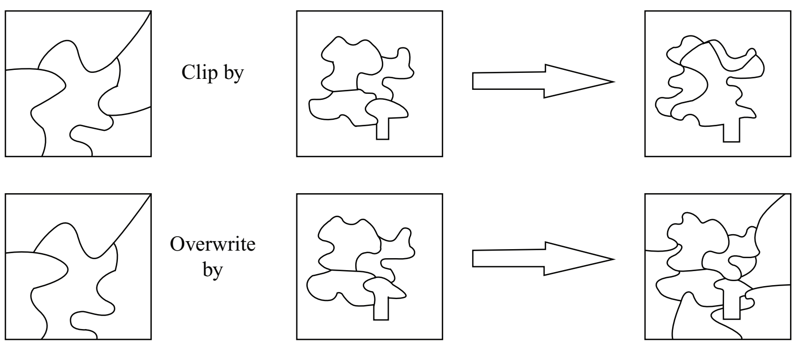


FIGURE 6.9 Two or more polygon overlay operators

polygon overwrite. The result of this binary operator is defined as a polygon layer with the polygons of the first layer, except where polygons existed in the second layer, as these take priority. Vector overlay of point in polygon and vector overlay of line in polygon are shown in Fig. 6.10a and 6.10b

Most GJSs do not force the user to apply overlay operators to the full polygon data set. One is allowed to first select relevant polygons in the data layer, and then use the selected set of polygons as operator argument. Vector overlays are also defined usually for point or line data layer. Different GISs use different names for these operators and the user is directed to check the documentation before applying any of these operators.

6.3.2 Raster Overlay

Raster overlay superimposes at least two input raster layers to produce an output layer. Each cell in the output layer is calculated from the corresponding pixels in the input layers. To do this, the layers must line up perfectly; they must have the same pixel resolution and spatial extent. If they do not align,

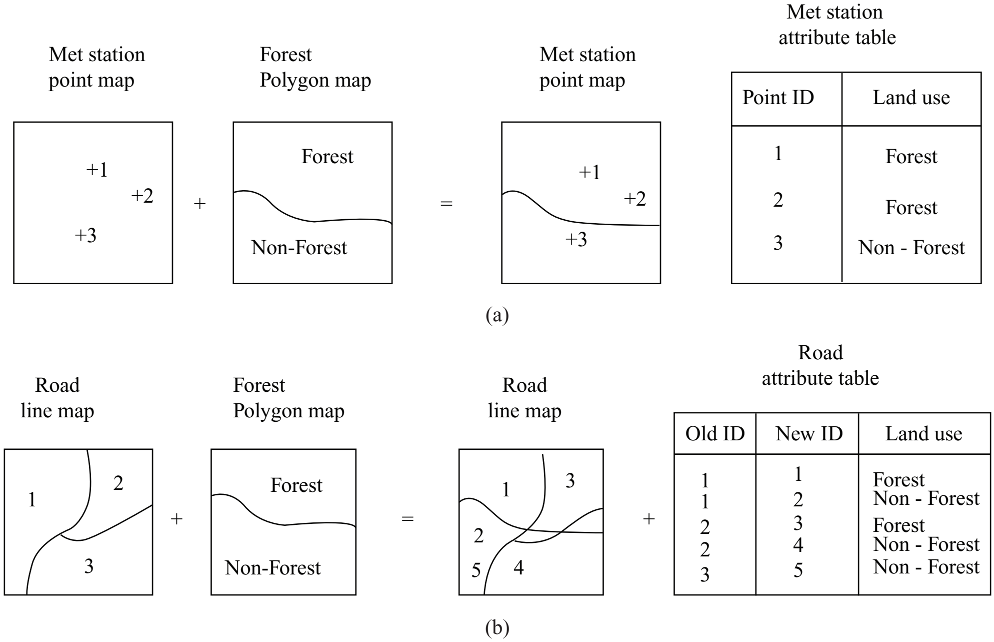


FIGURE 6.10 Example for vector overlay (a) point and (b) line in polygon

they can be adjusted to fit by the preprocessing functions. Once preprocessed, raster overlay is flexible, efficient, quick, and offers more overlay possibilities than vector overlay Vector overlay operators are geometrically complicated and this sometimes results in poor operator performance Raster overlays do not suffer this disadvantage, as most of them perform their computations cell by cell and thus they an fast.

Raster overlay frequently called map algebra is based on calculations which include arithmetic expressions and set and Boolean algebraic operators to process the input layers to create an output layer. The most common operators are addition, subtraction, multiplication and division, but other popular operators include maximum–minimum, average AND OR, and NOI. In short, raster overlay simply uses arithmetic operators to compute the corresponding cells of two or more input layers together, uses Boolean algebra like AND or OR to find the pixels that fit a particular query statement, or executes statistical tests like correlation and regression on the input layers (see Fig. 6.11).

6.3.3 Arithmetic Operators

Arithmetic operators allow the addition, subtraction, multiplication and division of rasters and numbers. The standard arithmetic operators used are multiplication, division, subtraction and addition (×, ÷, –, +,). These arithmetic operators have to be used only on appropriate data values and not on classification values. In Fig. 6.12, the result = DIV1 + DIV2.

Other arithmetic operators are modulo division (MOD), and integer division (DIV). In modulo division the reminder of division 13 MOD 4 will return 1 as 13 – 12 = 1, and 13 DIV 4 will give 3. Trigonometric operators are sine (*sin*), cosine (*cos*), tangent (*tan*) and their inverse functions *asin, acos* and *atan,* which provide radian angles as real values.

LAYER 1

3	2	0
1	2	3
7	1	0

LAYER 2

2	3	5
2	0	0
1	1	0

Add

5	5	5
3	2	3
8	2	0

Subtract

1	–1	–5
–1	2	3
6	0	0

Maximum

3	3	5
2	2	3
7	1	0

Layer 1>2 AND Layer 2>1

1	0	0
0	0	0
0	0	0

Minimum

2	2	0
1	0	0
1	1	0

Layer 1>2 OR Layer 2>1

1	1	1
1	0	1
1	0	0

Average

2.5	2.5	2.5
1.5	1.0	1.5
4.0	1.0	0.0

FIGURE 6.11 Raster overlay

DIV 1

1	1	0	0
	1	2	2
4	0	0	2
4	0	1	1

+

DIV 2

0	1	1	0
3	3	1	2
	0	0	2
3	2	1	0

=

RESULT

1	2	1	0
	4	3	4
	0	0	4
7	2	2	1

FIGURE 6.12 Arithmetic operators

In addition, there are five arithmetic functions. The *Abs* function gives the absolute value of the values in an input raster. The rounding functions, such as Round Up and Round Down, convert the decimal point values into whole numbers. *int* and *Float* convert values from and to integer and floating point values. The *Negate* (in Map Algebra, Unary-) function changes the sign of the input values by multiplying them by –1.

6.3.4 Logical Operators

The logarithmic functions perform exponential and logarithmic calculations on input rasters or numbers. The *base e* (*Exp*), *base 2* (*Exp 2*) and *base 10* (*Exp 10*) exponential capabilities and the natural (*Ln*), base 2 (*Log 2*), and *base 10* (*Log 10*) logarithmic capabilities are available. Raster calculus allows to compare rasters, cell by cell. For this we use the standard comparison operators

$$(= \leq \geq < > < >).$$

For example a sample comparison assignment is *"C := D<>E"* this will provide the truth values, either true or false, in the output cluster of *D*. Logical connectors are also supported in many raster assignments, such *as AND, OR, NOT, and exclusive OR* (*XOR*). The logical connectors or logical operators produce raster with the truth values such as *true* and *false.*

Condition expressions are used in overlays using decision tables, where multiple criteria must be taken into account. Some GIS accommodate setting up a separate decision table that will guide the raster overlay process.

6.3.5 Neighbourhood Function

The guiding principle of overlay operators is to compare or combine the characteristic value of a location from two data layers, and to compare or combine the value for all locations. There is another principle in a spatial analysis, where the principle is to find out the characteristic of the vicinity, called neighbourhood of a location. Many suitability questions depend not only on what is at the location, but also on what is near the location. Hence, the neighbourhood function allows to look around locally, and for neighbourhood analysis one must,

1. Define how to determine the neighbourhood for each target
2. Define which characteristics must be compared for each neighbourhood
3. Define which target locations are focused and what is their spatial extent

For example, if the target location is a police station, then its neighbourhood can be defined as an area within 5 km distance or an area within 2,500 m travel distance or an area within 15 minutes travel time or all residential areas, for which the police station is the closest one. The condition, that an area within 2,500 m travel distance is the one which indicate the characteristics to find out the neighbourhood and this can be the spatial extent. Hence one way of determining the neighbourhood around a target is by making use of the geometric distance function. Pollution and floods are defined by spread function. Spread functions are based on the assumptions that the phenomenon spreads in all directions, although it is not necessary to spread equally in all directions.

6.4 NETWORK ANALYSIS

Computations on network are completely different set of analytic functions. A network is a connected set of line, representing some geographic phenomenon. For example, take the case of transportation type, transported can be almost anything, people, cars and other vehicles along a road network, commercial goods along a logistical network, phone calls along a telephone network and so on. Network analysis can be done using either raster or vector data layers. Network analysis using vector data layers are most commonly used as line features and are usually associated with a network. The crucial aspect of a network is whether the network lines are considered directed or not. Directed networks associate with each line a direction of transportation whereas undirected networks do not. Planar network is one that can be shown in a two-directional plane. Many networks are planar, like river/road networks. A large scale networks are easy in computations as they have simpler topological rules. Various classical spatial analysis functions on networks are supported by GIS packages. The important ones are optimal path finding and network partitioning. The optimal path finding is the one which generates a best cost path on a network between a pair of predefined locations using both geometric and attribute data. Network partitioning is one which assigns network elements (nodes and segments) to different locations using predefined criteria.

REVIEW QUESTIONS

1. Define spatial analysis. What are the aims of spatial analysis.
2. Explain about the main analytic functions of a GIS.
3. Write a brief note on basic vector data sets.
4. How the raster data are measured.
5. Define classification and explain about the different methods of classification.
6. Define overlay and explain about vector overlay and raster overlay.
7. What you meant by arithmetic operators in spatial analysis?
8. Write short notes on arithmetic and logical operators.
9. Define neighbourhood function.
10. Write short notes on network analysis.

APPLICATION OF GIS 7

Chapter Outline

7.1 INTRODUCTION

A geographic information system (GIS) can be defined as a computer system capable of assembling, storing, manipulating and displaying geographically referenced information. Geographic data can be classified into two main classes, spatial data and attribute data. The two most popular types of spatial data are raster data and vector data. Raster data references spatial data to according to a grid of cells (or pixels), whereas vector data references spatial data to a series of coordinates. Attribute data are descriptions, measurements and classifications of the geographic features. Attribute data can be classified into four levels of measurement: nominal, ordinal, interval and ratio. The nominal level is the lowest level of measurement in which the data can only be distinguished qualitatively, such as vegetation type or soil type.

Data at the ordinal level can be ranked into hierarchies, but such differentiation does not show any magnitude of difference. Examples of this ordinal data include stream order and city hierarchies. The interval level of measurement indicates the distance between the ranks of measured elements, but a starting point is arbitrarily assigned. Ratio measurements of the highest level of measurements include an absolute starting point. Data of this category includes property value and distance. Time can also be considered as data element, because geographic information often changes over time, for example, changes over a particular piece of land due to construction and planning activities.

The three main components of a GIS are hardware, software and human resources. GIS hardware includes computers, computer configuration/networks, input devices, printers and storage systems. GIS-related software includes both the GIS programme and special application packages, such as digital terrain modelling and network analysis. Human resources used to operate a GIS typically include operational staff, technical professional staff and management personnel.

7.2 SOME APPLICATIONS OF GIS

7.2.1 GIS in Environmental Fields

Environmental fields have long used GIS for a variety of applications that range from simple inventory and query, to map analysis and overlay, to complex spatial decision-making systems. Examples include: forest modelling, air/water quality modelling and monitoring, environmentally sensitive zone mapping, analysis of interaction between economic, meteorological, and hydrological and geological change. Typical data input into an environmental GIS includes elevation, forest cover, soil quality and hydrogeology coverage. In many cases environmental GIS are used so that environmental considerations can be better incorporated into socioeconomic development enabling a balance between the two.

Biodiversity is receiving the attention of various scientists, planners and decision makers due to its importance as a natural reservoir with tremendous economic potential. Conservationists have focussed attention on this fast depleting resource. In situ conservation using the ecosystem approach is popular, which also protects various ecological services offered by the forest ecosystem. The emphasis is on identifying the most valuable biodiversity spots that harbour non-timber forest species such as endangered flora and fauna, medicinal plants and wild relatives of cultivated crops. While identifying such spots, it is also important to take into account the land use and human activities around the forest. Landscape ecology has benefited the most with the availability of spatial analysis tools like GIS.

Landscape ecology considers vegetation as a mosaic of patches of vegetation with unique landform, species composition and disturbance gradient and focuses on parameters such as patch sizes, patch shapes, patch isolation, relative importance of adjacent patches, and fragmentation. All these parameters have a direct bearing on the status of biodiversity within the forest ecosystem. Spatial analytical capabilities of GIS allow quantifying all the above-mentioned parameters with the remote sensing (RS) based vegetation type map. GIS was used to characterize the habitat of endangered animals using evaluating principles of landscape ecology.

7.2.2 GIS in Forestry

The use of GIS applications in the field of environmental research has grown exponentially in the last decade. Originally focusing on forestry, the tools have been used for a wide variety of environmental projects, from environmental research to forest fire prevention planning, to protecting endangered species. With the tools available today, and the increasing amount of data being compiled, the potential for new uses grows daily.

Forestry involves the management of a broad range of natural resources within a forested area. In addition to timber, forests provide such resources as grazing land for animals, wildlife habitat, water resources and recreation areas. The forest service is responsible for the management of forest harvesting, grazing leases, recreational areas, wildlife habitat, mining activities, as well as protecting endangered species. To balance the competing resource conservation and resource use, activities must be accommodated. Accessing the feasibility of these multiple uses is greatly enhanced by the use of GIS techniques. For example, the GIS for the Flathead National Forest in Montana includes digital terrain data, vegetation associations from Landsat satellite data, timber compartments, land types, precipitation, land ownership, administrative districts and the drainage network. GIS has been utilized for such analyses as timber harvesting, habitat protection and planning the location of scenic roads.

Forest fires have an important influence on the vegetation cover, animals, plants, soil, stream flow, air quality, microclimate and even general climate. The loss of timber is obvious and so is the damage to life and property. The loss of recreation value of the forest and the destruction of wildlife habitat are also the consequences of forest fires.

Researchers and scientists have long been trying to predict the behaviour of a forest fire. Computer modelling using high-resolution remote-sensing satellite imagery, powerful software and GIS has been the effort of many scientists. In order to model a forest fire, the techniques for obtaining, analyzing and displaying spatial information in a timely and cost-effective manner are needed. As forest fires are spatial, GIS is used as a tool for modelling. A fire simulation programme called FIRE has been developed using ARCINFO. The model puts the power of comprehensive fire behaviour prediction into the hands of qualified ground resource managers, where it can be most effectively applied.

7.2.3 GIS in Hydrology

Over the last two decades, the dramatic increase in computer power available to the hydrologist has led-to significant developments in the way that hydrological research and operations are conducted. Water resource applications of GIS are concerned with the hydrologic cycle and related processes. They are multi-faceted because

1. Many of the problems involve interactions between the hydrosphere, atmosphere, lithosphere and biosphere.
2. Solutions must serve competing groups of users.
3. Many of the important hydrologic processes have local, regional, national and global dimensions.

GIS technology allows for swift organization, quantification and interpretation of large quantities of geo-hydrological data with computer accuracy and minimal risk of human error.

7.2.4 Military Application

GIS plays a pivotal role in military operations, as they are essentially spatial in nature. The concept of command, control, communication and coordination in military operations is largely dependent on the availability of accurate information to arrive at quick decisions for operational orders. Military forces use GIS in a variety of applications including cartography, intelligence, battle field management, terrain analysis, RS, military installation management and monitoring of possible terrorist activity.

The use of GIS in the management of military bases facilitates maintenance and the tackling of all stores, which may be found on the base. GIS allows military land and facilities managers to reduce base operation and maintenance costs, improve mission effectiveness, provide rapid modelling capabilities for analyzing alternative strategies, improve communication and to store institutional knowledge. Global positioning system (GPS) provides the means of determining the position at sea.

7.2.5 GIS in Health Management

Public health management needs information on various aspects like the prevalence of diseases and facilities that are available to take decisions on either creating infrastructure facilities or for taking immediate action to handle the situation. These decisions have to be taken based on the observations made and the available data.

Most epidemiological data have a location and time reference. Advanced spatial analysis includes the combination of different data layers. Health authorities, for example, may be interested in the estimates of the number of children in a certain age group that may be exposed to malaria.

Climatic and topographic data can be used to determine the range of malaria mosquitoes. This range is unlikely to follow the *panchayat* union boundaries, but in GIS the two data layers can be combined to derive the number of children living within the affected areas in a particular *panchayat* union.

In short, the availability of statistical and other information in spatially referenced form and the functions provided by a GIS could allow analyses that were previously too expensive or impossible to perform. GIS is an innovative technology, ideal for generating data suitable for analysis, both with respect to space and time.

7.2.6 GIS in Geology

Another field where the GIS is useful is the branch of geology. Geological interpretation in one area, whether it is used for mineral and oil exploration or for engineering geological studies, is principally a compound process of various data. The geologist or explorer, linking different geological data, will seek for useful geological structure in an area. All geological data, to be fruitful and productive, should be interpreted considering their geographic locality.

The GIS provides presentation facilities and simultaneous data interpretation, enables the geologist or explorer to prepare mineral potential maps with various data more quickly and precisely, which was impossible to be carried out with analogue and traditional methods. The cartographic application of GIS in collecting, digitalizing and preparing geological maps in different scales and on the national scale (geological, air borne geophysics, mineral, geomorphology maps) are among the main applications of GIS for the numerous geoscientists.

7.2.7 GIS in Business

The use of GIS in business has greatly enhanced the efficiency in a number of areas, especially marketing research. Examples of the use of GIS in business include locating potential competitors, mapping market thresholds for retailers, providing computerized hazard information classifications, aiding risk management decisions in insurance companies, and enabling real estate agents to handle property data more efficiently.

Delivery services also utilize GIS in aspects such as navigation and monitoring of their fleets, routing optimization for shipping and deliveries, geocoding address matching and location searches. Typical data input in this category include road networks, street addresses, business profiles and socio-economic profiles.

7.2.8 GIS in Infrastructure and Utilities

GIS technologies are also widely applied to the planning and management of public utilities. Organizations dealing with infrastructure and public utilities find GIS a powerful tool in handling aspects such as planning, decision making, customer service, regulatory requests, standardization of methods and graphics display.

Typical uses include the management of the following services: electric, gas, water, roads, telecommunication, storm sewers, TV/FM transmitting facilities, hazards analysis and dispatch and emergency services. The typical data input includes the street network, topographic data, demographic data and the local government administration boundary.

7.2.9 GIS in Land Information

GIS has aided the management of land information by enabling easy creation and maintenance of data for land records, land planning and land use. A number of municipal governments, in particular, have started to implement GIS to help manage their land information. GIS makes input, updates, and retrieval of data such as tax records, land use plans, and zoning codes much easier than during the paper map era.

Typical uses of GIS in land information management include managing land registry for recording titles to land holdings, preparing land use plans and zoning maps, and cadastral mapping. Input of data into a land information GIS includes: political and administrative boundaries, transportation and soil cover.

7.2.10 GIS in Computer Cartography

The growth of computer-assisted cartography has been largely dependent on the development of vector-based GIS. With the help of GIS, cartographic tasks such as thematic overlays of information, map projections and map sheet layouts can be performed much more conveniently.

Continually updated geographic databases provide an easy way to produce new map editions. Automated map-making and virtual map images have replaced traditional paper maps in many applications. Web-based maps have made general purpose navigation far more accessible to the public. However, manually digitized paper maps remain the primary form of data input in an automated cartography GIS. Scanned maps are also often used.

7.2.11 GIS in Agriculture

Agriculture resources are among the most important renewable, dynamic natural resources. GIS software can help to integrate the data and can be used as a tool in precision agriculture applications. Agricultural survey is the backbone of planning and allocation of the limited resources to different sectors of the economy. Soil survey is an integral part of an effective agricultural research and advisory programme. It provides complete information about soils and is an inventory of the soil resources of the area. It gives the information needed for planning land use and soil management programme. RS and GIS are viewed as efficient tools for irrigation water management.

7.2.12 GIS in Archaeology

Archaeological applications of GIS have shown little inclination to address time-depth and historical factors. Several of the contributions are concerned with the use of GIS to manage and then analyze archaeological data that is not collected explicitly for GIS. Anthropology and archaeology are disciplines that collect and use geospatial data. The field of archaeology, in particular, is now quite advanced in its use of GIS-related technologies.

Human adaptation and living patterns are affected by a variety of phenomena occurring at all levels of observation, from local environmental and social conditions at the micro-scale to continent-wide phenomena at the mega-scale. GIS provides the potential to analyze affordable, continental-scale environmental data to study human adaptation patterns over large regions. A primary benefit of the GIS approach is that the data are available for a variety of research users, and the database is cumulative over time.

7.2.13 GIS in Fisheries

The dynamic nature of the marine environment presents many challenges to officials in charge of managing sustainable fisheries. Traditionally, attempts to extrapolate spatial structure from aquatic environments have relied on very limited samples, based primarily on catch information. With the advent of technologies such as GISs and RS, fisheries managers and commercial operations alike have access to information that helps them achieve their respective goals.

As these technologies develop, it will be necessary to frequently assess the resulting positive and negative effects for both the fishing industry and aquatic ecosystems. At present, the majority of studies involving GIS applications to fisheries have focussed on minimizing costs for the fishing industry. However, it is equally important that these technologies should be applied during stock assessments and policy determination. As the accuracy and availability of data sets increases, so do the potential number of GIS applications to fisheries.

The increase in applicable data is largely due to developments in the field of RS and satellite technology. Several countries have already incorporated these systems into GIS-based management plans. Buoys, satellites and other RS technologies produce information on a large geographic scale. Satellite systems such as the advanced very high resolution radiometer (AVHRR) and coastal zone colour scanner (CZCS) have produced data regarding the ocean surface temperature and colour (chlorophyll distribution), respectively. More recently, satellites carrying the sea-viewing wide field of view sensor and moderate resolution imaging spectro-radiometer (MODIS) have dramatically improved the spectral and spatial resolution of available imagery. Numerous fisheries applications for these data sets have been implemented or proposed for development.

7.2.14 GIS in Civil Engineering

GIS plays a vital role and serves as a complete platform in every phase of infrastructure life cycle. Advancement and availability of technology have set new marks for the professionals in the infrastructure development areas. Now civil engineering professionals are seeking help of these technologically smart and improved information systems like GIS for infrastructure development.

Planning: The major contribution of GIS in planning is to give us with an organized set of data which can help professionals to combat complex scenarios relating to the selection of site, environmental impact, study of ecosystem, managing risk regarding the use of natural resources, sustainability issues, managing traffic congestion, routing of roads and pipelines, etc.

Data Collection: Precise and accurate data is a very important factor of any successful project. GIS is equipped with almost all those tools and functions that enables user to have access to the required data within a reasonable time.

Analysis: Analysis is one of the major and most influential phases of infrastructure life cycle. Analysis guides us about the validity or correctness of design or we can say that analysis is a method which supports our design. Some of the analyses that can be performed by GIS are:

- Water distribution analysis
- Traffic management analysis
- Soil analysis

- Site feasibility analysis
- Environment impact analysis
- Volume or area analysis of catchment
- River or canals pattern analysis
- Temperature and humidity analysis

Construction: It is the stage when all layout plans and paper work design come into existence in the real world. The GIS helps the professionals to understand the site conditions that affect the schedule baseline and cost baseline. To keep the construction within budget and schedule GIS guides us about how to utilize our resources on site efficiency by:

- Timely usage of construction equipment
- Working hours
- Effects of seasonal fluctuations
- Optimizing routes for dumpers and concrete trucks
- Earth filling and cutting
- Calculation of volumes and areas of constructed phase thereby helping in estimation and valuation

Operations: Operations are controlled by modelling of site data and compared by the baselines prepared in planning phase. Modelling of site may be in the form of raster images or CAD drawings. These can help us to keep track of timely operations of activities. GIS can help to make a record of work that has been completed and can give us visualization in the form of thematic maps that will guide us about rate of operations, completed operations and pending operations.

7.2.15 GIS in Transportation Engineering

The application of GIS to a diverse range of problems in transportation engineering is now well established. It is a powerful tool for the analysis of both spatial and non-spatial data and for solving important problems of networking. Shortest path analysis is an essential precursor to many GIS operations. Engineers have worked on this and explored the use of fast shortest path algorithm on extensive road networks. They have evaluated the possibilities of optimization, in which the optimum routes, travel time, travel distance and cost for defined paths and for the optimum paths were determined for few transport services. Employing GIS techniques for route optimization is a good example of GIS application in transportation engineering. Finding out the optimal route for emergency services and that for time taken for service is on top most priority. Transportation engineers developed a genetic algorithm to optimize a bus transit system serving an irregularly shaped area with a grid street network. The total cost function is minimized subject to realistic demand distribution and street pattern.

7.2.16 GIS in Traffic Engineering

Road traffic managers normally organize this material by compiling the data into tables used in statistical analysis. However, as the volume of data is huge and multifaceted relations of data are involved, it becomes very difficult to administer them consistently. And here's where applying GIS technology to the structure of the database helps in handling the positions and properties of clients visually.

Applying GIS in transportation systems enables a broad collection of potential applications that are as varied as the area of transportation itself. Be it cars and trucks going down a road, trains moving along a track, ships sailing over the sea or planes soaring high above us in the sky, all have one thing in common, and that is they are objects moving alongside others in space. GIS can provide a valuable tool for handling these targets with regard to a spatially referenced framework, screening the routes as a transportation electronic network. As transportation direction is a spatial process, it can be handled efficaciously using GIS engineering science. Below are listed the gains of efficient transportation management:

- Effortlessness movement of traffic
- Time spent on roads by traffic can be reduced to minimum
- Improved mood of clients while driving
- Altered personal protection/roadway safety
- Efficient transport planning and designing

GIS has proved to be a constructive adjunct when envisioning the effect that traffic on roadways and for proposed new construction has on the environment. Using GIS for three-dimensional (3D) visual imaging may also facilitate solving disagreements that frequently take place when people from different engineering fields work together on a large project. Alterations in design can be made earlier than when the actual trouble evidences itself on-the-spot. A further growth of the application is the use of GIS in machines for educating drivers, parallel to simulators which are being used for planes and ships. This could prove to be particularly supportive for urgent situations where vehicle drivers are still new to a particular area or region.

The GIS can prove to be of immense significance in traffic operations. This is because incident and traffic direction organizations established on GIS offer numerous advantages for traffic operations managers. In reality GIS is valued significantly for its capability to function with high competence. GIS is competent at incorporating data feeds and apportioning "dashboard" or on-screen positions quickly. This characteristic makes it perfect for acquiring a complete picture of present traffic positions.

For example, traffic operations directors can monitor traffic blockages and other such data so that they are able to rapidly react to changes. They can also contribute their opinions with the public through internet sites and traffic information channels and modify the drivers' behaviour relative to information on road conditions, travel time, construction areas, lane or exit closures, adverse weather impeding traffic, etc.

7.3 GIS APPLICATION AREAS AND USER SEGMENTS

GIS are now used extensively in government, business and research for a wide range of applications including environmental resource analysis, land use planning, location analysis, tax appraisal, utility and infrastructure planning, real estate analysis, marketing and demographic analysis, habitat studies and archaeological analysis.

One of the first major areas of GIS application is in natural resources management, including the management of:

1. Wildlife habitat
2. Wild and scenic rivers

3. Recreation resources
4. Flood plains
5. Wetlands
6. Agricultural lands
7. Aquifers
8. Forests

One of the largest areas of GIS application has been in facilities management. Uses for GIS in this area includes locating underground pipes and cables, balancing loads in electrical networks, planning facility maintenance and tracking energy use.

Local, state and federal governments have found GIS particularly useful in land management. GIS has been commonly applied in areas such as zoning and subdivision planning, land acquisition, environmental impact policy, water quality management and maintenance of ownership.

More recent and innovative uses of GIS have used information-based street networks. GIS has been found to be particularly useful in address matching, location analysis or site selection and development of evacuation plans.

Globally, the GIS market is moving in several directions simultaneously. The GIS technology is becoming a little easier for the mainstream information technology (IT) to incorporate GIS into day-to-day applications. In the past, GIS has been, and even now largely continues to be, the preserve of the few who can afford it. Conventional products have been expensive and extremely hard to use for common users. In India, there are some segments that include a number of GIS application areas and business GIS, for example, telecom, water resources and the government sector.

7.4 CUSTOM GIS SOFTWARE APPLICATION

GIS technology is a cost-effective solution that when integrated effectively, will increase an organization's productivity, efficiency, and overall savings or profitability. But, in order to insure the success of any project, the necessary planning and resources must exist.

7.4.1 Custom GIS

Clients can understand the benefits and limitations they will experience in terms of the service to their customers, job efficiency and data accuracy by developing a pilot project to explore all aspects of implementing a GIS. The latest GIS packages continue to provide greater functionality and flexibility to the professional user, but oftentimes at the expense of the casual user.

If a client has a series of repetitive GIS tasks that need to be repeated accurately, then the client can save time and money and improve accuracy by having a custom interface designed. This makes the task easier for the user and overcomes the problems of intensive GIS training that may be necessary without a custom interface. In addition, a custom interface can be designed to integrate with spreadsheets, databases and graphics programs in ways that are not possible with GIS alone. GIS Integrated Solutions is a consulting and software development firm, specializing in GISs solutions. GIS Integrated Solutions has successfully provided consulting, software development, data conversion, training, digitizing and systems integration services for commercial firms and government agencies.

7.4.2 User Interface

It is often only after selection and installation of the GIS that users, who may not have been consulted in the process of choosing the GIS, observe that the GIS would perform very quickly, if only somebody knew the commands to type! Unfortunately, this insight comes too late. There is a need for a research for developing methods for formally assessing usability so that it can be introduced into the selection process; thus, forcing GIS vendors to pay attention to this aspect of their products.

Software for GISs is not normally sold as a one-size-fits-all solution, such as MS Word, but rather as a collection of spatial data processing tools; some GIS packages contain over 2,000 such tools. The absence of a concrete mental model of how the system functions, and without proper training, a user of a non-customized GIS will likely find use for only 5% of the available functionality, as has been demonstrated with the use of word processors. These scattered tools need to be assembled into a specific GIS application by experienced technicians/programmers. Application generation involves, among other things, grouping individual tools into functional subsystems, menus or macro commands. The GIS often comes with an internal macro language for this customization, and the most recently released versions allow customization in standard Windows development tools such as Visual Basic and PowerBuilder. Many of the tools in the GIS are interface-related tools: menus, dialogue boxes, pointing and data input device support, etc. It is the experienced technician's job to translate end-user needs into workable solutions given the GIS toolbox selected.

7.5 USABILITY ENGINEERING IN THE GIS DOMAIN

Usability engineering in general is highly developed. Usability testing originated for safety reasons, that is, how can a fighter plane perform optimally and be safely piloted; same for nuclear reactors, cars, etc. Today it is common practice to perform usability tests even with products and services, which are not safety critical, that is, office software. Appropriate methods are available, which can be applied to the GIS domain.

The discussion about GIS-specific usability issues has started in the last few years. However, end-user and customer requirements are not sufficiently taken into account in the development process of GIS applications. Usability engineering in the GIS domain can be best described as reactive rather than being proactive, for example, a GIS version 1.0 is released, and then subsequent versions are changed according to user complaints. Representatives from GIS developers and companies that customize GIS applications (systems integrators) were interviewed. Some of the interesting results include:

1. GIS users were interviewed for their opinions on the current state of usability engineering within their organizations. The amount of time they claimed to spend on this activity during GIS project development varied from 10% to 80%, depending on how the term was defined and extended. They responded that little time was spent on usability issues.
2. The objectives of GIS evaluation considered were cost/benefit for the customer, conformance with minimum requirements, user acceptance and comparison with competitive solutions.
3. Success factors for GIS user interface development and customization were user satisfaction, ease of learning and training, predictability, sales, aesthetic and minimalist design.
4. Quality factors considered important for the users were:
    a. User problems, efficiency of task performance, robustness, error frequency

b. Learning cost and information content
c. Workload visibility of system status, users' opinions.

These results justify the necessity of these best practice guidelines. Often, quality-of-use evaluation and assessment starts during customization, or when the GIS application comes into use. Usability is simply not being considered sufficiently before the initial system is built and delivered to the end-users.

7.6 IMPORTANT GIS USER INTERFACE ISSUES

This section describes the major issues of GIS user interfaces. The focus here is on quality of use, not functionality. However, the quantum of functionality is a usability issue. More functionality puts additional burden on the user, because the user may need to select functions from a larger list of items. More functionality may force the user to search out tricks and shortcuts during the execution of tasks, if one is able to execute these tasks at all.

7.7 GEOGRAPHIC VISUALIZATION

As the name indicates, geographic data are graphical. But the prefix *geo-* is added not necessarily because many GIS experts are geographers/geologists, but rather because the graphics managed by GIS depicts data of a certain spatial scale, geographic scale. This leads to interesting problems or situations not encountered in the neighbouring worlds of CAD or CAM, and others that are common to the two. Some of these are enumerated below.

1. *GIS displays wide regions on a small screen.*
2. *GIS allows navigation in large spaces.* Unlike the bird's eye, overall view map of an area, the user often deals with only a part of a large-scale space. This is related to the point above. It is not uncommon for the user to get 'lost' when zoomed into a small area without reference text (e.g., place names).
3. *GIS may support various methods of zooming and panning.* About 10% of time is spent on these tasks during data entry, which are related to the navigation of the previous point. Furthermore, it was found that optimization of zooming and panning can save a good deal of time (money) over the course of the average data entry, (digitizing) project.
4. *GIS permits different views with scale change.* The cartographic display should be able to display certain graphic elements at certain scales or levels of resolution (or zoom). Of all the objects in a GIS database, the questions from a database manager may be: 'which are to be shown at 1:10,000 (town) scale and which at 1:10,00,000 (nationwide) scale?'
5. *GIS may include reference maps.* This reference or digital landmark is now common in www.navigation.
6. *GIS allows customization of the view.* Users have certain control over the changing of views, including parameters to hide/show information, create perspectives (3D), etc.
7. *GIS permits transformation from 2D to 3D representation.* The high-end GIS packages generally offer perspective views, but these may require different interfaces. Most users of GIS

do not need true 3D, because they do not deal with solids. Terrain models are not true 3D (2.5D), because one does not look inside the mountain, but only at the surface. The transition from 2D to 3D is a very complex procedure and thus costly.

7.8 GEOGRAPHIC QUERY LANGUAGES

The interface encompasses more than just the graphic look and feel of a system following the windows, icons and mouse and pointer (WIMP) model. Users must also be able to express ad hoc queries and other orders, and must be able to receive information from the system using a system specific language. These languages are usually modelled after, or are standard relational database manager languages (based on SQL) and, as such, are not at all optimized to geographic elements or problems. Despite the presence of several prototypes, a true spatial SQL still does not exist, though some useful extensions are present in SQL3. SQL is designed for one-dimensional (tabular) operations, in what might be called 'name space'. Queries such as select all the cities *near* this spot (<point with the mouse>) are not handled well, or not at all, in most GIS, because of the query language's lack of support for fuzzy concepts such as near, and for x, y coordinate input from peripheral devices.

Another often-overlooked characteristic of the GIS user interface (query language, commands and messages) is that the interface normally is in English, primarily for commercial reasons. Although different languages express spatial relations differently, in most cases, a GIS user must communicate with the system using English terminology. The customization process can remove this impediment. However, some user sites prefer to keep English terms, so that the system coincides with the user manuals and training materials. A yet unanswered question is whether or not the technical language of GIS (command names) is really understood by non-native English speakers, or if it is simply memorized and repeated mechanically, and whether translation of all system concepts to native languages would significantly improve usability.

7.8.1 Compatibility and Portability of Systems

Apart from internal interface concerns, one should also consider system-to-system interfacing. This is more an integration and a database problem than a direct manipulation problem, but it figures centrally in the workflow of many users. Today's GIS normally cannot be thought of as an island, isolated from the rest of the organization; it must function as a node in larger systems.

The GIS customization process must account for this integration, and make sure that the proper gateways are included, to connect, for example, to a corporate mainframe system running a different operating system, file system, data encoding system and DBMS. From the user's point of view, any significant time spent on manually making and adjusting these system-to-system connections is the time that is stolen from his/her real professional tasks (i.e., terrain modelling, flood control, urban planning).

7.8.2 Future GIS User Interfaces

Fifteen years ago, very few computing or GIS experts predicted the graphical user interface revolution brought about by the X-Window and MS-Windows environments. Equally, we cannot predict with certainty what the next generation of GIS user interfaces will be, but we can address the strong trends that we see today. There are several technological trends to be expected to continue in the near future.

7.8.3 Internet Use

This has wide-ranging implications, because it is already transforming the architecture of GIS software, where instead of a dedicated workstation and GIS package, some users are able to utilize www (world wide web) browsers (at zero cost) to access server-based geographic information. These distributed applications require simple user interfaces to support only visualization and interaction (query) tasks, and they assume that the processing will be handled at the server side.

7.8.4 Object Orientation

Together with the Internet, object orientation (as a new software development paradigm) is changing the way GIS software is developed, customized and used. The user interface of the future will need to greatly facilitate the customization process, allowing users to pick and choose the software modules (objects) needed at any given time. Note that this topic is directly related to the standards of the Open GIS Consortium, especially their distributed object model.

7.8.5 Portable Computing

The distributed architecture and object orientation described above is leading to GISs, which are more portable. The former allows for the various modules and data to reside in various sites on the network, and the latter allows for a portable computer to load and run only the specific modules needed at any moment. Portable software applications will require compact and simple interfaces, due to the small size and processing capability of the (hand-held) hardware units.

7.8.6 Real Time Access to High-Resolution Satellite Data

Increased processor speed, higher disk volumes and the immanent availability of near-real time satellite imagery will allow future GIS to monitor quickly changing conditions, such as in traffic or forest fire applications, instead of simply displaying historical data (maps produced months earlier). This will require interfaces capable of filtering huge amounts of information and perhaps allowing users to view only the changes as the last dataset was downloaded. Resuming, the future of GIS seems to be in increasing portability, modularity and flexibility. User interfaces must adapt to this trend, and allow better and faster user-customization.

7.9 GUIDELINES FOR THE PREPARATION OF A GIS

Preparing a GIS involves the following:

1. Conduct a needs assessment, define proposed applications and objectives
2. Execute an economic analysis for GIS acquisition
3. Select among alternative systems and equipment
4. Establish a database

Benefits of a GIS may be so compelling, that the decision to acquire a system can be made with little hesitation. In most cases, however, the decision can only be reached after a thorough analysis. The following section introduces a systematic process for reaching a decision about acquiring a GIS.

Potential users must remember that a GIS is not always the right tool for a given situation, and it may not necessarily pay for itself.

Conduct a Needs Assessment, Define Proposed Applications and Objectives

Before deciding whether to acquire or use a system, planners need to make a meticulous evaluation of their GIS needs. This must include a definition of how their planning activities and decisions will be assisted by using a GIS. Specific objectives and applications of the GIS should be defined. Answers to the questions outlined below can help.

The following questions needs to be formulated to assess the need of a GIS:

a. What planning decisions are needed to be made?
b. Which decisions involve the use of mapped information and information susceptible to map display?
c. What information cannot be managed efficiently with manual techniques?
d. What information management activities will the proposed GIS support?
e. What are the number and types of decisions that will be supported with a GIS?
f. Is the GIS principally for analysis?
g. Is cartographic quality output needed?
h. To what extent will a GIS help achieve the desired objectives?
i. Who will be the users of the information generated with a GIS?
j. How many user groups will there be?
k. In terms of information, time, and training needs, what is required to obtain the desired results?
l. Are there budget and staff support?
m. What agencies are participating in similar projects?
n. To what extent would a GIS help to attract the interest of other agencies and facilitate cooperation?

The following questions help in evaluating the suitability of an available GIS:

a. What kind of system is it?
b. What hardware and software are used?
c. Are its capabilities compatible with the needs of the new users?
d. Is the in-house technical expertise capable of serving the new users?
e. What are the institutional arrangements that would allow the use of this GIS?
f. Who are the current users? To what extent is the current user network compatible with the network: envisioned?
g. What data does it contain? To what extent does the data presently in the system cover identified needs?

If this preliminary investigation indicates that obtaining and using a GIS is a good option for an agency, it should seek the most cost-effective method of doing so. A frequently neglected option is to determine if an existing system is available. If the existing GIS is underutilized, the current owner might find a time-share offer attractive, particularly if the new agency brings data and analyses to the partnership. If no suitable GIS exists, another alternative is a group of agencies to establish a GIS that meets their common needs. Obviously, the trade-off in both these options is lower cost versus independence of action, but if the partnership also brings improved working relationships and compatible data to a group of agencies that work on common problems, these benefits may exceed the independence cost. The questions above offer planners some guidance as to whether an existing system is suitable to their needs.

Another opportunity for reducing investment cost is the use of existing equipment. If a computer is available, is it compatible with the GIS envisioned? What are the economic and institutional costs of time-sharing and inconvenience?

The key elements needed for the planning of a GIS lie in the following factors:

a. The software purchase cost.
b. The hardware configuration needed to fit the software requirements.
c. The need of a new computer and the options has to be included or the cost of acquiring a new computer versus upgrading an existing one.
d. The anticipated hardware repair and maintenance, and software support costs.
e. The personnel requirements for the installation and operation of a GIS.
f. The question of whether existing personnel can be used or new personnel need to be hired. The need of a computer programmer and the anticipated cost for training.
g. The cost of allocating personnel to hardware and software maintenance.
h. The expected cost for the data input process. The number of staff needed to be hired or assigned to digitize the information.
i. The cost involved in maintaining the data generated for and by the system.
j. The availability of secure facility suitably equipped for protection of computers and data files.

The key elements needed for the benefit calculation depends upon the following factors:

a. The production or revenue losses mostly associated with lack of information. The comparison of the above with the information that would be available if a GIS was present.
b. The costs savings from substituting labour-intensive drafting process with a GIS.
c. The benefits of integrating more timely information in the decision-making processes, and the ability to perform sensitivity analysis development plan options.

Once an agency has reached tentative decisions to acquire GIS capability, alone or in partnership, it should undertake an economic analysis of the proposition.

Execute an Economic Analysis for GIS Acquisition

Acquiring a GIS is a capital investment that may represent several thousand US dollars. As contended by Sullivan (1985), standard investment appraisal methods can be applicable to information technologies such as GIS. The questions mentioned above will help planners to roughly estimate and compare the major cost and benefits associated with a GIS acquisition.

The cost of maintenance and repair of all components of a GIS must also be considered in the investment analysis. The more sophisticated the system, and the more remote the home base of the operation, the higher its maintenance cost. Software demands maintenance, and arrangements should be made to subscribe to effective support from the provider of the software. The hiring of expertise to modify the software according to the project should be expected. A GIS is a dynamic tool; there will always be new data and new capabilities to be added, requiring additional efforts and expenses.

Select among Alternative Systems and Equipment

When a new system must be established, planners must carefully select the appropriate hardware and software. The system should be simple and must, of course, fit the budget and the technical constraints of the agency. Large digitizers and plotters, which are capable of producing maps of cartographic quality, are expensive and difficult to maintain. Small equipment, which can be as effective as the larger models for map analysis, is becoming increasingly available at affordable prices.

There are many GIS packages available, some more expensive and more powerful than others. Some cheaper software have good analytical capabilities, but lack computer graphics. Based on objectives, budget and personnel constraints, planners should investigate the alternatives for GIS software with a simple interface, strong analytical and graphical capabilities and an affordable price. Regardless of the selection, GIS software must be tested, and its claims must be verified against the needs of the user. As the software for GIS projects can cost more than the hardware it is designed to run on, the testing should be done on the hardware configuration to be used.

The systems are ranked by cost, and information is provided on the type of operating system, type of output device supported (directly related to the kind of output maps produced, raster or vector) and other capabilities such as area measurement, statistical analysis and geo-referenced overlaying.

Establish a Database

A database is established for:

a. Determination of proposed applications of the system
b. Determination of data needs and sources for the applications selected
c. Design of the data files

Once the GIS has been acquired, an information system must be designed. Typically, first-time GIS users tend to put lots of seemingly appropriate data into the system, trying to develop some application immediately. Usually, systems designed on a data supply rather than on an information-demand basis result in a disarray of data files and a chaotic and inefficient database.

A systematic approach to building an efficient and practical database includes a careful determination of users' needs, defining intended applications of the needs, and, if possible, a design evaluation and/or testing in a pilot study (see the GIS design procedure outlined in Fig. 7.1).

Determination of Proposed Applications of the System Small planning agencies or specific hazard mitigation projects may need a simple analysis of what has worked elsewhere to define what the GIS will be used for and what products it is expected to produce. Large organizations or more comprehensive projects, however, need to develop a standard and systematic approach, usually requiring interviews with management, users and existing system support staff. Answers to the questions boxed in Section 7.9 can orient planners in identifying potential applications.

Determination of Data Needs and Sources for the Applications Selected Data on natural hazards, demographic data and location of population are the prime concerns of natural hazards management and should be defined very early in the process.

Infrastructure and settlement sites provide the logical links that make a GIS useful in identifying population locations. When this information is combined with recent data detailing changes in land use, a clear understanding of where the people are located and the kind of activities they are undertaking and how they may be affected by natural hazards can be obtained. With this information, disaster prevention and preparedness actions can be initiated.

Criteria to be considered when planning for a GIS acquisition are the hardware, software, cost and vendor support:

1. **Hardware:** The hardware criteria consist of CPU/system unit and features and peripherals. The CPU/system unit includes microprocessor, memory capacity (RAM), disk drives, backup system, expansion capacity of memory and other peripherals, I/O channels, communication ports, compatibility standards and the warranty terms of hardware. Features and peripherals include keyboards, monitors (terminals), printers, power supply and networking capacity.
2. **Software:** The software criteria consist of system software, utilities software and application software. System software criteria consist of capability of system software, flexibility and compatibility of system software, expandability of system software, special features and documentation of system software.

   Utilities software criteria consists of ease of use and integration of utilities software with the system, diagnostics and peripheral control of utilities software and languages. Applications software criteria consists of appropriateness of application software to needs, performance such as speed, capacity and flexibility of application software, interface capability and upgrade potential of application software and the documentation and training with other user services available for the application software.
3. **Cost:** This criterion consists of initial investment for the CPU, monitor and printers, cost of additional components such as digitizers, adapters and other peripherals, transportation and delivery charges, installation charges, upgradation costs and training costs.
4. **Vendor support:** This criterion consists of maintenance and training. The maintenance criterion consist of maintenance of staff, the existing customer base, the service facilities, the inventory of components, the guaranteed response time and the capacity to deal with the entire system. The training criterion consists of experience of staff, the facilities available and the range of courses offered.

   Once the information requirements are identified, sources that will provide this information should be distinguished. Usually, a number of first-hand sources of information already exist, including maps and other documents, field observations and remote sensors. Table 7.1 lists the usually available natural hazard information that can be incorporated into a GIS data file.

   In concept, GIS programmes should be developed to accept all kinds of data that will eventually be needed. Data may be available in the form of satellite images, weather satellite data, aerial photographs, generalized global or regional topographic or soils maps or population distribution maps. Data such as these are sufficient to build an in0itial GIS. Once the framework is developed, new items can be added at any time.

Design of the Data Files The next step is to design the cartographic layers to be entered into the system, and the spatial attributes to be assigned to them. In this regard, detail of the database, input scale and resolution must be considered. Cartographic layers are the different 'maps' or 'images' that will be read into the system and later overlaid and analyzed to generate synthesis information. For example,

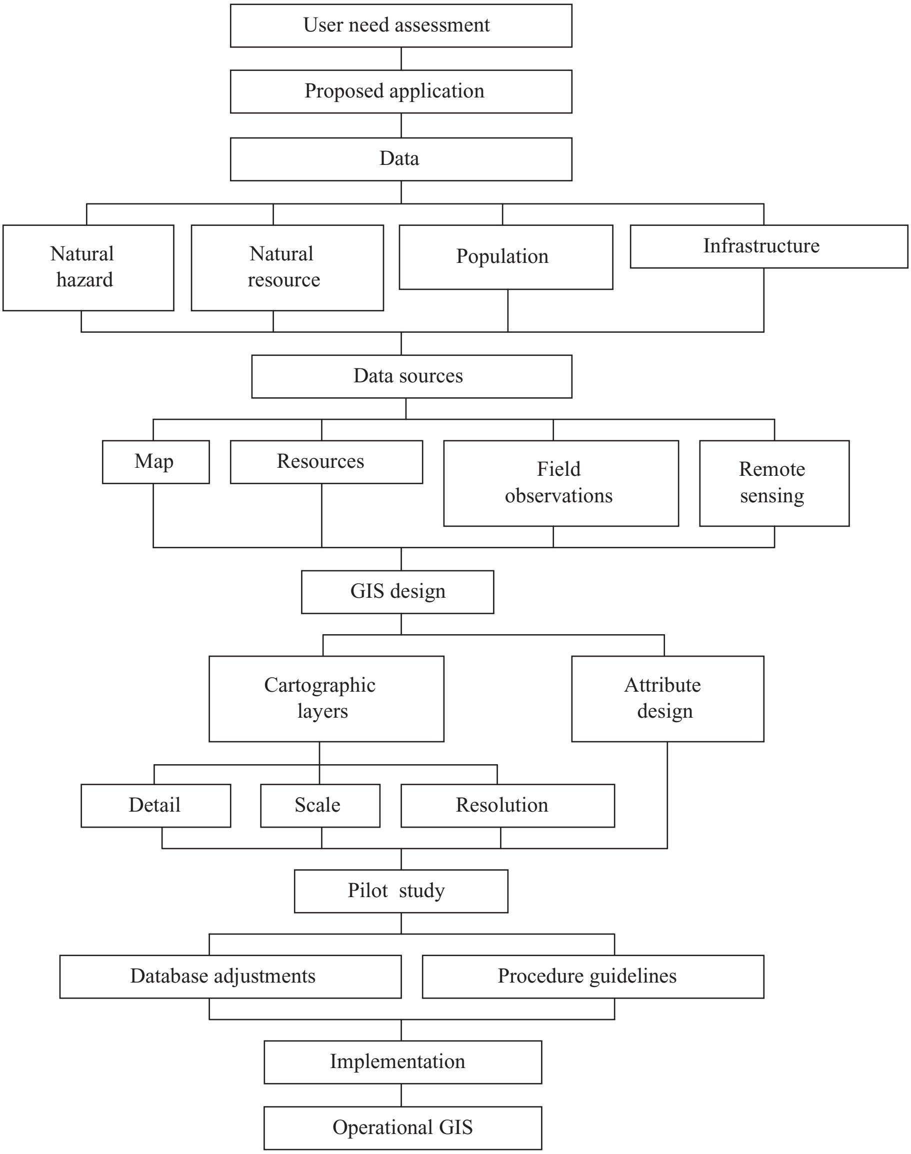


FIGURE 7.1 GIS design procedures

cartographic layers depicting past landslide events, geological characteristics, slope steepness, hydrology, and vegetation cover were entered and overlaid in a GIS to create a landslide hazard map, as described in Table 7.1.

There are three basic types of layers, and many different possible combinations among them: polygons (flood plains, landslide hazard areas), lines (fault lines, rivers, electrical networks) and points (epicenters, well locations, hydroelectric facilities). Selection of the correct layer type for a database depends on anticipated uses and on the scale and resolution of the source data. A volcano,

TABLE 7.1 Natural hazard information to be used in a GIS

Natural Hazard	Baseline Data	Intermediate Thematic Information	Synthesis Information
Earthquake	Epicentres	Maximum recorded intensity, magnitude	Seismic zoning (strong ground, motion data, maximum expected intensity or magnitude, recurrence interval)
	Fault lines	Frequency distribution and gap data	
	Plate boundaries		
Volcano	Volcano location	Previous event impact	Potential affected area (due to lava, pyroclastic flow)
		History of eruptions	
Hurricane	Landfall map	Previous event impact	Design event (surge tide elevation and flood elevation)
	Precipitation	Landfall frequency distribution	
	Wind		
	Coastal infrastructure		
Landslide	Bedrock geology	Previous event impact	Hazard susceptibility
	Slope		
	Vegetation		
	Precipitation		
Flood	Precipitation	Previous event impact	Design event (flood elevation and recurrence interval)
	Stream flow	Maximum stream elevation	
	Flood plain boundaries		
Desertification	Soils	Zones	Hazard zoning
	Precipitation	Aridity	
	Evapotranspiration	Erosion	
	Biomass production	Population density	
	Vegetation cover	Animal density	
		Land use	

for example, may be represented as a point at 1:250,000 scale, but it could well be a polygon at 1:20,000. Similarly, flood-prone areas may be represented as lines bordering rivers at scales smaller than 1:50,000, but as polygons on 1:10,000 scale maps. Planners must keep in mind that point and line representations may well be used for depicting variable locations, but they are seldom used for GIS operations involving cell measurement.

Spatial attributes are identifiable characteristics of the resource information assembled for the GIS. For example, attributes considered for infrastructure can include roads, bridges and dams. For land use, the different land use map units can identify the attributes.

All GIS input data are filed as attributes and can be recovered as individual items or aggregated into groups. A soil map provides a good illustration of attribute designation. One attribute in the soils 'layer' of data would be sand. All occurrences of sand would be located on the map. Once the attribute has been recorded, relevant descriptive material from the accompanying text should be included in the database, not just the legend. This greatly expands the usefulness of the information available to planners.

This same procedure, when used to prepare data for more than one point in time, provides the user with the information needed to measure changes overtime. The most frequent failure of time sequence data is due to the lack of details on the description of the attribute for the different time periods. Thus, it is important to include that information in text form within the GIS system.

Many attributes in some of the well-known and frequently used mapped information sources can provide ample information for hazard management in the typical GIS. Six particularly useful sources are:

1. Landuse and soil surveys
2. Climatic data
3. Location of volcanoes, landslide areas, and major geological faults
4. Natural features (rivers, flood plains)
5. Human features (infrastructure, population),
6. Topographic information (which provides elevation data, terrain complexity and watershed information)

Natural hazard management decisions based only on the above six sources of data can serve the GIS requirements in many situations. As an example, soil information can provide saturation and run-off characteristics; topography provides watershed area and topographic relief, and combined with soil data can help to identify flood plains; climatic records are particularly useful when combined with run-off characteristics from the soil survey to provide information on flooding and erosion; and life zone maps are useful in assessing desertification hazards. The number of people residing in a flood plain, what urban support centres exist, the location of roads, airports, and rail systems, can all be put into the system and analyzed in map form. This information is also useful in the preparation of emergency response plans.

The correct combination of attributes for particular decisions based on a GIS may call for a surprisingly small number of data input sources. Almost all natural hazard situations will be strongly influenced by one or two combined features. Mud slides, for example, usually occur in areas having steep terrain and soils high in clay content. New volcanic eruptions are most likely to occur in areas of historically high seismic activity. Planners or GIS users must understand that the purpose of a GIS is not to procure and incorporate all possible data. This is costly, time consuming, and provides users with an overabundance of mapped data that can be counter-productive. What is important is the acquisition of an appropriate amount of data that provides the necessary information for rapid, effective decision making for natural hazard management.

Too much detail may unnecessarily add to the cost of the GIS. If a data source is detailed beyond the point of usefulness, then generalized data should be used. If, for example, topographic data are mapped at 5 m contours, but some basic decisions will be reached using 50 m contours, then input and retrieval of topographic complexity can be reduced by a factor of 10. Careful study of the classification systems of the input data, combined with analysis of critical points of differentiation in the physical data sources can reduce the volume of data input without affecting the utility of the analysis.

Detail of the database must be directly correlated with the planning team's needs and it should be dynamic in nature. A planning team assigned to assess vulnerability to natural hazards could begin by looking at hazards at the national level, then shifting to more detailed studies in local areas of high risk. On the other hand, if an area is selected for regional development planning, the study of hazards can begin at the regional or local level. For example, if the development study is concerned with the transportation sector of a city and the area suffers frequent losses to landslides, the database established should obviously reflect this issue.

Regarding scale, planners or GIS users can take advantage of the flexibility some GIS offer by entering data at various scales and later requesting the system to adjust the scale to fit the particular purpose or stage of planning: small to medium scales for resource inventory and project identification; medium scales for project profiles and pre-feasibility studies; and large scales for feasibility studies, hazard zone mapping, and urban hazard mitigation studies.

Resolution or spatial accuracy of the database will be reflected in the number of cells (columns and rows or Xs and Ys) making up the database. The greater the number of cells used to cover a given area, the higher the resolution obtained. However, high resolution is not always necessary, and the trade-off between what is gained in terms of analytical capacity and what is lost in terms of consumption of the computer's memory and input time must be considered. The type of graphic adaptor, the size of the computer's memory, and the user's preference as to whether a full or partitioned screen should be used are determining factors in this respect.

Finally, the design of the database should be tested for performance. Following a pilot test, it is not uncommon to obtain a sizable set of database design rectifications. Guidelines are usually not only directed at the spatial accuracy of data and layer design, but also at the identification of possible obstacles for final system implementation, and the development of procedures or a methodology for performing tasks under normal operational conditions.

The wide array of GIS applications presented illustrates the value of GIS as a tool for natural hazard management and development planning. As demonstrated, GISs can improve the quality and power of analysis of natural hazard assessments, guide development activities and assist planners in the selection of mitigation measures and in the implementation of emergency preparedness and response actions.

As enticing as GIS may look, it is not a suitable tool for all planning applications. Most of the benefits of such an automated system lie in the ability to perform repeated spatial calculations. Therefore, before making the decision to acquire a GIS, planners need to determine what planning activities could be supported with the system and carefully assess if the amount of spatial calculations and analysis to be performed justifies automating the process. If only a few calculations are foreseen, it will probably be more cost-effective to rely on local draftsmen to draw and overlay maps and calculate the results.

PC-based GIS are the best option for a planning team. Even so, planners will have to select between scores of available hardware configurations and software capabilities, prices and compatibilities. Given the typical financial and technical constraints that prevail in Latin America and the Caribbean, the hardware configuration must be simple and affordable. For IBM-compatible systems, for example, a standard central processing unit (CPU), a high-resolution monitor, a small digitizer and an optional colour printer are usually effective enough for a development planning agency's needs, and can be easily purchased at affordable prices in most countries of the region. Large and sophisticated equipment requires more technical skills, is difficult to maintain and repair locally, and the added capabilities may not be significant for the planning agency's needs.

Similarly, there are many GIS software packages to choose from and, accordingly, a wide variety of capabilities and prices are available. Usually, the more expensive the software, the more powerful the analytical capability and sophisticated the output options. However, added capability, particularly in the area of cartographic quality output, is not always necessary, and may not pay for itself. Prices range from one hundred to more than fifty thousand US dollars. Although inexpensive systems lack certain features present in more expensive ones, they have functional capabilities sufficient to meet the basic analysis needs of natural hazard management activities. It is wise to start with some of these modest systems and later expand them according to the agency's needs.

Other aspects that should be considered are data availability and institutional support. For a GIS to be effective as a planning tool, any problems and difficulties in obtaining data from institutions with different mandates and interests must be resolved. A good understanding for sharing information between the different agencies involved in collecting, generating, and using data must be established to insure the dynamic nature of a GIS.

One last issue planners will have to face is the difficulty they will encounter in implementing GIS results. When it comes to translating GIS results into planning guidelines or mandates, it is not uncommon to see them rejected for political, economical or other reasons. This may become more complicated at the local level. When local data needs are generalized and included in a GIS for a larger area, conflicts due to people's detailed knowledge of the area may arise.

Natural hazard management requires cooperation at all levels to be successful. Convincing the decision makers that the GIS can provide timely, cost effective and correct information is a critical step that needs support and attention for every programme addressing natural hazard management issues.

Use of a Geo-Referenced Database

A geo-referenced database (GRDB) is a microcomputer-based program that combines data management with map display, allowing planners and emergency managers to graphically display hazard impact areas and relate them to people and property at risk.

Although a GRDB also uses points, lines, and polygonal symbols to represent data, it differs from a GIS in that it does not have overlaying capabilities. However, GRDB's ability to manage and combine large databases with map display, text relating displayed elements (hazard impact areas, location of shelters, health centres, fire stations, police stations, etc.) to their respective descriptive information, makes it suitable for emergency planning and post-disaster rehabilitation and reconstruction work.

7.10 APPLICATION OF GIS FOR LAND USE AND HOUSING MANAGEMENT

The ability of GIS to store, manage and manipulate large amounts of spatial data provides urban managers with a powerful tool. GIS's ability to link tabular, non-spatial data to location information is likewise a powerful analytic capability. Many different facets of government use GIS technology. GIS also provides ways of viewing and analyzing data that was previously impossible or impractical. With the aid of a GIS, a local planning and community development office can track zoning and site design plans that help to form and shape a city.

7.11 APPLICATION OF GIS IN THE ASSESSMENT OF PHYSICAL TRANSFORMATION OF AN URBAN AREA

Urban areas are complex multi-dimensional systems evolving out of an interaction of multiple agents at several levels. At any given singular moment of time, several transformations may occur simultaneously, which every human being perceives differently and comprehends individually. These individual experiences result from the perceptual and cognitive processes in the human brain also determine the meanings that we derive from our surroundings. These perceptual processes also determine the image

TABLE 7.2 Examples of GIS applications for natural hazard management at the local level of planning

Function	Potential Application	Examples
Data display	Aid in the analysis of spatial distribution of socioeconomic infrastructure and natural hazard phenomena	What lifeline elements lie in high-risk areas?
	Use of thematic maps to enhance reports and presentations	What population could be affected?
	Ink with other databases for more specific information	Where are the closest hospitals or relief centres in case of an event?
Land information storage and retrieval	Filing, maintaining and updating land-related data (land ownership, previous records of natural events, permissible uses, etc.)	Display all parcels that have had flood problems in the past. Display all non-conforming uses in this residential area
Zone and district management	Maintain and update district maps, such as zoning maps or flood plain maps	List the names of all parcel owners of area within 30 m of a river or fault line
	Determine and enforce adequate land use regulation and building codes	What parcels lie in high and extreme landslide hazard area?
Site selection	Identification of potential sites for particular use	Where are the hazards free vacant parcels which have at least z-bed-hospitals within 10 km radius
Hazard impact assessment	Identification of geographically determined hazard impacts	What units of this residential area will be affected by a 20-year flood
Development/land suitability modelling	Analysis of the suitability of particular parcels for development	Considering slope, soil-type altitude, drainage and proximity to development, what areas are more likely to be prioritized for development? What potential problems could arise?

that is created in the human brain related to the environment around us, and is a selective process influenced by our cultural and social positions. This process of formation of mental images through individual experiences and recollections is always rooted in the spatial-temporal context, and by forming connections between the past and the present, it can help us in assigning values to the remains from the past and justifying the need for their continuation into the future.

The urbanization process in India and its pace of growth has accelerated over the past 90 years. The process of urban development is guided and coordinated by the development plan of the city. Most of the major cities of India do have development plans, but these plans are often vitiated at various stages of its implementation, mostly in qualitative terms and provision of basic facilities. In most of the sectors of development, the intent of the proposals is not adequately translated into the envisaged physical framework, thus creating differentiated, disjointed and undesirable urban growth. When the demands are not met in a guided manner, illegal occupation of sites and services, land use transformation and unauthorized construction takes place. This is where the degradation of urban environment starts off, and the issue of assessing urban transformation assumes importance.

7.11.1 Land Use and Activities of an Urban Area

The land use variation shows the need to integrate the preservation process with the planning and development process. An integrated planning process needs to be defined, which considers the reality as perceived by the user, such that the 'process' is designed with the user's expectations providing the 'pointers'. This interface of planning has to be flexible and transparent to accommodate the change in perceptions or values through time, rather than setting out rigid parameters or design constraints, which predetermine the future behaviour and aspirations of the community. An understanding of the underlying spatial patterns and social systems is essential before any future interventions are carried out. This is important to avoid the dissonance that may occur in the resultant urban structure because of conflicting patterns superimposed over one another. A management plan for a city, especially for one that has strong links to the past, has to then necessarily start with understanding the transformations that have taken place in the city through time. Any planning has to evolve out of this spatial-temporal context of the city, which further emphasizes the need for the integration of a research aspect to a statistical process such as urban planning.

7.11.2 Application of GIS Possibilities and Limitations

GIS is a useful tool, particularly because of its capacity to support both spatial and non-spatial attributes and to combine purely representational techniques with analytical techniques. It can also be useful for handling data from diverse sources and forming links and interconnections between them. With a number of agencies and organizations involved in planning, the integrated process can well be a 'participatory process', where GIS can serve as a common platform and interface that permits data exchange and collaborative decisions.

Although most data in GIS has to be geo-referenced non-commercial solutions such as those in the environmental context are now looking at ways to integrate nongeo-referenced information in GIS. However, increasing reliance on rigid, cartographic renditions makes these maps extraneous, which can otherwise be a very useful resource for lending an insight into how perceptions of people have evolved over time. Although commercial GIS packages are still incapable of applying statistical analysis to such 'loose' representations, there have been a few recent efforts to integrate 'perceptual maps' in the process of understanding of our environs and such integrations could be made more effective by developing analytical techniques that need to be and could be applied to such cognate models.

Whether visual renditions can be converted into networks for analytical purposes in the urban context would depend on the kind of information that we seek out of them in the process. It can be highly useful if such statistical analytical packages can be linked with GIS, allowing the interchange of data that is mapped as network structure and as visual spatial representations. GIS allows an immense possibility of data storage and retrieval. In the Bhopal urban centre, the level of complexity is huge and the involvement of multiple agents that influence the urban landscape demands data collection on several levels and across several dimensions. When this data needs to be manually processed, spatial and non-spatial information can be linked only by limited options, such as keys next to maps or by the use of graphical techniques such as colours and symbols.

Databases for managing large data sources are now being widely used, but the correlation of data from more than one source is still mostly limited due to data protection policies that exist between various organizations. GIS can provide a base for the spatial and non-spatial data to be interlinked, and developing techniques such as relational databases or object-oriented databases in GIS can realize an added advantage of linking non-spatial data across several levels.

Research in the field of 'multiple views' is working towards the creation of parallel views where the same datum can be viewed across several different maps or layers of spatial information. In this instance, GIS provides the advantage of linking databases to information from maps that may be created in other software packages such as AutoCAD. GIS allows data input from such diverse sources as remote sensing, traditional cartographic maps, aerial photographs and other photographic images. It can be hoped that the data dissemination policy in India will soon be defined for less restricted data exchange, and data from remote sensing and other satellite information would be easily available for commercial purposes. Most European countries have relaxed their data protection rules, which allows for better exchange of data at a global level. If historic cities are being seen as global resources and their preservation is to be seen as a global responsibility, then it is fair to hope for information to be much more conveniently accessible at a global level. With the Internet forming the prominent interface where most global communities interact, more and more data resources are being made available on the World Wide Web, and any GIS application in the Indian context will benefit from a flexible national policy for data dissemination allowing for greater exchange.

7.12 APPLICATION OF GIS—CASE STUDIES

Case Study 1 *Agricultural Development Options Review in Cambodia—Land Cover Mapping*

Objective

Agriculture plays a central role in the economy of Cambodia. Nearly 85% of the country's population depends on agricultural activities. Since the early 1970s, due Distribution of information on world forest systems and people–forest relationships, specifically:

- Extent and location of different forest types.
- Types that are protected, the effectiveness of that protection.
- Relationship between people and forest distribution, the changes in forest cover over time.
- Current management practices and their degree of success.

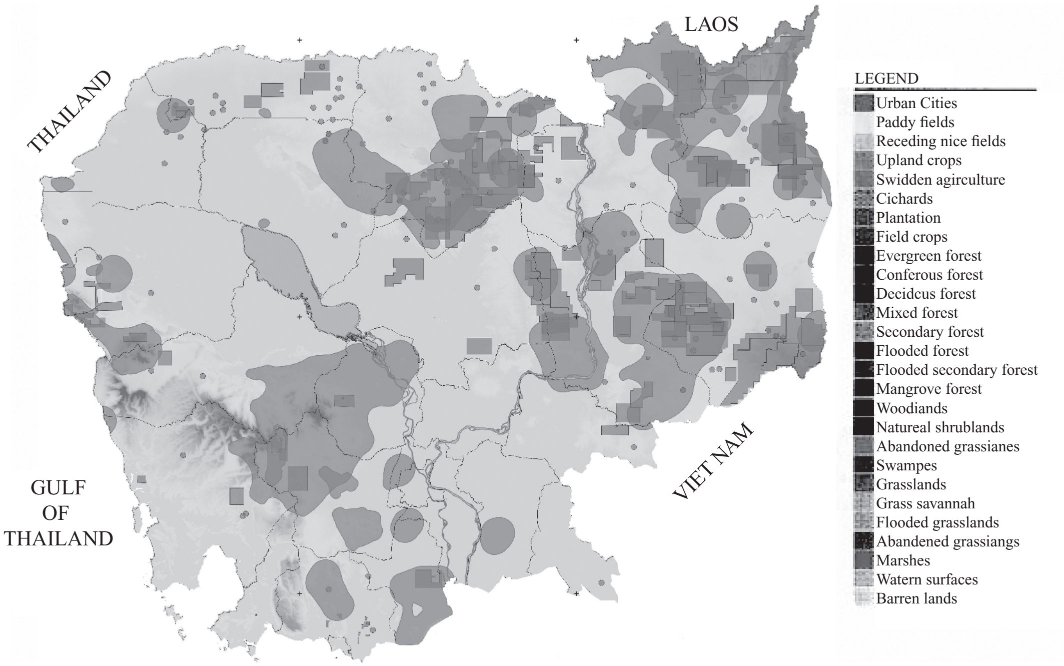


FIGURE 7.2 Land coverage and land utilization of Cambodia

Targeted users are institutions with national, regional and global mandates: governments, national and international non-governmental organizations (NGOs), donor organizations, development planners and research institutes (See Fig. 7.2).

Method

Collection and integration of data and information from existing sources at the two collaborating institutions, in particular:

- WCMC Biodiversity Map Library (BML).
- Analysis of the amount of tropical forest under protection worldwide, based on the ecofloristic zones adopted by FAO, carried out by WCMC.
- CIFGR studies on the relationship between forests and people in a number of tropical countries.

Case Study 2 *A Systems Analysis of the World's Forests*

Objective

Distribution of information on world forest systems and people–forest relationships, specifically:

- Extent and location of different forest types.
- Types that are protected, the effectiveness of that protection.
- Relationship between people and forest distribution, the changes in forest cover over time.
- Current management practices and their degree of success.

Targeted users are institutions with national, regional and global mandates: governments, national and international non-governmental organizations (NGOs), donor organizations, development planners and research institutes.

Method

Collection and integration of data and information from existing sources at the two collaborating institutions, in particular:

- WCMC BML.
- Analysis of the amount of tropical forest under protection worldwide, based on the ecofloristic zones adopted by FAO, carried out by WCMC.
- CIFGR studies on the relationship between forests and people in a number of tropical countries.

The forest coverage has been expanded to include all temperate and boreal areas of the globe, so that the coverage is now global. New datasets for forests from national sources have been incorporated into the forest coverage, ensuring that they are as accurate as possible.

Result

Outputs from the project include accurate maps of forests for the world and protected areas information, and periodic reports highlighting the most important features. The forest information is available on the Internet. The GIS forest and protected area coverage for the tropics have been published on CD-ROM and disseminated widely.

Case Study 3 *Collaborative Planning and Monitoring of Watershed Resources Management*

Objective

Shared Control of Natural Resources Project (SCOR) is a participatory action research project implemented as a result of collaborative effort of the Government of Sri Lanka, USAID and IIMI aimed at increasing productivity of natural resources in watersheds by combining technology, organisation, resources and policy in a collaborative mode to induce changes in use of resources. The role of GIS procedures is to link participatory rural appraisal tools with GIS for the extraction and dissemination of information from spatial data for collaborative planning and monitoring.

Method

SCOR mission requires the characterization of a watershed resource base, visual presentation of resource degradation, the analysis of constraints and actions, and the selection and application of indicators for measuring change and effects. Steps undertaken in a pilot sub-watershed of the Upper Nilwala watershed in the wet zone of Sri Lanka were:

1. Mapping the use of land and water resources with available secondary data
2. Updating land use information through participatory resource use survey
3. Establishing a computerized spatial database
4. Participatory constraint analysis and planning

5. Mapping the desirable future development
6. Collecting data by plot to evaluate the attainment of management levels
7. Contiguity analysis and spatial statistics to measure the adoption of better practices, analysing the clustering of plots achieving higher management levels, explaining, and predicting changes, and providing warning on the need of corrective actions.

ARC/INFO is used for digitizing, and IDRISI for analytical work.

Result

A series of maps used as a basis for analysis and recommendations. The project has planned to reach 20,000 farmer families in two pilot watersheds.

Case Study 4 *Diversity of Wild Potato Species in Bolivia*

Objective

Wild potatoes are relatives of the cultivated potato (mainly *Solanum tuberosum*). They occur in the American continent, from Colorado (USA) to Argentina and Chile, and are most abundant in the Andes of Peru and Bolivia. Wild potatoes are used in breeding programmes to improve the cultivated potato. To conserve and use wild potatoes, they have been collected and stored in ex situ gene banks. However, additional in situ conservation may be desirable. To guide in situ conservation in a meaningful way, the spatial distribution of wild potato diversity needs to be assessed. Some results of a case study on the diversity of wild potatoes in Bolivia are presented.

Method

The primary data used are the databases of the main potato gene banks. Much time was spent checking and correcting the geo-references of the accessions. Plotting the locations where the wild potatoes were found shows that the database suffers from spatial biases. (For example, the collection expeditions mainly travelled over the (main) roads, and this introduced a strong road bias.) Most of the work was carried out during a 6-month period by an MSc student. The work was carried out on a PC using IDRISI and ARCINFO software.

Result

Some of the data biases, and ways to correct them to produce less biased diversity maps, have been identified and quantified. The results enrich our understanding of wild potato diversity and may improve efforts on in situ conservation.

Case Study 5 *Environmental and Sustainability Indicators for Latin America and the Caribbean*

Objective

To ensure better access to information on sustainable development by devising relevant indicators for simplifying, quantifying and analyzing technical information and communicating it to various groups of users. GIS is a tool for integrating economic, social and environmental indicators in a spatial framework, which allows for more powerful analyses than conventional non-spatial methods. In this way, the cause–effect relationships alluded in indicator models and frameworks, such as the pressure-state-impact-response model, may be identified and analyzed more accurately and realistically. Besides spatial analysis, GIS provide a mean for organizing large datasets: the UNEP-CIAT Indicators Project has over 100 indicators stored in its GIS database.

Method

Comprehensive data search and data validation is required to ensure reliable and up-to-date sources. The datasets not yet available in spatial format have to be related to spatial datasets or have a spatial element introduced, so that they can be visualized in the GIS. The software has to be customized to enable users with little or no computer experience to get the most from the indicator datasets. Finally, land use models have to be incorporated to allow users to develop 'what if?' scenarios. This extends the applicability of the product from a visualization tool to a spatial decision support system.

Result

CD-ROM containing the indicator database and the GIS software are used to visualize and analyze the indicators.

Case Study 6 *High-Resolution Remote Sensing: Detailed Information for Participatory Research*

Objective

High-resolution imagery, either airborne or from space, together with advanced software, open the way to precision mapping, whereby 1:1,000 to 1:5,000 scale images can be routinely processed and provide detailed information on land use and high precision digital elevation models (DEMs). This can be compared to a very efficient alternative to land surveying, and can also be used for participatory research. For example, unlike a Landsat TM imagery, high-resolution images may serve as a practical information input for a local community to recognize and localize such key elements as farm boundaries, group of houses, trees and trails. High-resolution imagery can also be useful for understanding the behaviour of models and the loss of information as one goes to smaller scales.

Method

Once a camera taking a series of images (e.g., air photos) is properly positioned in space, orthoimages and DEMs can be produced. Using precise ground control points (determined with GPS), along with bridging techniques to tie overlapping images, the software matches the features in the overlap area and computes elevation at each match. A DEM typically has 5 m pixels with 1 m elevation precision. The orthoimage is derived from the original image by transforming it geometrically, using a DEM, to perfectly overlay a map. Its precision is typically 1–3 m, with a 1-m pixel size. Value-added products such as chromo stereo images can be produced to increase the usability of orthoimages for participatory research, as alternatives to physical models. In the case of Rio Tascalapa Watershed in Honduras illustrated below, ground control points were taken with decimetre accuracy with a Dual Frequency Leica 3323 GPS. ERDAS Orthomax v. 8.2 on a sun space station 10 was used for processing.

Result

In the past, 1:1,000 scale styrofoam models have been built and used in the field to stimulate discussions on watershed management by the community. With the new technique, users can delineate features (e.g., soil units, land use, micro-watersheds) directly on a chromo stereo image; therefore, producing geo-referenced information that can be readily transferred to a GIS. Detailed maps showing transportation networks, buildings and topography can be used to validate smaller scale maps, as well as for cross-scale modelling.

Case Study 7 *Land use GIS for the Cajamarca Area in Peru*

Objective

CIP's research on Andean natural resources is mostly embedded in CONDESAN, the Consortium for Sustainable Development in the Andes. Most CONDESAN related research is located at its 'benchmark sites', Cajamarca being one of them. To better support various related research activities at this site, there is a need for both sound databases to characterize the area, and for tools that can be used to integrate different types of new data and knowledge.

Method

GIS is used for general benchmark site characterization, as well as for specific research projects.

Result

Research topics supported by the Cajamarca GIS include the relations between land use, land use change and land degradation, and the effect that policy may have on these processes. Research is carried out at different scales, from parcels to larger units. One of the central themes of research is how GIS and mathematical models can be used to improve extrapolation of research results in heterogeneous environments like the Andes.

Case Study 8 *Mountain Environment and Natural Resources Information Service: Gorkha District Database*

Objective

The Hindukush-Himalayan region (HKH region) presents a wide range of both ecological and development problems. Isolated solutions have proved to be counterproductive, because most of the problems are interconnected. An integrated approach to solving the problems needs a strong database. In 1990, ICIMOD established the Mountain Environment and Natural Resources' Information Service as a resource centre for the HKH region for the study and application of GIS technology. One of the pilot activities was to build a GIS of the Gorkha district.

Method

The established database of the Gorkha district is based mainly on secondary data. The population figures were extracted from the 1991 Census, and all other socioeconomic figures were retrieved from a ward-level baseline survey. The database currently covers such themes as:

1. Elevation, water bodies and meteorological data.
2. Land utilization and land resources.
3. Settlements, roads, population and agriculture.

The database is based on ARC/INFO software. A PC platform was used for data input and digitizing, while the geographic analysis was carried out using IBM RISC System/6000TM.

Result

The database is available in a PC format. Among its applications have been:

1. The agro-climatic zoning of the district territory.
2. The analysis of livestock feed situation, including feed supply, feed requirements and livestock carrying capacity.
3. The assessment of the potential of horticultural development as a function of climate, land use and access to market infrastructure.
4. The correlation of land use and climatic factors.
5. The analysis of appropriate locations for potato production during optimal growing periods.

REVIEW QUESTIONS

1. Discuss the various applications of GIS in civil engineering.
2. Briefly explain the application of GIS in:
   i. Forestry
   ii. Hydrology
   iii. Geology
3. How is GIS helpful in controlling epidemic diseases?
4. What are the major areas of GIS application in natural resource management?
5. What is custom GIS?
6. Discuss in detail the specific usability issues of GIS.
7. What are the situations encountered during GIS visualization?
8. What is a geographic query language?
9. What are the guidelines for the preparation of a GIS?
10. What are the points to be formulated to assess the need of a GIS?
11. What are the factors needed for the planning of a GIS?
12. List out the key elements needed for the benefit calculation of a GIS?
13. What are the attributes that can provide information for hazard management in a GIS?
14. Explain the application of GIS in the assessment of physical transformation of an urban area.

8 BASICS OF TOTAL STATION

Chapter Outline

8.1 INTRODUCTION

Total stations are designed for measuring of horizontal and slope distances, horizontal and vertical angles and elevations in topographic surveys and geodetic works, tachometric surveys, as well as for survey solutions. The measurement results can be recorded into the internal memory of the total station and can be transferred to a personal computer. The basic properties are unsurpassed range, speed and accuracy of measurements. Total stations are developed in view of the maximal convenience of work of the user. Total station measures angles and distances to various points under survey and the coordinates of the observed points (*X*, *Y* and *Z* or northing, easting and elevation) relative to the total station position are calculated using trigonometric/geometric functions

The most commonly used surveying instrument nowadays is the total station (Fig. 8.1). A total station is a combination of an electronic theodolite, an electronic distance measuring (EDM) device and a microprocessor with memory unit. The electronic digital theodolite, first introduced in the late 1960s by Carl Zeiss Inc., helped to set the stage for modern field data collection and processing. When the electronic theodolite was used with a built-in EDM unit, the birth of the new concept in fully automated surveying started. The original name for an instrument of this type was electronic tacheometer, but Hewlett-Packard introduced the name total station over 30 years ago and the name immediately caught on with the profession (see Figs 8.2 and 8.3). With this device, one can determine angles and distances from the instrument to the points to be surveyed. With the aid of trigonometry, the angles and distances may be used to calculate the actual positions (*x*, *y* and *z* or northing, easting and elevation) of surveyed points in absolute terms.

A standard transit is basically a telescope with cross hairs for sighting a target. The telescope is attached to scales for measuring the angle of rotation of the telescope (normally relative to north as 0 degree) and the angle of inclination of the telescope (relative to the horizontal as 0 degree). After rotating the telescope to aim at a target, one can read the angle of rotation and the angle of inclination from a reference point. All total stations have an EDM device and electronic angle scanning.

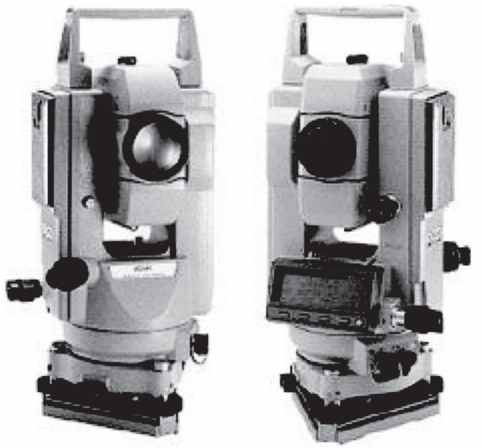


FIGURE 8.1 An electronic total station

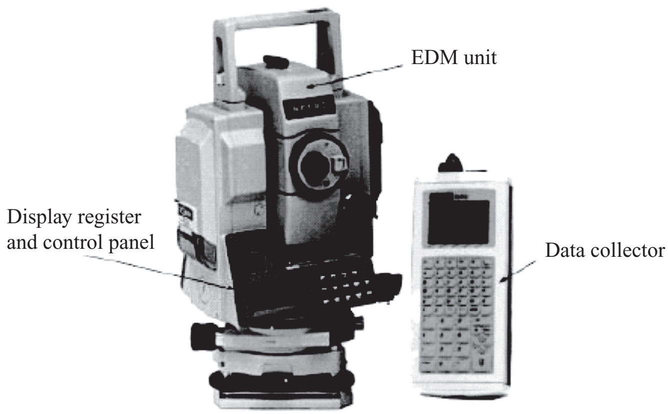


FIGURE 8.2 An electronic total station with data collector (old model)

The coded scales of the horizontal and vertical circles are scanned electronically, and then the angles and distances are displayed digitally. The horizontal distance, the height difference and the coordinates are calculated automatically and all measurements and additional information can be recorded.

The electronic transit provides a digital read-out of those angles instead of a scale, which is more accurate and less prone to errors arising from interpolating between marks on the scale or from recording error. The read-out is also continuous, and, hence, angles can be checked at any time.

In most of the total stations a modulated near infrared light emitting diode, which sends a beam from the instrument to the prism. The prism reflects this beam back to the total station. The portion of the wavelength that leaves the instrument and return is assessed and calculated. Or in other words, the EDM instrument fitted inside the telescope of total station transmits an infrared beam, which is reflected back to the unit with the help of a prism (after total internal reflection), and the EDM uses timing measurements to calculate the distance travelled by the beam. With few exceptions, the EDM instrument requires that the target be highly reflective, and a reflecting prism is normally used as the target.

FIGURE 8.3 An old model total station

Most of the total stations include data recorders. The raw data (angles and distances) and the coordinates of points sighted are recorded along with some additional information (usually codes to aid in relating the coordinates to the points surveyed). The data thus recorded can be directly downloaded to a computer at a later time. The use of a data recorder further reduces the recording errors and eliminates the need for a person to record the data in the field.

The determination of angles and distances are essentially separate actions. One aims at the telescope with great care first. This is the part of the process with a real potential for human error. When the telescope has been aimed, the angles are determined. The observer then initiates the reading of the distance to the target by the EDM. This takes only a few seconds and the calculations are performed immediately.

Total stations are widely used in many construction sites (Fig. 8.4). In many cases, it is not fully used because the users are unaware of its full operational capability. Total station used for levelling comes under the classification indirect levelling. It can maintain considerable accuracy and is hence used for many public works such as construction of roads, airports and harbours.

For field work, the total station is mounted on a tripod and levelled before use (Fig. 8.5). Meanwhile, the prism is mounted on a pole of known height. The mounting bracket includes aids for aiming the instrument. The prism is mounted so that its reflection point is aligned with the centre of the pole on which it has been mounted (Fig. 8.6). Although the tip of the pole is placed on the point to be surveyed, the instrument must be aimed at the prism. So, it will calculate the position of the prism and not the point to be surveyed. As the prism is directly above the tip, the height of the pole may be subtracted to determine the location of the point. That can be done automatically. The pole must be held upright, and a bubble level attached to give the technician holding the pole a check (Fig. 8.7).

It is not as easy as one might expect to hold the pole upright, particularly if there is any wind, and as a result, multiple readings may be required. Because of this problem, the sighting method

FIGURE 8.4 Different models of total station

FIGURE 8.5 Wooden and aluminium tripods for total station

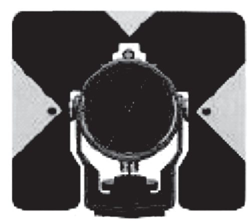


FIGURE 8.6 Typical standard prism and sight

FIGURE 8.7 Typical standard rod (with bubble level)

chosen at such occasions is, if possible, not to begin by sighting on the prism itself, but on the tip of the pole where it touched the ground. The angle from north would then be fixed and unaffected by the movement of the pole. Then the aim of the telescope could be raised to the level of the prism, adjusting only the angle of inclination.

8.2 ADVANTAGES OF TOTAL STATION

The advantages of total station include:

1. Quick setting of the instrument on the tripod using laser plummet
2. On-board area computation programme to compute the area of the field
3. Greater accuracy in area computation because of the possibility of taking arcs in area computation
4. Graphical view of plots and land for quick visualization
5. Coding to do automated mapping. As soon as the field jobs are finished, the map of the area with dimensions will be ready after data transfer
6. Enormous plotting and area computation at any user required scale
7. Accuracy of measurement is high
8. Manual errors involved in reading and recording are eliminated
9. Calculation of coordinates is fast and accurate
10. More work can be accomplished within short time
11. Relatively quick collection of information (data)
12. Multiple surveys can be performed at one set-up
13. Digital design data can be uploaded to total station for setting out of structures to be constructed

8.3 DISADVANTAGES OF TOTAL STATION

1. Their use does not provide hard copies of field notes. Hence, it may be difficult for the surveyor to look over and check the work while surveying.
2. For an overall check of the survey, it will be necessary to return to the office and prepare the drawings using appropriate software.
3. Total station should not be used for observations of the sun, without special filters, if not, the EDM part of the instrument will get damaged.
4. The instrument is costly, and for conducting surveys using total station, skilled personnel are required.

8.4 MEASURING ANGLES

The total station measures horizontal/vertical angles opto-electronically. An absolute circle scanning system is used to read the coded graduations of a glass circle as shown in Fig. 8.8.

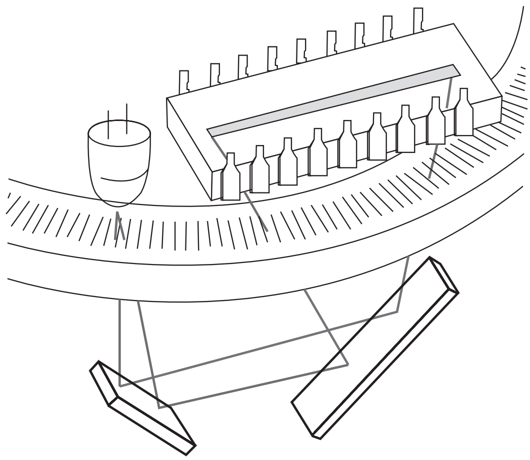


FIGURE 8.8 The circle scanning principle of measuring angles

The circle contains only one coded graduation track, which contains all the positional information. The code is read by a linear CCD array and an analog-to-digital converter, and displays position with maximum accuracy. (CCD stands for *charge-coupled device, and* is a silicon-based multichannel *array detector* of ultraviolet, visible and near-infrared light.)

Determining of the centres of the individual code lines on the array, and then using appropriate algorithms to find the mean, produces a fine measurement. A single measurement involves around 60 code lines, improving the interpolation accuracy, the redundancy and the reproduction.

The value, measured for the horizontal direction, is corrected before being displayed or recorded. The correction is calculated from the following parameters, as a function of the vertical angle measured:

1. The latest collimation error and tilting-axis error, which is to be determined and stored in the instrument
2. The momentary component of the vertical-axis tilt, transverse to the line of sight

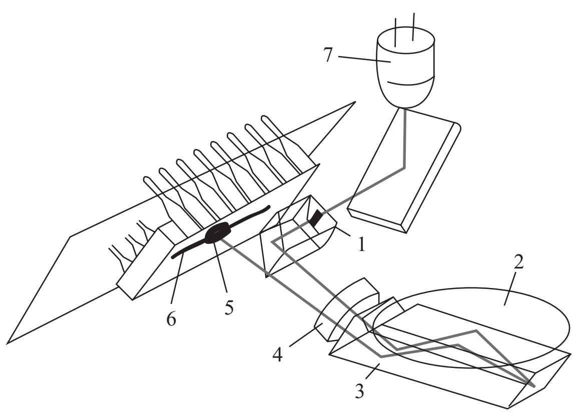


FIGURE 8.9 Tilt sensor of total station

The vertical angle is corrected by the amount of the stored index error and by the component of the vertical-axis tilt in the direction of the line of sight. A tilt sensor monitors both components of the vertical-axis tilt. Figure 8.9 shows the principle of the tilt sensor, in which a liquid mirror forms the reference horizon.

The reticle located on the prism is illuminated and is imaged on the linear CCD array by way of the imaging lens after double reflection at a liquid surface (oil). The triangular line pattern of the reticle makes it possible to capture both tilt components. Longitudinal tilt alters the spacing between the differentially oriented lines. Transverse tilt shifts the centres of the entire line pattern along the CCD array. This arrangement enables the tilt sensor to be made so small that ideally it could be placed centrally over the vertical axis.

8.5 TYPES OF TOTAL STATIONS

In the early days, three classes of total stations were available—manual, semiautomatic and automatic.

Manual Total Stations It was necessary to read the horizontal and vertical angles manually in this type of instrument. The only value that could be read electronically was the slope distances.

Semiautomatic Total Stations The user had to manually read the horizontal circle for these instruments, but the vertical circle readings were shown digitally. Slope distances were measured electronically and the instruments could, in most cases, be used to reduce the values to horizontal and vertical components.

Automatic Total Stations This type is the most common total station used nowadays. They sense both the horizontal and vertical angles electronically and measure the slope distances, compute the horizontal and vertical components of those distances, and determine the coordinates of observed points. To compute the coordinates of observed points, it is necessary to properly orient the instrument to some known directions such as true north, magnetic north or to some known bearing. The coordinate information obtained can either be stored in the total station's memory or by using an external data collector.

Manual total stations and semiautomatic total stations are obsolete now. At present, it is the age of fully automatic total stations and robotic total stations. The price range of an automatic total station comes around US $3,000–8,000, depending on its special features.

Almost all total stations in the market use infrared as the carrier for distance measurement. The less expensive unit with a single prism reflector can measure distances up to 2,000 m. Those in the higher price range are capable of measuring distances up to 4,000 m, when single prisms are used. The accuracies of measurements with the less expensive instruments probably run about 3 mm + 3 ppm over a distance of up to 1,500 m, [$\pm$ (3 + 3 ppm $\times$ D) mm] and the expensive total stations can run about 2 mm + 2 ppm over a distance of up to 2,000 m, [$\pm$ (2 + 2 ppm $\times$ D) mm] when taking measurements with a prism. The accuracy of prism-less measurement is usually $\pm$ (3 + 2 ppm $\times$ D) mm and the range of prism-less measurement varies from 200 to 500 m depending upon various models.

8.6 ADVANCEMENT IN TOTAL STATION TECHNOLOGY

Servo-Driven and Robotic Total Stations Refinements to an existing technology are the classes of servo-driven and robotic optical total stations. Their importance in the last few years is seen to be steadily increasing. Their added functionality makes them suitable for intense mapping. Because of their capacity to improve the surveying operation significantly, they can be classified into a separate group.

Servo-Driven Total Stations Servo-driven instruments are particularly appealing where automatic pointing is desired. This is done by using motors to aim and position the instrument. In the case of setting out, it makes it feasible to set control points for surveying with very little sighting through the telescope. When used with the data collection software, the pre-determined coordinates of the point, which have been selected after setting up the instrument and making an observation to the back sight point, are used to automatically set the horizontal and vertical angles of the instrument. In the case of traversing or other control survey functions, the servo drives can be used to point the instrument in the direction of the next target of the observing programme, requiring only fine-pointing adjustments by hand. When these instruments are used manually, because they are servo-driven, they have friction clutches that afford great speed in pointing, as there are no locks to be adjusted. Furthermore, fine pointing is aided by having unlimited travel in the tangent screws. Again, because of the servo-driven design, the limit stops of the fine motion screws no longer can exist.

The servo-driven instrument has the disadvantage of data collection and coding occurring at the instrument. It is also mandatory that at least two people be on the crew. Still, tremendous productivity gains have been reported.

Auto-Tracking Servo-Driven Total Station A further enhancement to the servo-driven instrument is the auto-tracking feature. This enables the servo-driven instrument to lock onto a target and follow it. By having a servo-driven instrument with auto-tracking feature, several operational improvements are obtained. The target is followed as it moves, the person holding the rod (rod person) seldom has to wait for the instrument person. Aiming and focussing are eliminated from the manual operations required to take a reading. Errors in observations due to parallax error are eliminated as well, and phenomenal increases in accuracy have been reported. When active targets are used with the trackers in traversing and other control surveying operations, standard deviations in angle measurement of 0.2 has been reported. Furthermore, in mapping or setting out operations, multiple rod persons can be used for higher productivity. A review of an auto-tracking servo-driven instrument compared with a manually operated total station showed a 107% increase in data collected.

Robotic Total Stations The robotic total station can be set up over a control point and left there while the surveyor carries the prism to the various points that are to be located. The instrument itself will track the prism when the surveyor presses a button at his position, and the instrument will quickly and accurately take and record the necessary readings. Here a communication link is added to facilitate the placement of a data collector/total station controller at the rod, so that one-person surveying is possible. Robotic instruments have the advantages of being controlled from the prism point; thus, making coding and quality control at the point being measured, which greatly improves the usability and quality of the data.

The equipment used for robotic surveying includes a pole with circular level, a reflector prism, telemetry equipment for communication with the robot and the keyboard. The instrument is set up over a control station and oriented in the usual manner. To do this, it is necessary to enter the coordinates of the station where the instrument is located and take a back sight along a line with a known azimuth. After this set up is completed, the surveyor carries the remote positioning unit (RPU) to the points for which data are desired (Fig. 8.10). If the instrument is used as a robot, the surveyor carries out all measurements located at the various target points using the RPU. When the RPU button on the instrument is activated, the instrument becomes a robot, permitting a one-person operation and the data recording and data storage taking place at the instrument. The surveyor takes the RPU to the point to be sighted and then presses the search button. The robot searches for and locates the RPU. Furthermore, the robot

FIGURE 8.10 The RPU (courtesy of Leica geosystems)

will follow the RPU as it is moved from point to point. In fact, the robot will track the prism even if it is placed flat on the ground while the surveyor is driving a stake or performing some other task. When the lock is lost, the surveyor should press the search button again to renew the search and the instrument will lock onto the RPU. The maximum time required is usually a few seconds.

The fact that the robotic instrument can remain unattended, and yet it can continually track the reflector, cuts back greatly on the time required in normal surveying for sighting, focussing and reading. To collect data using a robotic total station, the only time required is for the surveyor to walk up to the point with the rod and to plumb the rod. The robot will make and record the desired measurement within 2–3 sec.

When the surveyor is on target, the measurement is initiated with the data collector/controller (RPU) mounted on the prism pole. To ensure that quality observations are taken, the operator can remotely initiate checks to the back sight point. Similarly, reports on the level of the instrument can be transmitted through the communication link to obviate walking back to the instrument periodically. It is more common for the target to be lost when the prism pole is placed on the ground, so that stakes can be driven, to dig for a sub-surface monument, etc. The robotic instrument of today is set up with sophisticated search algorithms, which can be preset by the operator in advance to make the search for the target as efficient as possible. It is possible to preset a search window, so that the search operation does not occur over the entire range of 360 degrees.

The instrument can be directed to a point at a certain direction on command from the instrument controller when all other automatic search methods do not work or consume too much time. During setting out, the problem of aligning the rod person with the pre-determined aim of the instrument is required. A feature called a track light that emits twin beams of light along the line of sight, generally facilitates this. The colour visible in the beam indicates to the rod person the direction in which he is required to move to be aligned with the cross hairs of the instrument. When the rod person is on light, both colours are simultaneously visible. In some configurations, a third colour is visible. When aligning for the observation of a point to be a measured, a robotic instrument enables more freedom from mistakes for the surveyor.

One concern of working with robotic instruments is that of protecting the instruments. There is the possibility of having the unmanned instruments stolen or run over by vehicles. Another problem may occur where heavy vehicular traffic is in the vicinity, as it may cause so much interferences that the robot may have difficulty in locking on the target.

8.7 AUTOMATIC TARGET RECOGNITION

Some manufactures have designed an instrument with automatic target recognition (ATR) (see Fig. 8.11), which uses an infrared light bundle sent coaxially through the telescope. In such types of instrument, the telescope must be pointed roughly at the target prism first, either manually or under software control, and then the instrument will do the rest.

FIGURE 8.11 Motorized total station with ATR and electronic guide light (courtesy of Leica geosystems)

ATR comes with a lock-on mode, where the instrument, once sighted at the prism, will continue to follow the prism as it is moved from station to station. To ensure that the prism is always pointed to the instrument, it is good to use a 360-degree prism, which assists the surveyor on keeping the lock-on over a period of time. If the lock-on is lost due to intervening obstacles, it is re-established after manually pointing at the prism. ATR recognizes a target up to 1,000 m. ATR will function in darkness, and requires no focussing or fine pointing. ATR works with all types of prisms, and maintains a lock on prisms moving up to a speed of 15 km/hr (at a distance of 100 m).

8.8 IMAGING SCANNING AND ROBOTIC TOTAL STATION

Imaging scanning and robotic total station are used for advanced imaging and high-accuracy surveying, incorporating real-time field imagery with spatial data. The powerful functionality of this kind of advanced total station is controlled by a software that produces ‘photography with dimension’, a revolutionary and cost-effective terrestrial photogrammetry, which is an alternative to laser scanning (terrestrial LIDAR scanning).

The most modern long-range non-prism (2,000 m) robotic instruments are able to scan long ranges as well. Mining and monitoring applications can be accomplished with this kind of total station.

This kind of total station comes with WLAN capability and can be remotely controlled. It is mostly equipped with touch the screen image to drive the cross hairs for highest degree of data accuracy, and clarity of image. Imaging scanning and robotic total station features dual digital imaging cameras which provides a colour, real-time image on the touch LCD display. To define a scan area, simply tap the image display, or for uniform measurement of an area, select the grid feature. The total station locks these points to their exact position on the image, even when the instrument is rotated. Once all points to be measured are chosen, instrument performs a reflectorless measurement of each point. Long-range scanning is best suited for quarries, mines, huge construction sites and landfills. Mid-range scanning is best for the construction of bridges, embankments and dams, levy and monitoring. Short-range scanning is suitable for interior of buildings and small worksites.

FIGURE 8.12 An imaging total station (courtesy of Sokkia)

FIGURE 8.13 An imaging total station (courtesy of Leica)

8.9 HYBRID ROBOTIC TOTAL STATION

Hybrid robotic total station is the combination of GNSS positioning, an optical total station and automated data workflow. By using both GNSS and optical measurements, any job site project can be completed faster, and with the precision needed. The precision and tracking ability of the robotic instrument provides precise measurements anywhere the prism can be seen. The GNSS positioning of the RTK rover pole can be used for measurements that are not in the line of site.

8.10 REFLECTORLESS MEASUREMENT

Reflectorless electronic distance measurement (EDM) is an advantage in total station technology, which enables the measurement of distances up to hundreds of metres without the prism at the target. Using reflectorless measurements inaccessible objects and objects located at dangerous sites can be mapped easily. In order to use this property properly, the user/surveyor should have the knowledge about the principle and characteristics of reflectorless measurement. Distance measurement without prism is also available on many instruments, by employing two different coaxial red laser systems. One laser is invisible and is used to measure long distances (up to 6 km with a single reflector), the other is visible which does not require a reflector, and has a limited range of about 200–500 m. A single key stroke allows one to alternate between the visible and invisible laser. Reflectorless measurements are useful for surveying the facades of buildings, tunnel profiling, cooling tower profiling, bridge components, dam faces, remote points and in any situation where the point to be observed is difficult to access. In this type of machine, extremely narrow laser is used to define the target points. Accuracy of reflectorless total station measurements are as good as ±(1 mm + 1 ppm). When no person is required to place a reflector at a point to be measured, total station surveying may become a one-person operation, saving labour costs. Reflectorless measurements can be further complimented by video capture, enhanced scanning capability, etc.

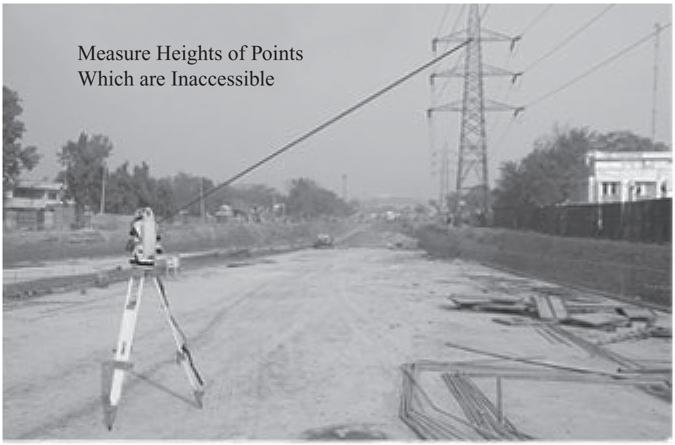


FIGURE 8.14 Reflectorless measurement

8.11 BUILT IN SOFTWARE

Various field operations in total station can be executed with the help of wide variety of programs integrated with microprocessor of total station. All these programs require clear definition of the instrument station and at least one reference point (orientation). All subsequent stations can be identified in terms of (X, Y and Z). Typical programs include the following functions:

1. Topographic surveying
2. Missing line measurement (MLM)
3. Slope reduction
4. Resection
5. Remote elevation
6. Remote distance
7. Offset measurement
8. Stakeout
9. Area computation
10. Tracking
11. Curve setting
12. Traversing
13. Triangulation

REVIEW QUESTIONS

1. Describe briefly about the classification of total stations.
2. Write short notes on:
   - i. Robotic total station
   - ii. Semiautomatic total station
   - iii. Automatic total station
   - iv. Servo-driven total station
3. What are the latest advancements in the total station technology?
4. What is the basic principle behind automatic target recognition (ATR) of a total station?
5. What is a tilt sensor in a total station? Explain its advantages.
6. How are angles measured using a modern total station?
7. What are the advantages and disadvantages of a total station?

ELECTRONIC DISTANCE MEASUREMENTS 9

Chapter Outline

9.1 INTRODUCTION

This chapter deals with the principles and procedures of separate electronic distance measurement (EDM*)* instruments that are mountable with optic/electronic theodolites. Original EDMs were individual units used only for distance measurements. EDM instruments used as a modular component in modern total stations also work under the same principle of separate EDM instruments.

EDM was first introduced in the 1950s and has undergone several modifications. The early versions of the instruments, though very precise over long distances, were very large, heavy, complicated and expensive. Electronic/optic theodolite mountable EDM was introduced first, and these are in use even today. Rapid advances in related technologies have provided lighter, simpler and less expensive instruments. Latest versions of EDM instruments are used as a modular component of total stations. Some earlier versions of EDM instruments used natural light for the calculation of distances, but the latest version of EDM uses infrared light, laser light or microwaves.

9.2 MEASUREMENT PRINCIPLE OF EDM INSTRUMENT

The principle of the measurement device in EDM, which is currently used in a total station along with electronic/optic theodolites, is that it calculates the distance by measuring the phase shift during the radiated electromagnetic (EM) wave (such as an infrared light, laser light or microwave) from the EDM's main unit, which returns by being reflected through the reflector that is positioned at a

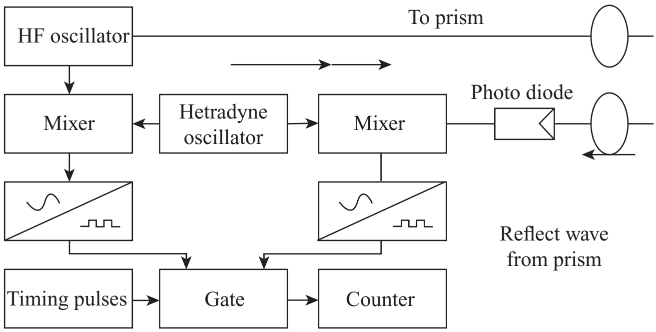


FIGURE 9.1 EDM's structure

measurement point (Fig. 9.1). This phase shift can be regarded as a part of the frequency that appears as the unit of time or length under a specific condition.

When the slope distance L and the slope angle ϕ are measured by EDM, if the elevation of point A is the reference point, we can find the elevation of point B by the following formula (Eqn. 9.1; Fig. 9.2):

$$\text{Elevation of point B} = \text{Elevation of point A} + \text{HI} \pm L \sin \phi - \text{HR} \tag{9.1}$$

Figure 9.3 shows a wave of wavelength λ. The wave is travelling along the X-axis with a velocity of 299,792,560.4 km/sec (approximate velocity of light in vacuum). The frequency of the wave is the time taken for one complete wavelength:

$$\lambda = \frac{c}{f}$$

where λ is the wavelength in metres; c the velocity in km/sec and f the frequency in hertz (one cycle/sec).

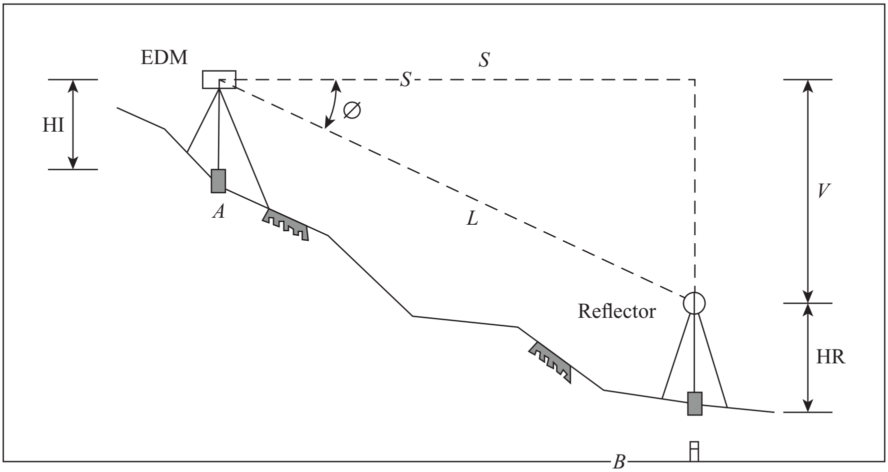


FIGURE 9.2 Slope distance (horizontal distance error) by incidence angle

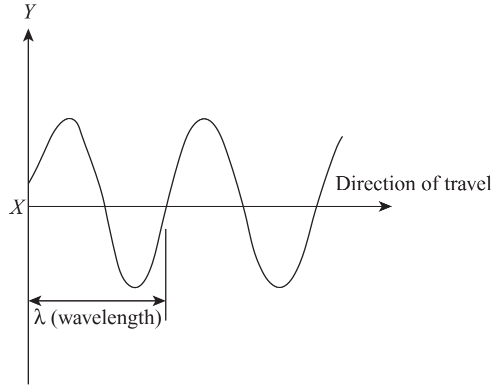


FIGURE 9.3 EDM wave

Figure 9.4 shows a modulated EM wave being emitted from an EDM instrument and being reflected back to the instrument. Here the double distance is taken as $2L$, which is equal to the total whole number of wavelength $n\lambda$ and the partial wavelength ϕ Therefore, the distance between the EDM instrument and the reflector (L) is calculated as follows:

$$L = \frac{n\lambda + \phi}{2} \text{ metres}$$

The partial wavelength (ω) is determined by calculating the phase shift required by correlating the transmitted and reflected waves, that is, by calculating the phase delay required to match precisely the transmitted and the reflected waves.

S = Station

r = Reflector component of addition constant

Z = Target

E = References plane within the EDM for phase comparison

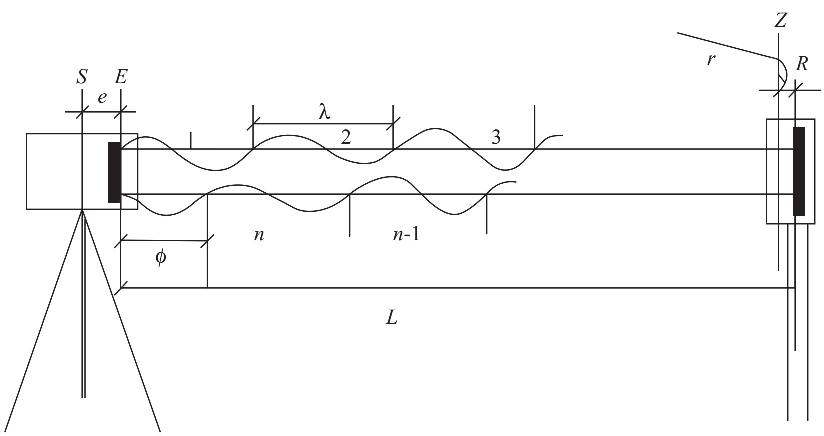


FIGURE 9.4 Principle of EDM

Note: The components of *e* and *r* are only auxillary quantities.

λ = Modulation of wavelength

ϕ = Fraction to be measured of a whole wavelength of modulation $\Delta\lambda$

e = Distance metre component of addition constant

R = Reference plane for the reflection of the wave transmitted by the EDM

The instrument transmits a series of three or four modulated waves at different frequencies. By substituting the resulting values of λ and ϕ in the above equation for three or four different frequencies, the value of n can be found. The instruments are designed to carry out this procedure within seconds, and display the value of L. Some EDM instruments use pulsed laser emissions, and these instruments determine the distance by measuring the time taken between the transmission of the signal and the reception of the reflected signal, by making use of the pulsed laser beam.

The velocity of light (EM waves) through the atmosphere can be affected by

1. Temperature
2. Atmospheric pressure
3. Water vapour content in the atmosphere

The corrections for temperature and pressure are determined manually by referring the monograms given with all total stations (Fig. 9.4), or the corrections are calculated automatically in the instrument itself by inputting the values for temperature and pressure.

For measuring short distances using EDM instruments, atmospheric corrections are relatively insignificant. For measuring long distances, atmospheric corrections become quite important.

9.2.1 Distance Measurement Technique

EDM employ two main techniques for measuring the distance:

1. Using timed pulse techniques such as those used in variety of radar instruments.
2. Using measurements of a phase difference, which may be equated to one part of a cycle expressed in units of time or length.

Pulse methods have advantages over the phase difference methods but their weight and power requirement are such that they cannot be classed lightweight portable instruments.

9.2.1.1 Timed Pulse Techniques

All such measurements incorporate a very precise measurement of time usually expressed in units of nanoseconds (10^{-9} sec), which an EM wave takes to travel from one station to another. In this method, a short, intensive pulse radiation is transmitted to a reflector target, which is immediately transmitted back to the receiver. As shown in Fig. 9.5, the distance (D) is computed as the velocity of light (c) multiplied by half the time ($\Delta t/2$) the pulse took to travel back to the receiver ($D = c \times \Delta t/2$).

9.2.1.2 Phase Difference Technique

The relationship between wavelength and associated phase difference can be illustrated by Fig. 9.6, which shows that for a given complete cycle of EM wave, the phase difference can be expressed both in terms of angular (degrees) and linear (fraction of wavelengths) units. In phase difference method used by majority of EDM instruments, the instrument measures the amount $\delta\lambda$ by which the reflected signal is out of phase with the emitted signal (see Fig. 9.7).

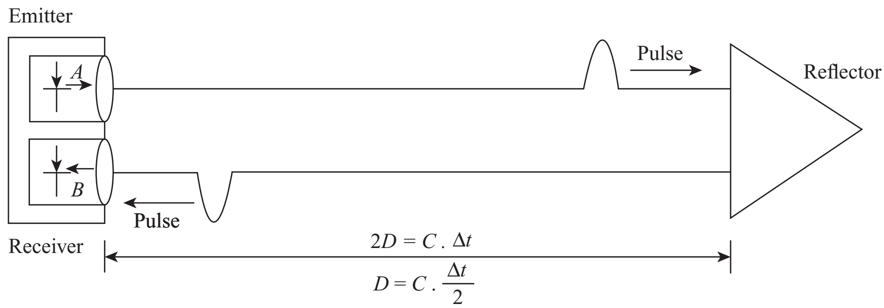


FIGURE 9.5 Principle of EDM instrument based on pulse measurement

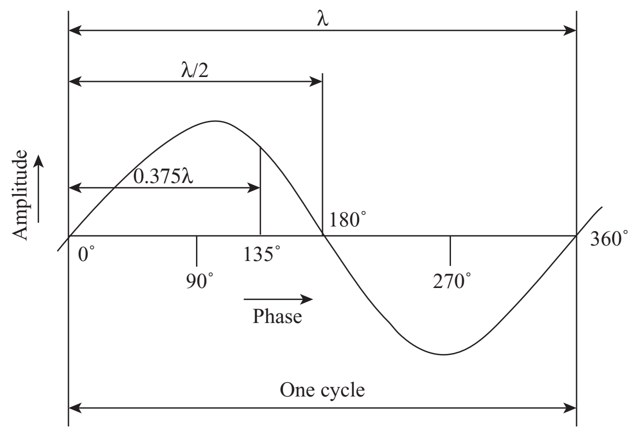


FIGURE 9.6 Relationship between wavelength and phase difference

9.2.2 Classification of EDM

EDM can be classified on the basis of three parameters: (1) wavelengths used, (2) working range and (3) achievable accuracy.

1. Classification on the Basis of Wavelength

Present generation EDM instruments use the following types of wavelengths: (a) infrared, (b) laser and (c) microwaves. The first two types of systems are also known as electro-optical whereas the third category is also called the electronic system.

Electro-Optical Systems

Infrared: Systems employing these frequencies of EM spectrum allow use of optical corner reflectors (special type of reflectors to return the infrared signal) but need optically clear path between two stations. These systems use transmitter at one end of line and a reflecting prism or target at the other end.

Laser: These systems also use transmitter at one end of line and may or may not use a reflecting prism or target at the other end. However, the reflector-less laser instruments are used for short distances (100–350 m). These use light reflected off the feature to be measured (say a wall).

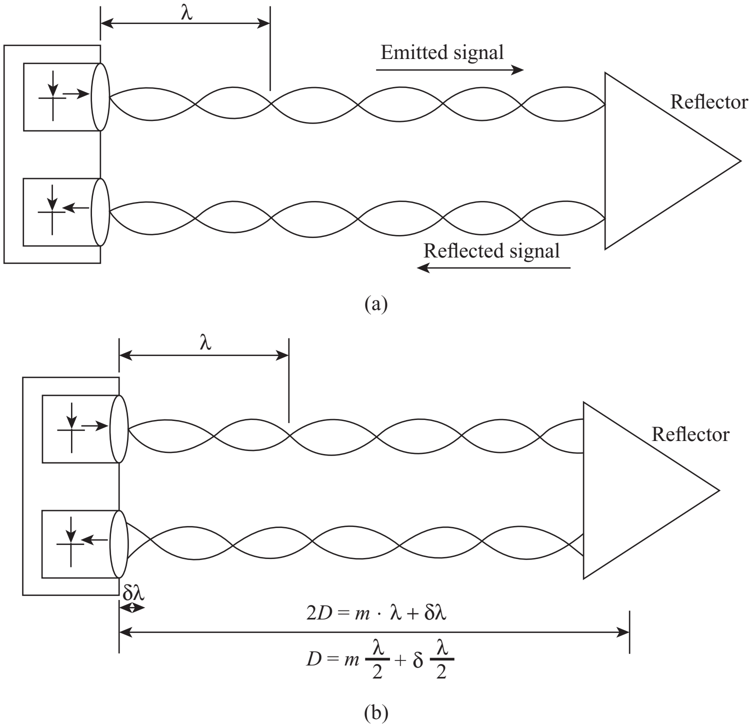


FIGURE 9.7 Principle of phase measurements. (a) Signal in phase and (b) signal out of phase

Electronic System (Microwave)

These systems have receiver/transmitter at both ends of measured line. Microwave instruments are often used for hydrographic surveys normally up to 100 km. At present, this type of EDMs has generally been replaced by Global Positioning System (GPS). These can be used in adverse weather conditions such as fog and rain, unlike infrared and laser systems. However, uncertainties caused by varying humidity over measurement length may result in lower accuracy and prevent a more reliable estimate of probable accuracy. Multipath effects at microwave frequency also add to slight distance error, which can be reduced by taking series of measurements using different frequencies.

2. Classification on the Basis of Range

EDMs are also available as:

a. Long-range radio wave equipment for ranges up to 100 km
b. Medium-range microwave equipment with frequency modulation for ranges up to 25 km
c. Short-range electro-optical equipment using amplitude modulated infrared or visible light for ranges up to 5 km

3. Classification on the Basis of Accuracy

Accuracy of EDM is generally stated in terms of constant instruments error and measuring error proportional to the distance being measured: ± (*a* mm + *b* ppm).

The first part in this expression indicates a constant instrument error that is independent of the length of the line measured. The second component is the distance-related error.

Here, a is a result of errors in phase measurements (θ) and zero error (z), whereas b results from error in modulation frequency (f) and the group refractive index (ng). The term group index pertains to the refractive index for a combination of waves—carrier wave and multiple modulated waves—in EDM. θ and z are independent of distance, but f and ng are functions of distance and are expressed as

$$a = \sqrt{\sigma_\theta^2 + \sigma_z^2}$$

$$b = \sqrt{\left(\frac{\sigma_f}{f}\right)^2 + \left(\frac{\sigma_{\text{ng}}}{\text{ng}}\right)^2}$$

In the above equations, σ indicates the standard error. Most of the EDMs have accuracy levels from ±(3 mm + 1 ppm) to ±(10 mm + 10 ppm). For short distances, part a is more significant; for long distances, part b will have large contribution.

Comparison of the main edm types:

EDM Type	Advantages	Disadvantages
Electro-optical	Less susceptible to atmospheric conditions. Less expensive: only a single transmitter needed.	Shorter range.
Microwave	Can penetrate fog and rain. Longer range. Transmitter at both ends allows voice communication	Atmospheric effects are greater. Susceptible to ground reflected signals. More expensive: requires two transmitters.

9.3 EDM INSTRUMENT CHARACTERISTICS

The characteristics of EDM instruments are as mentioned below:

Distance Range: Distance can be measured up to 1 km using a single prism under average atmospheric conditions. Short-range EDM instruments can measure up to 1,250 m using a single prism. Long-range EDM instruments can be used for the measurement up to 15 km using 11 prisms.

Accuracy: For short-range EDM instruments: ± 15 mm + 5 ppm. For long range EDM instruments: ±3 mm + 1 ppm

Measuring Time: The measuring time required is 1.5 see for short-range measurements and up to 4 sec for long-range measurements. Both accuracy and time are considerably reduced for tracking mode measurements.

Slope Reduction: Manual or automatic in some models. The average of repeated measurements is available on some models.

Battery Capability: 1,500–5,000 measurements depending on the power of the battery and the temperature.

Non-Prism Measurements: Non-prism measurements are available with some models. They can measure up to 100–350 m.

9.3.1 Different Wavelength Bands Used by EDM

Usually, EDM uses three different wavelength bands and their characteristics are:

Microwave Systems

Range up to 150 km
Wavelength 3 cm
Not limited to line of sight
Unaffected by visibility

Light Wave Systems

Range up to 5 km (for small machines)
Visible light, lasers
Distance reduced by visibility

Infrared Systems

Range up to 3 km
Limited to line of sight
Limited by rain, fog, other airborne particles

9.4 ERRORS IN EDMs

The distance measured by EDM is expressed by the formula:

$$S = U + \frac{m\lambda}{2} \tag{9.2}$$

where U is the phase shift of the reflected light wave; λ the wavelength and m the number of transmitted wavelength.

We know that wavelength λ is a function of frequency f and electric wave's velocity v

$$\lambda = \frac{v}{f} \tag{9.3}$$

The velocity of the electronic wave v is 2,99,792.5 km/sec, that is, the same as the velocity of light under vacuum. It is always slower than the velocity of light c in the atmosphere, and we can correct the influence of the atmosphere and calculate it by the following formula:

$$v = \frac{c}{n} \tag{9.4}$$

Here, we have n as the refractive index of air. We should measure the temperature and humidity in the air according to the measurement line, then to find the exact value n, we substitute formula in Eqn. 9.4 for the formula in Eqn. 9.3. Now the value λ of the transmitted signal becomes:

$$\lambda = \frac{c}{nf} \tag{9.5}$$

When the wavelength under a specific atmospheric condition is assumed as λ_1, then it can be described as

$$\lambda_1 = \frac{c}{n_1 f} \tag{9.6}$$

Now the EDM's distance can be expressed as $S_1 = U_1 + m\frac{\lambda_1}{2}$, and U_1 can be expressed by the phase shift of $\frac{1}{2}\lambda_1$. If $n_2 = n_1 = n$ during the measurement, the corrected value of λ is

$$\lambda_2 = \frac{c}{n_2 f} \tag{9.7}$$

Therefore, the real distance S is

$$S = U_2 + m\frac{\lambda_2}{2} \tag{9.8}$$

where m is the number of transmitted wavelength.

From Eqns 9.6 and 9.7, we get:

$$\lambda_2 = \lambda_1 \frac{n_1}{n_2} \tag{9.9}$$

Hence, the corrected distance under a specific atmosphere condition is expressed as the following:

$$S = U_1 \frac{n_1}{n_2} + \frac{m\lambda n_1}{2n_2} = S_1 \frac{n_1}{n_2} \tag{9.10}$$

In order to find the final corrected distance, we should correct the error of the EDM's zero point Z_0, and add the earth curvature, slope and the corrected error of the average sea level ΔS. Consequently, the formula for the final corrected distance is as shown below:

$$S_0 = S_1 \frac{n_1}{n_2} + Z_0 + \Delta S \tag{9.11}$$

where S_1 is the measured distance; n_1 the refractive index under a specific atmospheric condition usually in a laboratory; and n_2 the refractive index at the moment of measuring

When Eqns. 9.6 and 9.10 are substituted for Eqn. 9.11, the corrected distance S_0 is obtained as follows:

$$S_0 = U_1 \frac{n_1}{n_2} + m \frac{c}{2n_2 f} + Z_0 + \Delta S \tag{9.12}$$

where m is the number of transmitted wavelength and c is the velocity of light in vacuum.

The variance of the distance S is as follows:

$$\sigma_{S_0}^2 = \sigma_u^2 + \left(\frac{m}{2nf}\right)^2 \sigma_c^2 + \left(\frac{m}{2nf^2}\right)^2 \sigma_f^2 + \left(\frac{m}{2n^2 f}\right) \sigma_n^2 + \sigma_{Z_0}^2 + \sigma_{\Delta S}^2 \tag{9.13}$$

And if Eqn. 9.13 is simplified by $2S \cong m\lambda \cong \frac{mc}{nf}$, the following result is obtained:

$$\sigma_{S_0}^2 = \sigma_u^2 + S^2 \left[\left(\frac{\sigma_c}{c}\right)^2 + \left(\frac{\sigma_f}{f}\right)^2 + \left(\frac{\sigma_n}{n}\right)^2 \right] + \sigma_{Z_0}^2 + \sigma_{\Delta S}^2 \tag{9.14}$$

Here, $\sigma_u = U_1 \frac{n_1}{n_2}$ is the standard deviation of the total value.

EDM's accuracy is expressed by the following general formula:

$$\sigma_S^2 = a^2 + b^2 S^2 \tag{9.15}$$

And if Eqn. 9.14 is expressed by the form of Eqn. 9.15, then a^2, b^2 can be obtained as follows:

$$a^2 = \sigma_u^2 + \sigma_{Z_0}^2 \; (\sigma_{\Delta S} \text{ is not considered}) \tag{9.16}$$

$$b^2 = \left(\frac{\sigma_c}{c}\right)^2 + \left(\frac{\sigma_n}{n}\right)^2 + \left(\frac{\sigma_f}{f}\right)^2 \tag{9.17}$$

where σ_c is the error of the electric wave's velocity, which is the same as light velocity under vacuum; σ_f the error of modulated frequency; σ_n the error of refraction coefficient; σ_u the error of phase-shift measurement; σ_{Z0} the error of zero point and $\sigma_{\Delta S}$ the geometric error not included in Eqn. 9.15.

9.5 ERROR CORRECTION IN EDMs

The device for measuring distance by light waves should always have the correction for the measured value. Distance measurement devices are generally influenced by the following factors:

1. Observation variable that we can directly observe.
2. The non-observation variable that we cannot directly observe.

The variables that are possible to observe and determine a specific condition are temperature, atmospheric pressure and steam pressure, and their effects are different. The refraction index is calculated by the variables of temperature, atmospheric pressure and steam pressure, and it determines the ratio between the velocities of EM waves under vacuum (C_V) and EM waves under the regular atmospheric condition (C_A). Therefore, it is expressed as $n = \frac{C_V}{C_A}$, and if we know the refraction index (which is always <1), we can compute the temporary spread velocity of the EM wave, and this equals to the spread velocity of the measurement signal.

Although the measurement signal travels the same distance with various velocities according to the weather conditions, the computer installed at the measurement device is programmed in advance to use a certain fixed wave velocity. This velocity deviation generates the signals in proportion to the measured distance.

Therefore,

$$n_L = \frac{n_{gr} - 1}{(1 - \alpha t)} \times \frac{p}{760} \tag{9.18}$$

where t is the measured temperature (°C); P the measured pressure (mmHg); n_{gr} the refraction index to the wavelength of the reflected frequency under 28°C and 760 mmHg and α = 0.003661 (1/28°C), expansion coefficient of air per l°C.

The variables that are not possible to be observed directly and influence the measurement process are rapid weather change, rainfall, heat haze and fog. Therefore, it is necessary to correct the factors that largely influence the measurement result as much as possible. The variables to be corrected are refraction of atmosphere, height of sea level, refraction by the projection method, difference of scale coefficient and so on.

$$D = Sm \cdot \sin Z$$

$$\Delta H = Sm \cdot \cos Z + E - R + \mathrm{i} - r \tag{9.19}$$

Here, we have, D as the horizontal distance; ΔH the difference of elevation; S_m the slope distance; Z the ceiling angle; E the earth refraction index; R the influence by refraction; i the height of horizontal, central axis of device reference point and r the earth radius (6,377 km).

The distance correction value, D_0 at sea level is as follows:

$$D_0 = D \cdot \left(1 - \frac{h}{r}\right) \tag{9.20}$$

where D is the horizontal distance from sea level to the reflector and h the reflector's height with respect to the sea level.

9.6 ZERO CORRECTION

Zero correction is the process of minimizing errors in EDM due to the factors discussed below.

9.6.1 Prism Integer

The reflection of light in a prism decreases when its slope angle increases with respect to the prism face, and it has much dispersion of light if the measurement distance gets longer. In such situation one should install a reflector facing uprightly to the measuring device.

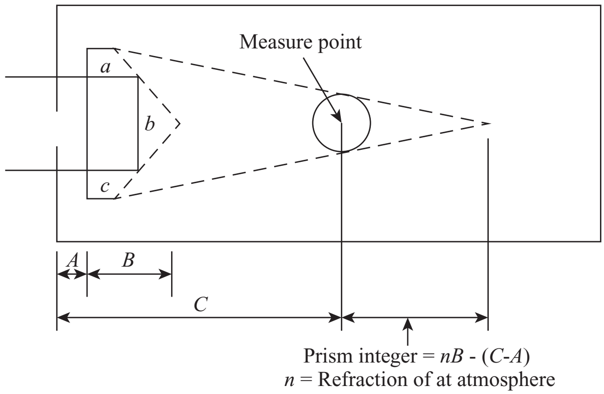


FIGURE 9.8 Prism integer

The light radiated from the distance measurement device is reflected through the prism, and if the inside length of the prism is converted into distance in the air, we get $B = \frac{a+b+c}{2}$ (as shown in Fig. 9.8).

As a prism has its own integer, it is better to use a prism that exactly fits into a distance measurement device by a light wave. If we use it under the condition that a prism's integer is not exactly adjusted, we get a large measurement error. In Fig. 9.8, the distance actually measured is Line B, but what we try to find is Line $(C - A)$. The difference between Line B and Line $(C - A)$ becomes the prism's integer.

As the refraction index of the prism is higher than the atmosphere, we should consider this refraction index, and the prism integer C is expressed as follows:

$$C = nl' - l \tag{9.21}$$

Here, we have, n as the refraction index of the atmosphere; l' the B in Fig. 9.8 and l the $(C - A)$ in Fig. 9.8.

9.6.2 Error by Incidence Angle

As the error occurs according to the light angle that enters into a prism, it is desirable to set up a prism side vertically and exactly to the measuring point when we set up a prism (Fig. 9.9). Table 9.1 shows the experimental value that TOPCON Co. executed at a 1 km section by the prism's incidence angle.

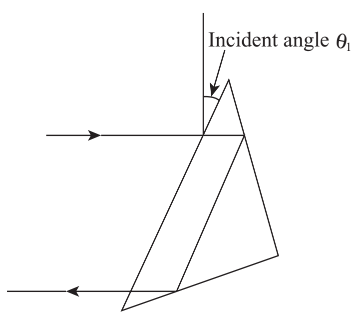


FIGURE 9.9 Distance error by incidence angle

TABLE 9.1 Distance error volume

Incident Angle (v_1)	Error (mm)
0°	0
5°	0.2
10°	0.6
20°	2.5
30°	5.7
40°	10.3

9.7 REFLECTOR USED FOR EDMs

Any surface capable of reflecting the electro-optical signal will allow distance measurement. However, the more efficient the reflector, the stronger the returned signal and the longer distance which can be measured. Efficiency includes amount of signal reflected along with the direction of its return path. For example, while a flat mirror reflects most of the signal, but if it is not perpendicular to the incoming path, the reflected signal will be deviated away from the EDM instrument.

To overcome this problem, a corner cube prism is used as a reflector for most Total Station Instruments. A corner cube prism is based on a 45-degree right angle prism. This type of prism has the property that any signal which intersects its long (hypotenuse) side will be reflected parallel to the incoming path even if the prism is not perpendicular to the signal path (See Fig. 9.10, 9.11 and 9.12).

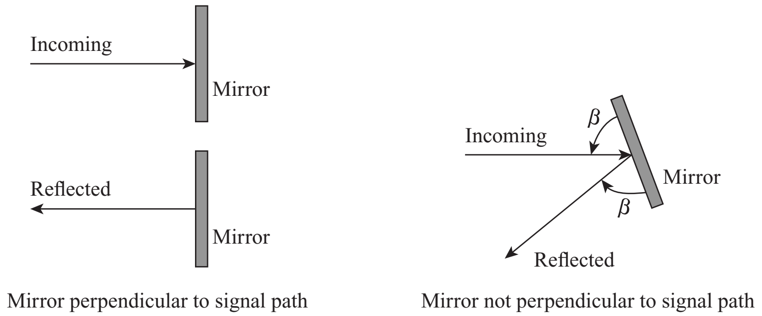


FIGURE 9.10 Mirror reflector ±(5mm + 5ppm)

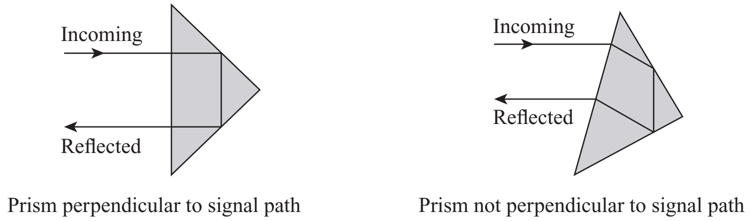


FIGURE 9.11 Prism reflector

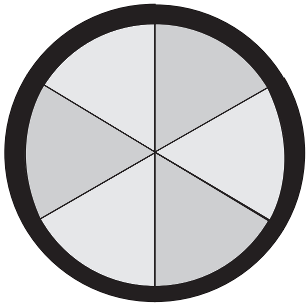


FIGURE 9.12 Prism, front view

A typical corner cube prism uses a glass cylinder having three 45-degree facets at one end. This creates three right angle prisms all sharing the glass cylinder's flat front as their hypotenuse. From the front the facets appear as six radial segments.

9.7.1 Prisms Used for EDMs

Prisms are used to reflect transmitted signals from an EDM instrument (see Figs. 9.13–9.16). A single reflector is a cube corner prism that has the characteristic of reflecting light rays precisely back

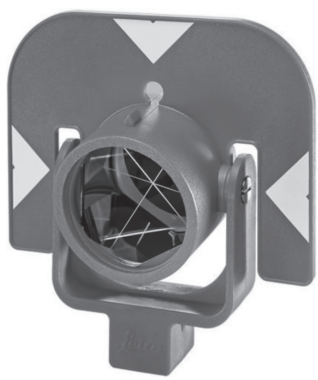


FIGURE 9.13 A high-precision single-prism target

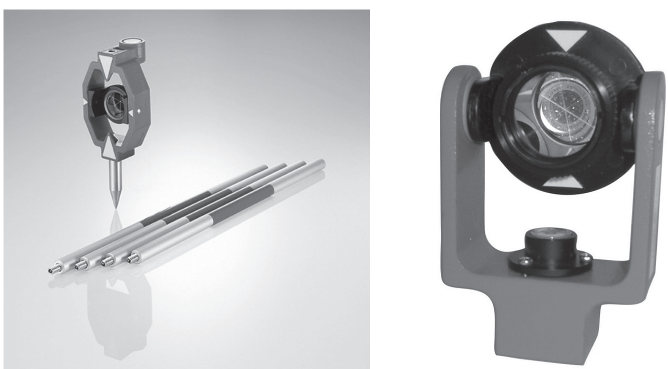


FIGURE 9.14 A mini prism

FIGURE 9.15 A single prism mounted on a tribrach and a 360 degree prism

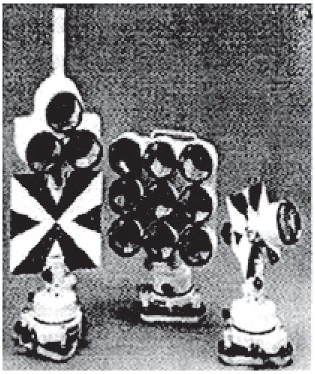


FIGURE 9.16 Different types of prism reflectors

to the emitting EDM instrument. This retro-direct capability means that the prism can be somewhat misaligned with respect to the EDM instrument and still be effective. Cutting the corners off a solid glass cube forms a cube corner prism. The quality of the prism is determined by the flatness of the surface and the perpendicularity of the 90-degree surfaces. Prisms can be tribrach-mounted on a tripod (Fig. 9.15), centred by an optical plummet, or attached to a prism pole held vertical on a point with the aid of a bull's eye level. If higher accuracy is required, then the prisms must be tribrach mounted. For control surveys, tribrach of prisms can be detached and then interchanged with theodolite (and EDM), similarly mounted at the other end of the line being measured. This interchangeability of prism and total station speeds up the work, as the tribrach mounted on the tripod is centred and levelled only once. Equipment that can be interchanged and mounted on a tribrach already set up is known as forced centering equipment.

Prisms mounted on adjustable-length prism poles are portable and are well suited for stakeout surveys and topographic surveys. It is very important that the prisms mounted on poles or tribrach be permitted to tilt up and down, so that they can be made perpendicular to EDM signals sent from much higher or lower positions. This characteristic is particularly important for short sights.

9.7.2 Reflector-less EDMs

The inherent advantage of reflector-less EDMs is the ability to measure distances to points not accessible with a prism. Their major drawback is their shorter range. A reflector-less EDM uses short pulses of high-energy laser light. This energy is considerably higher than that used by phase-shift total stations to get a return signal off low-reflection surfaces. The instrument measures travel times of the laser pulses and from that can determine the total instrument–surface instrument distance. As the laser pulses reflect off different surfaces, care must be exercised when pointing the instrument. This is especially critical when there are multiple surfaces at various orientations near the measurement point. Many total station instruments feature a built-in laser pointer which provides the operator a visual indication of where the measurement will be made.

9.8 ACCURACY IN EDMs

EDM instrument accuracies are stated in terms of an instrumental error and a measuring error, which are proportional to the distance measured. Typically, accuracy is given as ±(5 mm + 5 ppm). Here, the first numerical ±5 mm is the instrument error, which is independent of the measurement, and the second numerical 5 ppm (i.e., 5 ppm = 5 mm/km) denoted the distance related error. High-end receivers are getting accuracies up to ±(3 mm + 1 ppm), whereas the low-end receivers are getting accuracies up to ±(10 mm + 10 ppm). The second part of the error, known as a proportional part error (l0 ppm), is insignificant for most works, and the constant part of the error ±(5 mm) is less significant as the distance being measured lengthens. At 100 m, an error of ±5 mm represents 1/20,000 accuracy, whereas for 1,000 m, the same instrument error represents 1/200,000 accuracy.

For most accurate works, both the EDM instrument and the prism reflectors must be corrected for off-centre characteristics. The measurement is recorded from the electrical centre of the EDM instrument to the back of the prism and back to the electrical centre of the EDM instrument. The manufacturer compensates for the difference between the electrical centre of the EDM instrument and the plumb line through the tribrach centre at the factory itself. The prism constant is eliminated either by the EDM instrument manufacturer at the factory or in the field.

9.9 FIELD PROCEDURE OF EDM

The field procedure of EDM includes the following:

Set up: EDM instruments are fixed on a tribrach (for tribrach-mounted EDM instruments) after the tribrach has been set over a required point on the surface of the earth (see Fig. 9.17). Centering should be carried out using an optical plummet. Theodolite yoke-mounted EDM instruments are simply attached to the theodolite before or after the theodolite has been set over a required station point. Prisms are set over the remote station points by either inserting the prism to a tribrach after centering, or by holding the prism vertically over the required point with the help of a prism pole. The EDM instrument is then turned on. A quick check for battery, display, etc. is made to ensure that it is functioning properly. The height of the instrument and the height of the prism centre are measured and recorded.

Sighting the target and aiming: The target is sighted by orienting the EDM instrument (orienting the telescope in the case of theodolite yoke-mounted EDM instrument). Telescope or yoke-mounted EDM instruments have the optical line of sight a bit below than the electronic signal. The surveyor can set the electronic signal precisely on the prism centre by adjusting the appropriate horizontal and vertical slow motion screws. This can be done with respect to the audible maximum signal intensity

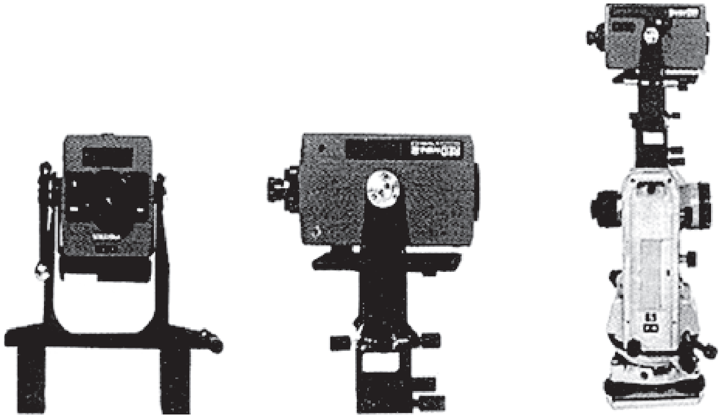


FIGURE 9.17 Two side views of an EDM with fixing clamps and fixed in an electronic theodolite

level from the EDM instrument, or according to the graphical display of the EDM instrument. Older versions of EDM instruments are provided with an attenuator that has to be adjusted for varying distances, in which the strength of the signal has to be reduced for short distances and increased for long distances to avoid signal overload.

Measurement: Before taking measurements, the vertical angle has to be fed into the EDM instrument. Slope distance measurement is accomplished by pressing the measure button of the EDM instrument. Within a couple of seconds, the result will appear on the display of the EDM instrument. Usually, the EDM instruments have a liquid crystal display (LCD). If no measurements appear in the display, then the surveyor should check the switch position, battery status, attenuation and target sight. EDM instruments with a built-in microprocessor will show the horizontal and vertical distances, coordinates, etc. The input data like vertical angle, ppm correction and prism constant are entered through the keyboard of the EDM instrument. Most of the EDM instruments have a tracking mode, which helps in the case of layout surveys and permits continuous distance update as the prism is moved closer to its final layout position. Surveyors may use hand-held radios for communication for long-distance measurements.

Record: The measured data can be recorded conveniently in field note format, or they can be manually entered into an electronic data collector. The distance data must be accompanied by all relevant atmospheric and instrumental correction factors. Total station instruments, which have automatic data acquisition capabilities, will record and store the measurement automatically.

9.10 GEOMETRY OF EDMs

The slope distance (S) is measured by the EDM as in Fig. 9.18. The slope angle (α) is measured by the accompanying theodolite. The figure illustrates the usage of EDM when the optical target and the reflecting prism are at the same height. The height of the EDM instrument and the theodolite (hi) are measured with a steel tape or by a graduated tripod-centering rod. The height of the reflector/target is measured in a similar way. If the height of the reflector is set in such a way that the height of the prism equals the height of the instrument (hi), it will simplify the computations. Nothing will happen even if

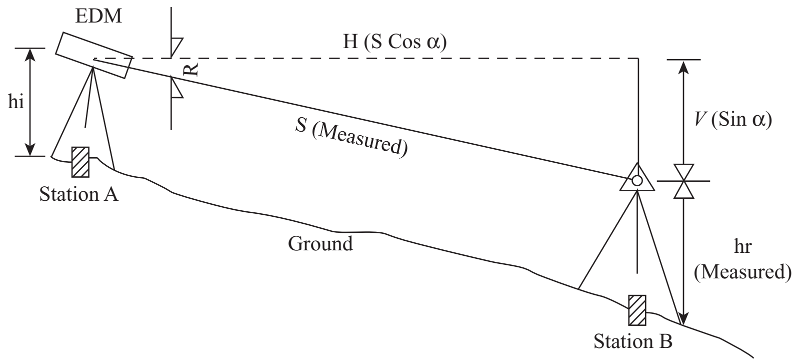


FIGURE 9.18 Geometry of an EDM

the target height and the instrument height are unequal. The elevation of the station is known and the elevation of station B is calculated as follows:

$$\text{Elevation of station B} = \text{Elevation of station A} + \text{hi} \pm \text{V} - \text{hr}$$

When the EDM instrument is mounted on the theodolite and the target is located below the prism, the geometric relationship will be as in Fig. 9.19. The additional problem encountered in this

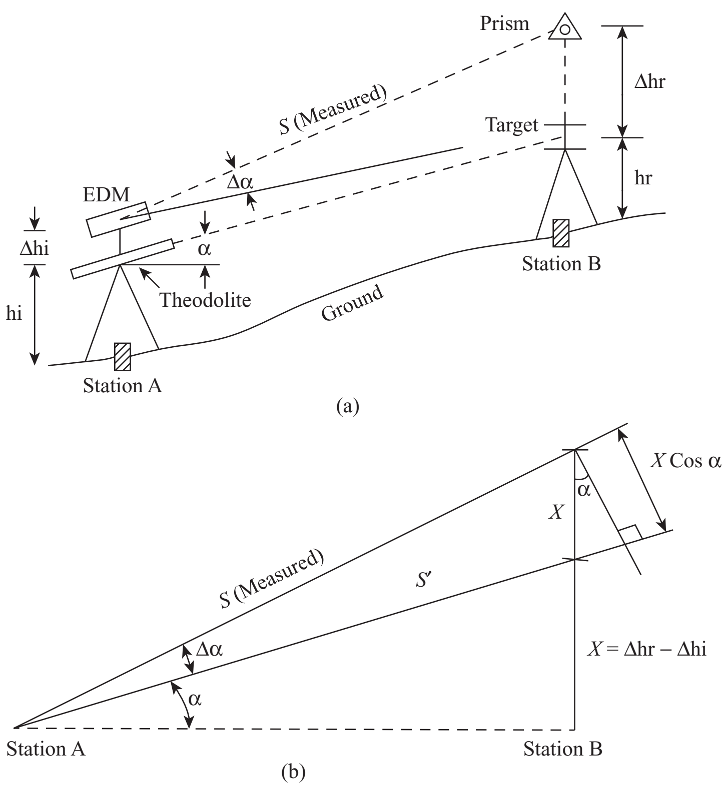


FIGURE 9.19 (a, b) Geometry of EDM when EDM is mounted on a theodolite

situation is the computation of the correction of the vertical angle ($\Delta\alpha$) that occurs when Δhi and Δhr are different. The precise size of the vertical angle is important because it is used in conjunction with the measured slope distance to compute the horizontal and vertical distances. The difference between Δhr and Δhi is '$X = \Delta hr - \Delta hi$'. The small triangle formed by extending S' has the hypotenuse equal to X and an angle of $\Delta\alpha$. This permits computation of the side $X \cos \alpha$, which can be used together with S to determine $\Delta\alpha$.

$$\frac{X \cos \alpha}{S} = \sin \Delta\alpha$$

PROBLEM 1 An EDM slope distance AB is determined to be 672.243 m. The EDM instrument is 1.815 m above station A and the prism is 1.995 m above station B. The EDM instrument is mounted on a theodolite whose optical centre is 1.700 m above the station. The theodolite was used to measure the vertical angle to a target on the prism pole, which is measured as 5°45'34". The target pole is 1.800 m above station B. Compute the horizontal distance AB and the elevation of station B, if the elevation of station A is 345.472 m.

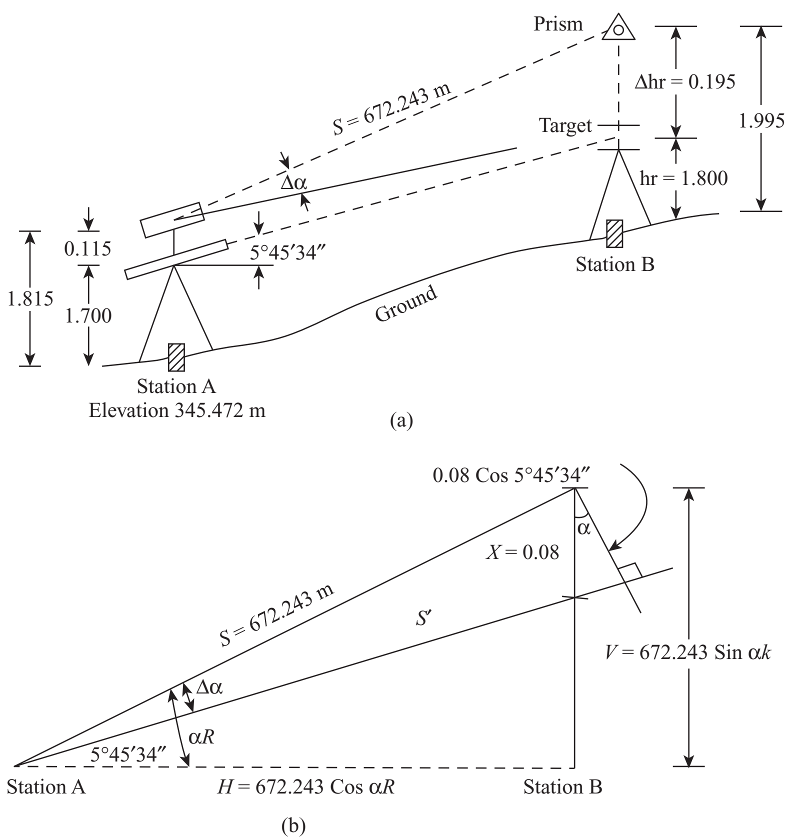


FIGURE 9.20 (a) and (b)

Solution: The given data are shown in Fig. 9.20(a) and the resultant figure is shown in Fig. 9.20(b). The value of X is calculated as follows:

$$X = (1.995 - 1.800) - (1.815 - 1.700) = 0.08 \text{ m}$$

$$\text{Here } \sin \Delta\alpha = \frac{0.08 \cos 5^\circ 45'34''}{672.243}$$

$$\text{Hence, } \Delta\alpha = 24.42'' \text{ and } \alpha_x = 5^\circ 45'9.58''$$

$$H = 672.243 \cos 5^\circ 45'9.58'' = 668.857 \text{ m}$$

$$\text{Elevation B} = \text{Elevation A} + 1.815 + 672.243 \sin 5^\circ 45'9.58'' - 1.995$$

$$= 345.472 + 1.815 + 67.382 - 1.995 = 412.674 \text{ m.}$$

9.11 EDM WITHOUT REFLECTING PRISMS (REFLECTOR LESS MEASUREMENTS)

Most modem EDM instruments can measure distances using reflecting prisms. The measuring surface itself is used as a reflector. Some models use time pulsed infrared signals, transmitted by a laser diode.

These EDM instruments can be used conventionally using reflecting prisms for distances up to 4 km. Their capability of measuring distances without prisms' limits to 80–250 m, depending upon the light conditions (cloudy days and night darkness provide better measurement of distances) and the smoothness of the reflecting surface. Usually, when measuring with prisms, the accuracy comes to around ±(3 mm + l ppm) for top end models. But the accuracy drops to 610 mm when measurements are taken without prisms. Targets with light coloured flat surfaces perpendicular to the measuring beam provide the best range accuracies.

EDM measurements without reflecting prisms provide quick results (within 0.8 sec in rapid mode and 0.3 sec in tracking mode), which means that applications for moving targets are possible. This type of application can be done in near shore hydrographic surveying and in many areas of heavy constructions. This technique is already being used with an infrared data collector to measure cross-sections in mining applications. In taking cross-sections of above-ground excavated works measuring dangerous or difficult access points, like bridge components, cooling towers and face of dams, this technique can be employed effectively.

Some manufacturers employ laser diodes for reflectorless measurements. These instruments permit direct acquisition of distances from the target. These instruments are used with an attached laser, which helps to positively identify the feature being measured. That is, the visible laser beam is set on the desired features, so that the surveyor can make sure that the correct surface is being measured and not the feature just beside it or just behind it. As the measurement is very fast care must be taken in not to measure by mistake some object(s) that may temporarily intersect the measuring signal, for example, the traffic or a branch of a tree moving due to the wind. The methodology adopted for reflectorless measurement using laser diode is as follows.

Two coaxially measuring distances are incorporated into the reflectorless distance measuring total stations as in Fig. 9.21. Both distance measurers operate on the well-known phase measurement principle. For reflectorless distance measurements, a laser beam is nominally used. The infrared laser beam for measuring distances to prisms has a wavelength of 780 nm, and is used for

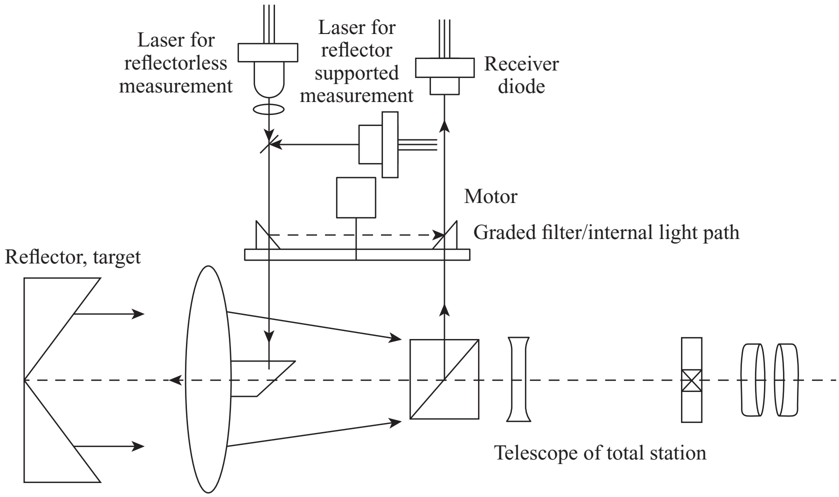


FIGURE 9.21 Optical design of a total station measuring without a reflector

range measurements of 3,000 m with a single prism at an accuracy of 2 mm + 2 ppm. The visible red laser beam has a wavelength of 670 nm and measures distances up to 80 m with an accuracy of 3 mm + 2 ppm without using a reflector. In modem instruments, distance measurements up to 12 km are done by using a special frequency system. The long-range measuring mode can be engaged for long distances. With it, the laser beam can also be used to measure the prism at distances of 1000–5000 m.

9.12 FOCUSSING AND SIGHTING

The reticule cross hairs are focussed by pointing the telescope at some other uniformly light surface (the sky), and by rotating the inner black portion of the eyepiece until the cross hairs are sharp and black.

The telescope is pointed towards the reflector by means of optical sight. Use of this sight will speed up acquisition of the reflector significantly. With both eyes open, one can have the image of the white cross hair in the optical sight in one eye, and the reflector in the other. When the two are superimposed, one is approximately on the target. This saves a significant amount of time searching for the reflector with the actual telescope.

The horizontal and vertical rotations of the theodolite are controlled in two ways. First, they can rotate the theodolite horizontally by hand by placing two hands on either side of it and rotating it. To rotate vertically, the EDM unit is rotated up or down by hand by nudging it up or down.

9.13 EDM ACCURACIES

EDM accuracies are stated in terms of constant instrumental errors and a measuring error proportional to the distance being measured.

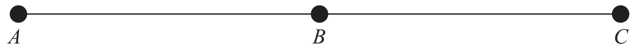


FIGURE 9.22 Method of determining the instrument-reflector constant

Typically, accuracy is claimed as 6(5 mm + 5 ppm). The 65 mm is the instrumental error that is independent of the length of the measurement, whereas the 5 ppm (5 mm/km) denotes the distance related error.

Most of the instruments now available are claiming accuracies in the range of 6(3 mm + 1 ppm) to 6(10 mm + 10 ppm). The preoperational part error (ppm) is insignificant in most of the work, and the constant part of the error is less significant when the distances being measured are more. At 100 m, an error of 65 mm represents 1/20,000 accuracy, whereas at 1,000 m, the same instrumental error represents 1/200,000 accuracy.

When dealing with accuracy, one should note that both the EDM and the prism reflectors must be corrected for off-centre characteristics. The measurement being recorded goes from the electrical centre of the EDM to the back of the prism (allowing for refraction through the glass) and then back to the electrical centre of the EDM. The EDM manufacturer at the factory or in the field compensates for the difference between the electrical centre of the EDM and the plumb line through the tribrach centre.

The EDM/prism constant value can be field checked in the following manner. A long line (>1 km length) is laid out with end stations and an intermediate station as in Fig. 9.22.

The overall distance AC is measured, along with partial lengths AB and BC. The constant value will be present in all measurements. Therefore,

$$AC - AB - BC = \text{Instrument/prism constant.}$$

Alternatively, the constant can be determined by measuring a known baseline, if one can be conveniently accessed.

9.14 DIRECT REFLEX EDM TECHNOLOGY

The latest technology in EDM is the direct reflex measurements and electro-optic distance measurement without cooperative targets or prisms. Direct reflex measurements enables the surveyors to measure remote points accurately without locating a physical target at each point. Hence, it makes the possibility of one-person surveying. When direct reflex combines with robotic technology, the concept of one person total station surveying increases even further. EDM measurement without a co-operative target can be achieved by either two methods: time of flight (pulsed laser) or phase shift.

9.14.1 Time-of-Flight (Pulsed Laser) Measurement

Time of flight employs a pulsed laser to provide the measurement. The time-of-flight method measures timing information to calculate a range measurement. The EDM generates many short infrared or laser light pulses, which are transmitted through the telescope to the target. These pulses reflect off the target and rerun to the instrument, where the inbuilt electronics determine the round trip time for each light pulse. As the velocity of light through the medium can be accurately estimated, the travel time can be used to compute the distance between the instrument and target. Although the time-of-flight method has longer range, it also meets the highest standards for eye safety, because the intervals between laser pulses prevent the accumulation of energy which can harm. Each pulse is a direct range measurement, so if thousands of pulses are sent in a second, a good average value in measurement can be taken quickly. Typically 20,000 pulsed laser measurements are taken every second. This is averaged to an accurate distance measurement value.

9.14.2 Phase-Shift Measurement

The direct reflex standard is a laser distance unit based on the phase comparison method. The EDM transmits a coaxial intensity modulated optical measuring beam that is reflected by a prism or scattered by a surface on which the beam is directed. The phase difference between the transmitted light and the reflected received light is detected and represents the distance. The direct reflex standard EDM works on prism mode as a fast and precise long-range EDM up to 3,000 m to a single prism. In direct reflex mode, the direct reflex standard EDM transmits a collimated visible red laser beam to the target point and the distance is calculated between transmitted and received light. The instrument measures a constant phase offset despite inevitable variations in the emitted and received signal. Only the phase offset is obtained through the phase comparison, initially a cycle ambiguity prevents the total distance from being directly estimated. The cycle ambiguity is resolved using multiple measurements modulation wave lengths, which provide unique integer number of cycles. Once the integer number is achieved, the distance to the target can be very accurately determined.

9.14.3 Comparison of the Two Methods

The time-of-flight method used light pulses to directly measure distances, while the phase-shift method uses modulated light to measure a phase shift, which provides distances once a cycle ambiguity is resolved. The pulses used for the time-of-flight method can be many times more powerful than the energy used for a phase-shift EDM. The time-of-flight method can therefore measure much longer distances, with or without a prism, than the phase-shift technique. The time-of-flight method has been slightly less accurate than the phase-shift method. The time-of-flight and phase-shift methods also differ in their tolerances of interruptions to the line of sight during measurement (e.g., interruptions such as traffic passing through a beam while measurements near roads).

9.14.4 Laser Safety Standards

The laser pulses used in time-of-flight method, although powerful enough to measure at ranges of hundreds of metres, are short in duration and therefore the laser beam does not accumulate energy. The continuous laser beams that are sometimes used to extend the range of phase-shift EDM instruments can lead to an accumulation of energy, which may be hazardous. The three laser classes relevant to most surveying instruments are Class 1, Class 2 and Class 3R.

Class 1 lasers are invisible lasers that meet the highest safety standards. Direct exposure of the measurement beam to skin or naked eye is unlikely to harm. It also do not cause danger if another surveying instrument is pointed into the source of the Class 1 laser beam.

Class 2 lasers emit visible laser radiations, which may be hazardous to the naked eye, if the beam is stared directly into the eye. Users must be very careful to avoid looking directly into the beam with optical instruments such as binoculars or other surveying instruments. Class 2 lasers are generally safe to use in public places without any special precautions, other than not staring directly into the laser beam. Regulations do not require the use of warning signs, audible warnings, etc.

It is possible to extent the range of phase-shift EDM by increasing the power of the light sources, typically from less than 1 mW to 5 nW. However, the higher powered continuous laser light emitted increases the health and safety risks of the laser beam, changing the classification of the laser to Class 3R. Only qualified and trained persons should be assigned to install, adjust and operate Class 3R laser equipment. The area in which these lasers are used should be posted with an appropriate laser warning sign. Care should be exercised to prevent the unintentional reflection of radiation. Only Class 1 and Class 2 laser instruments should be used for demonstration.

REVIEW QUESTIONS

1. Define electronic distance measurement (EDM) used for surveying and explain its importance in advanced surveying.
2. Explain the measurement principle in electronic distance measurement.
3. What are the characteristics of EDM instrument?
4. Write short notes on different wave bands used for electronic distance measurements?
5. What are the errors in electronic distance measurements?
6. Derive a formula to calculate the final corrected distance (S_0) of an EDM.
7. Derive a formula for the calculation of the variance of the final corrected distance (σS_0) of an EDM.
8. How are errors corrected in electronic distance measurements?
9. What is a zero correction in an EDM, and how is it carried out?
10. What is a prism integer? How it is calculated?
11. Describe different types of prisms used for electronic distance measurements.
12. Describe the field procedures of an EDM instrument for the measurement of a linear distance.
13. What you mean by geometry of an EDM? Explain with a neat diagram.
14. How are electronic distance measurements taken without reflecting prisms?
15. Write a short note on EDM accuracy.

SURVEYING USING TOTAL STATION 10

Chapter Outline

10.1 INTRODUCTION

Conventional surveying has used analogue methods of recording data. The present trend is to introduce digital surveying equipment into the field for a fully digitized work. Electronic total stations now perform the fastest digital data collection methods. Total stations have drastically increased the amount of topographic data that can be collected during a day. Figure 10.1 shows a flow diagram of the digital information from field to finish. The method is well suited for topographic surveys in urban landscapes and in huge construction sites. Modern total stations are also programmed for construction

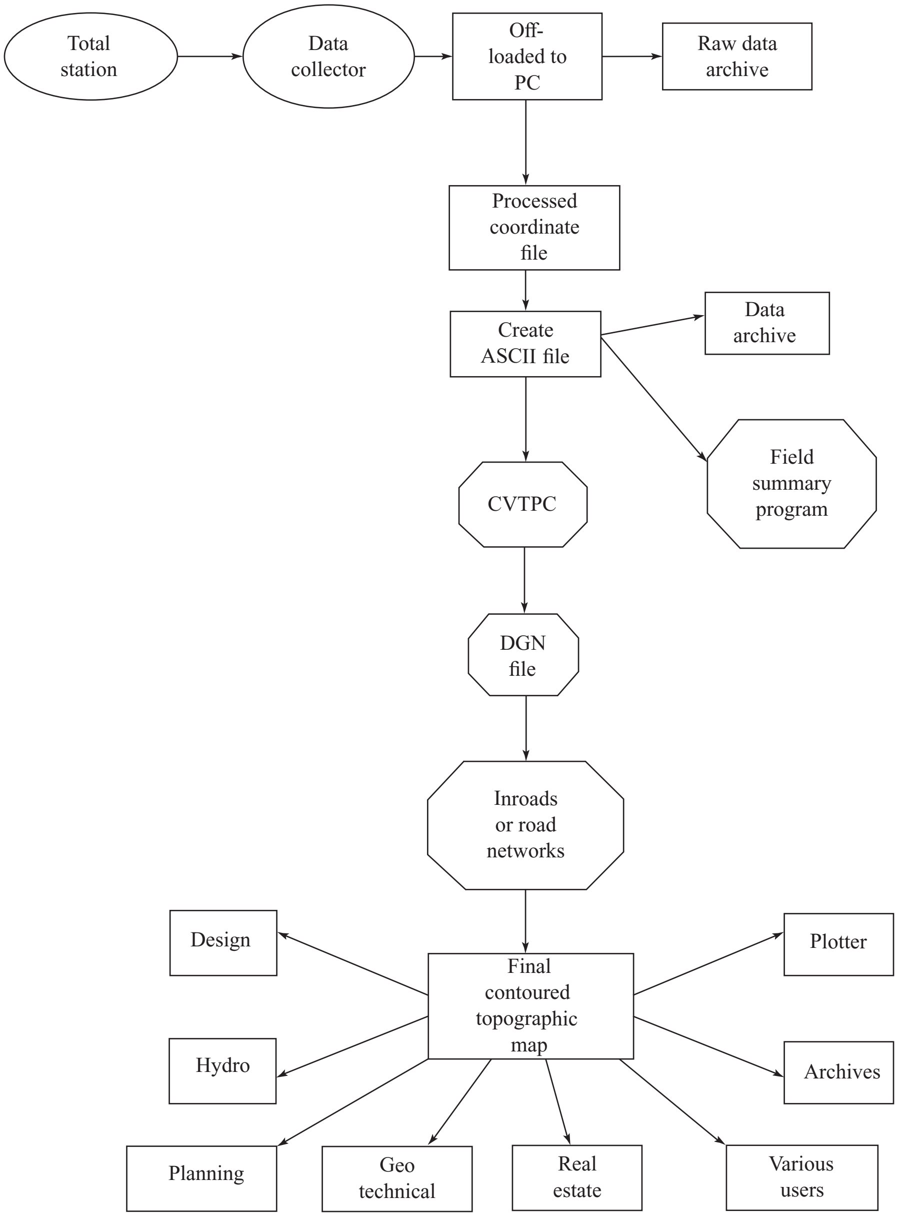


FIGURE 10.1 Total station data—from field to finish

stakeout and highway centreline surveys. Total stations have made trigonometric levels as accurate as the different differential level techniques in areas possessing large relief landforms. These instruments can quickly transfer three-dimensional (3D) coordinates and are capable of storing unique mapping feature codes and other parameters. In the past, this could only be recorded on paper media such as field books. One of the best features of the total station is the ability to download data directly into a computer without human errors.

Before the introduction of total station, a plane table has long been regarded as the best way to map a small area. The output of a plane table is just a low precision analog drawing. As a map is drawn directly on a sheet in a fixed scale, there was no way to improve the quality.

The advent of the total station survey has made it possible to accurately gather enormous amounts of survey measurements quickly. Even though total stations have been in use for more than 30 years, they are only now beginning to become popular among the surveying and engineering community. Over the last 25 years, total stations have become common field equipment.

In the early 1980s, surveying instrument manufacturers introduced what has become a true total station, redefining the term by creating an entirely electronic instrument. The read out on the display panels of a total station and the readout from the EDM are in a digital form. This feature eliminated reading errors, which can occur while using an optical theodolite. Due to the versatility and the lower cost of electronic components, future field instruments will be more like total stations that measure angle and distance simultaneously having all capabilities of theodolites, having electronic recording of horizontal and vertical angles and storage capabilities of all relevant measurements (spatial and non-spatial attribute data) for manipulation with computer. Using a total station, one can measure a distance to a suitable range with an accuracy of better than 5 mm (plus 1 ppm), and angles can be turned with the accuracy of 0.5 sec, all accomplished electronically.

A vast increase in productivity resulted due to the introduction of this electronic equipment. In most land surveying situations, the normal crew size can be reduced to two when equipped with an electronic total station. As the data acquisition time is so fast, in some situations, more prism poles are employed for the fast execution of survey works. This often results in an overall reduction in man-hours spent on the job. Figure 10.1 depicts the capability of an electronic total station to perform direct field-finish mapping products. In the late 1990s, total station became more popular and affordable.

10.2 FUNDAMENTAL PARAMETERS OF TOTAL STATION

The fundamental parameters of the total stations are the parameters for calculation and the correction factors and constants.

10.2.1 Parameters for Calculation

A total station is a digital theodolite with an EDM and a microprocessor. The theodolite measures the horizontal angle (Hz) and the vertical angle (V) of the line of sight from the centre of the total station to the centre of a target on a point to measure. The centre of the total station is at the intersection of the rotation axes of the horizontal and vertical circles. The centre of a target is at the intersection of the axis of the centering rod and the axis of tilting. The EDM measures the slope distance between the centre of the total station and the centre of a prism (not the centre of the target). The CPU calculates the coordinates (E, N, H) in a rectangular system of the point under the target, with reference to the total station coordinates ($E0$, $N0$, $H0$) using the measured polar coordinates (SD, Hz, V), instrument height, target height and several correction factors and constants (see Fig. 10.2).

10.2.2 Correction Factors and Constants

The scale factor (in ppm) is a sum of atmospheric correction ($\Delta D1$ in ppm), in reduction to mean sea level ($\Delta D2$ in ppm) and the projection scale factor ($\Delta D3$ in ppm). The atmospheric correction factor is a function of temperature, atmospheric pressure and humidity. Input of atmospheric correction is usually required. Most user manuals have diagrams to estimate atmospheric corrections from

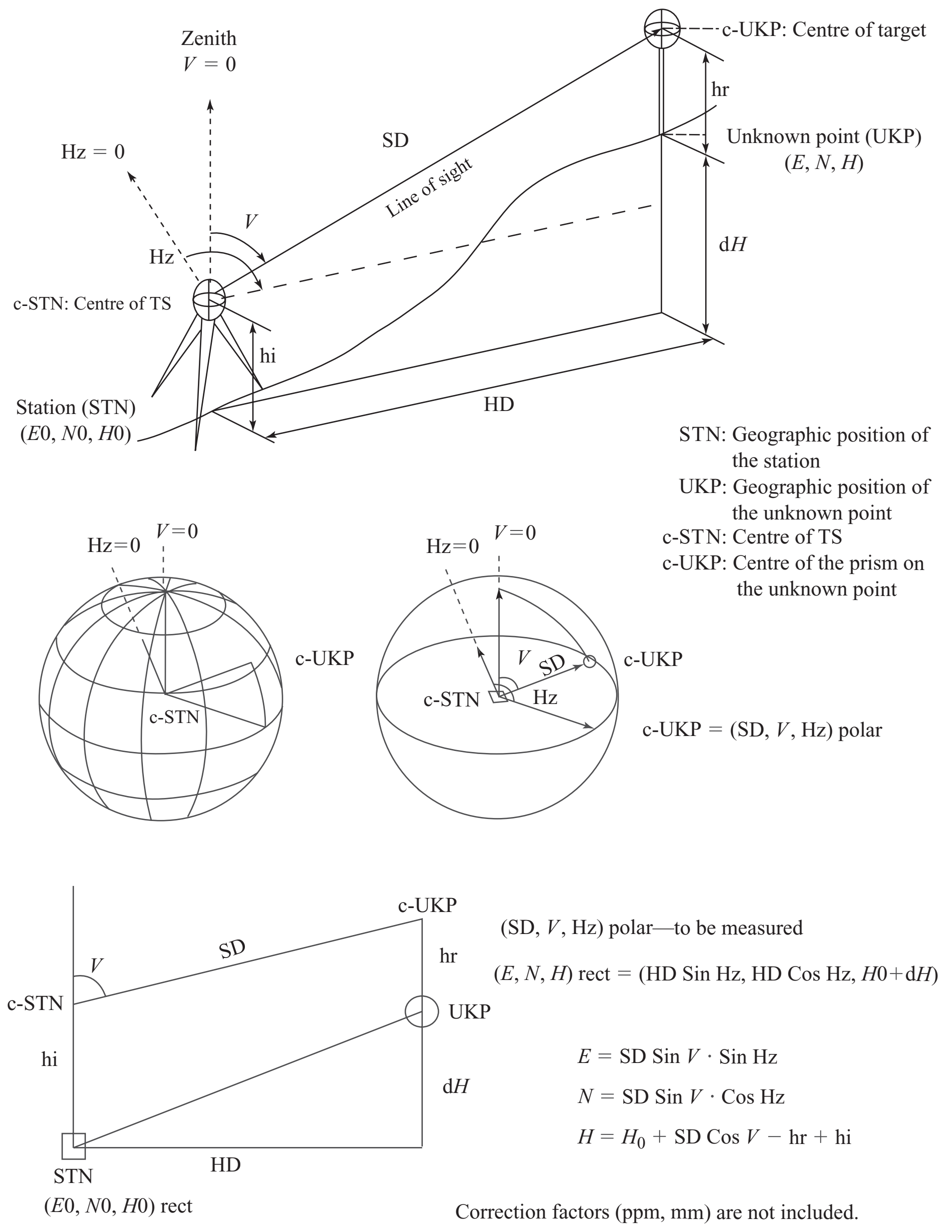


FIGURE 10.2 Fundamental parameters and formulae

the three parameters. New total stations can calculate atmospheric corrections. Using the formulae narrated below, the atmospheric correction can be calculated very easily. In any case, a thermometer and barometer are necessary to estimate atmospheric correction.

$$\Delta D1 = 281.8 - \left[\frac{0.29065 \cdot P}{(1 + a \cdot t)} - \frac{4.126 \cdot 10^{-4} \cdot P}{(1 + a \cdot t)} 10^{x} \right]$$

where $\Delta D1$ = the atmospheric correction (ppm);

P = the atmospheric pressure (hPa);

t = the temperature (°C) at the time of observation;

and h = the relative humidity (%).

$$\alpha = 1/273.16$$

$$x = \frac{7.5t}{273.3 + t} + 0.7857.$$

Reduction to mean sea level ($\Delta D2$) and projection scale factor ($\Delta D3$) are necessary to map in scale on the earth's spheroid surface at a particular geographic locality. If site accuracies are considered rather than geographic accuracy, reduction to mean sea level and projection scale factor need not be considered. Practically, the amounts of correction derived from these factors are less than 100 ppm.

The prism constant is to be added to the distance measured by the EDM. The path of the laser between the source in a total station and the optical centre of a prism (reflector, target) (Fig. 10.3) is not same as the true distance. The prism constant indicates the distance in millimetre to be added to (positive) or subtracted from (negative) the measured distance. Typically, the constant ranges from 0 to 50 mm. The mistake in setting these consonants' results in much larger error than other factors. For example, '0 mm' Leica prism constant is not applicable for the Sokkia total station.

For exact calculation of the distance and the height difference, the curvature of the earth's surface and the refractive index of the atmosphere are to be taken into account. The user's manual is to be referred for the reduction formulae employed for each model total station. The CPU is programmed to use these formulae to give coordinates of measured points. When one recalculates rectangular coordinates from polar coordinates, one should use these formulae, although the difference is not very big.

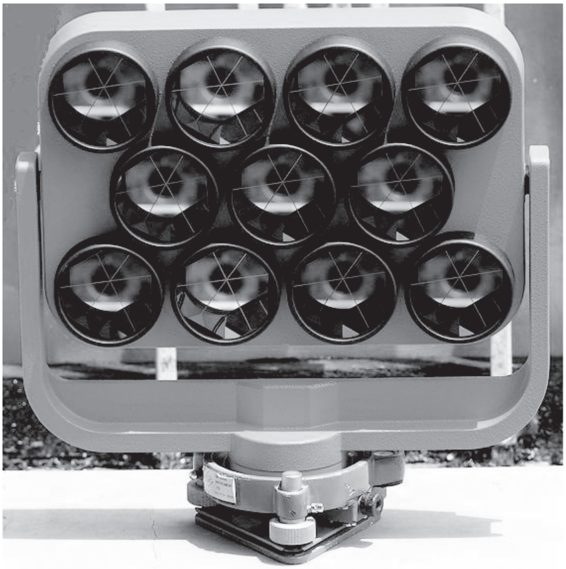


FIGURE 10.3 Different prisms used for total station surveys

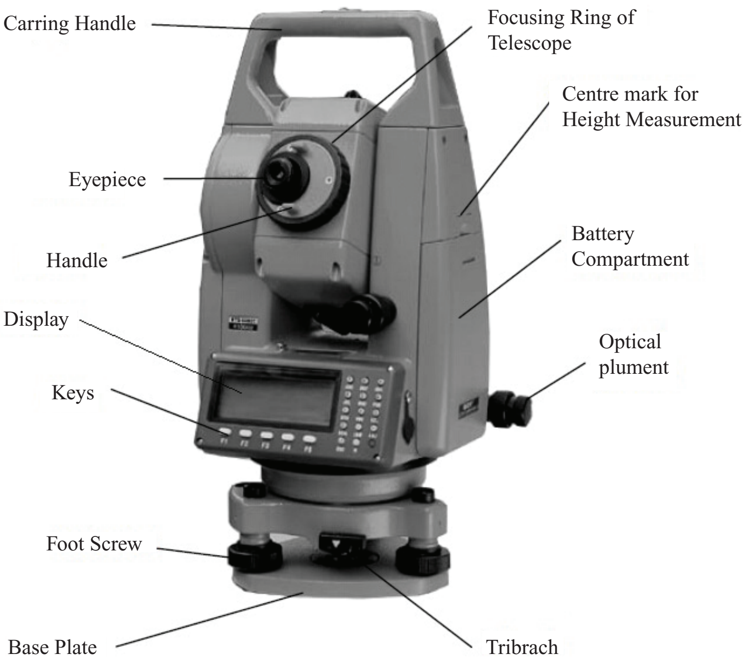


FIGURE 10.4 Parts of a total station

10.3 PRECAUTIONS TO BE TAKEN WHILE USING A TOTAL STATION

The following precautions need to be taken while using a total station:

1. Always carry a total station in a locked hard case even for a very short distance. Take the total station out of the hard case only for fixing it firmly on a tripod for taking observations.
2. Do not move or carry a tripod with the total station fixed on it, except for centering.
3. Use both hands to hold the total station handle.
4. Never release the handle before the total station is fixed with the tripod's fixing screw.
5. Set up the tripod as stable as possible.
6. Always keep the top of the tripod, the bottom and top of the tribrach and the bottom of the total station clean and away from any shock and impact.
7. Take maximum care when the tribrach is removed from the total station.
8. Do not make the total station wet.

10.4 FIELD EQUIPMENT

Modern electronic survey equipment requires surveyors to be more maintenance conscious than they were in the past. They have to take care about power sources, downloading data and the integrity of data, including whether or not the instruments and accessories are accurately adjusted and in good form. When setting up a crew to work with a total station and a data collector, it is helpful to supply the party chief with a checklist to help the crew maintain its assigned equipment and handle the collected data upon returning to the office. It is also important that each crew should be supplied with all necessary equipment and supplies. These should be stored in an organized and easily accessible manner.

Preparing an equipment list carefully will assure the survey crew (two-person crew, consisting of a party chief/rod person and a notekeeper/instrument person) a sufficient equipment inventory to meet the general needs of boundary, layout and topographic surveys. This procedure will confine what is needed to maximize productivity when using a total station with a data collector.

The minimum equipment inventory required is as follows:

1. Total station set
   a. Total station instrument in a hard case
   b. Battery charger
   c. Extra batteries
   d. Memory module/card, serial cable
   e. Rain cover
   f. User manuals
   g. Tripod
   h. Tape measure
2. Prism set
   a. Prism
   b. Prism holder c. Centering rod
3. Back sight set
   a. Prism
   b. Prism holder
   c. Prism carrier (to be fixed on tribrach, with optical/laser plummet)
   d. Tribrach (to exchange prism carrier and total station)
4. Data processing
   a. Laptop computer with serial port or USB port
   b. Serial cable or USB-serial adaptor
   c. Terminal application
   d. Application programme: MS Excel, Adobe illustrator, Coordinate Converter, etc. e. Data backup device and media (zip, memory card, etc.)
5. Survey tools
   a. Stakes, nails, paint, marker b. Hammer
   b. Thermometer, barometer/altimeter
   c. A pair of radio (with hand-free headsets)
   d. Clipboard, field note, pen f. Compass
   e. GPS

The total station set, one set of the prism set and survey tools are indispensable throughout. The data can be processed back in the office, but it is better to arrange all data and make a backup as soon as possible in the field itself, before forgetting mistakes and details. With one back sight set, measurement of a back sight and traverse will be very easy and accurate. To occupy a new point, simply exchange the part above the tribrach. The centre position of the total station and the target centre above the tribrach are the same after the exchange.

With this equipment inventory, a two-person field crew will be able to handle most of the survey tasks that are routinely encountered in day-to-day operations. An additional tripod, plumbing pole, carrier, tribrach and reflector would give the crew even greater flexibility, and allow them to handle many projects more efficiently. It is also helpful for the field crew to have a convenient place to store their assigned equipment. The crews should be equipped with briefcase-sized cases that will hold three tribrachs, four reflectors with holders, three carriers and four target plates. A hard camera case or pistol case works well for this purpose. With all the components stored in one place, it makes the inventory of the equipment easy and reduces the chance of the equipment being left at the job site. This also allows for proper equipment maintenance.

10.5 TOTAL STATION SET UP

The following steps are followed for the set up of a total station:

1. Choose an adequate instrument station. Make sure that an observer can safely operate the instrument without knocking it over. It is necessary to have the centre of the instrument, which is the point of intersection of the transverse axis and the vertical axis of the instrument, directly over a given point on the ground (the instrument station).
2. Remove the plastic cap from the tripod, and leave the instrument in the case until the tripod is nearly level. Stretch the tripod legs 10–15 cm shorter than their maximum length.
3. Open the legs of the tripod to set the tripod head at the level of the operator's upper chest. When the total station is set up on the head, the operator's eye should be slightly above the eyepiece. The instrument height is important for an effective and comfortable survey. It differs in the looking-down position and the looking-up position. One should not touch or cling to the tripod during the survey.
4. At a new station without a reference point on the ground, level up the total station at an arbitrary point, where a stake can easily go in and be steady, and put down the stake at the centre using the plummet.
5. To occupy an existing station above a reference point, first roughly level up the tripod head right above the point. For levelling up, a small level is useful. To find out the position, use a plumb bob or drop a stone through the hole in the tripod head.
6. Once roughly levelled and centred, push each tip of the tripod leg firmly into the ground, applying full weight of the observer on the step above the tip. Apply the weight along the tripod leg without bending it.
7. Check the level and centre it again. Adjust the level by changing the leg length.
8. Fix a tribrach with a plummet, a tribrach and a prism carrier with a plummet or a total station with a built-in plummet on the tripod head.
9. Adjust the three screws of the tribrach to centre the bubble of a spirit level with the following steps:

a. Release the lock of the horizontal circle.
b. Rotate the instrument to set the plate level parallel to AB at the first position.
c. Turn the foot screws A and B in the opposite direction, the same amount to centre the plate level. This will adjust the tilt on the aa axis.
d. Rotate the instrument 90 degrees to set the plate level at the second position.
e. Turn the foot screw C to the centre plate level, adjusting the tilt along cc.
f. Rotate the instrument 90 degrees to set the plate level in the third position.
g. Turn the foot screws A and B in the opposite direction, the same amount to eliminate half the centering error.
h. Rotate the instrument 90 degrees to set the plate level at the fourth position.
i. Turn the foot screw C to eliminate half the centering error.
j. Repeat b to i until the plate level is centred in all directions (give a little time for slow movement of the bubble in viscous fluid).

10. Pull out the optical plummet and use the optic ring to focus at the graticule and then focus at the mark on the ground. Or, turn on the laser plummet. Rotate the plummet or the total station to check it is centred within 1 cm from the reference point. If not, estimate the amount of offset and carefully translate the entire tripod as much as the offset. Return to 4 and try to level and centre again. The total station on the tripod head can be translated 1 cm from the centre; therefore, rough centering within 1 cm is necessary. Be careful to see that the centre of the optical plummet or the laser point is on an axis perpendicular to the horizontal circle of the total station. If the total station is not level, the plummet line does not coincide with the plumb line.
11. Put the total station on the tribrach if it is not there.
12. Use the plate level for the final levelling of the total station. Follow the instruction given in Fig. 10.5.
13. When the total station is finely levelledup, use the plummet to check centering. If the plummet centre is off the reference point, slightly loosen the fixing screw below the tripod head and

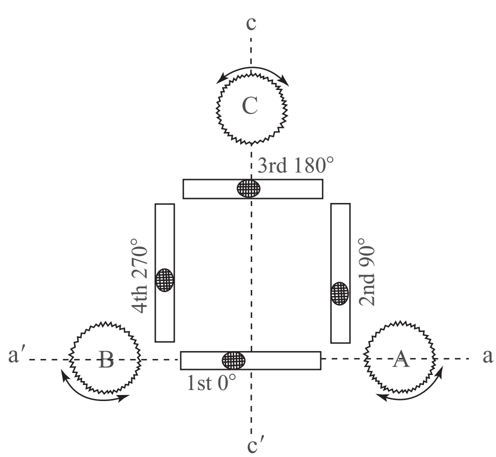


FIGURE 10.5 Fine levelling procedure

translate the tribrach to place the plummet centre on the exact point. Do not rotate. When the translation is done, tighten the fixing screw moderately. If any portion of the base of the tribrach goes outside the tripod head, return to 4.

14. Rotate the total station by 180 degrees. If the plummet centre goes away from the point, slightly loosen the fixing screw and slide the total station halfway to the centre.
15. Repeat the steps 12 and 13 until the plummet centre stays exactly on the centre of the mark.
16. Tighten the fixing screw firmly without applying too much pressure. Never loosen the screw until all the measurements are finished.
17. Measure the instrument height. The centre of the total station is marked on the side of the alidade. The vertical distance between the mark and the ground is the instrument height.
18. Check the plate level from time to time during measurement before the total station tilts beyond the automatic correction.

10.6 SETTING UP A BACK SIGHT

A back sight is a reference point for the horizontal angle. At the beginning of a new survey, a back sight can be set at an arbitrary point and marked. The best way to set up a back sight is to use a prism carrier and a tribrach on a tripod. The procedure for levelling up and centering of the prism is the same as that for the total station. If there is no plummet in the tribrach and the prism carrier, use the plummet of the total station and then exchange the total station above the tribrach with a prism carrier. A prism should be put right on the reference point when sighting is possible from the total station.

Measure the target height at the back sight. This height is the vertical distance between the centre of the target (prism) and the ground beneath. When the station and back sight are ready, measure the azimuth from the station to the back sight using a compass. The azimuth is between 0 and 360 degrees measured clockwise from north. Correct the magnetic declination to get the true azimuth and record the true azimuth. If the geographic coordinate or grid coordinates of the point occupied by the total station and the target at the back sight is known, then the total station will automatically calculate the true azimuth, provided the station values are fed into the total station manually.

By using a surveyor's compass or a GPS receiver, very accurate orientation can be done. Carrier phase GPS receivers easily attain less than 10 mm accuracy on a 50-m baseline. GPS is useful to fix both orientation and geographic location of the total station measurement. Measure the station, back sight and the furthest points more than 200 m apart in four directions, both with a GPS receiver in UTM coordinates and with a total station. When we process the data after measurement, rotate and translate the total station coordinates to best fit with the GPS coordinates and then plot on a UTM grid (by this method, comparison or check can be done both for GPS and total station measurements).

10.7 AZIMUTH MARK

An azimuth mark is a back sight without a prism. Only the azimuth is measured from the station. An azimuth mark is a distal point or an object with a sharp and clear vertical edge to be taken as a reference of orientation from the station. It should be a geometric point or a vertical lane to aim at, with precision. At the same time, it must be so distinct that no confusion among other objects should happen. Once a nice azimuth mark is found on the telescope, keep a detailed sketch and comments in the field book.

An azimuth mark may substitute for a back sight for certain total stations, in case it is not necessary to define the errors. Some CPU and data processing applications do require back sight measurements both for angles and distances. Even when one sets up a back sight, setting up of additional azimuth marks beside the back sight is useful to check whether the configuration has not gone wrong.

10.8 MEASUREMENT WITH TOTAL STATION

When both the total station and back sight are finely levelled and centred, the hardware setup is over and the software setup is to be started. The software setup of a total station differs from one make to another. One has to follow the user's manual of each instrument. The list below gives common important settings for most instruments. Most total stations memorize these settings, but it is better to check through the setup menu to avoid a false setting.

System: Choose appropriate existing interface for data output

Angle Measurements: Tilt correction/Tilt compensator (2 axis)

Horizontal angle increments: At right angles (clockwise)

Unit setting: Angle in degrees/min/sec, distance in metres, temperature in degree centigrade and pressure in hectopascal.

EDM settings: Select IR laser, fine measuring mode, use RL with caution. Set appropriate value for the prism constant (from the user's manual of the equipment)

Atmospheric parameters: Get parts per million for the diagram from the manual of the equipment or let the total station calculate from the atmospheric pressure in hectopascal and temperature in degree centigrade.

Communications: Set all communication/interface parameters same for a total station and data logger/PC. They are baud rate, databits, parity, end mark and stop bits. Refer to the manual of each device before data downloading.

10.9 TOTAL STATION INITIAL SETTING (GENERAL SETTING REQUIRED FOR ALL MODELS)

The following are the steps for the initial setting of a total station:

1. Turn on the total station.
2. Release both horizontal and vertical locks.
3. Some total stations require rotating the telescope through 360 degrees along the vertical and horizontal circles to initialize angles.
4. Adjust the telescope to best fit to the observer's eye. Using the inner ring of the eyepiece, make the image of the cross-hair sharp and clear.
5. Rotate the alidade until the Hz angle reading is equal to the azimuth to the back sight measured by the compass (for Sokkia models only). Push the HOLD key once. The Hz angle will not change until the next hold.
6. Aim at the very centre of prism at the back sight. For the coarse aiming, rotate the alidade and the telescope by hand using optical sight. Adjust focus using the outer ring of the eyepiece.

When the prism comes into the sight and close to the centre, lock the horizontal and vertical drives. Then use dials to aim at the exact centre of prism.

7. For Sokkia models, push HOLD button again. The horizontal reading will now change according to the rotation of the telescope in the horizontal direction. For Leica models, input the azimuth of the back sight manually in the measurement setup window.
8. If a station ID and back sight ID are required, use a two- or three-digit serial number like 101, 102, . . . for each reference point. Use a four-digit number for unknown points.
9. Input station parameters like hi (height of the instrument), *E*0, *N*0 and *H*0 (easting, northing and RL of the point where the instrument is set up). Use 1,000, 1,000, and 1,000 for *E*0, *N*0 and *H*0 to avoid negative figures. If the coordinates are known, manually input the data.
10. Input the target height (hr).
11. Check the pointing at the prism again.
12. Using the distance calculation key, make the back sight measurement. From the LCD display of the total station, note the horizontal angle, the vertical angle, slope distance, easting, northing and height, and record them in a field book with a sketch of the plan. Here the horizontal angle, vertical angle and the slope distance are the raw data.
13. Create a new job or open an existing job. A job is a block of data sets stored in the memory like a file. One can create a new job or append data to an existing job. A job name is used as an output file name in a new Leica total station with .gsi extension.

10.10 FIELD BOOK RECORDING

The observer can record all numerical data and a little text data in the total station, but descriptive information and graphic information should be recorded in the field book. The following is a suggested list for the survey records:

1. Place, date and time
2. Surveyor's name
3. Temperature, atmospheric pressure
4. Station coordinate (*E*0, *N*0, *H*0), UTM by GPS and height of the instrument
5. Back sight coordinates (slope distance, vertical and horizontal angels), (*E*, *N*, *H*), UTM by GPS and height of the reflector.
6. Azimuth mark Hz, sketch of the telescope view.
7. Sketch map of the sight and measured objects.
8. Description of measurement. Point ID number (from-to), object, height of the reflector. Repeat this for each discrete object, or group of points measured with different prism height. This must be the input to the total station each time it changes.
9. Back sight coordinates measured again at the end.

10.11 RADIAL SHOOTING

From a station, we measure as many objects as possible within the sight. This method is called radial shooting. The objects are classified as points. Each point is recorded as coordinates with a point ID number. A group of points on a discrete line should be measured consecutively for the sake of plotting. The pair of the first and the last point ID numbers is necessary to separate from the other ones. Smaller intervals of points result in more accurate records of the shapes. The balance between the scale of the map, importance of the objects and time and the purpose of the measurement determines the interval needed.

Regularly distributed points are essential for reliable representation of the surface geometry. A 4- or 5-m mesh is usually good for a wider area. During the measurement, repeat back sight measurement or azimuth check regularly, for example, 50 points to see there is no change in geometric configuration. Loose fixing screws or tripod bolts, shocks to the tripod or tripod on soft ground may result in the shift of the total station and back sight. An error should be within a few millimetres. Discard the data possibly affected by the troubles and measure the back sight carefully.

10.12 TRAVERSE

When it is not possible to view the entire mapping area from the first station, we traverse to a new station and repeat radial shooting. Adjusting the coordinates and orientation of the second station, measured coordinates from multiple stations will be in a unique system. Most total stations have a programme for traverse.

1. Set up a prism on a tripod, tribrach and prism carrier after centering on a mark on the ground. The back sight point may be used as a new station.
2. Measure the new station and record the $E1$, $N1$, $H1$ and horizontal angle, record the angular value in the memory and in a notebook. Turn off the total station.
3. Leaving the tribrach on the tripod, exchange the total station above the tribrach with the prism on the prism carrier.
4. The exchanged total station and prism should be levelled and centred. Carefully apply small adjustments for fine levelling and centering.
5. Turn on the total station at the new station and point at the prism.
6. Run a traverse programme.
7. Input the station coordinate ($E1$, $N1$, $H1$) and the new height of the instrument (previous height of the prism)
8. Pointing the centre of the prism, set Hz0 (horizontal angle zero) as Hz1 + 180 or Hz1 − 180 (Hz1 > 180). Use the previous station as the new back sight.
9. Input the new hr (height of the reflector) and measure. The coordinate of the first station must be ($E0$, $N0$, $H0$). The error must be less than a few millimetres.
10. To define errors and evaluate accuracy, follow the standard procedures for surveyors.

10.13 SURVEY STATION OR SHOT LOCATION DESCRIPTION USING CODES

Each survey station or shot location (point), must be described with respect to surveying activity, station identification, and other attribute data. Total stations that come equipped with their own data collectors (e.g., Sokkia, Lieca, Topcon, Pentax) will, in many cases, prompt for the data entry (e.g., BS, FS) and then automatically assign appropriate labels, which will then show up on the survey

TABLE 10.1 Operation codes of survey points according to AASHTO

Code (Numeric)	Code (Alpha)	Descriptions
1	BM	Bench mark
2	CM	Concrete monument
3	SIB	Standard iron bar
4	IB	Iron bar
5	RIB	Round iron bar
6	IP	Iron pipe
7	WS	Wooden stake
8	MTR	Coordinate monument
9	CC	Cut cross
10	N&W	Nail and washer
11	ROA	Roadway
12	SL	Street line
13	EL	Easement line
14	ROW	Right of way
15	CL	Centreline

TABLE 10.2 Operation codes of topography according to AAHSTO

Code (Numeric)	Code (Alpha)	Descriptions
16	EW	Edge walk
17	ESHLD	Edge shoulder
18	C7G	Curb and gutter
19	EWAT	Edge of water
20	EP	Edge of pavement
21	RD	CL road
22	TS	Top of slope
23	BS	Bottom of slope
24	CSW	Concrete sidewalk
25	ASW	Asphalt side work
26	RW	Retaining wall
27	DECT	Deciduous tree
28	CONT	Coniferous tree
29	HDGE	Hedge
30	GDR	Guide rail
31	DW	Driveway
32	CLF	Chain link fence
33	PWF	Post and wire fence
34	WDF	Wooden fence

printout. Point description data can be entered as alpha or numeric codes. This descriptive data will also come up on the printout and can be tagged to show up on the plotted drawing if necessary. Some of the data collectors are now equipped with barcode readers, which when used with prepared code sheets permit instantaneous entry of descriptive data.

Some data collectors are designed to work with all total stations in the market. These data collectors have their own routines and coding requirements. As this technology continues to evolve, and continues to take over many surveying functions, standard procedures will hopefully develop.

A positive indication of the trend towards standardization surfaced with the 1990 Survey Data Management System (SDMS) developed by the American Association of State Highway and Transportation Officials (AASHTO). The SDMS identifies and defines all highway-related surveying tasks, activities and data tags. SDMS can be used for manual as well as automated collection and processing. As the tasks and activities in highway surveying are similar to or identical to all surveying tasks and activities, this new system could have particular application for general surveying situations in which third party data collectors are being used, or in which the nature of the data processing and plan preparations are such that the coding standardization would provide demonstrable benefits.

In addition, new data collectors and computer programmes not only permit the entry of the point code, but also permit the surveyor to enter the various levels of attribute data for each coded point. For example, a tie-in to a utility pole could also tag the pole number, the use such as electric or for telephones, the material such as wood, or concrete, connecting poles and year of installation. The type of expanded attribute data is typical of the data collected for a geographical information system (GIS). Tables 10.1 and 10.2 give the operation codes according to AASHTO.

10.14 OCCUPIED POINT (INSTRUMENT STATION) ENTRIES

An example of occupied point entries with codes are explained below (see Fig. 10.6). The entries are as follows:

1. Code 10 (see Table 10.3)
2. Height of the instrument (the measured value is entered into total station)

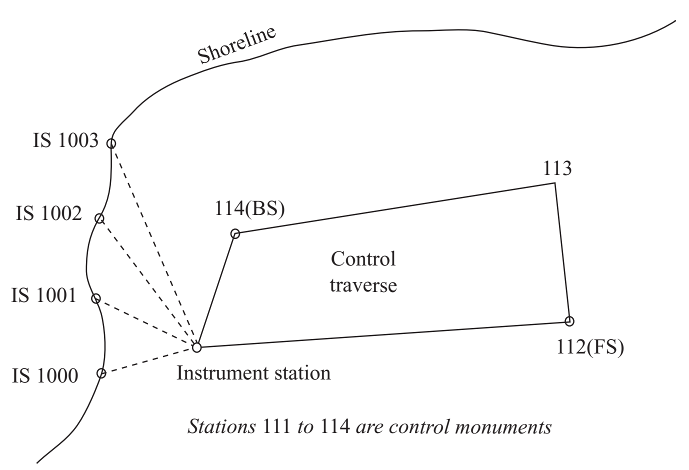


FIGURE 10.6 Sketch showing intermediate shoreline ties to a control traverse

3. Station number entered (111)
4. Station identification code entered (alpha code 10)
5. Coordinates of occupied station entered (either grid or geographical)
6. Coordinates of back sight (BS) station, or reference azimuth to BS station entered

TABLE 10.3 Typical alpha and numeric descriptors for survey activity

Activity	Code (Alpha)	Code (Numeric)
Occupied station	OCC	10
Back sight	BS	20
Fore sight	FS	30
Intermediate sight	IS	40

Note: In some data collectors, the coordinates of the above stations may instead be entered in the system computer file just prior to the field data reductions and adjustments.

Sighted point entries are coded as below:

1. Operation code 20, 30 or 40—for BS, FS or IS (for some models)
2. Height of the prism/reflector (HR)—measured and value entered
3. Station number, for example, 114(BS)
4. Station identification code (e.g., BM for bench mark)

The procedure for the control traverse is detailed below (Fig. 10.6):

1. Enter the initial data and occupied station data as detailed above.
2. Sight at station 114; zero the horizontal circle (any value can be set instead of zero).
3. Enter code 20(BS)—or respond to the data collector prompt.
4. Measure and enter the height of the prism/reflector (hr).
5. Press the appropriate measure buttons, for example, slope distance, horizontal angle and vertical angle.
6. Press the record button after each measurement; most instruments measure and record these three measurements when they are in the automatic mode, by pressing just one button.
7. After the station measurements have been recorded, the data collector will prompt for the station point number (e.g., 114) and the station identification code (e.g., 02).
8. If appropriate, as in traverse surveys, the next sight is to the FS station (code 30); repeat steps 4, 5, 6 and 7 using correct data.
9. While at station 111, any number of IS (code 40) can be taken to define the topographic features being surveyed. The prism/reflector is usually mounted on an adjustable-length prism pole with the height of the prism (hr) being set to the height of the total station (hi), The prism pole can be steadied with a brace pole to improve accuracy for more precise sightings. Some softwares permit the surveyor to identify, by a further code number, points that will be connected on the resultant plan (shoreline points in this example). This connection and off features permit the field surveyor to virtually prepare the plan (for graphics terminal or plotter) while performing

the actual field survey. Alternatively, the surveyor can connect the points while in edit mode on the computer. Clear field notes are essential for this activity.

10. When all the topographic details in the area of the occupied station (111) have been collected, the total station can be moved to the next traverse station (to 112 or so), and the data collection can be proceeded in the same manner as described earlier. That is, BS @station 111, FS @ station 113 and take all relevant IS readings.

10.15 DATA RETRIEVAL

Use a serial port (RS232C for PC), or serial cable adaptor or serial-USB adaptor to connect to the total station. Release the serial port from other drivers such as the modem, printer and scanner. Set the parameters of serial data transportation on the total station and the operating system of the PC. Use the communication setup in the total station and in the termination application. Typical parameters are 9,600-baud rate, 8 bits, no parity, 1-stop bit and X-on/X-off. Both ends of the connection should be set similarly for communication.

Make the PC application ready for receiving data and save the log, input and send command to the total station. The received character strings will be scrolled in the log window. When the ID of the last point is shown and scrolling stops, save the log as a text file.

The received data will be like a text file. Using MS-Excel point ID numbers, the measured parameters and coordinates are to be cut out. Spending a few minutes with MS-Excel will solve the problem. Unless recalculation from observed data is carried out, slope distance, horizontal angle, vertical angles needed, point ID, easting, northing and height are necessary. The prism height is usually recorded. Compare the data file with the field book to check that all the changes in the height of the reflector are correctly recorded. The missing adjustment results in sudden increase of height.

Given below is the Sokkia printing format:

OBS F12 003-6000	S.Dist 220.610	V.obs 93-49’50”	H.obs 225-08’30” <CR + LF>
	Code BS <CR + LF>		
POS TP 6000	North 1844.739	East 1843.969	Elev 1984.297 <CR + LF>
	Code BS <CR + LF>		

Here we need 6000, 220.610, 225.1416667, 93.835056, 1843.969, 1844.739, 1984.297 (No, SD, Hz, V, E, N, H).

Leica GIS format:

110005 + 0001001 21.324 + 09554200 22.324 + 10144350 31..00 + 00022985 81...00 + 01022385
82.00 + 00997685 83..00 + 00995514 51..1. + 0003 + 000 87...10 + 00001300 88..10 + 00001492
<CR + LF>

This format is much smarter than Sokkia’s (81, 82, 83) (E, N, H), 51 : ppm, 87 : hr, 88 : hi,

We need 1001, 22.985, 101.7430556, 95.9116667, 1022.385, 997.685, 995.514.

Modern total stations have data stored on board, eliminating the costly data collectors. Most of the modern total stations have the data storing devices such as PCMCIA cards, which can be directly read into the computer through PCMCIA card readers.

If a topographic data has been tied to a closed traverse, the traverse closure is calculated and then all adjusted values for northings, eastings and elevations (X, Y, Z) are computed.

10.16 FIELD GENERATED GRAPHICS

Many surveying software prog rammes permit the field surveyors to identify field data shots so that subsequent processing will produce appropriate computer graphics. For example, MicroSurvey International Inc. software has a typical 'description to graphics' feature that enables the surveyor to join field shots, such as curb line shots (see Fig. 10.7) by adding a 'Z' prefix to all but the last 'CURB' descriptor. When the programme first encounters the 'Z' prefixes, it begins joining the points with the same descriptors; then when the programme encounters the first curb descriptor without the 'Z' prefix, the joining of points is terminated. Rounding (curved curb) can be introduced by substituting an 'X' prefix for the 'Z' prefix (see Fig. 10.7).

Other typical graphics prefixes in the above example include the following:

1. 'Y' joins the last identical descriptor by drawing a line at a right angle to the established line (see fence line in Fig. 10.7).
2. '. . .' causes a dot to be created in the drawing file, which is later transferred to the plan. The dot on the plan can itself be replaced by inserting a previously created symbol like a tree, a manhole or a hydrant (see MH in Fig. 10.7). If a second dash follows the first prefix dash, the ground elevation will not be transferred to the graphics file (refer HYD in Fig. 10.7), as the feature elevation may not be required.
3. '.' Instructs the system to close back on the first point of the string of descriptors with the same characters (see BLDG and BUS SHELTER in Fig. 10.7).

Other software programmes create graphics stringing by similar techniques. For example, the Sokkia software gives the close itself a stringing capability (e.g., FENCE, CURB1, CURB2), which the surveyor can easily turn on/off (see Fig. 10.8).

In some cases, it may be more efficient to assign the point descriptors from the computer keyboard after the survey has been completed. For example, the entry of descriptors is time consuming on some electronic field books/total stations, particularly in automatic mode. In addition, some topographic features like the edge of water in a pond or lake can be captured in sequence, thus permitting the surveyor to edit in these descriptors efficiently from the computer in the processing stage. If the point descriptions are to be added at the computer, clear field notes are indispensable.

Refer to Fig. 10.9 for an illustration of this stringing technique.

The pond edge has been picked up (well defined) by 10 shots, beginning with 58 and ending with 67. Using an appropriate application programme (MicroSurvey International), the point description edit features can be selected from the pull-down menu and the following steps may occur:

1. 'Point to be described?' 58 . . . 66 (enter)
2. 'Description?' ZPOND (enter)
3. 'Points to be described?' 67 (enter)
4. 'Description?' POND (enter)

After the point descriptions have been suitably prefixed (either direct field coding or by editing techniques shown here), a second command is accessed from another pulldown menu that simply

Point	Description
1	ZC1
2	ZC2
3	ZFENCE
4	HYD
5	ZBUS
6	ZBUS
7	ZBUS
8	ZBUS
9	ZC1
10	XC1
11	ZC1
12	C1
13	ZC1
14	ZC1
15	XC1
16	MH
17	MH
18	ZFENCE
19	YFENCE
20	ZC2
21	ZC1
22	ZBLDG
23	ZBLDG
24	ZBLDG
25	BLDG
26	FENCE
27	C2
28	C1

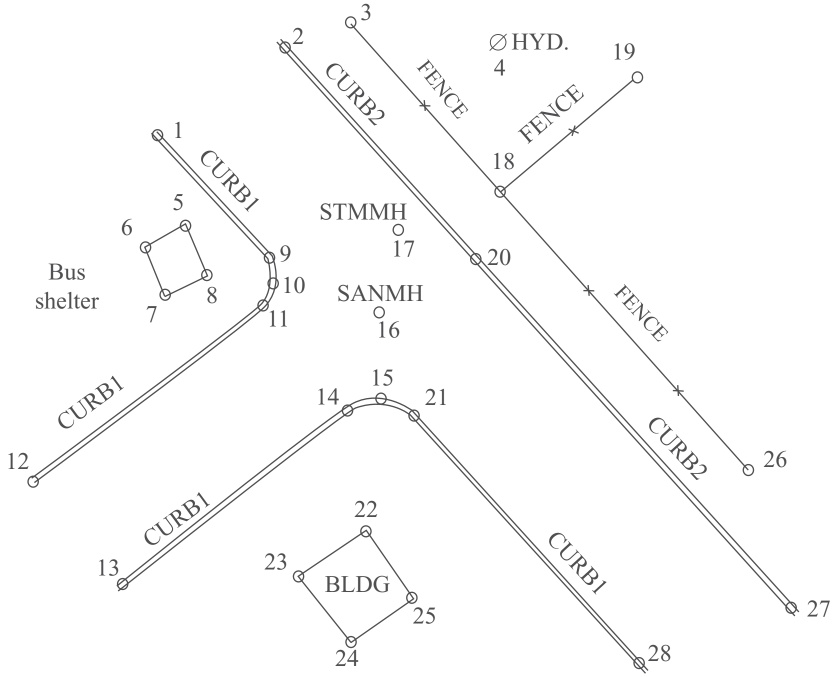


FIGURE 10.7 Field notes for total station graphic description—MicroSurvey International codes

converts the prefixed point description file so that the shape of the pond is produced with graphics. The four descriptor operations described here are of less work than what would be required to describe each point using field entries for the 10 points shown in this example. Larger features requiring many more field shots would be even more conductive to this type ofpost-survey editing of descriptors.

It is safe to say that most projects requiring graphics to be developed from total station surveys utilize a combination of point description field coding and post-survey point description editing. It should be noted that the success of some of the modern surveys still depends on a significant degree on the old-fashioned and reliable survey field notes.

It is becoming clear that the 'drafting' of the plan of survey is increasingly becoming the responsibility of the surveyor, either through direct coding techniques or through post-survey data processing. All newer versions of surveying software programmes enable the surveyor to produce a complete plan of the survey.

10.17 CONSTRUCTION LAYOUT USING TOTAL STATIONS

Total stations are well suited for collecting data in topographical surveys and the collected data could readily be downloaded to a computer, and with the help of an application software, the

Point	Description
1	CURB1
2	CURB2
3	FL1
4	HYD
5	BLD1
6	BLD1
7	BLD1
8	BLD1 CLOSE
9	CURB1 PC
10	CURB1
11	CURB1 PT
12	CURB1
13	CURB1 ST
14	CURB1 PC
15	CURB1
16	SAN
17	STM
18	FL1
19	FL2 JP
20	CURB2
21	CURB1 PT
22	BLD2
23	BLD2
24	BLD2
25	BLD2 CLOSE
26	FL1
27	CURB2
28	CURB2

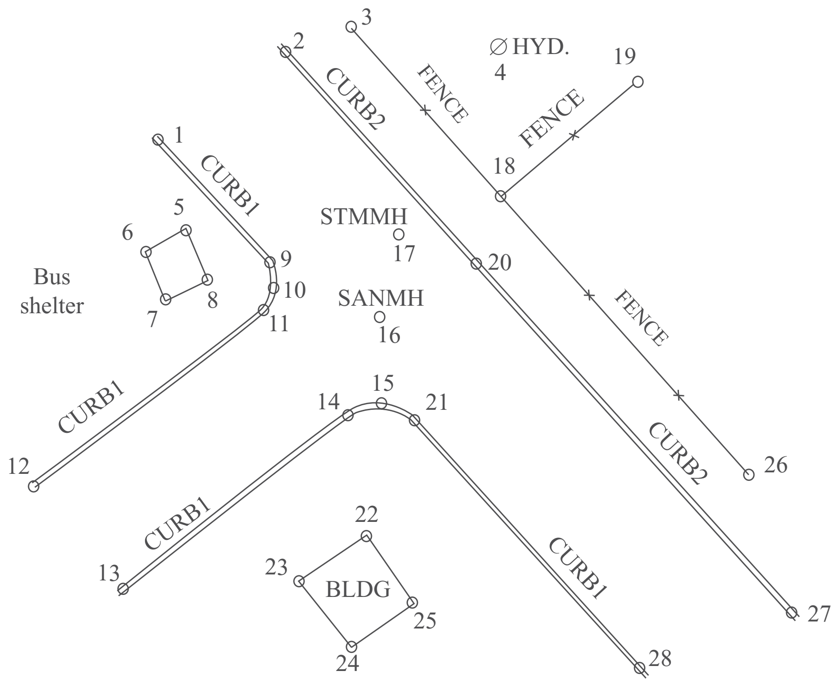


FIGURE 10.8 Field note for total station graphic description (Sokkia codes)

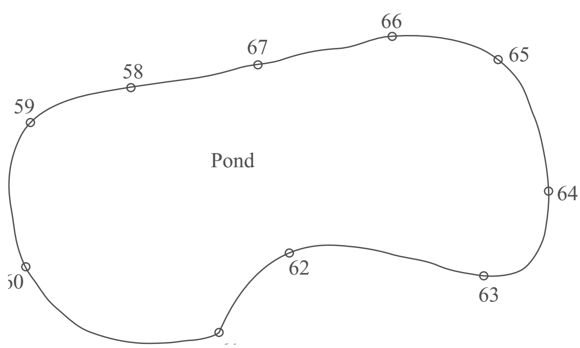


FIGURE 10.9 In sequence field coding defining topographic survey

downloaded data can be processed into print coordinates— northings, eastings and elevations (X, Y, Z)—along with point attribute data.

The significant increase in efficiency made possible that the total station topographic surveys can be transferred to layout surveys when the original point coordinates exist in computer memory or in a PCMCIA card, together with the coordinates of all the key design points. To illustrate this, consider the following example of a road project.

First of all, the topographic details are collected by using a total station setup at various control points, and the details transferred to a computer, processed, adjusted if necessary and converted into *X*, *Y* and *Z* coordinates. Various coordinate geometry and road design programmes can then be used to design the proposed road. When the proposed horizontal cross-section and the profile alignments have been established, the proposed coordinates (*X*, *Y* and *Z*) for all key horizontal and vertical (elevation) features can be computed and stored in the computer files. The point coordinates will include the top-of-curb and centreline positions at various regular stations, as well as all changes in the direction of slope, catch basins and traffic islands and will have all curved and irregular road components.

The computer files now include coordinates of all control stations, all topographic details, and finally, all design component points. The layout will be accomplished by setting up of the instrument at a control point, sighting another control point with the correct azimuth on the horizontal circle and then turning the computed angle to locate the desired layout point; the computed distance is, at the same time, measured out with the total station. When the total station is set to tracking mode, the surveyor will set the prism on the target by rapid trial and error movements. The drawing will give an aid to the surveyor in the field in laying out the works properly. All the layout prints listed on the printout will be usually shown in drawings, together with curve data and other explanatory notes.

Modern total stations offer an even more efficient technique. Instead of having the layout data only on a printout, the coordinates of all layout points can be uploaded into the total station's memory. The surveyor can then identify the occupied control point in the field and the reference back sight points for orienting the total station. The desired layout point number is then entered with the required layout angle and the distance being measured from the stored coordinates and displayed. The layout can be processed by setting the correct azimuth and then, by trial and error, the prism is moved to the layout distance (the total station is set to tracking mode for all but the final measurements); with

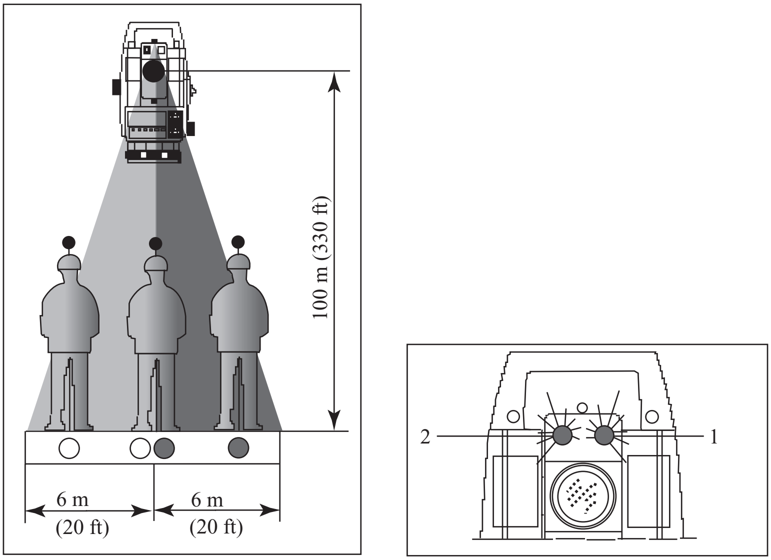


FIGURE 10.10 EGLI guide light

some total stations, the prism is simple tracked with the remaining left/right (±) distance being displayed alongside the remaining near/far (±) distance. When the correct location has been reached, both displays show 0.000 m (0.00 ft). If the instrument is set up at an unknown position (free station), its coordinates can be determined by sighting control stations whose coordinates have been previously uploaded into the total station's memory. This technique is known as resection and is available on all modern total stations. Sighting on two control points can locate the instrument station, although additional sightings (up to a total of four) are recommended to provide a stronger solution and an indication of the accuracy level achieved.

Some modern total stations are equipped with a guide light, which can move the prism holder on-line very quickly. It comes equipped with internal storage of more than 5,000 points and an ELGI guide light that is useful in layout surveys, as the prism holder can quickly place the prism on-line by noting the coloured lights sent from the total station (see Fig. 10.10).

The flashing lights (yellow on the left and red on the right, as viewed by the prism holder), which are 12-m wide at a distance of 100 m, enable the prism holder to place the prism on-line with final adjustments as given by the instrument operator. With automatic target recognition (ATR), the sighting-in process is completed automatically.

In an ATR equipped total station, the telescope must be roughly pointed at the target prism, either manually or under software control, and the instrument then does the rest. The ATR module is a digital camera inside the total station that notes the offset of the reflected laser beam, permitting the instrument to move automatically until the cross hairs have been electronically set precisely on the point. After the point has been precisely sighted, the instrument can then read and record the angles and distances. ATR also comes with a lock-on mode in which the instrument, once sighted at the prism, will continue to follow the prism as it is moved from station to station. To ensure that the prism is always pointed to the instrument, latest 360 degrees prisms are manufactured and supplied with modern stations, which greatly assists the surveyor in keeping the lock-on over a period of time.

10.18 OVERVIEW OF COMPUTERIZED SURVEY DATA SYSTEMS

Advances in computer science have had a tremendous impact on all aspects of modern technology. The effects on construction and engineering surveying have been significant. In advanced surveying, tremendous changes are taking place in the field of data collection, processing and map preparation. To appreciate the full impact of this technology, one has to view the overall operations, that is, from the field to the computer, computer processing and data portrayal in the form of maps and plans. Figure 10.11 gives a schematic overview of an integrated survey data system.

10.19 DATA GATHERING COMPONENTS

The manual entry of field lacks the speed associated with interfaced equipment, but after the data have been entered, all the advantages of electronics techniques are available to the surveyor. The raw field data, collected and stored by the total station, is transferred to the computer through a standard RS-232 interface connection, and the raw data download programme, which is supplied by the manufacturer, is run properly to form properly formatted field data required for mapping.

At this stage, coordinate geometry programmes can be used to calculate traverse closures and adjust all acceptable data into X, Y and Z values. If only topography was taken, there may be no need

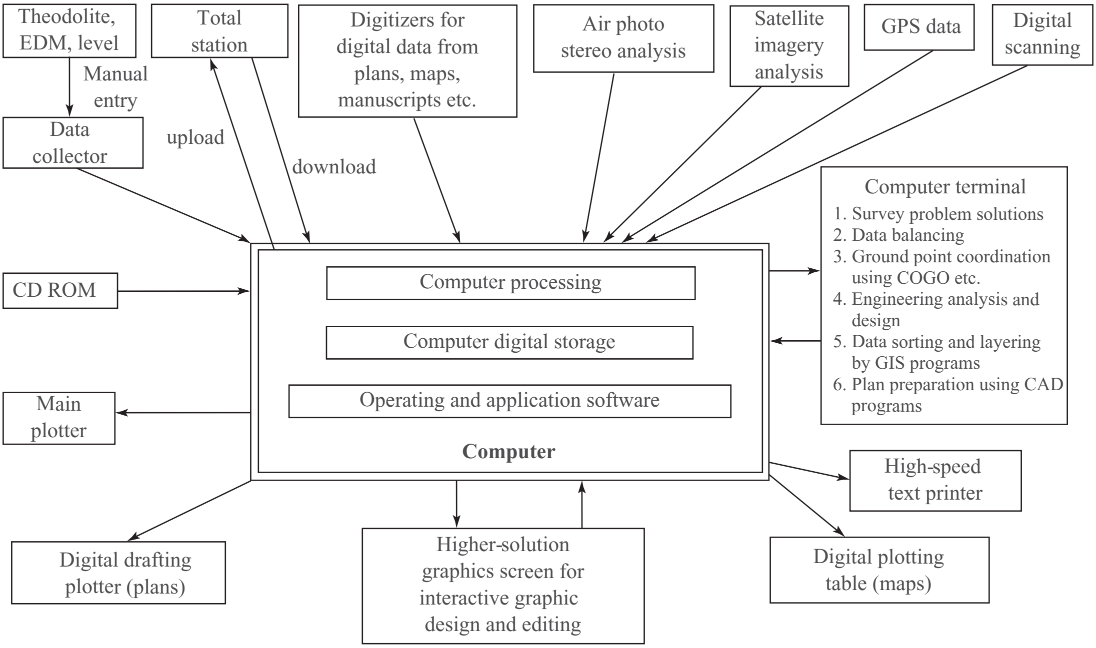


FIGURE 10.11 Computerized survey data system

of adjustments, and the programme can directly compute the required coordinates. Also at this stage, additional data points such as inaccessible ground points can be computed and added to the data file.

The instruments location (OCC) coordinates and the back sight orientation can be entered into the data collector/total station at the commencement of the survey, or they can be entered in a computer file, where they can be directly accessed for data processing. Existing maps and plans have a wealth of data that may be relevant for an area survey. If such maps and plans are available, the data can be digitized on a digitizing table or by digital scanners, and added to the *X*, *Y* and *Z* coordinate files. In addition to distances and elevations, the digitizer can provide codes, identification and other attribute data for each digitized points. One of the most important features of the digitizer is its ability to digitize maps and plans which are plotted at various scales, and store the distances and elevations in the computer at their ground (grid) values.

The stereo analysis of aerial photos is a very effective method of collecting topographic ground data, particularly in high-density areas where the cost of a conventional survey would be high. Many municipalities conduct routine surveys of all major roads to develop plans and profiles that can be used for design and construction. With the advent of computerized surveying system, the stereo analyzers can coordinate all horizontal and vertical features and transfer these *X*, *Y* and *Z* coordinates to computer storage.

Satellite imagery is received from remote sensing satellites such as Land Sat and can be processed by a digital image analysis system that classifies terrain into soil and rock types and vegetation cover. This data can be digitized and added to the computer storage. Finally, precise position location can be determined by satellite observations. The NAVSTAR system of positioning (GPS) will revolutionize the way control surveys are performed.

10.20 DATA PROCESSING COMPONENTS OF THE SYSTEM

The control portion of the schematic diagram (Fig. 10.10) depicts the data processing components of the system. Total station data can be closed and adjusted by means of various coordinate geometry programmes. The missing data positions can be computed by using various intersections, resection and interpolation techniques, with the resultant coordinates being added to the database. If the data are to be plotted, a plot file may be created that contains plot commands including symbols and joint commands for straight and curved lines; labels and other attribute data are also included.

Design programmes are available for most construction endeavours. These programmes can work with the stored coordinates to provide a variety of possible designs, which can then be quickly analyzed with respect to cost and other factors.

Some design programmes incorporate interactive graphics, which permit a plot of the survey to be shown to scale on a high-resolution graphic screen. Points and lines can be moved, created, edited and so forth, with the final position coordinated right on the screen and the new coordinates added to the coordinate files.

10.21 DATA PLOTTING

Once the plotting files have been established, data can be plotted in a variety of ways. Data can be plotted on to a high-resolution graphic screen. Then the plot can be checked for completeness, accuracy and so forth. If interactive graphics are available, the plotted features can be deleted, enhanced, corrected, crosschecked, crosshatched, labelled, dimensioned and so on. At this stage, a hard copy of the screen display can be printed using a plotter. The resultant plan can be plotted to any scale, limited only by the paper size.

10.22 EQUIPMENT MAINTENANCE

The following checklist will aid each crew in properly maintaining and keeping an inventory of their assigned equipment. At the end of each workday, the party chief should check that the following duties have been performed:

1. Clean all reflectors and holders. A cotton swab dipped in alcohol should be used on the glass surfaces. A crew member can do this during their trip back to office.
2. Clean the tribrachs. They should be dusted daily.
3. Remove dust from all instruments. A soft paintbrush or a shaving brush works well. If an instrument has been exposed to moisture, thoroughly dry it and store in an open case. Download the data collector to the computer.
4. Backup all files generated from the download and check the integrity of the backup files before erasing the field data from the data collector.
5. Clean the batteries and connect them to the charger. Some batteries require a 14- to 20-hour charging, so one set of batteries may have to be charged while a second set is in operation.

10.23 MAINTAINING BATTERY POWER

One of the biggest problems faced by the users of total stations with data collectors is maintaining an adequate power supply. There are several factors that should be considered when assessing power needs:

- A topographic survey entails more data than a boundary survey. Normal production in a topographic mode is 200–350 measurements per day. A boundary survey can entail making 16 measurements or so from each traverse point and occupying 10–15 points per day. Determine the number of measurements that would normally be made in a day and consult the manufacturer's specifications to determine the number of shots expected from a fully charged new battery.
- Keep in mind that the batteries will degrade over a period of time. This means that a new battery, with sufficient power for 500 measurements when new, may only be capable of 300 measurements after a year of use.
- Some batteries take up to 10 hours to get fully charged. If the work schedule will not permit 10 hours for charging a second set of batteries, a battery with adequate power to supply the instrument for more than 1 day should be purchased.
- Modern total station instrument may take 3 days to use one battery. As newer instruments use lesser power than those in the market 15 years ago, this should be considered in determining power needs.

In addition to the proper assessment of the power need, a record of the history and current status of the power supply should be made readily available. When batteries begin to get weak, there is generally a rapid deterioration in their performance. To monitor the performance of a particular battery, record the serial number in a battery log book. If problems arise with a particular unit, check the log to see when the battery was purchased or when it was last replaced. Then again discharge and recharge the battery. If performance is still not up to speed, have it checked to determine the weak cell and replace it. The cost of having a battery replaced is minimal when considering the cost of lost work time due to power failure.

Also record the date on which the battery was charged on the shipping label that is attached to the battery box. When the battery is fully used, simply cross out the date, thus eliminating the confusion of not knowing which battery needs to be charged. Check the time of previous recharge of the battery, and if it exceeds 15 days, recharge it again. This keeps the power supply at peak performance. Always consult the operator's manual for recharge specifications.

It is always a good idea to have backup power available that lasts 15 min of work. Most manufacturers provide cabling for power backup for the instrument with an automobile battery. Some can even supply a quick charge system that plugs into the automobile cigarette lighter.

Power Pointers

- Assess power needs for the particular job
- Assess power usage of the equipment
- Monitor performance of each battery
- Monitor battery age, usage and replacement information
- Have one day's worth of backup power readily available

10.24 TOTAL STATION JOB PLANNING AND ESTIMATING

To estimate project standards expected of field crews,

1. Most crews will make and record 200–300 measurements per day. This includes any notes that must be put into the system to define what was measured. When creating productivity standards, keep in mind that a learning curve is involved. Usually, it takes a crew four to five projects to become familiar with the equipment and the coding system to start reaching the potential productivity of the system.
2. A two-person crew is most efficient when the typical spacing of the measurements is less than 15 m. When working within this distance, the rod person can acquire the next target during the time it takes the instrument operator to complete the measurement and input the codes to the data collector. The instrument operator usually spends 20 sec (+ or –) in sighting a target and recording a measurement, and another 5–10 sec coding the measurement.
3. When the general spacing of the data exceeds 15 m, having a second rod person will significantly increase productivity. A second rod person allows the crew to have a target available for measurement while the first rod person is moving. If the distance of the move is 15 m or greater, the instrument will be idle with only one rod person.
4. When dealing with strip topographic situations, data must be acquired every 1 m along the length of the job. The rule of thumb of one measurement for every 1 m of linear topography works very well for estimating purposes. Using this estimate, the typical field crew will make and record between 350 and 500 measurements or 300 and 500 m of the strip topography per day. Typically, a two-person crew equipped with a recording total station and data collector picks up 400 m a day. Depending on the office or field reduction software being used these data can produce both the planimetric and contour maps as well as transfer the data to an engineering design package with very little additional manipulation.

To execute a fast economic and accurate strip topography, the following tips are to be remembered:

1. Estimate one measurement for every 1 m of project.
2. If the shots are greater than 15 m, a second rod person adds up the efficiency of the crew.
3. Expect a two-person crew equipped with a recording total station and data collector to pick up at least 400 m of strip per day.

Conventional location or topographic surveying often requires a three-person field party. The party chief will be working at the instrument, recording the measurements and other information in the field book. The party chief is responsible for gathering data and he must pay close attention to the movements of the rod man. When using the power of field-to-finish data collection, the experience and judgement of the rod man is an important factor. Most organizations have the party chief or senior field technician to run the rod and allow the less experienced person to operate the instrument. The rod man communicates the codes and other instructions to the instrument operator, who enters them into the instrument or data collector and takes the measurements. The ease of operation of total stations is fast and hence it is very easy to train an instrument operator. This free more experienced personnel to control the pace of the job and to concentrate on gathering the correct data.

Multiple rod men can be used for increased productivity. Data collection provides a tremendous increase in the speed of performing fieldwork by eliminating the need to read and record measurements

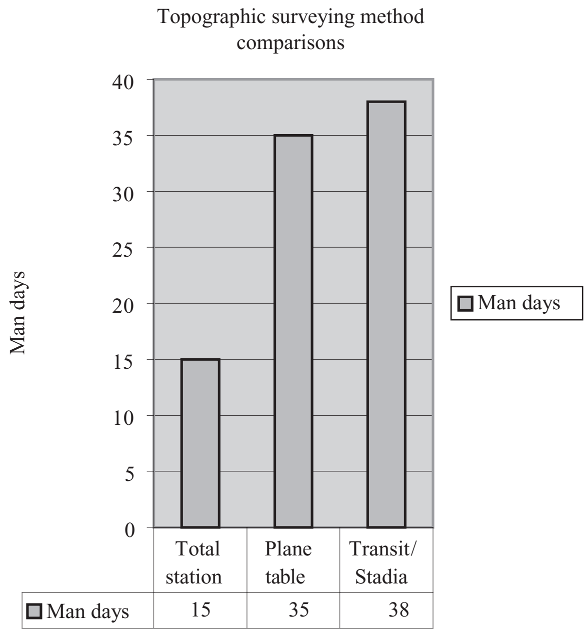


FIGURE 10.12 Topographic surveying method comparisons

and other information. Many organizations have reduced the size of their field crews by eliminating the field book recorders. On projects where a large number of shots are needed, the use of two (or more) rod men has resulted in excellent time and cost savings. The rod men can work independently in taking ground shots or single features. When more than one rod man is used, crew members should switch jobs throughout the day. This helps to eliminate fatigue in the person operating the instrument. As an extension of the concept discussed above, it is a good idea to have an experienced person running one rod and directing the other rod men. If possible, each rod man and the instrument operator should have a radio or other means of reliable communication.

Electronic data collection (EDC) has proved to be an extremely cost-effective means of gathering data. When competing against a grid or baseline and the offset style of surveying, EDC often results in time savings in the field. However, deriving a horizontal and vertical position on the located points is only part of the process. The ultimate goal is usually a map, showing either planimetrics, or contours or both. The productivity of total stations compared with other methods is shown in Fig. 10.12, which depicts the enhanced productivity of a total station relative to traditional plane table or transit stadia methods. The time savings in design/construction layout are shown in Table 10.4.

10.25 ERROR SOURCES OF ELECTRONIC THEODOLITE

All theodolites measure angles with some degree of imperfection. These imperfections result from the fact that no mechanical device can be manufactured with zero error. In the past, very specific measuring techniques were taught and employed by surveyors to compensate for minor mechanical imperfections in theodolites. With the advent of electronics, the mechanical errors still exist, but are related to in a different way.

TABLE 10.4 Total stations estimated field time saved over other methods

Application on Project Site	%Time in Field Saved
Topographic surveys	25
Control traverse	5
Design layout	30
As-built	35
X-select	20

One must clearly understand the concepts behind the techniques and the adjustments for errors that electronic theodolites now make. The following paragraphs provide the major sources of error when using a theodolite, and also the particular method employed to compensate for that error.

Circle Eccentricity Circle eccentricity exists when the theoretical centre of the mechanical axis of the theodolite does not coincide exactly with the centre of the measuring circle. The quantum of error corresponds to the degree of eccentricity and the part of the horizontal circle being read. When represented graphically, circle eccentricity appears as a sine wave. Circle eccentricity in the horizontal circle can always be compensated for by measuring both faces (opposite sides of the circle) and taking the mean as a result. Vertical circle eccentricity cannot be compensated for in this manner, because the circle moves with the telescope. More sophisticated techniques are required for this.

Some theodolites are individually tested to determine the sine curve for the circle error in that particular instrument. Then a correction factor is stored in ROM that adds or subtracts from each angle reading, so that a corrected measurement is displayed. Other instruments employ an angle-measuring system consisting of rotating glass circles that make a complete revolution for every angle measurement. Fixed and moving light sensors scan the angles. The glass circles are divided into equally spaced intervals, which are diametrically scanned by the sensors. The amount of time it takes to input a reading into the processor is equal to one interval, thus only every alternate graduation is scanned. As a result, measurements are made and averaged for each circle measurement. This eliminates scale graduation and circle eccentricity error.

Horizontal Collimation Error Horizontal collimation error exists when the optical axis of the theodolite is not exactly perpendicular to the telescope axis. To test for horizontal collimation error, point to a target in face one and point back to the same target in face two; the difference in horizontal circle readings should be 180 degrees. Horizontal collimation error can always be corrected by means of the face one and face two alignment of the instrument.

Most electronic theodolites have a method to provide a field adjustment for horizontal collimation error. Again, the manual for each instrument provides detailed instructions on the use of this correction.

In some instruments, the correction stored for horizontal collimation error can affect only measurements on one side of the circle at a time. Therefore, when the telescope is passed through the zenith (the other side of the circle is being read), the horizontal circle reading will change by twice the collimation error. However, when this happens, these instruments are functioning exactly as designed.

When prolonging a line with an electronic theodolite, the instrument operator should turn a 180-degree angle and turn the horizontal tangent so that the horizontal circle reading is the same as it was before.

Height of Standards Error In order for the telescope to plunge through a truly vertical plane, the telescope axis must be perpendicular to the standing axis. There are no such things that are most perfect in the physical world. All theodolites have a certain degree of error caused by imperfect positioning of the telescope axis. Generally, a qualified technician should accomplish determination of this error because horizontal collimation and height of standards errors interrelate and can magnify or offset one another. Horizontal collimation error is usually eliminated before checking for height of standards. Height of standards error is checked by pointing to a scale the same zenith angle above a 90-degree zenith in face one and face two. The scales should read the same in face one as in face two.

Circle Graduation Error In the past, circle graduation error was considered a major problem. For precise measurements, surveyors advance their circle on each successive set of angles, so that circle graduation errors are compensated. Present technology eliminates the problem of graduation errors. This is accomplished by photo-etching the graduations onto the glass circles. Next make a very large precise master circle and photograph it. An emulsion is applied to the circle and a photo-reduced image of the master is projected onto the circle. The emulsion is removed and the glass circle is etched with very precise graduations.

10.26 TOTAL SURVEY SYSTEM ERROR SOURCES AND HOW TO AVOID THEM

In every survey there is an accuracy that must be attained. The first step in using field and office time most effectively is to determine the positional tolerance of the points to be located. After this has been accomplished all sources of errors can be determined and analyzed. Some sources of errors are pointing errors, prism offsets, adjustment of prism pole, EDM alignment, collimation of the telescope, optical plummet adjustment, instrument/EDM offsets, curvature and refraction, atmospheric conditions, effects of direct sunlight, wind, frozen ground and vibrations. The accuracy required for each survey should be carefully evaluated. Each of the following factors can accumulate and degrade the accuracy of measurements:

Pointing Errors Pointing errors are due to both human error to point the instrument and environmental conditions limiting clear vision of the observed target. The best way to minimize pointing errors is to repeat the observation several times and use the average as the result.

Uneven Heating of the Instrument Direct sunlight may heat up one side of the instrument, which is enough to cause small errors. For the highest accuracy, pick a shaded spot for the instrument.

Vibrations Avoid instrument locations that vibrate. Vibrations can cause the compensator to be unstable.

Collimation Errors When sighting points a single time (e.g., direct position only) for elevations, check the instrument regularly for collimation errors.

Vertical Angles and Elevations When using total stations to measure precise elevations, the adjustment of the electronic tilt sensor and the reticule of the telescope becomes very important. An easy way to check the adjustment of these components is to set a baseline. A line close to the office with a large difference in elevation will provide the best results. The baseline should be as long as the longest distance that will be measured to determine elevations with intermediate points at 30–60 m intervals. Precise elevations of the points along the baseline should be measured by differential levelling. Set up the total station at one end of the baseline and measure the elevation of each point. Comparing the two sets of elevations provides a check on the accuracy and adjustment of the instrument.

Atmospheric Corrections Instruments used to measure atmospheric temperature and pressure must be correctly calibrated.

Optical Plummet Errors The optical plummet or tribrachs must be periodically checked for misalignment.

Adjustment of Prism Poles When using prism poles, precautions should be taken to ensure accurate measurements. A common problem encountered when using prism poles is the adjustment of the levelling bubble. Bubbles can be examined by establishing a check station under a doorway in the office. First mark a point on the top of the doorway. Using a plumb bob, establish a point under the point on the doorway. If possible, use a centre punch to make a dent or hole in both the upper and lower marks. The prism pole can now be placed into the check station and easily adjusted.

Recording Errors The two most common errors associated with fieldwork are reading an angle incorrectly and entering incorrect information into the field book. Another common (and potentially disastrous) error is an incorrect rod height. Although electronic data collection has all these errors eliminated, it is still possible for the surveyor to identify an object incorrectly, make a shot to the wrong spot, or input a bad target height or hi. For example, if the surveyor normally shoots a fire hydrant at the ground level, but for some reason shoots it on the top of the operating nut, erroneous contours would result if the programme recognized the fire hydrant as a ground shot and was not notified of this change in field procedure.

Angles As a rule, a surveyor will turn a doubled angle for move-ahead, traverse points, property corners or other objects that require greater accuracy. On the other hand, single angles are all that are required for topographic shots.

10.27 CONTROLLING ERRORS

A set routine should be established for a survey crew to follow. Standard operating procedure should require that control be measured and noted immediately on the data collector and in the field book after the instrument has been set up and levelled. This ensures that the observations to controlling points are established before any outside influences have had an opportunity to degrade the setup. In making observations for an extended period of time at a particular instrument location, observe the control points from time to time. This ensures that any data observed between the control shots are good, or that a problem has developed and appropriate action can be taken to remedy the situation. As a minimum requirement, survey crews have to observe both vertical and horizontal control points at the beginning of each instrument setup and before the instrument is shifted.

One of the major advantages of using a total station equipped with data collection is that errors previously attributed to blunders (i.e., transposition errors) can be eliminated. Even if a wrong reading is set on the horizontal circle in the field, or a wrong elevation is used for the benchmark, the data itself may be precise. To make the data accurate, many software packages will allow the data to be rotated and adjusted as it is processed. The only way to assure that these corrections and observations have been accurately processed is to compare the data to control points. Without these observations in the magnetically recorded data, the orientation of that data will always be in question.

The use of a total station with a data collector can be looked upon as two separate and distinct operations. The checklists for setting up the total station and data collector are as follows:

Total Station

a. If the EDM is modular, mount it on the instrument
b. Connect the data collector
c. Set up and level the instrument
d. Turn on the total station
e. Set the atmospheric corrections (ppm). This should be done in the morning and at noon
f. Set the horizontal circle
g. Set the coordinates
h. Observe the back sight (check whether the azimuth to the back sight is 180 degrees from the previous reading)
i. Observe the back sight benchmark (obtain difference in elevation). This may require factoring in the height of the reflector above the benchmark
j. Compute the relative instrument height (benchmark elevation/difference in height). Note the height of the rod and note the computations in the field book
k. Input the Z (elevation) value in the instrument or the data collector
l. Observe the back sight benchmark (check elevation)
m. Invert and repeat (check elevation)

Data Collector

a. Record the date and job number
b. Record the crew number and the instrument serial number
c. Record the field book number and the page number
d. Record the instrument location (coordinates)
e. Record the back sight azimuth
f. Record the standard rod height
g. Record the height of the instrument

Note: All the above information should also be recorded in the field book.

h. Observe and record the measurement to the back sight benchmark
i. Enter the alpha or numeric descriptor of the above point into the data collector
j. Observe and record the measurement back sight benchmark or check benchmark (if setting benchmark, note in the field book and repeat with instrument inverted)
k. Enter the alpha or numeric descriptor of above point into the data collector

l. Observe and record the measurement to back sight
m. Enter the alpha or numeric descriptor of the above point into the data collector
n. Invert and repeat steps 'l' and 'm'
o. Observe and record the measurement to foresight
p. Enter the alpha or numeric descriptor of the above point into the data collector
q. Invert and repeat steps 'o' and 'p'
r. Observe and record the measurement to side shot
s. Enter the alpha or numeric descriptor of the above point into the data collector (repeat steps 'r' and 'm' as needed)
t. When the setup is complete, or at any appropriate time, repeat shots on vertical and horizontal control. Observe the display values and record it in the data collector

10.28 FIELD CODING

Whether the data are recorded by hand or electronically, the most time-consuming survey operation is the recording of a code or description to properly identify the point during processing. For example, in a topographic or planimetric survey, identification of points, which locate the position of curbs, centrelines, manholes, and other similar features are essential for their correct plotting and contour interpolation.

In spite of this slow coding process encountered when using today's data collectors, the advantages heavily outweigh the disadvantages. These advantages include the collection of error-free numeric data from electronic total stations virtually at the instant they are available, and the error-free transfer of this data to an office computer system without the need for manual entry.

Field coding allows the crew to perform the drafting and provide a more logical approach. As the field crew can virtually produce the map from the field data, this eliminates the need for many field book sketches. They can also eliminate office plotting and editing by connecting the dots to produce a final product. Total station users who gather data to be processed with other systems, typically record descriptive information with each point measured, and gather 200–300 points per day with the total station. Users report 300–700 points per day if descriptive information is kept to the necessary minimum. The coding scheme is designed so the computer can interpret the recorded data without ambiguity to create a virtually finished product.

10.29 FIELD COMPUTERS

The greatest advantage of survey reduction in the field is uncovering a mistake, which can easily be corrected if the crew and equipment are on the site. Laptop and notebook computers are popular field items. These computers are used to download GPS and total station data. Once the files are stored in the computer, data reduction can be done easily with programmes stored in these machines.

Listed below are some software considerations to install on topographic field computers:

1. Interface with field data collector
2. A system of predefined codes for most common objects and operations in a database

3. User-defined codes for site-specific requirements in a database
4. Survey adjustment programmes such as:
   a. Compass rule adjustment
   b. Transit rule adjustment
   c. Least squares
   d. Angle adjustment
   e. Distance adjustment
5. A programme which can assign an alphanumeric descriptor field for each survey point
6. A full-screen editor to examine and edit ASCII data files
7. An interface programme to convert files into common graphic interchange formats such as IGES or DXF
8. A programme to connect features which were not recorded in order such as fence, curb and gutter, edge of pavement and waterline
9. An operating system, which will be compatible with post-processing machines with CADD programmes such as Intergraph, Microstation and AutoCAD
10. Custom programmes which can use all the features available to the total station or the data collector
11. Select software which provides training, if possible
12. A coordinate conversion software

The requirements for a field computer in data collection are:

1. Portable computer with a good processor at least 1.0 GHz speed
2. Tough enough for field use
3. Memory: 1 GB (minimum), 250 gigabyte internal hard disk
4. USB 2.0 terminal
5. Math coprocessor needed
6. Serial port (RS 232) is an added advantage
7. Modem: 2400 baud and above
8. VGA or Super VGA graphics
9. Portable printer
10. Back up battery for at least 2 hours

10.30 MODEM FOR DATA TRANSFER (FIELD TO OFFICE)

A modem is a device that modulates and demodulates binary data transmission over a telephone network. This device MODulates the carrier for transmission and DEModulates for reception; hence the term MODEM. The carrier is simply a tone with three characteristics, any one of which can be varied or modulated to impose a signal on the carrier. They include amplitude, frequency and phase. The exact method of variation must be the same for two or more modems to be termed as ‘compatible’.

10.31 TRIGONOMETRIC LEVELLING AND VERTICAL TRAVERSING

Trigonometric levelling is the process of determining the differences of elevations of stations from observed vertical angles and known distances. Trigonometric levelling is the single most important new application brought into widespread use by the increasing acceptance of the total station. It is a fair statement that the error sources and types which affect trigonometric levelling measurements are among the least understood of the commonly done surveying procedures. Knowledge of the limitations of trigonometric levelling, together with means (instrumentation and procedures) to account for such limitations, is essential in using and supporting the use of modern surveying technology.

Total station trigonometric levelling can achieve accuracies similar to those reached using a spirit level. Third-order accuracy should be easily obtainable. First-order accuracy has been done, but the procedures are involved and not commonly followed.

10.32 TRIGONOMETRIC LEVELLING—FIELD PROCEDURES

The trigonometric levelling can be done in two ways viz., observations taken for the height and distances and geodetic observations. In the first method, one has to measure the horizontal distance between the given points if it is accessible. Then take the observation of the vertical angles and then compute the distances using them. If the distances are large enough then provide the correction for the curvature and refraction linearly to the distances that are computed.

In geodetic observations, the distances between the two points are geodetic distances and hence the principles of the plane surveying are not applicable here. The corrections for the curvature and refraction are to be applied to the angles directly.

To obtain third- or second-order vertical accuracies with a total station, the following field procedures should be rigorously followed:

1. Careful setup and levelling
2. Use Face I and Face II observations
3. Reciprocal measurements
4. Take multiple observations
5. Protect instrument from sun and wind
6. Use proper target based on instrument/EDM configuration
    a. Tilting target if necessary
    b. Good quality reflectors
    c. Correct prism offsets
    d. Unambiguous target
    e. Maintain target in good adjustment
7. Limited sight distances
    a. 300 m max
    b. Reduce atmospheric-related error
    c. Improves vertical angle accuracy
8. Accurately measure temperature and pressure

   a. At least twice a day
   b. If long steep line measurements are taken at both the ends, use averages
9. Watch for adverse refraction

10.33 TRIGONOMETRIC LEVELLING—ERROR SOURCES

The following error sources impact the accuracy of trigonometric levelling with electronic total stations:

1. Instrument
   a. Distances
   b. Vertical angle accuracy
   c. Vertical compensator, dual axis compensation
   d. No boost to vertical angle accuracy
2. Nature

TABLE 10.5 Combining sources of error (for 150 m line)

Errors in Vertical Angle Measurement		
Source	Type	Nominal Amount
Instrument accuracy	Random	-/ 3 sec
Collimation	Systematic	-/ 3 sec
Measure hi and HT	Random	-/ 0.0015–0.0320 m
C and R	Systematic	0.0015 m
Hand-held prism pole	Random	-/ 0,0015 m
30 mm prism offs error	Random	-/ 3 mm Heat wave
	Random	-/ 0.003 m
Unshaded instrument	Random	-/ 12 cm to 10 sec

Combined angular and linear error is between 0.0072 and 0.0144 m. Vertical angle precision for the 150 m line is therefore in the range 1:10,000 and 1:210,000.

Table 10.5 shows the precision resulting from horizontal distance and vertical angular measurements as needed to resolve differences in elevations from trigonometric observation.

10.34 APPLICATION OF TOTAL STATION

Total station can be used for the following surveying purposes

a. Data collection for topographic surveying
b. Control survey (traverse or triangulation)
c. Height measurement (remove elevation measurement (REM)).
d. Stakeout or setting out

e. Resection or trilateration
f. Area calculations, etc.
g. Calculation of distance between two remote points (remote distance measurement (RDM) or missing line measurement (MLM))
h. Curve setting
i. Setting of super elevation

REVIEW QUESTIONS

1. What are the advantages of a total station over conventional surveying instruments?
2. What are the fundamental parameters of a total station?
3. What are the precautions to be taken while using a total station?
4. Write short notes on:
   i. Atmospheric correction factors of a total station
   ii. Scale correction factors of a total station
5. What is the principle behind the linear distance measurement of a total station?
6. List out the minimum equipment inventory required for a total station survey.
7. How is a total station set up over a point during the fieldwork?
8. How is a back sight measured while the setup of a total station over a field point?
9. How is an azimuth mark provided for a total station survey setup?
10. Describe the required steps for the initial setting of a total station for a fieldwork.
11. What is radial shooting?
12. How is a traverse survey performed using a total station?
13. What is the relevance of codes for the data collector?
14. Highlight the procedure for a traverse survey for the observation of at least 10 field points using a total station.
15. How is data retrieved from a total station and plotted for getting a final result?
16. Describe a construction layout using a total station.
17. Describe the field-generated graphics with a neat sketch, when a total station is used for a topographical survey.
18. How is the total station equipment maintained?
19. Describe the job planning required to carry out a topographic survey using a total station.
20. What are total station error sources?
21. What are the remedies to avoid the errors in a total station?
22. What is field coding?
23. Define trigonometric levelling.
24. What are the field procedures for trigonometric levelling?

DATA COLLECTION PROCEDURES 11

Chapter Outline

11.1 GENERAL

In this chapter, the procedures for data collection using a data collector (provided with old model total stations) are described. Modern total stations have an inbuilt data collector, which also follows some of these procedures.

In the first step of this process, the field survey, the vertical and horizontal angles are measured along with slope distances using the total station. The angle and distances are stored with a point number and description in the data collector. The survey data are then transferred to the microcomputer via a cable connection for data processing and field data storage. The microcomputer is either an in-office desktop system or a laptop model that can be used on site.

The data is then processed in the microcomputer to produce a coordinate file that contains the point number, point code, X–Y–Z coordinate values and a point descriptor. Once the data are on the workstation, it is converted into a graphics design file for use in a CADD programme such as the microstation, or AutoCAD files into intergraph design files. Level, label, symbol and line definitions are assigned to each point based upon the point code. The programme can transform data into a two-dimensional (2D) or three-dimensional (3D) design file.

The 3D file is used to create the digital terrain model (DTM), which is used to produce the contours. The resulting topographic data are then plotted for review. The final editing and addition of notes are completed, yielding topographic data in a digital format or as a plotted map. Uniform operating procedures are needed to avoid confusion when collecting survey data. The use of proper field procedures is essential to prevent confusion in generating a map. Collection of survey points in a meaningful pattern aids in identifying map features.

11.2 FUNCTIONAL REQUIREMENTS OF A GENERIC DATA COLLECTOR

The field note is an important factor in surveying with a total station. Some surveyors require field notes to be kept, while others use a data collector to replace field notes. Total stations calculate coordinates in the field itself and can continuously store coordinates, either in their own memory or in a data collector. If field notes are required, only specific items are considered in the transfer of data from a data collector to an office computer. The advantage of this method is that a check is provided on field notes. Most field note errors are made by transcription (e.g., writing 12 instead of 21 in the field book). Data transmitted to an office computer, through an RS-232C port, can be listed on an office printer to provide a check for transposition errors in the field notes. If the data collection is bidirectional, then it must receive data from the office computer for stakeout purposes as well as transmit data to the computer. Field notes can again be considered or ignored.

Four types of notes are kept in practice: (1) sketches, (2) tabulations, (3) descriptions and (4) combinations of these. The most common type is a combination form, but an experienced recorder selects the version best fitted to the job at hand. The location of a reference point may be difficult to identify without a sketch, but often a few lines of description are enough. Benchmarks are also described. In note keeping this axiom is always pertinent: when in doubt about the need for any information, include it and make a sketch. It is better to have too much data than not enough. Observing the suggestions listed below will eliminate some common mistakes in recording notes:

1. Letter the notebook owner's name and address on the cover and first inside page, in India ink.
2. Use a hard pencil or pen, legible and dark enough to copy.
3. Begin a new day's work on a new page.
4. Immediately after a measurement, always record it directly in the field book, rather than on a sheet of scrap paper for copying it.
5. Do not erase recorded data.
6. Use sketches instead of tabulations when in doubt.
7. Avoid crowding.
8. Title, index and cross-reference of each new job should be written.
9. Sign surname and initials in the lower right-hand corner of the right page on all original notes.

Topographic locations are numbered according to data record numbers. Data record numbers (point numbers) depict the type of location and where locations are measured. This helps office personnel improve digital field drawings into final design drawings. More importantly, blunders and mislabelled feature codes may be caught before costly design errors are made. The finished map and the sketch should be similar. Sketches are not required to be at any scale.

Electronic files are sufficient for submittal without identical hand entries from a field book. Video and digital cameras can be used to supplement the field sketch and provide a very good record of the site conditions for the CADD operator, design engineer and user of the topographic map.

11.3 DATA COLLECTION OPERATING PROCEDURES

Uniform operating procedures are needed to avoid confusion when collecting survey data. The use of proper field procedures is essential for preventing errors in generating a map. Collection of survey points in a meaningful pattern aids in identifying map features. Experience has resulted in the following steps for collection of field data:

1. Establish horizontal and vertical control for radial survey. This includes bringing control into the site and establishing setup points for the radial survey. Primary control is often brought into the site using the GPS satellite receivers. The traverse through radial setup points can be conducted with a total station as the radial survey is being performed. Experience indicates that a separate traverse is preferable. A separate traverse results in less opportunity for confusion of point identification, and allows the quality of the traverse to be evaluated before it is used. Elevations are established for the radial traverse points using conventional levelling techniques instead of the trigonometric values determined from the total station.
2. Perform radial surveys to obtain information for mapping:
    a. Set the total station over control points established as described above.
    b. Measure and record the distance from the control point up to the electronic centre of the instrument, as well as the height of the prism on the prism pole.
    c. Maintain accuracy. To prevent significant errors in the map elevations, the surveyor must report and record any change in the height of the prism pole. For accuracy, use a suitable prism and target that matches optical and electrical offsets of the total station.
3. Collect data in a specific sequence.
    a. Collect planimetric features (roads, buildings, etc.) first.
    b. Enter any additional data points needed to define the topography.
    c. Define break lines. Use the break lines in the process of interpolating the contours to establish regions for each interpolation set. Contour interpolation will not cross break lines. Assume that features such as road edges or streams are break lines. They do not need to be redefined.
    d. Enter any additional definition of ridges, vertical fault lines and other features.
4. Draw a sketch of planimetric features. A sketch or video of planimetric features is an essential ingredient of proper deciphering of field data. The sketch does not need to be drawn to scale and may be crude, but must be complete. The sketch helps the CADD operator, who has probably never been to the jobsite, confirm that the feature codes are correct by checking the sketch.
    a. A detailed sketch is critical to the design engineer. Detail sketches can be used to communicate complex information directly to the engineer without lengthy discussions.
    b. Miscellaneous descriptive notes can also be shown on the sketch for later addition to the design file. These notes are usually clearer and contain more information when shown on the sketch than when entered into the data collector.
5. Obtain points in sequence. The translation to CADD programme will connect points that have codes associated with linear features (such as the edge of a road) if the points are obtained in sequence. For example, the surveyor should define an edge of a road by giving shots at intervals on one setup. Another point code, such as natural ground, will break the sequence and will stop the formation of a line on the subsequent CADD file. The surveyor should then obtain the opposite road edge.

6. Use proper collection techniques. Using proper techniques to collect planimetric features can give an automatic definition of many of these features in the CADD design file. This basic picture helps in operation orientation and results in easier completion of the features on the map. Improper techniques can create problems for office personnel during analysis of the collected data. The function performed by the surveyor in determining which points to obtain and the order in which they are gathered is crucial. The party chief often does this task. Cross training in office procedures gives field personnel a better understanding of proper field techniques.
   a. Most crews will make and record 250–400 measurements per day. This includes any notes that must be put into the system to define what was measured. A learning curve is involved in the establishment of productivity standards. It usually takes a crew for five to six projects to become confident enough with their equipment and the coding system to start reaching the system potential.
   b. A two-person crew is most efficient when the typical spacing of the measurements is less than 50 ft. When working within this distance, the rod person can acquire the next target during the time it takes the instrument operator to complete the measurement and input the codes to the data collector. The instrument operator usually spends about 20 sec in sighting a target and recording a measurement, and another 5–10 sec in coding the measurement.
   c. When the general spacing of the measurements exceeds 50 ft, having a second rod person will increase productivity. A second rod person allows the crew to have a target available for measurement when the instrument operator is ready to start another measurement coding sequence. Once the measurement is completed, the rod person can move to the next shot, and the instrument operator can code the measurement while the rod people are moving. If the distance of that move is 50 ft or greater, the instrument will be idle if you have only one rod person.
   d. Data collection provides a tremendous increase in the speed of the fieldwork by eliminating the need to read and record measurements and other information.
   e. On jobs where a large number of shots are needed, the use of two (or more) rod persons has resulted in excellent time and cost savings. Communication between the rod person and the instrument person is commonly done via T/R radio. The rodmen can work independently in taking ground shots or single features; or they can work together by leapfrogging along planimetric or topographic feature lines. When more than one rod person is used, crew members should switch jobs throughout the day. This helps to eliminate fatigue in the person operating the instrument.

11.4 RESPONSIBILITY OF THE FIELD CREW FOR DATA COLLECTION AND PROCESSING

The responsibilities of the field crew for data collection and processing include the following:

1. Make a backup copy of the raw data before the file transfer. Once file transfer is successfully completed, data in the data collector can be deleted.
2. Print a copy of the formatted data and check it against the field notes. Check the field input of data against the field notes. Specifically, check the instrument locations, azimuths to back sights

and the elevation of benchmarks. In addition, scan the data for any information that seems to be out of order. Check the rod heights.

3. Edit the data. Eliminate any information that was flagged in the field as being in error. In the system, make a record of any edits, insertions, deletions, the person responsible for making them and when they were made.
4. Process the control data. Produce a short report of the data that was collected in the field. Check the benchmark elevation to be certain that the given elevation is the calculated elevation and that the coordinates of the back sights and foresights are correct.
5. To assure that good data are being supplied by the field, make certain that the field crew fully understands the automated processes that are being used and that they take care to gather data appropriately. It is much easier and more productive for the field crew to get a few extra shots where they know there will be difficulty in generating a good contour map than it will be for those in the office to determine where certain shots should have been made and add them to the database. In addition, make sure they pick up all break lines necessary to produce the final map.
6. The field crew will need to become educated about the contouring package used for the data processing. The crew should always observe the product of the contouring programme. This will help them to understand where and what amount of data may be needed to get the best results.
7. The office staff needs to be aware that in some circumstances, the field staff will have difficulty in getting some information (terrain restrictions, traffic, etc.).
8. The persons responsible for the fieldwork should be involved in the initial phase of editing, because they will most likely remember what took place. Preferably, the editing should be done on the same day the data is gathered, while the field person's memory is still fresh. If it is not possible for someone to walk through the site to ensure that the final map matches the actual conditions, then the field person should be the one to review the map.

11.5 INTERFACING THE DATA COLLECTOR WITH A COMPUTER

For many surveying operations, electronic data collection is the routine work. To produce a map showing planimetrics and contours, most software systems require a large amount of post-processing, which includes:

Computer Interfacing Many of the benefits of automated data collection are lost if the data stored cannot be automatically transferred to a computer system.

Hardware Compatibility Most micro- and minicomputers in the market today are supplied with (or have as an option) a serial interface board. The serial interface typically supports communications at different baud rates (speed of transmission) and with different parity settings. To control the flow of data, either a hardware or software handshake is used. Cables are connected to the serial interface board using a standard 9-pin connector. Occasionally, non-standard connectors with a different number of pins are used. Every data collector stores data in a different format and the problem is to translate the data from the format used by the data collector into a file with a standard ASCII format. Data standardization will become more important in the future and surveyors should be searching for methods that make system integration easier.

11.6 DIGITAL DATA

The fact that survey data collected by computers is in digital form has until recently been of interest only to surveyors themselves. As the final product delivered to the clients were drawings, surveyors needed to invest only for the computer equipment and the software they needed to get the digital data collected and plotted as a scaled drawing. Now the situation is changing. The proliferation of computer graphics used by architects, engineers and developers has meant that surveyors are asked, even required, to deliver survey information in digital format. These demands can pose thorny technical problems for those who do not consider this eventuality when they acquire their computer systems. The time and expense to work out the technical details of digital data delivery can be prohibitive to those who consider themselves as surveyors, not computer experts.

11.7 DIGITAL TRANSFER OF THE DATA TO APPLICATION SOFTWARE

There are two ways to transfer survey information digitally: as numeric data or as graphic files. The first is simpler from the surveyor's point of view. It begins with a text file—the sort of data that can be produced using a word processor. Text files are easiest to transfer between computers, but clients want data that computer software can interpret to produce drawings, not raw field notes.

1. Again, if the surveying software permits the output of the appropriate information in a text file, reformatting that information is, at worst, a minor programming task and may be possible simply through the global replacement feature of a word processing package. However, surveyors who use word processors to edit text files should be sure to use the 'ASCII' output option that is available in most word processors. This creates a 'generic' text file without embedded control or formatting characters.
2. Most CADD systems require digital deliverables and graphic files compatible with their particular system. This is a more problematic request because every CADD system has a unique and proprietary graphic data format. This means, for example, that graphic data produced in AutoCad cannot be loaded onto an intergraph system without some sort of intermediary 'translation'. Thus, even when data collectors are interfaced with a major CADD package, the diversity of CADD systems being used worldwide today virtually guarantees that there will be clients using different systems and unable to load the graphic file directly.
3. Translation of graphics data can be handled in two ways—by direct translation or through a neutral format. A direct translator is a computer programme that reads graphic data in one specific CADD system's format and outputs the same graphic information in a second CADD system's format. Although this is generally the quickest and most foolproof way to perform translation, it is often the most expensive. As direct translation programmes only address the problem of translation between two specific systems, several different translation programmes may be necessary to provide data that meet the compatibility requirements of all the surveyor's clients.
4. As all CADD vendors regard their data formats as proprietary, this process generally requires programmers who are intimately familiar with both CADD systems to write translation software.
5. It may be hard to locate all the programmes required. Software prices are high because there is little competition in this market. And there is a limited number of buyers who need to

communicate between any two specific CADD systems. Finally, most CADD vendors release at least one, and sometimes two, new versions of their software each year. Many releases include changes in graphic data format, so direct translation software can have a life of a less than a year.

6. Users may purchase software maintenance contracts. Like hardware maintenance, these generally charge a monthly fee to guarantee users that the software will be upgraded when either CADD system changes its data format. Users can purchase each updated version as it becomes available.
7. The second way to tackle graphic data translation is through neutral format translators. A neutral format is a non-proprietary graphic data format intended to facilitate the transfer of graphic information between CADD systems. Documentation is made available to the public. One such format, the Initial Graphic Exchange Specification (IGES), is an ANSI standard, and documentation is available through the National Technical Information Service in Washington. Other neutral formats have been designed by specific CADD vendors to facilitate data exchange with their systems. The two most frequently used are Auto Desk's Drawing Interchange Format (DXF) and Intergraph's Standard Interchange Format (SIF). The neutral format most commonly used a few years ago was SIF, but DXF now appears to be more generally accepted, particularly among PC-based CADD users. (The other format in which graphic data are sometimes transferred between CADD systems is a plot format, typically Cal-Comp or Hewlett-Packard.)
8. Neutral format translation requires two steps. First, the originator of the data, in this case the surveyor, translates the graphic information from a CADD system's proprietary format into the neutral format. This is the format in which the data are delivered to the client. Second, the client must then translate the data from the neutral format to a CADD system's proprietary format.
9. A major inconvenience of this approach is that it takes at least twice as long as direct translation. With a large survey, it can eat up time on both the surveyor's and the user's systems. Another problem is that users may need to purchase translation programmes between the neutral format and their CADD systems if vendors do not provide them as part of the CADD software purchases.
10. Finally, programmes do not always execute properly. This can be due to an error in the software or a mistake on the part of the user. Translation programmes, whether direct or neutral format, are no exception. The added difficulty with the neutral format approach is that it is difficult to pinpoint where the failure occurred—at the surveyor's end or at the user's end. The situation is particularly frustrating when a client who has had painful experiences with unsuccessful and costly graphic data translations may demand that the data be delivered in their CADD system's format.
11. New computer products are being made available every day. Often there is a trade-off between the enhanced degree of functionality in state-of-the-art software packages and their limitations in translation capability. If a software package proves to be truly exceptional and finds a large number of users, translation software will almost surely follow. If the software has limited appeal, either because it is extremely special purpose or because it is not well marketed, compatibility problems will most likely persist.
12. Requests for digital data deliverables will certainly become more frequent. Surveyors currently looking at new computer systems or considering an upgrade should make data exchange capability a major criterion. They should contact major clients to determine their CADD preference and quiz prospective software vendors about their translation software capabilities.

If the necessary translation software is available but too expensive, the vendor may be able to recommend service bureaus to provide translation services.

13. Surveyors who have computer systems and are generally pleased with their software's functionality, but who are encountering requests for digital deliverables, should do a quick survey of their major clients to determine what CADD equipment they are using. They should then contact their software vendor to see what solutions they suggest. There may, in fact, be a translation programme already available, either through the vendor or through a third party. If not, the more the requests the vendor gets for translation capability, the more viable a translation programme will appear as a new software product.
14. Another good source of information is a software users group, if one exists. Finally, the surveyor can contact CADD service bureaus in the area to see what data translation services they are able to perform. Fortunately, many service bureaus have invested heavily in translation software and are becoming expert in CADD data translation into a number of formats.
15. A final concern in the delivery of digital data is the media on which the data will be transferred. CADD files are relatively large and extremely cumbersome to transfer via modem. Much preferable and more reliable is the physical transfer of a diskette, magnetic tape or tape cartridge. When surveyors discuss CADD output with clients, they should explore the question of which media the clients use. Those with large computers probably prefer 1/2-inch 9-track magnetic tape. Those using micros will want diskettes or cartridge tapes. Whereas the large magnetic tape specification is standard, both diskettes and cartridges come in a variety of sizes and formats: high density, double density, etc. The surveyor may decide to forgo a 3-1/2-inch floppy drive if existing clients use 5-1/4-inch high-density diskettes.

11.8 REQUIREMENTS OF A DATA COLLECTOR

The data collector is vital to large surveys using the total station. Assumptions or oversights made at the time of equipment purchase can force a survey operation into equipment problems on the job for the economic life of the equipment. Listed below are some options which should be considered while purchasing a data collector:

1. Weatherproof, designed for rugged/durable field use
2. A non-volatile memory, which ensures data safety
3. Allow the storage of at least 1,000 points
4. Full search and edit routines immediately on the spot
5. Automatic recording with electronic theodolites
6. Formatting must be very flexible for manual entry, even for various CADD levelling tasks
7. Capability to use two files in the collector: one file for collection, the other file for processed data for stakeout tasks
8. The data collector must communicate with the electronic theodolite
9. All the features of the total station should be usable with the data collector purchased
10. The data collector must be compatible with the software one purchases or plans to use
11. Mixing brands should not cause a service problem

11.9 CODING OF FIELD DATA WHILE USING A DATA COLLECTOR

Whether data are recorded manually or electronically, one of the most time-consuming survey operations is the recording of a code or description to properly identify the point during processing. For example, in a topographic or planimetric survey, identification points which locate the position of curbs, gutters, centre lines, manholes and other similar features are essential for their correct plotting and contour interpolation.

1. Especially in topographic or planimetric surveying, surveyors have often wished for some way to speed up the process. For the most part, surveyors tolerate the time-consuming coding process, because it is the only way of ensuring an accurate final product.
2. In spite of this slow coding process, when using data collectors available today, the advantages heavily outweigh the disadvantages. These advantages include the collection of blunder-free numeric data from electronic total stations virtually at the instant they are available, and the error-free transfer of these data to an office computer system without the need for manual entry.
3. Field coding allows the crew to become the drafter and provide a more logical approach, as the field crew can virtually produce the map from the field data and eliminate the need for many field book sketches. They can also eliminate office plotting and editing by connecting the dots, to produce a final product. The coding scheme is designed so that the computer can interpret the recorded data without ambiguity to create a virtually finished product.
4. Additional codes may be required from time to time when surveying new features.
5. Either numerical point codes or alphanumeric point codes can be entered into the total station. This identification will vary from district to district, but the descriptor should be standardized throughout the corps of engineers.
6. Whenever districts require specialized point codes, then the attribute file may be edited to include these changes.

11.10 SUMMARY OF DATA COLLECTOR FIELD-TO-FINISH PROCEDURES

The following is a summary of data collector field to finish procedures:

1. Gather field data and code the information.
2. Off-load the data to the computer and process the information using equipment-specific software.
3. Create the ASCII coordinate file containing point number, *X*-coordinate, *Y*-coordinate, *Z*-coordinate, standardized descriptor and any additional notes.
4. Import the ASCII coordinate file into a CADD programme and create a graphic file.
5. Use the CADD programme to develop a final map with topographic, planimetric information, including contours, and utility information.
6. Edit map.
7. Plot map.

11.11 DATA COLLECTION IN MODERN TOTAL STATIONS

Modern total stations are having enough internal electronic data storage to record distance, horizontal angle, vertical angle and coordinates of about 10,000 points measured/computed. When the stored data is downloaded from a total station onto a computer, suitable application software is used to compute the results and to generate a map of the surveyed area. The latest generations of total stations are capable of storing more than 10,000 points and are capable to show the map on the touch-screen of the instrument immediately after measuring the points.

REVIEW QUESTIONS

1. What are the functional requirements of a generic data collector?
2. What are the data collection procedures while using a data collector?
3. What are the responsibilities of field crews for saving the data collected and processing of the data?
4. How is a data collector interfaced with a desktop computer?
5. Describe the methods of transferring digital data of a survey to an application software.
6. What are the requirements of a data collector?
7. Describe the coding of the filed data while using a data collector.

AUTOMATIC LEVEL, DIGITAL LEVEL AND OPTICAL THEODOLITES 12

Chapter Outline

12.1 AUTOMATIC LEVEL

The automatic level employs a gravity-referenced prism or mirror compensator to orient the line of sight (line of collimation) automatically. The instrument is quickly levelled when a circular spirit level is used. When the bubble has been centred (or nearly centre), the compensator takes over and maintains a horizontal line of sight, even if the telescope is slightly tilted (see Fig. 12.1).

Automatic levels are extremely popular in present-day surveying operations and are available from most of the survey instrument manufacturers (Fig. 12.2). They are easy to set up and use, and can be obtained for use at almost any required precision. A word of caution: All automatic levels employ a compensator referenced by gravity. This operation normally entails freely moving prisms or mirrors, some of which are hung by fine wires. If a wire or fulcrum breaks, the compensator will become inoperative, and all subsequent rod readings will be incorrect. Automatic levels use a magnetic damping system to control the swing of the compensator (see Fig. 12.1).

The operating status of the compensator can be verified by tapping the end of the telescope or by slightly turning one of the levelling screws (one manufacturer provides a push button), causing the telescopic line of sight to veer from the horizontal. If the compensator is inoperative, the cross hair will appear to deflect momentarily before returning to its original rod reading. Constant checking of the compensator will avoid costly mistakes caused by broken components.

Most of the levels (most new surveying instruments) are now equipped with a three-screw levelling base. Whereas the support for a four-screw levelling base is the centre bearing, the three-screw instruments are supported entirely by the foot screws themselves. Adjustment of the foot screws of a three-screw instrument effectively raises or lowers the height of the instrument line of sight. Adjustment of the foot screws of a four-screw instrument does not affect the height of the instrument line of sight, because the centre bearing supports the instrument. The surveyor should be aware that the

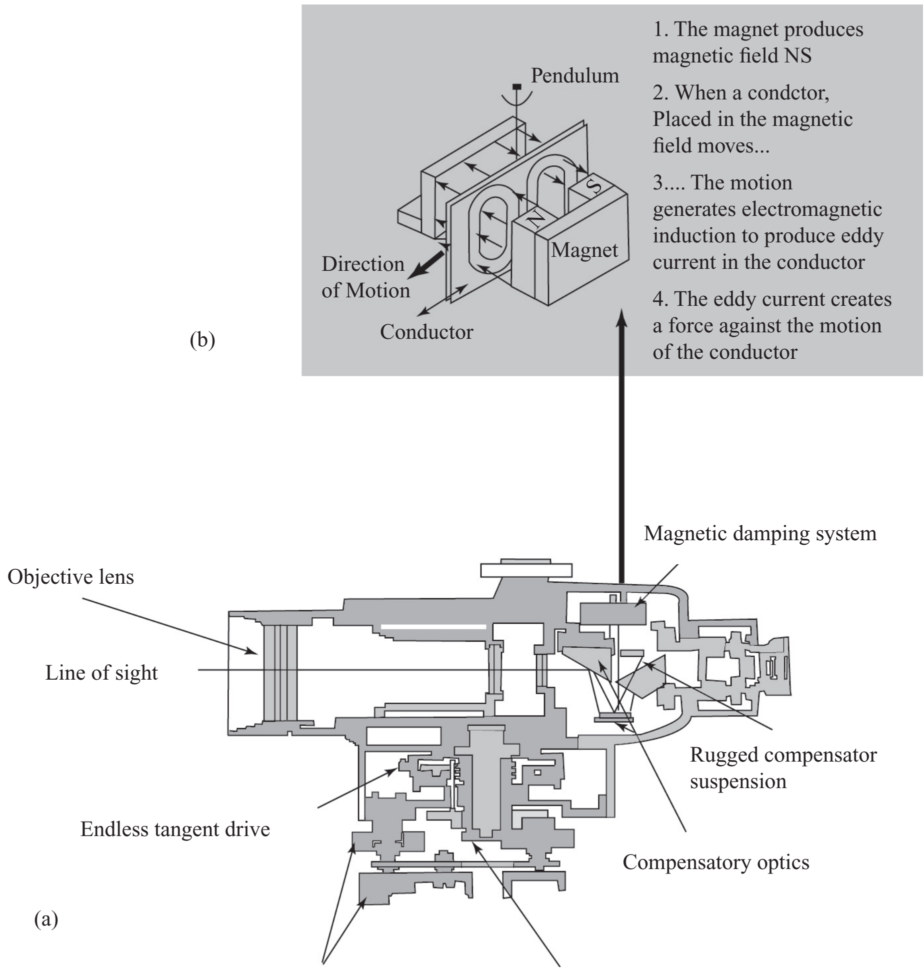


FIGURE 12.1 (a) Schematic diagram of an automatic level, and (b) magnetic damping system used in automatic level

adjustments made to a three-screw level in the midst of a setup operation will effectively change the elevation of the line of sight and could cause significant errors on very precise surveys (e.g., benchmark levelling or industrial surveying).

Levels used to establish or densify vertical control are designed and manufactured to give precise results. The magnifying power, setting accuracy of the tubular level or compensator, quality of optics, and so on are all improved to provide precise rod readings. The least count on levelling rods is 0.01 ft or 0.001 m. Precise levels are usually equipped with optical micrometres, so that readings can be determined one or two places beyond the rod's least count.

Many automatic levels utilize a concave base which, when attached to its domed-head tripod top, can be roughly levelled by sliding the instrument on the tripod top. This rough levelling can be accomplished in a few seconds. If the bull's eye bubble is nearly centred by this maneuver and the compensator is activated, the levelling screws may not be needed at all to level the instrument.

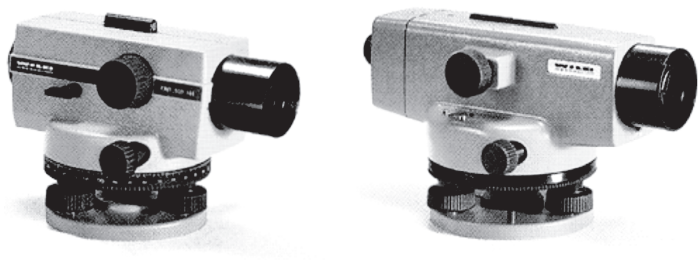


FIGURE 12.2 Different models of automatic levels (wild NAK2 and NA2)

12.2 DIGITAL LEVEL

The digital level is an automatic level (equipped with a pendulum compensator) capable of normal optical levelling with a rod graduated in feet or metres. It can also take readings automatically using a bar code rod. Figure 12.3 shows a digital level and bar code rod.

The digital level features digital, electronic image processing for determining heights and distances, with the automatic recording of data for later transfer to the computer. When used in electronic mode with the rod face graduated in bar code, this level will, with the press of a button, capture and process the image of the bar code rod. This processed image of the rod reading is then compared with the image of the whole rod, which is stored permanently in the level's memory module, to determine height and distance values. The rod shown in Fig. 12.3 is a 4.05-m long rod graduated in bar code on one side or either side.

After the instrument has been levelled, the observer must focus the image of the bar code properly. After that the observer should press the measure button to begin the image processing, which

FIGURE 12.3 A digital level with bar code rod (bar-coded on one side and normal graduation on the other side)

usually takes about 4 sec. The heights and distances are automatically determined along with the horizontal angles. In some models horizontal angles must be read and recorded manually in this case.

Preprogrammed functions include level loop survey, two-peg test, self-test, set time, and set units. Coding can be the same as that used with total stations, which means that the processed levelling data can be transferred directly to the computer database. The bar code can be read in the range of 1.8–100 m away from the instrument: optically, the rod can be read as close as 0.5 cm. If the rod is not plumb or is being held upside down, an error message will flash on the screen. Other error messages include ‘instrument not level’, ‘low battery’ and ‘memory almost full’. Rechargeable batteries are said to last for about 2,000–5,000 measurements. Distance accuracy is in the range of 1/2,500–1/3,000, whereas levelling accuracy is stated as having a standard deviation for a 1-km double run of 1.5 mm for electronic measurement and 2.0 mm for optical measurement.

12.2.1 Advantages of Digital Levels

A digital level offers the following advantages compared with the conventional levelling and recording procedures:

- Fatigue-free observation as visual staff reading by the observer is not required.
- User friendly menus with easy to read, digital display of results.
- Measurement of consistent precision and reliability due to automation.
- Automatic data storage eliminates booking and its associated errors.
- Automatic reduction of data to produce ground levels, thereby eliminating arithmetical errors.
- Fast, economic surveys resulting in saving in time of up to 50%.
- Data on the storage medium of the level can be downloaded to a computer enabling quick data reduction for various purposes.
- Digital levels can also be used as conventional levels with the help of dual marked staff (bar coded on one side of the staff for automated reading and conventional graduation on other side of the staff) in case it is difficult to record readings digitally (e.g., for long distances).

12.2.2 Components of Digital Level

Main components of digital level consist of two parts: hardware, which includes the digital level and levelling staff, and software.

The digital level and its levelling staff are manufactured in such a way that they can be used for both conventional and digital operations. In most cases, the digital levelling staves have dual marking. One side it is marked with binary bar codes for digital recording. The other side is marked with linear staff readings similar to conventional staff. The staff is made up of glass-fibre-strengthened synthetic material with low coefficient of thermal expansion for high accuracy. For highest precision work, Invar bar coded staves are used.

A digital level has the same optical and mechanical components as a normal automatic level. However, for the purpose of electronic staff reading a beam splitter is incorporated which transfers the bar code image to a detector diode array. Figure 1.1 shows components of a typical digital level. The light, reflected from the white elements only of the bar code, is divided into infrared and visible light components by the beam splitter. The visible light passes on to the observer, the infrared to diode array. The acquired bar code image is converted into an analogous video signal, which is then compared with a stored reference code within the instrument. The image correlation procedure then obtains the height relationship by displacement of codes, whereas the distance from instrument to staff is dependent on the scale of code.

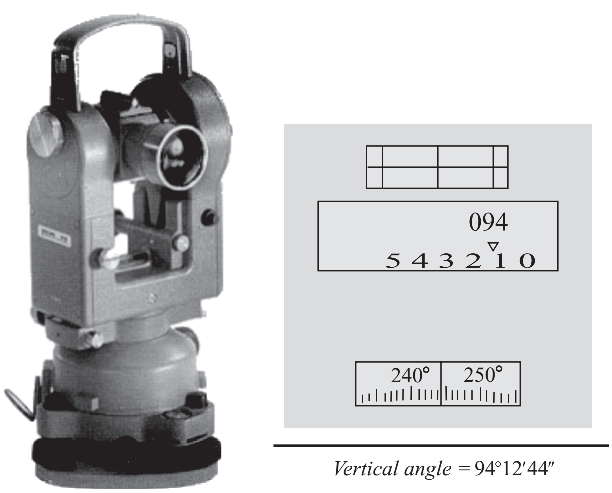
The data processing is carried out on a microprocessor and the results are displayed on matrix display. The measurement process is initiated by an interactive keypad and data can be stored onboard. Data from digital levels are stored onboard REC module or on PCMCIA/Micro-SD cards and can be transferred to computer for further processing. Various capabilities of digital levels are as follows:

- Measuring elevation
- Measuring height difference
- Measuring height difference with multiple instrument positions
- Levelling
- Slope setting
- Setting out with horizontal distance
- Levelling of ceilings

12.3 MICRO-OPTICAL THEODOLITES (MICROMETRE THEODOLITE)

Theodolites are generally classified into vernier theodolites and micro-optical theodolites, depending upon the method by which the horizontal and vertical circle readings are read. Micrometre theodolites with optical reading are generally more precise than vernier theodolites. In a micrometre optical theodolite or micro-optical theodolite (also known as micro-optic theodolite), both the vertical and horizontal angular readings are taken with the help of a micrometre and a direct angle reading microscope.

Micro-optical theodolites are generally more precise. For example, the micro-optical theodolite (wild T2 theodolite) shown in Fig. 12.4 reads directly to 0.1, and by estimation to 0.5.

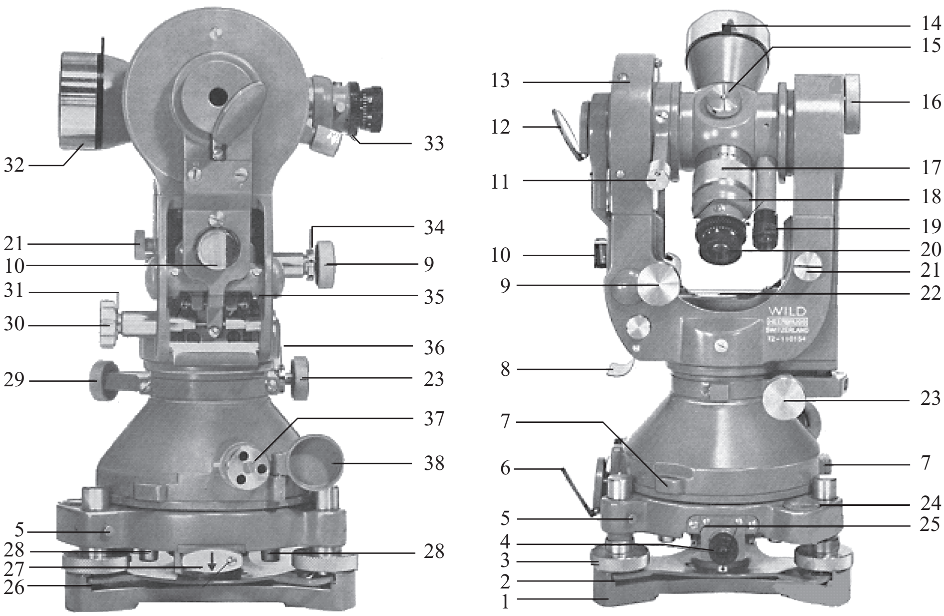


Vertical angle = 94°12′44″

FIGURE 12.4 Wild T2 one-second optical theodolite (micro-optical theodolite)

12.3.1 General Description of a Micro-Optical Theodolite

The tribrach of the instrument has 3-ft screws (3) (see Fig. 12.5(a) and 12.5(b)) used for setting the standing axis of the instrument vertical. The base plate (1) has a central thread to clamp to the tripod. The spring plate (2) presses the foot screws into the base plate. There is a swivel-locking knob (27) on the tribrach that can be used for detaching the theodolite from the tribrach by turning the locking knob. When the arrow mark on the turning knob is in the downward direction, the instrument is locked to the tribrach (arrow down, locked; arrow up, unlocked). An optical plummet is provided in the tribrach for centering the instrument over ground points. The circular level provided in the tribrach allows appropriate levelling. The circular level, together



(a) (b)

1. Tribrach
2. Spring plate
3. Foot screw
4. Optical plummet
5. Adjustment screw for 3
6. Illumination mirror for the horizontal circle
7. Retaining lug
8. Reflector for index level
9. Vertical drive knob
10. Index level viewing prism
11. Vertical clamp
12. Illumination mirror for the vertical circle
13. Vertical circle housing
14. Foresight
15. Back sight, centering pin and regulating knob for reticle illumination
16. Micrometre drive knob
17. Focussing sleeve
18. Reticle adjustment screws
19. Reading eyepiece
20. Telescope eyepiece
21. Selector knob for circle reading
22. Plate level
23. Horizontal drive knob
24. Circular level
25. Adjustment screws for 4
26. Locking screw for 27
27. Tribrach swivel knob locking device
28. Holding studs of the instrument
29. Horizontal clamp
30. Index level setting screw
31. Adjustment ring for 30
32. Telescope objective
33. Bayonet ring for telescope eyepiece
34. Adjustment ring for 9
35. Adjustment screws for index level
36. Adjustment ring for 23
37. Circle drive knob
38. Protective cover (open) for 37

FIGURE 12.5 (a) Side view of a micro-optical theodolite (part numbers explained along with (b)); (b) Straight view of a micro-optical theodolite with the optical plummet, is used for levelling and centering the tribrach, when the instrument is not attached to the tribrach.

The conical lower part of the theodolite contains the cylindrical standing axis system and the horizontal circle. The axis sleeve is rigidly connected to the centering flange and studs (28), which fit into the tribrach. The axis stem, to which the alidade is screwed, rotates inside the sleeve. For setting the circle, the horizontal circle carrier can be rotated around the axis sleeve by turning the circle drive knob (37), which is protected against the unintentional use by a cover (38). A mirror (6) is opened for the illumination of the horizontal circle.

For night observations, this can be replaced by a plug-in lamp, usually provided with advanced optical theodolites like the wild T2 model (see Fig. 12.4). To the left of the mirror is a socket, input point for the internal wiring of the instrument, to which a battery box can be connected for illuminating the horizontal and vertical circles and the reticle for night observations.

The alidade is the upper rotatable part of the instrument. Its main parts are the standards, the vertical circle (13), the tilting axis with the telescope, the circle reading optics (see Fig. 12.6) and plate level (22) for setting the standing axis vertical and the index level. By tightening the horizontal (29) and vertical (11) clamps, the instrument can be clamped so that the telescope points in any direction, and after tightening the clamps, the horizontal (23) and vertical (9) drives are used for fine positioning

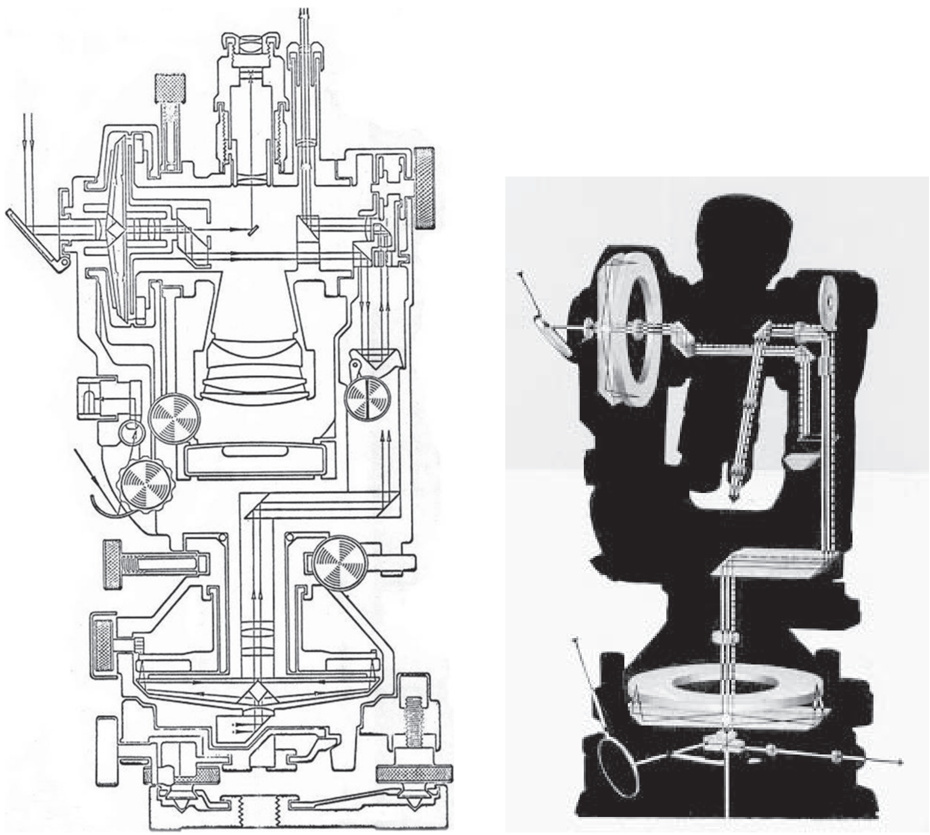


FIGURE 12.6 Cross-section of a micro-optical theodolite showing the paths of rays through the instrument

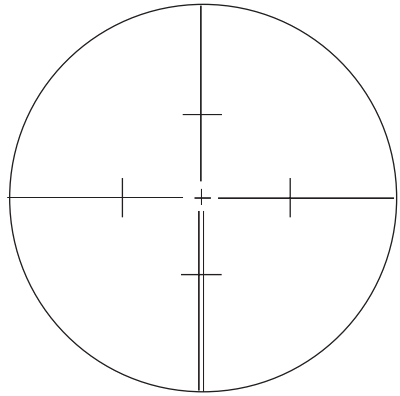


FIGURE 12.7 Telescope reticle with 1:100 stadia hairs

the telescope to a target. (In old model instruments, one standard hold the vertical circle (13) and its index level, which is centred by turning the index level setting screw (3) until the ends of the split bubble are seen in coincidence in the prism (10)). Coincidence indicates that the reading system for the vertical circle is plumb. An illumination mirror (12), which can be replaced by a pug in lamp for night observation, is located on the outside of the vertical circle housing. The other standard holds the micrometre knob (16), which is turned to set the images of the diametrically opposed circle graduations in coincidence, and the selector knob (21) for selecting which circle is to be read in the eyepiece (19) of the microscope.

The telescope is generally short and transit at both ends. The eyepiece (20) can be rotated for focussing the reticle and usually has a dioptric scale to allow immediate setting to suit an observer's eyesight. A bayonet ring (33) usually holds the eyepiece in position. After a slight left turn of the ring, the eyepiece can be removed and can be replaced by a diagonal eyepiece, auto collimation eyepiece or eyepiece lamp. One half of the vertical hair of the reticle is usually a single hair for splitting fine target, and the other half is a double hair for straddling a target, as in Fig. 12.7.

For optical distance measurement, stadia hairs with a 1:100 constant can be used. The telescope can be focused by turning the sleeve (17), which usually has engraved arrows to indicate the turning direction to infinity. When using electric illumination for night or underground work or tunnel work, the knob (15) on top of the telescope is turned, so that light from the vertical circle lamp is reflected from a mirror inside the telescope towards the reticle.

The brightness of the field of view illumination is varied according to the position of the knob. For daylight work, the position of the knob is of no importance. The image taken is seen in a micro-optic theodolite is inverted and laterally reversed.

12.3.2 Centering the Theodolite with the Optical Plummet

The optical plummet is used to centre the micro-optic theodolite over a ground point to about 0.5 mm when the instrument is the usual 1–2 m above the ground. Set the tripod over the ground mark and centre roughly by eye, by dropping a pebble. Attach the instrument to the tripod. Centre the circular bubble (24) and turn the eyepiece of the optical plummet (4) until its cross hairs are in focus. Slacken the tripod fixing screws, move the instrument over the tripod head until the cross hairs coincide with the ground mark (when doing this, do not rotate the tribrach in relation to the tripod head or the circular bubble will be disturbed) and finally retighten the fixing screw. Level up the instrument with the plate level as in the case of general setting of theodolites, adjust the centering if necessary and check the plate level again. The centering is correct when the cross hairs of the optical plummet are in coincidence with the ground point, and the plate level remains in the same position for all directions of the alidade. The tribrach without the theodolite can be set up on a tripod and centering and levelling can be carried out by means of the circular level and the optical plummet.

The optical plummet can also be used as follows. Set the tripod over the ground mark by eye estimation and tread the tripod shoes into the ground. Attach the instrument. Look through the optical plummet and bring the cross hairs onto the ground mark by turning the foot screws. Centre the circular bubble by extending or retracting the tripod's legs. The cross hairs will still be on the ground mark. Level up with plate levels if necessary, and adjust the centering by moving the instrument over the tripod head.

12.3.3 Focussing and Sighting

The telescope is aimed at the sky or towards a uniformly lighted background such as a piece of white paper or a wall. The milled black dioptric ring on the eyepiece (20) is turned until the cross hairs are sharp and black. To ensure that the cross hairs are focused correctly, the ring should again be turned anticlockwise until the image starts to go out of focus. A small clockwise rotation will again focus the hairs, so that they are sharp and black once more. This setting is constant for one observer, and the dioptric scale number should be noted so that the eyepiece focussing can be set quickly and correctly at all the future instrument stations.

The horizontal clamp (29) and the vertical clamp (11) are loosened and the target roughly sighted along the open sights of the telescope. Both clamps are then tightened, the focussing sleeve (17) turned until the target is seen and the drive screws (23 and 9) turned to bring it near the intersection of the cross hairs.

The focussing sleeve (17) is then turned until the object is seen in the telescope, clearly and without parallax. Arrows on the focussing sleeve indicate the correct direction of focussing to infinity. The observer must move his eye slightly, sideways and up and down, to ensure that there is absolutely no apparent movement of the image in relation to the cross hairs. If such a movement is detected, it means that parallax still exists between the reticle and the image, and this must be removed by a small re-adjustment of the focussing. Focussing does not affect the sharpness of the reticle.

After the correct focussing has been obtained, the telescope is ready for pointing. By turning the horizontal drive screw, the single vertical cross hair can then be made to split the target, or if the double vertical hair is to be used, the target may be straddled. For fine pointing, to obtain vertical angles, the vertical cross hair should first be moved a little to the left or right of the target and the horizontal cross hair then brought on to the target by means of the vertical drive screw. The last turn of a drive screw should always be made clockwise to avoid the possibility of a backlash error.

12.3.4 Reading Angles

For daylight work, the illumination mirrors of the instrument (6 and 12) are opened and turned towards the light so that the circles are evenly illuminated. At night and inside tunnels or underground, it will be necessary to use the electric illumination for taking angular measurements. The eyepiece (19) on the reading microscope is turned until the images of the circle and the micrometre scale are in focus. The selector knob (21) is used to select the required circle. The colour seen in the reading eyepiece also indicates which circle is being read. Usually, two distinct colour backgrounds will be given for horizontal and vertical circles. There is a line on the selector knob which is usually red in colour, and when this line on the selector knob is horizontal the three windows in the reading microscope are yellow (for T2 theodolite) and denotes horizontal circle readings. When the red line is vertical, the windows are white and the vertical circle is read.

In the top window of the reading microscope, images of *graduation lines* of diametrically opposite parts of the circle are seen. The two images appear to be separated by a fine line (see Fig. 12.8(b)). The actual graduation interval of the circle is 20, but coincidence between the graduation lines occur every 10. In the upper part of the central window are seen the whole degree (see Fig. 12.8(a)). Each number has a triangular index beneath it. Below this index is the row of numbers for tens of minutes having numbers 0 to 5 (00, 10, 20 . . . 50).

In the bottom window is the micrometre scale. Since the graduation lines in the top window can be brought into coincidence every 10’, the range of the micrometre is 10’. The value of one scale interval is 1. Reading should only be taken to the nearest second, although it is permissible to estimate half an interval (0.5”) by estimation. Estimating the tenth of an interval is pointless, because in the time needed for this the micrometre can be reset and a second reading can be taken. The mean of two such readings will be more realistic than estimating tenths, and in any case, there is really no point in trying to estimate the tenth of a second if the inherent accuracy of the theodolite is considered.

Before taking a reading, the graduation lines in the top of the window should be made to coincide. If the graduation lines in the top of the window are not coincided (see Fig. 12.8(b)), the micrometre knob (16) must be turned until the graduation lines coincide exactly as in Fig. 12.8(c).

The angle reading principle is simple (see Fig. 12.8(a)). The coincidence is obtained by turning the micrometre knob and the readings are noted as follows: The number of degrees is 94. The 1 under the triangular index gives 10 min. The micrometre reading is 2’44 . The total is 94° + 10’ + 2’44”,

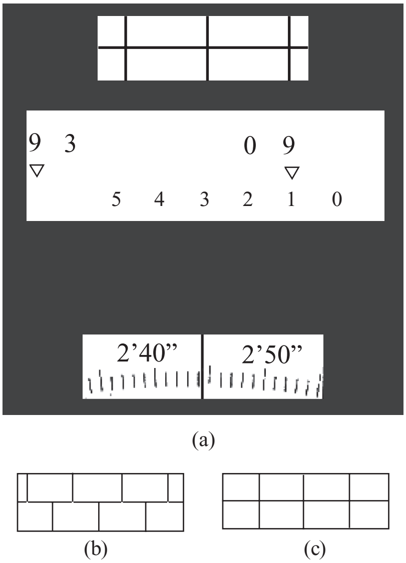


FIGURE 12.8 (a, b, c) Reading example, horizontal or vertical circle, 94°12’44”

which in practice is read directly as 94°12’44”. The vertical and horizontal circles are read exactly the same way as described.

When measuring angles of sets of directions, it is convenient to set the horizontal circle so that the initial reading to the reference object (R.O.) is about zero or the known bearing from the instrument station to the R.O. After fine pointing the telescope to the R.O., the required reading is set in the reading microscope. As an example, assume the pointing to R.O. should be set to a reading of 94°12’44”. The 2 44 are set against the index of the micrometre scale by turning the micrometre knob (16). The cover (38) is then opened and the circle drive (37) turned until 94° appears and its triangle is over the number 1 (10). Finally, still by using the circle drive, the graduations are set as accurately as possible in coincidence. The reading 94°12’44” has now been set (see Fig. 12.8(a)).

The cover (38) is now closed to prevent any accidental displacement of the circle during subsequent measurements. In the case of ‘zeroing’, when readings are to be taken on both faces, it is recommended to set the face left reading to slightly more than zero, so that the possibility of having a face right reading less than 180°, which tends to make the reduction slightly more complicated, is avoided.

12.3.5 Measuring Single Angles

The measurement of a single angle between two targets is the most accurate form of angle measurement. As only two targets have to be sighted, measuring time is short. Thus systematic errors arising from residual changes in the verticality of the standing axis and twisting of the tripod can practically be avoided. The usual sequence for measuring a single angle is:

Face Left	Swing Clockwise	Left Target (R.O.)
Face Left	Swing Clockwise	Right Target
Face Right	Swing Anticlockwise	Right Target
Face Right	Swing Anticlockwise	Left Target (R.O.)

As described earlier, it is convenient to set the circle to about zero for the initial pointing to the left target (R.O.). If greater accuracy is required, the measurement may be repeated with additional sets and the mean of them all taken to obtain the accepted value. If two sets are to be observed, the setting for the second one should be about 90° greater than for the first, and with four sets, 45°, 90°, 135°, should be added (i.e., for *n* sets, the setting should be changed by 180° ÷ *n*). In all cases, the minutes and seconds should also be altered. When only one set is being observed, it is good to change the setting of the circle at the end of the first half. After transmitting to the face right position, the circle drive (37) is used to change the settings by about 90°. Each half set is reduced to the reference object separately, and the mean result is taken afterwards.

12.3.6 Measuring Sets of Directions

In triangulation it becomes necessary to measure the direction to several targets from one station. The measuring routine is the same as that described above (Section 12.3.5), except there are more targets. A distinct target, which is likely to remain clearly visible, is chosen as the R.O.. With the instrument in face left position, the alidade is turned clockwise and target 1 (R.O.) is finely pointed. The circle drive knob is turned to set the circle to about zero or to any other required value. The reading for target 1(R.O.) is now taken. The alidade is turned clockwise and targets 2, 3, 4, . . ., *n* are observed. After the last target, *n*, the telescope is transmitted to face right, and with an anticlockwise swing the targets are observed in the reverse order, that is, n, , . . ., 4, 3, 2, 1 (R.O.). If additional accuracy is required more sets can be taken with different circle settings as described in Section 12.3.5. It may be desirable to close each half set to the R.O. In this case, the first half set, in face left and clockwise swing will be 1(R.O.), 2, 3, 4, . . ., n, 1(R.O.). The telescope is then transmitted to face the right anticlockwise swing for the second half set, 1(R.O.), n, . . ., 4, 3, 2, 1(R.O.). The values are averaged and any difference between the R.O. readings can be adjusted through the set.

If more than six targets are to be observed, it may be preferable to split the observations into at least two different sets, with each set containing not more than six targets. The R.O., and preferably one other target, should be common to each set.

12.3.7 Measuring Vertical Angles

With the telescope in the face left position, bring the horizontal hair on to the target by means of the vertical drive (9). [For old model micro-optical theodolites, centre the index level by turning the setting screw (30) until the two ends of the split bubble are seen in coincidence in the prism (1).] Take the vertical circle reading. Transit the telescope and repeat the observation in the face left position. The reading AL in face left is the zenith angle ϕ, the reading AR in face right is $(360 - \phi)$. The vertical angle β (elevation (+) or depression (–)) can be derived from the vertical circle reading as follows:

$$\beta_L = 90° - A_L$$
$$\beta_R = AR - 270°$$
$$\beta = ½\,(\beta_L + \beta_R)$$
$$\phi = ½\,(A_L - A_R)$$

(360° has to be added to A_L)

Example Solution

$$\begin{aligned} A_L &= 94°12'49" & \beta_L &= (-)\ 4°14'49" \\ A_R &= \underline{265°44'51"} & \beta_R &= (-)\ 4°15'09" \\ A_L + A_R &= 359°59'40" & \beta &= (-)\ 4°14'59" \\ A_L + A_R &= 188°29'58" & \phi &= 94°14'59" \end{aligned}$$

Reduction by this method is self-checking. The sum $A_L + A_R$ should always be constant within ±5". The difference of $A_L + A_R$ from 360 is twice the vertical collimation error.

12.3.8 Measuring Vertical Angles with the Three Wire Method

If the vertical angle is to be measured several times to increase the accuracy and to expose gross reading errors, the two horizontal stadia hairs as well as the horizontal cross hair are used in both face left and face right observations. Let U_L (face left) and U_R (face right) be the vertical circle reading for the upper stadia hair, A_L and A_R for the centre hair as before, and L_L and L_R for the lower hair. In this example, the number in brackets indicates the sequence of observing.

$$\phi = ½\,(FL - FR)$$

(360° has to be added to FL)

$$\beta = 90 - \phi$$

Example

Let the angle obtained using the three wire method be:

$U_L = 85°\ 44\ 34$ (1)	$A_L = 85°01'47''$(2)	$L_L = 86°18'56''$(3)
$U_R = \underline{273°41'19''}$(6)	$A_R = \underline{273°58'28''}$(5)	$L_R = \underline{274°15'40''}$(4)
$2\phi = 172º03'15''$	$2\phi = 172º03'19''$	$2\phi = 172º03'16''$
Mean = $172°03'19''$	$\phi = 86°01'38''$	$\beta = 03°58'22''$

In the face left and face right observation, one always starts to observe with the upper hair in the field of view.

12.3.9 Tacheometric Observation

For tacheometric work, it is usually sufficient to use the theodolite in face left only. The telescope has stadia lines (multiplication constant 100x) for measuring the distance of a vertical staff. The vertical staff is read and is cut by the lower stadia hair and the upper stadia hair. The difference between the staff readings is called the intercept (S), and for a horizontal sight, 100 times the intercept is the horizontal distance from the centre of the instrument to the staff. If the line of sight is inclined, the vertical angle (β) must be measured (reading on face left is usually sufficient) and the horizontal distance D is calculated using:

$$D = 100 \cdot S \cdot \cos^2 \beta \ (\beta = \text{vertical angle})$$
$$= 100 \cdot S \cdot \sin^2 \phi \ (\phi = \text{zenith angle})$$

If the difference in heights is also needed, the reading (z) where the staff is cut by the centre hair is taken. The height (i) of the theodolite's tilting axis above the ground point is also measured. The difference in heights (H) between the ground at the instrument station and the foot of the staff is:

$$\Delta H = 100 \cdot \text{S} \cdot \sin \beta \cos \beta + (i - z)$$
$$= 100 \cdot S \cdot \sin \phi \cos \phi + (i - z)$$

The sign of the first term ($100 \cdot S \cdot \sin \beta \cos \beta$) is positive for an elevation and negative for a depression. To simplify the calculations, it is useful to sight the staff so that $z = i$.

12.3.10 Horizontal Collimation Error and Its Adjustments

Any deviation from a right angle is known as the horizontal collimation error, c. Horizontal collimation error adjustment is needed to make the line of sight perpendicular to the tilting axis. If observations are made on both faces and the mean of the reading is taken, the horizontal collimation error is eliminated. Therefore, there is no need to try to adjust the horizontal collimation error to zero, because of the following reasons:

1. It is neither possible nor necessary
2. If the adjustments screws are not set correctly (i.e., they are either too tight or loose), the theodolite will never hold again its adjustment

For angle measurements, if the collimation error is less than 1° or 30 (i.e., the difference between the face left and face right readings is less than 1° or 2°), adjustment is not usually recommended.

However, for setting out very long straight lines by transiting the telescope and taking the mean position, it may be convenient to keep the collimation error smaller than this.

For determining the collimation error, the instrument is levelled up, put in the face left position (vertical circle left of the line of sight) and a sharp target, which is at least 100 m away and at about the same height as the instrument, is sighted with the vertical hair. The horizontal circle is read (A_L). The telescope is transited, the vertical hair is brought on to the target again, and the horizontal circle reading in face right (A_R) taken.

The following example shows how to determine the horizontal collimation error, *c*.

	360°	*c* = Correction	Corrected Value
A_L	48°14'53"	–33"	48°14'20"
A_R	228°13'47"	+33"	228°14'20"
$A_L - A_R$	180°01'06"	1'06"	180°00'00"
		$2c$ = 1'06"	

The telescope is still pointed to the target in the face right position. To adjust the horizontal collimation error, turn the micrometre knob (16) and set the correct value 4'20 against the index of the micrometre scale. Now turn the horizontal drive (23) until the graduation lines giving the full corrected value of 228°14'20", are in exact coincidence. The reading is now correct, but the vertical hair will no longer be on the target.

Now the horizontal capstan headed screws (18) in the eyepiece (normally four numbers) are adjusted. If the vertical hair is on the left of the target, unscrew the adjustment screw on the left side of the eyepiece by a small amount and immediately screw the other one by the same amount (if the hair is right of the target, the reverse will apply). Turn the screws in this way, by small and equal amounts, until the vertical hair coincides with the target. Do not over-tighten the screws or leave them loose. Finally, repeat the test to see if the horizontal collimation error is within acceptable limits. This adjustment is so delicate that this should only be carried out when it is absolutely essential.

12.3.11 Vertical Collimation Error (Index Error) and Its Adjustments

With a horizontal line of sight and the index level centred, the vertical circle reading should be 90° 00' 00" in face left. Any discrepancy in the above value is known as the vertical collimation error or index error, *i*. By measuring vertical angle in both faces and by taking the mean, the effect of this error is eliminated. The error should normally be adjusted only if it is more than 30".

To determine the error, level up the theodolite, set the telescope in the face left position (vertical circle left of the line of sight), bring the horizontal hair on to a well-defined target which should be at least 100 m away, set the ends of the split bubble in coincidence by turning the index level setting screw and then take the vertical circle reading (*A*L). Repeat in the face left position and take the vertical circle reading (*A*R). The following example shows how to determine the vertical collimation error, *i*.

	360°	*i* = Correction	Corrected Value
A_L	86°14'35"	+ 44"	86°15'19"
A_R	273°43'57"	+ 44"	273°44'41"
$A_L - A_R$	359°58'32"	1'28"	360°00'00"
		$2i$ = 1'28"	

The theodolite is still in face right with the horizontal hair on the target. To adjust the vertical collimation error, turn the micrometre knob (16) and set the corrected value 4’41 against the index of the micrometre scale. Then turn the index level setting screw (30) until the circle graduations coincide to give the full correct value of 273°44’41”.

The reading is now correct, but the ends of the bubble will no longer be in coincidence. The adjustment screws (35) must now turn to bring the ends back into coincidence. The adjustment is made step-by-step, slackening one screw by a small amount and immediately tightening the other by the same amount until the ends of the split bubble coincide. Do not over-tighten the screws or leave them loose. Repeat the observations to determine if the vertical collimation error is now within the acceptable limits.

12.4 DIGITAL PLANIMETER

Planimeters are area measuring tools. A planimeter is a measuring instrument used to determine the planar area of a material/plan/figure. It is used to measure areas on maps of any kind and scale. To measure the area of a drawing/object with a planimeter, first to set the scale of the drawing object and then push, pull or roll the planimeters head around along the perimeter or outline of your drawing to get the area. There are mainly two kinds of planimeteres, mechanical planimeter and digital planimeter.

Digital planimeters require initial settings for units and scale (see Figs. 12.9–12.10). It gives a direct reading of the area traced as square centimetres, with some reading directly in any unit of area including acres, square metres, square kilometres, etc. The most advanced units will also store data for downloading into a personal computer. Most digital planimeters have various memory functions which enable you to add areas, accumulate measurements and average multiple measurements. Digital planimeters are available with pole arms or rollers. Best results are obtained by measuring over smooth surfaces. Photos (glossy surface) or laminated maps increase the chances of the measuring wheel slipping of planimeters, resulting in inaccurate measurements. Folds, seams, tape or any irregularity on the surface can cause errors in measurements.

FIGURE 12.9 Digital planimeter (electronic rolling planimeter Courtesy M/S.Sokkia)

FIGURE 12.10 Digital planimeter (courtesy M/S.Sokkia)

A digital planimeter have a high friction roller with a diamond-inlaid wheel to provide stable and precise measurements. There is a tracer arm, which is to be moved while measuring lengths and areas. There is a power switch fixed to this arm, which will also release the arm for movement. A magnifying lens is provided over the tracing point of the arm, for getting a clear view of lines and points. The tracing point is kept over the points or moved over curves to be traced. A mode switch enables the shift from the point mode to continuous mode.

Digital planimeter is very simple to use. First of all, the mode has to be selected (a point mode or continuous mode). In the case of a straight-line boundary, the point mode is used and in point mode, the tracer point is kept over the vertices or the end points of the lines one after the other and the lengths of the lines are measured. But in the case of curved boundaries, the continuous mode is used. The tracer point has to be carefully moved along the boundary line.

The procedure for measuring lengths and areas using a planimeter is as follows:

i. Switch on the power supply of digital planimeter (by lifting the fixing lever of the tracer arm).
ii. Set the units, vertical scale and horizontal scale with the help of the set key and number keys. Press the set key repeatedly to set the units and the horizontal and vertical scales.
iii. Set the tracer point on the starting point and press the start switch to start operation.
iv. Move the tracer point to the other point of the line in the case of a straight line. For curved lines, move the tracer point carefully over the boundary.
v. After the measurement press the end key to display the length and areas in terms of the units and scales already set. Feed the results in to the instrument memory. Download the results to a computer.
vi. Switch off the instrument after measurement.

General specification of a planimeter are as follows:

Measurement functions: coordinate, area, line length, side length, radius, angle

Measurement mode : Line(point), Curve (stream), circular arc (Arc)

Measuring units: Metric (mm, cm, m, sq.m), foot (in., ft, mi, acre), user-defined unit

Measuring range 380mm × 10m

Accuracy :±0.1%

12.5 LASER DISTANCE METRE (LASER RANGE FINDER)

A laser distance metre uses a laser beam to determine the distance to an object. The most common form of laser rangefinder operates on the time of flight principle by sending a laser pulse in a narrow beam towards the object and measuring the time taken by the pulse to be reflected off the target and returned to the sender. Due to the high speed of light, this technique is not appropriate for high precision sub-millimetre measurements, where triangulation and other techniques are often used. A laser distance metre sends a pulse of laser light to the target and measures the time it takes for the reflection to return. For distances up to 30 m, the accuracy is ±3 mm. On-board processing allows the device to add, subtract, calculate areas and volumes and to triangulate. Laser distance metre is also known as a laser tape. Laser tape measures are alternatives to traditional metal tape measures; they are used to calculate lengths, widths and heights of up to 200 m. Laser tape measures are now commonly used by contractors, surveyors, architects, GIS professionals and other people who make a lot of field measurements (See Fig. 12.11).

To use a laser tape measure, place the device on one end of what one want to measure, and then aim the laser beam so that it hits an object at the other end. If there's no wall, pole or anything like that, a target can be put at point, one want to measure up to. The process is similar to using a conventional tape measure, except they use a laser beam instead of metal tape. Once laser hit at the right spot, press the button, and the laser distance metre measures and calculates the distance and displays it on its screen. The calculation is done through precision optics and laser physics using the phase-shift method, in which a laser hits an object and compares its reflection with the beam sent out, or using the time-of-flight method in which the time it takes for an optical pulse to reflect back is calculated. Some laser distance metre can measure multiple distances and add them together automatically.

FIGURE 12.11 A laser distance metre (courtesy of M/S.Leica)

REVIEW QUESTIONS

1. What is the principle of an automatic level? Describe the function of a compensator in an automatic level.
2. What is the function of a magnetic damping system in an automatic level?
3. What is a digital level? How does the digital level measure the levels and horizontal distance using a bar code rod?
4. How does an optical theodolite measure horizontal and vertical angles?
5. Describe the centering of a micro-optic theodolite.
6. Describe the method of focussing and sighting while using a micro-optic theodolite.
7. Describe the method of measuring vertical angles using the three-wire method with a micro-optic theodolite.

AERIAL SURVEYING 13

Chapter Outline

13.1 GENERAL BACKGROUND

Photogrammetry is the science of making measurements from aerial photographs. It is defined as the art of obtaining reliable measurements from photographs for the determination of geometric aspects such as size, form, angle, distance and coordinates of the objects photographed. Photogrammetry is the technique of measuring two-dimensional (2D) or three-dimensional (3D) objects from photographs or imageries stored electronically on tape or disk taken by a video or charge-coupled devices (CCD) cameras or radiation sensors such as scanners. The important use of photogrammetry is the construction of accurate planimetric and topographic maps by radial triangulation and photogrammetric procedures.

The final results from photogrammetry are:

1. Coordinates of the required object points.
2. Topographic and thematic maps.
3. Rectified photographs and orthophotos.

The most important feature is the fact that the objects are measured without being touched. Therefore, instead of photogrammetry, some authors use the term remote sensing. Remote sensing is rather a young term, which was originally confined to working with aerial photographs and satellite images.

Principally, photogrammetry can be divided into:

1. Depending on the lens-setting
   a. Far range photogrammetry (with camera distance setting to indefinite).
   b. Close range photogrammetry (with camera distance settings to finite values).
2. Depending upon the platform used for taking photographs
   a. Aerial photogrammetry (which is mostly far range photogrammetry).
   b. Terrestrial photogrammetry (mostly close range photogrammetry).

Aerial photogrammetry is mainly used to produce topographic or thematic maps and digital terrain models. In aerial photogrammetry, a camera mounted in an aircraft flying over the required area takes aerial photographs. Aerial photogrammetry is good for reconnaissance and preliminary surveys for the construction of roads, railways and transmission lines. Terrestrial photogrammetry is a branch of photogrammetry wherein the photographs are taken from a fixed position on or near the ground. Hence in terrestrial photogrammetry, ground photographs are used. The users of close-range photogrammetry are architects and civil engineers, and they use it for the supervision of buildings, documentation of their current sites, and for the study of deformation and damages.

13.2 TERRESTRIAL PHOTOGRAMMETRY

The principle of terrestrial photogrammetry is similar to plane table surveying. When the directions of an object photographed from two points of a measured base are known, their position can be located by the intersection of two rays drawn to the same object. But in plane table surveying, most of the work is executed in the field, and in terrestrial photogrammetry it is done in the office. It is used for taking photographs in terrestrial photogrammetry. A photo-theodolite is a combination of a one-second theodolite and a terrestrial camera.

13.3 AERIAL PHOTOGRAMMETRY

In aerial photography, also known as remote sensing, information about the environment is recorded from a distance, usually from sensors carried in an aircraft or spacecraft. An aerial photograph is an image, at an instant of time, detected by the photographic film (sensors) due to the light reflecting properties of the terrain on the top of earth surface. The nature of the image depends upon the quality and nature of the film, the reflectance characteristics of the terrain and atmospheric conditions at the time of exposure. The details recorded vary with the scale of the photography and the resolving power of the camera film system. A photographic analogue of the landscape attracts a variety of users, because landscape is the main topic of geographical, geological and topographical studies. Aerial photography is mainly used for photo interpretation and image interpretation. Photo interpretation is a qualitative aspect concerned with the identification of features and their significance. Photogrammetry is a quantitative aspect concerned with the accurate measurement of features recorded by photography or infrared or microwave images recorded from a satellite. Image interpretation is a general term used for the interpretation of infrared or microwave images.

13.4 PHOTOGRAPHING DEVICES

A photographic image is a central perspective. This implies that every light ray, which reached the film surface during exposure, passes through the camera lens, which is mathematically considered as a single point, the so-called perspective centre. In order to take measurements of objects from photographs, the ray bundle must be reconstructed. Therefore, the internal geometry of the used camera, which is defined by the focal length, has to be precisely known. The focal length is called the principal distance, which is the distance of the projection centre from the image plane's principal point.

Modern digital cameras have revolutionized photography. Both digital and film-based cameras use optical lenses. The digital camera records image data with electronic sensors such as CCD and complementary metal oxide semiconductor (CMOS) devices. The advantage of a digital camera is that the image data can be stored, transmitted and analyzed electronically. Cameras on board satellites capture image data and transmit it back to earth along with all sensed data for processing.

Film-based photographic technology still dominates the major part of aerial imagery systems. A 23 cm × 23 cm format is used for majority of the film-based photographic cameras. The film-based cameras capture topographic features in detail, and these photos can be scanned effectively for electronic processing and interpretation of the images.

There are two types of camera systems, which are commonly used for aerial photography. The first one is a single camera, which uses a fixed focal length of a large format negative (negative film) usually 23 cm × 23 cm size. This camera is used only for aerial photography and is equipped with a highly corrected lens and vacuum pressure against the film to minimize distortion, as the aerial photographs used for photogrammetric purposes must satisfy the required accuracy parameters. The second type of camera system consists of one or more cameras that use a smaller negative size, like the most common 35-mm films used for general purposes. These smaller format camera systems do not have a high-quality lens that is required to meet the normal measurement accuracies for the production of standard maps using photogrammetric methods. These smaller format camera systems are very useful and inexpensive for updating land use changes and for the acquisition of special type of photography to enhance terrain aspect, such as vegetation growth.

13.4.1 Metric Cameras

Metric cameras have stable and precise internal geometries and very low lens distortions (see Fig. 13.1). They are distinguished by the complexity and accuracy of the lens assembly. Therefore, they are very expensive devices. The principal distance of the lens assembly is constant, which means that the lens cannot be sharpened when taking photographs. As a result, metric cameras are usable within a limited range of distance towards the object. Aerial metric cameras are built into aircrafts, mostly looking straight downwards.

Almost all metric aerial cameras have an image format of 23 cm × 23 cm. A metric aerial camera must have the following features:

1. A lens with a low distortion.
2. Focal length of lens calibrated and corrected to 0.01 mm.
3. A film-flattening device to ensure a flat-film surface at the time of exposure.
4. Fiducial marks to locate the principal axis (at least four fiducial marks to orient exposure to the camera calibration).

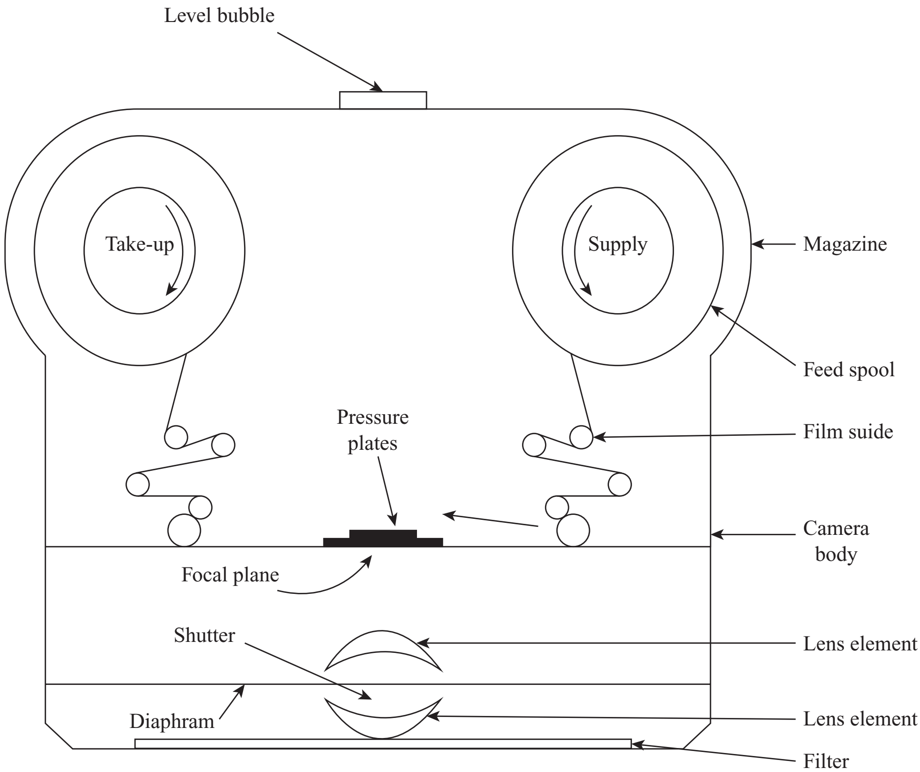


FIGURE 13.1 Components of an aerial survey camera (metric camera)

As aerial photographs are used for the production of small- and medium-scale topographic maps, the aerial survey cameras are equipped to minimize all errors such as distortion of lens. The drive mechanism is housed in a camera body as shown in Fig. 13.1. This mechanism is motor driven, and the time between exposures to achieve the required overlap is set based on the photographic scale and the ground speed of the aircraft. The film is thus advanced from the feed spool to the take-up spool at automatic intervals. The focal plane is equipped with a vacuum derive to hold the film flat at the instant of exposure. The camera has at least four fiducial marks built in, such that each exposure can be oriented properly to the camera calibration. (Fiducial marks are a set of four marks located in the corners or edge centred, or both, of a photographic image. These marks are exposed within the camera onto the original film and are used to define the frame of reference for spatial measurements on aerial photographs. Opposite fiducial marks, if connected, intersect at approximately the image centre of the aerial photograph.) As the lower and air pressure inside the aircraft tends to pull the film away from the focal plane and towards the lens resulting in poor focussing, the focal plane is equipped with a vacuum drive.

If the axis of the aerial camera is tilled by 20-degree from the vertical, the resultant photograph is termed as an oblique photograph (high oblique if the horizon is recorded and low oblique if it is not). Usually, in aerial photography for surveying purposes, the axis of the camera is maintained vertical or near vertical. Although oblique photographs may give a conventionally more familiar view than vertical photographs, they are not suitable for detailed interpretation.

In order to avoid blurring of the image, image movement compensation (IMC) is achieved by moving the film across the local plane in step with the aircraft's speed over a terrain. The terrain is photographed continuously on the moving film, and thus there is no single centre of perspective.

13.4.2 Stereo Metric Camera

If an object is photographed from two different positions, the line between the two projection centres is called a base. If both photographs have viewing directions that are parallel to each other and in a right angle to the base, then they have similar properties like the two images of human retinas. Therefore, the overlapping area of these two photographs, which are called a stereo-pair, can be viewed in 3D, simulating human stereoscopic vision. In practice, a stereo-pair can be produced with a single camera photographed from two positions or using a stereo metric camera. A stereo metric camera, in principle, consists of two metric cameras mounted at both ends of a bar. This bar, which functions as the base, has a precisely measured length of 40 cm or 120 cm. When both cameras have the same geometric properties, stereo pairs are created easily.

13.5 AERIAL PHOTOGRAPHS

Aerial photographs used for mapping and photo-interpretation can be divided into:

1. According to the Direction of Camera Axis

Vertical photographs Vertical photographs are taken with the axis of the aerial camera vertical or nearly vertical. A vertical photograph taken using an aerial camera resembles a planimetric map. These photographs can be obtained with low-tilt cameras used for photogrammetry.

Terrestrial photographs Terrestrial photographs are usually taken using a photo theodolite in terrestrial photogrammetry. This is used for the survey of structures and monuments of archaeological importance.

Oblique photographs Aerial photographs taken with the optical axis of the aerial camera tilted from the vertical are called oblique photographs. These photographs cover larger ground area, but the quality of details fade out towards the far end of the photograph.

2. According to the Type of Exposure

Black and white (*panchromatic*) Black and white photographs are commonly used for aerial photogrammetry. The image on these photographs is similar to the natural view, but the colour is restricted to shades of grey. To eliminate haze, a yellow filter is used for black and white photographs.

Colour photographs (*true colour*) True colour photographs present the objects as they appear in their natural colour. For good colour contrast, a scale larger than 1:25,000 is used. As compared to black and white aerial photographs, they have better interpretation capability and are better for photogrammetric studies. Filters are to be selected considering the presence of haze and flying altitude. Colour photographs are expensive and great care is required to preserve them.

Infrared The film used for infrared is of a special type and has advantage of recording information, which panchromatic film is not able to gather. Infrared photography, can be black and white or in colour, depending on the type of film used. In black and white infrared photographs, the vegetation appears in lighter tones and water appears in darker tones, as water has high-absorption characteristics of incident infrared energy and vegetation usually has a high infrared reflectance. Infrared colour photographs show vegetation in various hues of red. The difference between types of healthy and unhealthy vegetation is brought out in distinctive colours. The false colour film is effective in the case of camouflage detection because the infrared reflections from natural foliage and from foliage painted in colour are different. In practice, the use of infrared false colour picture in combination with panchromatic photography enhances the interpretation considerably.

13.5.1 Information Recorded on Photographs

The information recorded on a typical photograph are

1. Principal distance for determining the scale of photograph.
2. Fiducial marks for determination of principal points.
3. Altimeter recording to find flying height at the moment of exposure.
4. Watch recording giving the time of exposure.
5. Level bubble recording indicating tilt of camera axis.
6. Principal distance for determining the scale of photograph.
7. Number of photograph, the strip and specification number for easy handling and indexing.
8. Number of camera to obtain camera calibration report.
9. Date and time of photograph.

13.6 PHOTOGRAPHIC SCALE

The features shown on a photograph are similar to those on a planimetric map, but with the difference that the planimetric map has been rectified through ground control, so that the horizontal scale is consistent in any point on the map The scale of a photograph is the ratio between the distance measured on the photograph and the ground distance between the same two points.

In photogrammetry, the representative factor is a fraction in which the numerator is unity and the denominator is the number of units on the ground represented by one unit on the photograph or map. The air photo contains scale variations, unless the camera is perfectly level at the instant of taking the photograph and also the terrain being photographed is level.

As the aircraft is subject to tilt and changes in altitude due to updrafts and down drafts, the chances of the focal plane being level at the instant of exposure are rare. Changes in elevation cause scale variation (see Fig. 13.2). The basic problem is transferring an uneven surface like the ground to the flat focal plane of the camera. In Fig. 13.2(a), points *B, O* and *C* are almost at the focal plane is level. Therefore, all scales are true on the photograph, as the distance $BO = B'O'$ and $OC = O'C'$. *B, O* and *C* are points on a level reference datum that would be comparable to the surface of a planimetric map.

Therefore, under these unusual circumstances, the scale of the photograph is uniform because the ratio *bo:BO* is the same as the ratio *oc:OC*. In Fig. 13.2(b), the focal plane is tilted and the topographic relief is variable. Points *B, C, O, D, E* and *F* are at different elevations. It can be seen visually that although $B'C'$ equal $C'O'$, the ration *bc:co* is not equal.

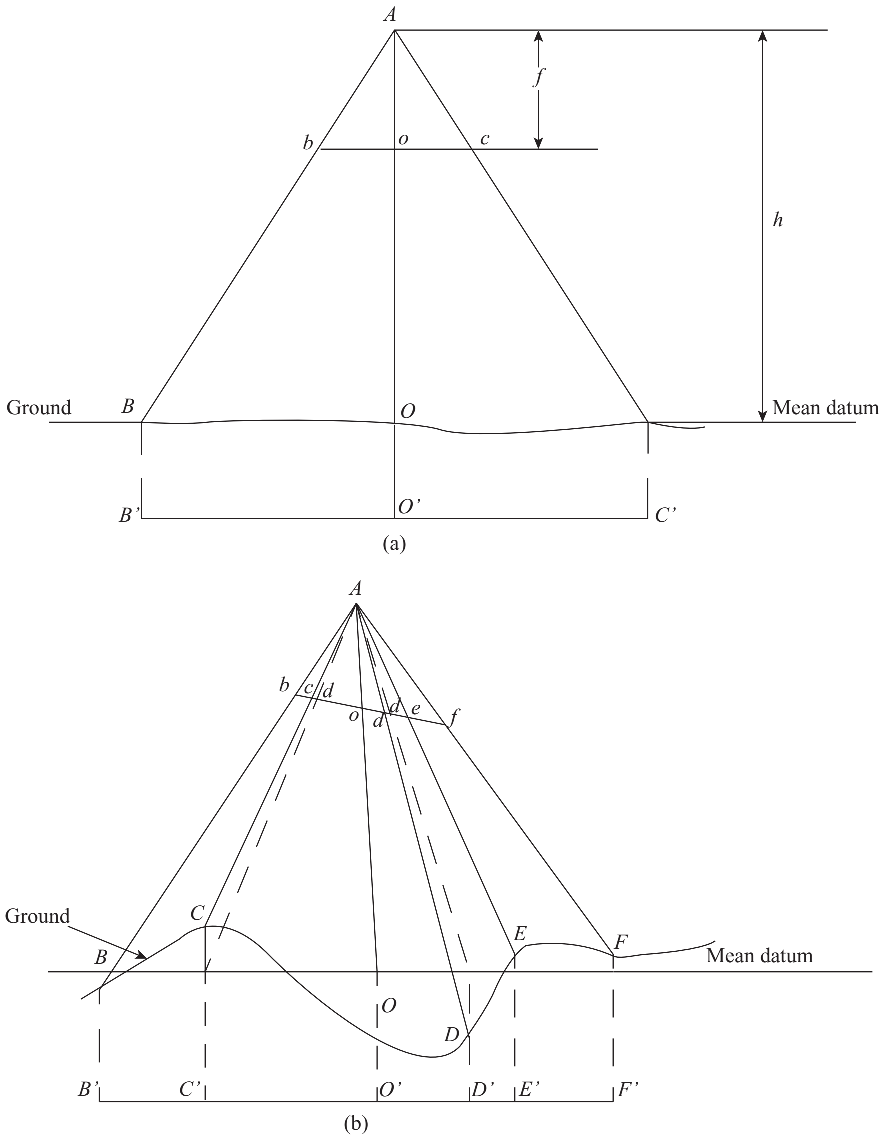


FIGURE 13.2 Scale difference caused by tilt and topography: (a) level ground and local focal plane and (b) tilted focal plane and undulated ground

Hence the photographic scale for points between *b* and *c* will be significantly different from that for points between *c* and *o*. The same variations in scale can also be seen for the points between *e* and *e*, and *f*. The overall average scale of the photograph is based partially on the elevations of the mean datum shown in Fig. 13.2(b).

The mean datum elevation is intended to be the average ground elevation. Examining the most accurate available contour maps of the area and selecting the apparent average elevation, this can be determined. Distances between points of the photograph that are situated at an elevation of the mean datum will be at the intended scale. Distances between points having elevations above or below the mean datum will be at a different photographic scale depending on the magnitude of the local relief.

The scale of a vertical photograph can be calculated from the focal length of the camera and the flying height above the mean datum. The flying height and altitude are different elevations. The altitude is the sum of the flying height and datum. That is,

Altitude = Flying height + Mean datum.

By the similarity of triangles shown in Fig. 13.2 (a) $\frac{bo}{Bo} = \frac{co}{CO} = \frac{f}{h}$

The ratio $\frac{BO}{bo}$ is the scale ratio between the ground and the photograph, f the focal length and h the flying height above the mean datum.

Hence, the scale ratio is SR = $\frac{h}{f}$

If h = 3,000 m and f = 150 mm, then the scale ratio SR = 3,000/0.150 = 20,000

Here the average scale of the photograph is 1:20,000.

13.7 PHOTO INTERPRETATION

Photo interpretation is the art of examining photographic images for the purpose of identifying objects for the judgement of their significance. This ability of good photo interpretation depends on good training and good experience of the interpreter. There are some aspects of photographic characteristics, which can assist in the adoption of a systematic approach to interpretation. Factors that are most relevant to photo interpretation are the shape, size, shadow, tone, texture, pattern, site, situation and association. Each factor may influence in the correct identification and classification of an object, but their relative importance will vary according to different factors.

Shape The outline configuration of an object helps in its identification. The plane view can be resolved from the photograph if the photographic scale is known.

Size The physical dimension of a feature may be a distinctive attribute and may assist where the shape is in conclusive. This requires knowledge of the nominal photograph scale, so that photograph measurements can be converted to the real ground size of a feature. Apparently this simple factor may be of crucial significance, for example, in trying to decide on the likely identity of animals without the aid of ancillary information.

Colour or Tone Features on a photograph are easily detectable due to the difference in tone or colour that results from the different spectral reflectance properties of the features of the topography. A sound knowledge of reflectance properties of different ground targets is an aid to interpretation. Sound knowledge of the variation in colour due to phenomena like ageing of plants, drought in a vegetation area and ill-affected vegetation enable the user to make more meaningful interpretation. Comparison of tones should generally be made only within a photograph or set of adjacent photographs, which are subjected to the same processing conditions, because variations in processing can affect the absolute tonal rendition of features on photographs. The idea is to have a set of dark and light targets of known ground reflectance included in each film sortie, to permit calibration of the resulting image tones.

Shadow When illumination conditions result in shadows on a photograph, the shadows provide a great help in defining objects such as bridges, towers and chimneys. When the photograph is taken in the morning or in the evening, a good shadow effect will be created due to sunlight.

Texture The rate of change of tone within the image is known as texture. Texture help in the identification of vegetation cover types. A large scale wooded area that has appearance of a course texture will become a smooth texture if the photograph is taken at a very high altitude. Hence, the scale of the photographs must be taken into account while assessing the texture.

Pattern Pattern is the spatial arrangement of a feature. It is a distinctive identifying characteristic. For example, the planting pattern of many crops is distinctive. The planting pattern of rice is different from a coconut plantation or an orange plantation. Regular geometric pattern are usually associated with human impact on the landscape. Rock pattern on the ground and folding pattern of the terrain are examples of natural pattern.

Site, Situation and Association The site is the piece of terrain occupied by the feature of interest. In the case of man-made features, the site will be carefully selected if such favoured sites can be of crucial importance in helping to identify a feature, especially historical or archaeological features. A three-dimensional image provided by stereo viewing is important in site recognition.

The situation of a feature refers to the wider setting within the photograph, or within the region depicted on the photograph. Knowledge of the geographic location of the photographed area acts as a filter on the interpreter's knowledge, screening out a large number of unlikely possibilities and concentrating attention on a smaller number of probabilities.

Association is the word used to describe the line of reasoning, whereby a feature may not be conclusively identified by itself. By encompassing a group of features associated within a particular function, more certain identification is possible when considering the association of a riverside site, railway lines, canal basin, coal heaps and a large building with smoke stacks conclude that there is a long-established heavy industrial plant, which was initially dependent on water for the processing of imported raw material and coal, which was initially brought in by canal and then transported by rail.

13.8 FLYING HEIGHTS AND ALTITUDE

While planning aerial photography, flying heights and altitude must be determined. Good planning is required for acquiring supplementary aerial photography using a small format camera. The flying height (h) is determined using the following relationship:

$$SR = \frac{h}{f}$$

Hence

$$h = SR \times f$$

If the desired scale ratio (SR) is 1:5000 and the focal length of the lens (f) is 150 mm, then $h = 5000 \times 0.150 = 750$ m

If the desired scale ratio (SR) is 1:10,000 and the focal length of the lens is 50 mm, then $h = 10{,}000 \times 0.050 = 500$ m.

The flying heights calculated are the vertical distances that the aircraft flies above the mean datum. Therefore, the altitude at which the aircraft must fly is calculated by adding the elevation of the mean datum to the flying height. If the elevation of the mean datum is 75 m, then the altitude in the above examples becomes 825 and 575 m, respectively. These are the readings for the aircraft altimeter maintained throughout the flight to achieve the desired average photographic scale. If the scale of the existing photograph is unknown, it can be determined by comparing a distance measured on the photograph with the corresponding distance measured on the ground or on a map of known scale. The points used for this comparison must easily be identifiable on both the photograph and the map, such as road intersections, building corners, and rivers or stream intersections. The photographic scale is found using the following relationship:

$$\text{Photo scale/map scale} = \text{photo distance/map distance}$$

As the above relationship is based on ratios, the scales on the left side must be expressed in the same units. The same applies to the measured distance on the right side of the equation. For example, if the distance between two identifiable points on the photograph is 16 cm and on the map it is 4 cm, and the map scale is 1:40,000, the photo scale is calculated as follows:

$$\frac{\text{Photo Scale}}{1:40,000} = \frac{16}{4}$$

$$\text{Hence photo scale} = \frac{16}{40,000 \times 4} = 1:10,000$$

13.9 MAPPING FROM AERIAL PHOTOGRAPHY

In aerial photogrammetry, a planimetric map of a ground area is prepared using aerial photographs. In ordinary topographic surveying, it is essential to establish a framework of ground control points before detailed surveys and mapping.

In aerial photogrammetry, ground control points are established and determined in the field. If a suitable and accurate map or plan of the area to be photographed is available in a suitable scale, it is used as a control base. If some good control points are given, then the measurements derived from almost any type of photography will have less number of sources of error.

The number of sources of error will be minimised and the task of transforming the positions of photo-images into planimetric map positions will be simplified if the following conditions are observed:

1. The photography is done with a metric camera.
2. The camera axis is kept vertical.
3. Successive photographs must overlap by about 60% along the flight direction.

In case of failure to satisfy the above ideal conditions, a near-vertical aerial photo image can have considerable plan errors. A map is an orthogonal projection of a portion of the surface of the earth on to a horizontal datum. But the aerial photograph is a perspective projection of the earth's surface. Hence the photo images of ground features suffer displacements whenever there is any ground

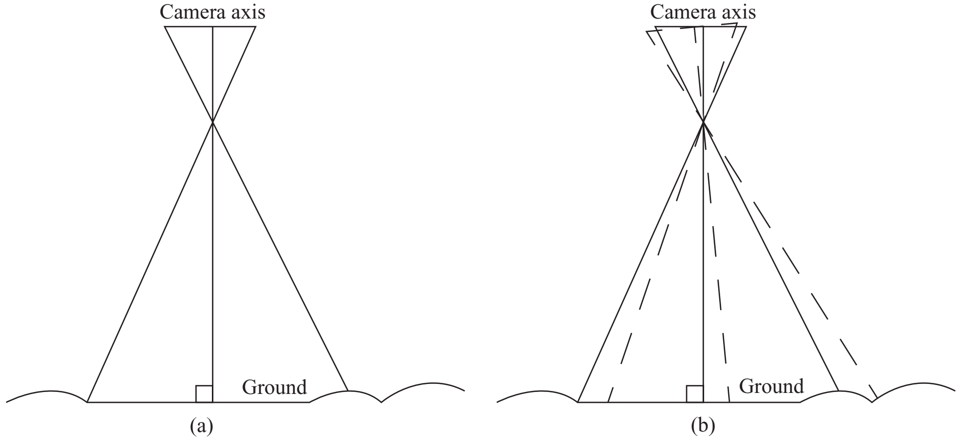


FIGURE 13.3 (a) Camera axis perpendicular to ground surface (true vertical). (b) Camera axis near vertical (deviates from vertical for a small amount

relief in the photographed area and whenever the aircraft movement causes the camera axis to be tilted away from a truly vertical position.

The photo images suffer no displacement from these effects only in rare situations when perfect vertical axis photography of flat surface is obtained. Hence all aerial photography contains some tilt and relief displacement errors (see Fig. 13.3(a) and (b)). Hence corrections for image displacement errors have to be given for the preparation of planimetric maps from aerial photographs.

13.10 RELIEF DISPLACEMENT (RADIAL DISPLACEMENT)

Relief displacement occurs when the point being photographed is not at the elevation of the mean datum. As the aerial photograph is a central projection, all elevations and depressions will have their images displaced from their original positions on the ground except the object at the nadir point or the principal point in vertical photographs. Relief displacement is the distance between the position on the photograph if it were on the reference plane and its actual position due to relief (see Fig. 13.4).

A vertical feature in the ground AA' will appear as a displaced linear image on the aerial photograph aa'. More precisely, any point on the ground that has location and elevation different from the ground plumb point C will show an image displacement on the photograph. From Fig. 13.4, by the properties of similar triangles.

$$\frac{bb'}{cb'} = \frac{BB'}{CB'} = \frac{BB'}{(H-h)}$$

Now bb' = de (the image displacement) and $cb' = r$ (the radial distance to the top of the image from the photograph centre).

Hence

$$\frac{de}{r} = \frac{\Delta h}{(H-h)'} \tag{13.1}$$

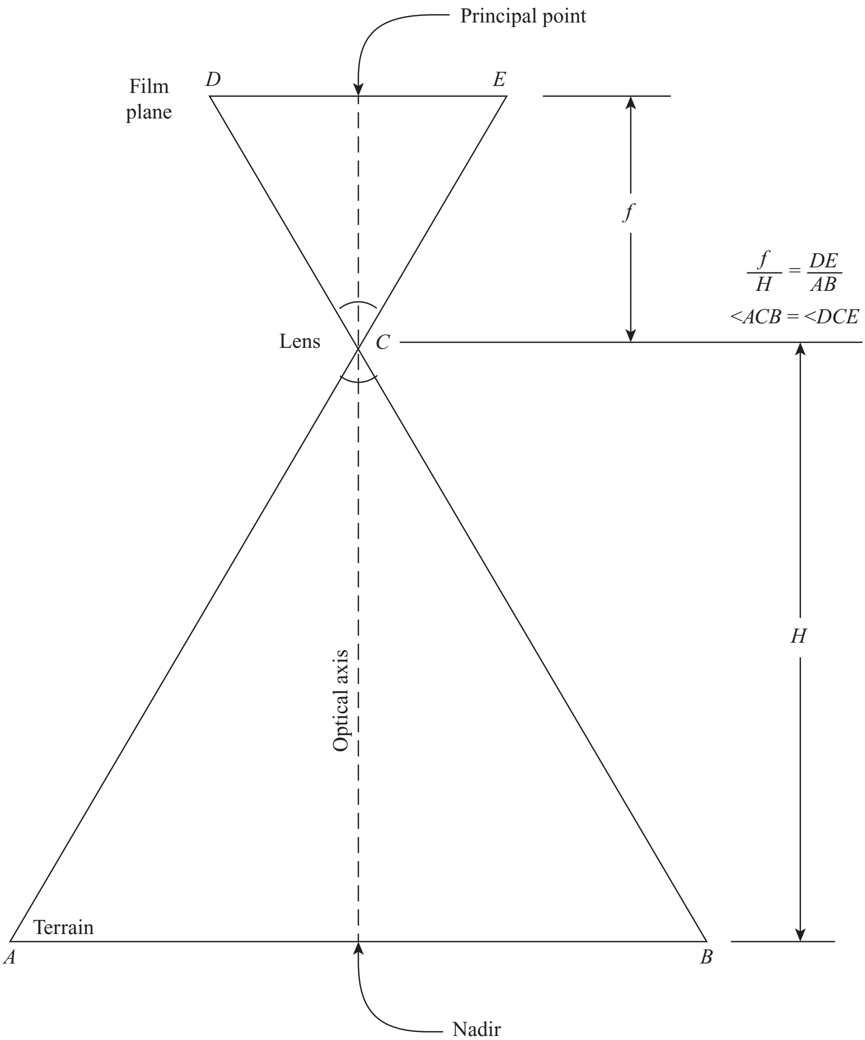


FIGURE 13.4 A diagram showing the geometry of a vertical photograph. The terrain distance *AB* subtends an angle at the camera lens, which is equal to that subtended by the image distance *DE*. this relationship remains constant regardless of focal length of cameral (f) and flight height (h)

Which can be written as,

$$de = \frac{r\Delta h}{(H - h)}$$

Hence

$$\Delta h = \frac{de(H - h)}{r} \tag{13.2}$$

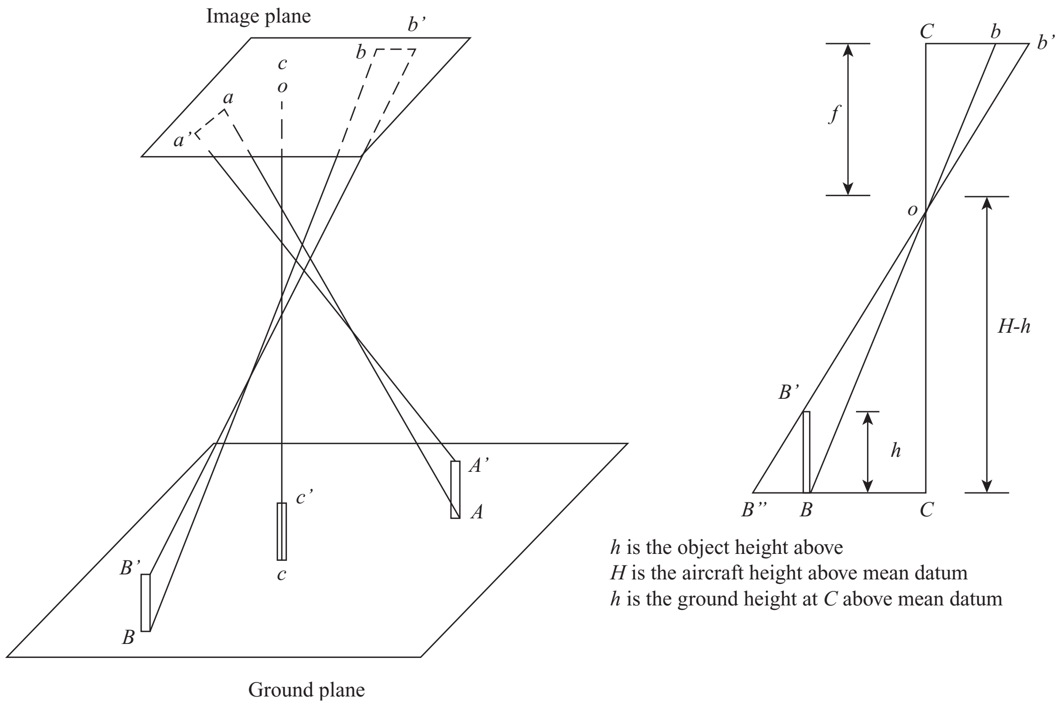


FIGURE 13.5 Linear image displacements on a photograph due to relief

From (13.1), it can be seen that the image displacement due to relief is directly proportional to r (the radial distance to the top from the centre of the photograph) and the ratio of the height of the feature above ground to the ground clearance of the aircraft, which is $\dfrac{\Delta h}{(H-h)}$. Also, the magnitude of image displacement can also be calculated using the formula, $de = \dfrac{r\Delta h}{(H-h)}$.

On aerial photography of urban areas where there is ground relief of buildings, the effect is very noticeable. On aerial photography of upland areas, however, there are seldom any simple vertical structures to draw the attention to this effect. In order to achieve good registration of continued lines and boundaries, it is necessary to compensate for the effect of the relief displacement. Nevertheless, relief displacements occur at any point that is higher or lower than ground point vertically below the aircraft.

The point on the photograph at which radial relief displacements originate is the nadir point (see Fig. 13.5).

13.11 TILT DISPLACEMENTS

All vertical aerial photographs are slightly tilted due to the deviation of the camera axis from its true vertical position. The tilt displacement results in displacement of images and scale variation on aerial photographs.

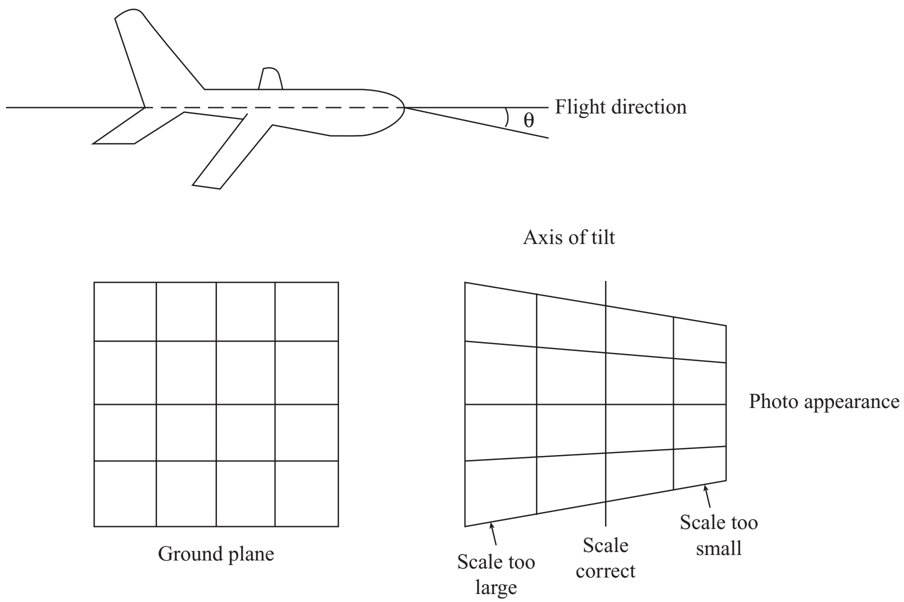


FIGURE 13.6 Effect of tilt on photo image scale

Tilt will tend to be of a known magnitude and will act in a random direction. Tilt displacement of a point is radial from the isocentre, and the magnitude dt of the displacement is given by

$$\mathrm{d}t = \frac{r^2 \sin^2 \theta}{f},$$

Where r is the radial distance on the photograph from the isocentre, θ is the angle of tilt and f is the focal length of the lens.

The most severe consequences of tilt displacements are the scale changes that are produced across the photograph (see Fig. 13.6). These effects may be appreciated most clearly by considering the appearance of some regular features, such as a grid, on a titled photograph. The tilt may also be considered to be acting only along the direction of flight and only at right angles to the flight.

13.12 CORRECTION OF RELIEF AND TILT

While preparing for a planimetric map, the correction for relief displacement and tilt displacement has to be given. The point of zero displacement will always be the nadir point for relief displacements and the isocentre for tilt displacement. Hence, in the correction for the above displacements, the origin should be referred to as the nadir or isocentre on an aerial photograph. Hence, in practice, the principal point that is readily located by fiducial marks is assumed to be the origin of these radial displacements. This is termed as radial line assumption and is reasonably correct for tilt less than 3 degrees and for ground relief less than 10% of the flying height.

13.13 FLIGHT PLANNING

Good planning is required in aerial photogrammetry. Various stages in the production of aerial photographs include proper planning of aerial photography, photography during flight and processing the negative for the production of positive prints. For planning aerial photography, the required parameters are:

1. Purpose of photography
2. Nature of the area to be photographed
3. Scale of photography
4. Flight direction
5. Time of photography
6. Nature of the atmosphere (season of photography)

The scale and type of photography shall depend on the purpose for which it is required, such as large-scale topographic or small-scale mapping, or for detailed photo interpretation for geological, geo-technical or forestrial investigations. The extent, layout and topography depend upon the area to be photographed. The selection of vertical, oblique or convergent, or normally vertical type of photography depends on the nature of data required for photographic interpretations and the area to be covered. The scale of photography depends on:

1. The type of details to be identified.
2. Planimetric accuracy desired.
3. Altimetry accuracy desired.

The usual flight direction will usually be in the east-west direction, unless it is changed due to some special consideration like the layout area of photography. The time of photography should normally be confined to the period when the sun's elevation is more than 30 degrees. The best season suited for aerial photography is the season that gives a cloud-free, fog-free and dust-free environment.

13.14 PLANNING FLIGHT LINES AND LAYOUT OF PHOTOGRAPHY

Once the photographic scale, flying height and latitude are calculated, the details of implementing the mission are carefully planned. The most significant factors in this planning process include:

1. Selecting a suitable aircraft and technical personnel.
2. The initial reconnaissance survey of the area to be photographed is used for planning the flight lines and for calculating the flight line bearing.
3. Observing the weather conditions.

To achieve photogrammetric mapping and to examine the terrain for air photo interpretation purposes, it is essential that each point on the ground must appear in at least two adjacent photographs along a line of flight, so that all points can be viewed stereoscopically.

Figure 13.7 illustrates the relative location of the flight lines and photograph overlaps, both along the flight line and adjacent flight lines. An area over which it is decided to acquire an air photograph is called a block. The block is outlined on the most accurate available photographic map.

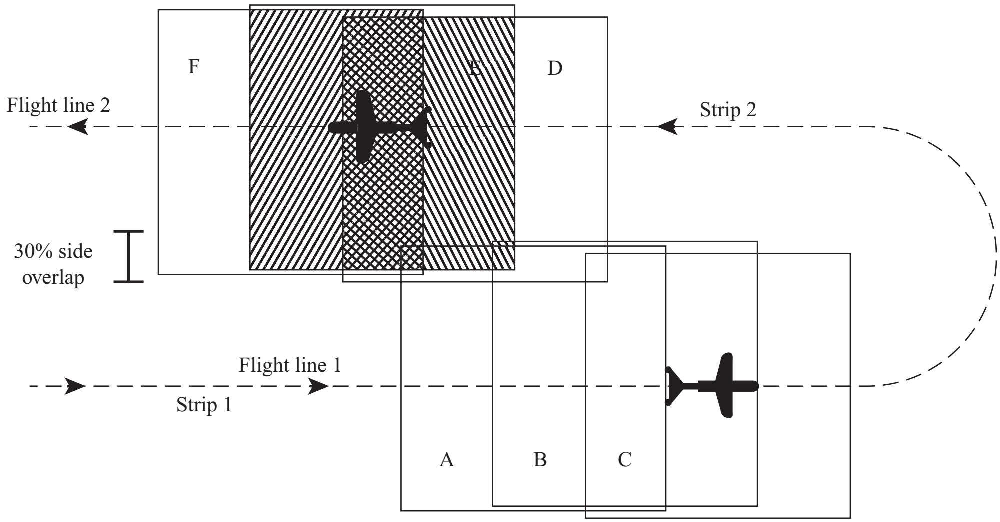


FIGURE 13.7 Overlap requirement for a good photographic coverage

The location of the flight lines required to cover the area properly are then plotted, such as flight line 1 and 2. The aircraft proceeds along the flight line 1 and takes air photos at predetermined time intervals to provide 60% forward overlap between adjacent photographs. The minimum overlap to ensure that all ground points show on two adjacent photographs is 50%. However, at least 60% forward overlap is standardized as the aircraft is subjected to altitude variations, tip and tilt as the flight proceeds.

The extra 10% forward overlap is given an allowance for these contingencies. The air photo coverage of two consecutive flight lines must have an overlap of at least 30%. Here also, 25% side overlap is standardized, but giving a 5% overlap as an allowance for altitude variations, tip and tilt and tilt of the aircraft, 30% side overlap is usually provided. This side lap ensures that the photographs will not have gaps of non-photographed ground, and it also extends control between flight lines for photogrammetric methods.

The following points highlight the reasons for keeping overlap in aerial photography:

1. Due to the distortion caused by tilt and relief displacements and due to the distortions caused by lens, the displacements are more pronounced in the outer part of the photograph than near the centre of each photograph. Hence, if an overlap of 60% in the forward direction of photography is provided, these distortions and displacements can be nullified.
2. To connect different control points together very precisely, it is desirable that the principal point of each photo should appear on the edges of as many adjacent strips (photos) as possible.
3. For the purpose of viewing the photographs stereoscopically, only the overlapped portion is useful. The object can be viewed from more than one angle during stereoscopic examination, if sufficient overlap is provided.
4. By giving sufficient overlap, each portion of the topographical area is photographed 3–4 times. Hence any pictures having maximum distortion due to excessive tilt can be rejected without the necessity of a new photograph.

Problems arising due to relief displacement during photo compilation are solved partially by increasing the forward overlap to as high as 80%. Although this results in roughly twice the total number of photographs at a similar scale, a smaller portion of the central portion of each photo, which is least affected by relief displacement, can be more effectively used to assemble the mosaic.

13.15 COVERAGE OF THE PHOTOGRAPH

The number of air photos required to cover an area is important. The approximate number of air photos required to cover a given area stereoscopically (i.e., every ground point) is shown on at least two adjacent photos along the flight line (see Fig. 13.8) can be calculated easily. The amount of forward overlap and side overlap to be used in flight planning depends upon the effective coverage of each photograph. The effective coverage of each photograph depends on the size of the format or the focal plane opening, the focal length and angular coverage of lens. The effective angular coverage of the lens with a 30-cm focal length is represented by a cone of apex, which lies at the front nodal point, and the apex angle of which is about 60 degrees. In general, the effective coverage with a 30-cm lens will embrace more than a 23 cm × 23 cm format size, and hence the entire photograph is useable.

The number of photographs required is calculated by dividing the total area to be photographed by the new area covered by a single photograph considering the overlaps. The amount of overlap and side lap to be used in flight planning depends upon the effective coverage of each photograph.

The relation between the separation of flight lines and the separation between photographs must be arranged to give the greatest area of each stereo pair. The effective coverage of the photograph depends on the size of the format or the focal plane opening, the focal length (f) and the angular coverage of the lens. A cone, the apex of which lies at the front nodal and point and the apex angle of which is about 60 degrees, represents the effective angular coverage of the lens with a 30-cm focal length (see Fig. 13.9).

The number of photographs required is calculated by dividing the total area to be photographed by the area covered by a single photograph.

Let A be the area to be photographed and l the length of the photographic distance of flight, w the photograph normal to the direction of the flight and s the scale factor.

The scale of the photograph $s = \dfrac{H}{f}$

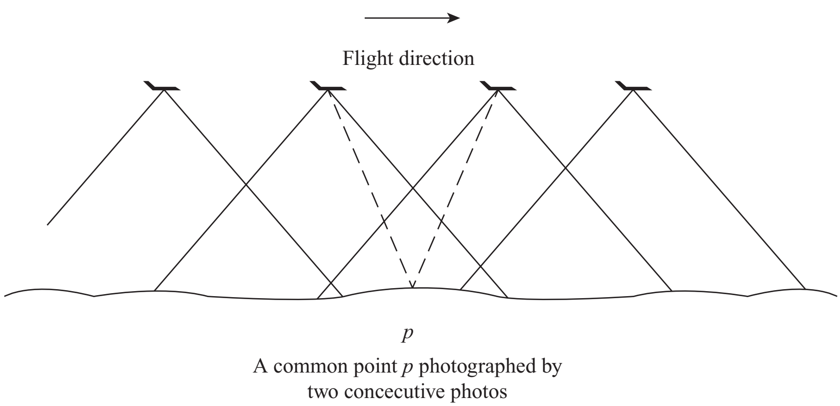


FIGURE 13.8 Overlap along a flight line

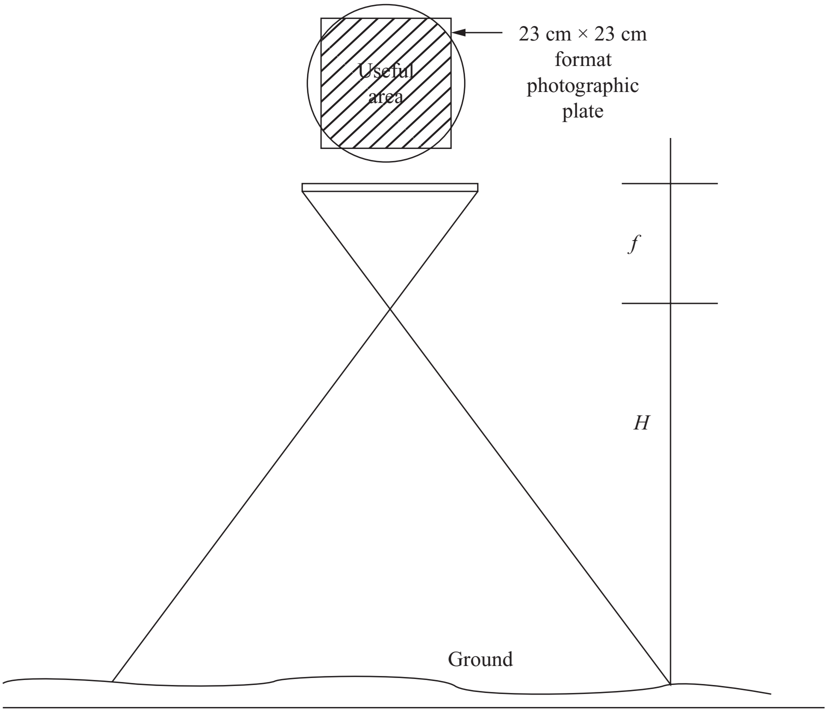


FIGURE 13.9 Angular coverage

In the above equation, H is measured in metres and f is measured in centimetres.

Let L be the net ground distance corresponding to l, and W the net ground distance corresponding to w.

Therefore, the net ground area covered by each photograph = $L \times W$.

Let Pl be the percentage overlap between successive photographs in the direction of flight expressed as a radio, and Pw be the side lap expressed as the ratio, as each photograph has a longitudinal lap of Pl,, the actual ground (L) covered by each photograph is given by:

$$L = (1 - Pl)\ sl \tag{13.3}$$

Similarly, the actual ground width (W) covered by each photograph is given by

$$W = (1 - Pw)\ sw \tag{13.4}$$

Hence the net ground area covered by each photograph is given by

$$a = L \cdot W = (1 - P_l)\ s \cdot l \cdot (1 - P_w) s \cdot w = l \cdot w \cdot s^2\ (1 - P_l)\ (1 - P_w) \tag{13.5}$$

The number of photographs (N) required is given by

$$N = \frac{A}{a} \tag{13.6}$$

If the rectangular dimensions of the ground are given instead of the total area, then the number of photographs required is computed by calculating the number of strips and the number of photographs required in each strip, and multiplying the two factors.

Let L_1 be the dimension of the area parallel to the direction of the flight, L be the dimension the area normal to the direction of the flight, N_1 the number of the photographs in each strip, N_2 the number of strips required and N the total number of photographs to cover the whole area. Now the net length covered by each photograph is calculated by the formula:

$$L = (1 - P_1) \cdot s \cdot l$$

Hence the number of photographs in each strip is calculated by

$$N_1 = \frac{L_1}{(1 - P_l) \cdot s \cdot l} + 1$$

Similarly, the net width covered by each photograph is calculated by

$$W = (1 - P_w) \cdot s \cdot w$$

Hence the number of strip required is given by

$$N_2 = \frac{L_2}{(1 - P_w) \cdot s \cdot w} + 1$$

Thus, the number of photographs required is $N = N_1 \times N_2$; that is,

$$N = \left\{ \frac{L_1}{(1 - P_l) \cdot s \cdot l} + 1 \right\} \times \left\{ \frac{L_2}{(1 - P_w) \cdot s \cdot w} + 1 \right\}$$

In addition, the time interval between exposure (T) is calculated by the following formula:

$$T = \frac{3{,}600L}{V}$$

where V is the speed of the aircraft in kilometre/hour, L the ground covered by the each photograph in the direction of the flight, which is equal to $L = (1 - P_1) \cdot s \cdot l \cdot$ in kilometre.

13.16 GROUND CONTROL FOR MAPPING

An aerial photograph is not perfectly level at the instant of exposure and the ground surface is not always flat. As a result, ground control points are required to manipulate the air photos physically or mathematically, before mapping can be done. There are two methods of adopting ground control points. The first one is establishing control points where existing photography is to be used for the mapping. In the second method, ground control points are established prior to aerial photography.

Ground control is required for each data point positioning. The accuracy with which the measurements must be made varies in each case, depending on the following requirements:

1. Preparation of topographic maps.
2. Measurement of distances and elevation of structures such as buildings, dams, etc.
3. For controlled mosaics.
4. For the construction of orthophotos.

Kinematic Global Positioning System (GPS) surveying methods can be adopted effectively for establishing control points on the ground. Ground control can be classified as targeted and photo-identifiable control points, and can also be classified as horizontal control, vertical, or as 3D control. Horizontal and vertical controls required different configurations to make them serve their intended purposes. The surveyor needs to know what type of control is called for when he or she attempts to pick or photo-identify the point. Accessibility for surveying should also be considered when selecting the locations for control points.

Targeting operations are an essential part of photogrammetric mapping to be considered prior to establishing a control survey. Preflight targeting is performed to make ground locations of control points visible on the photographs. Easy identification and clear image of the control points on the photographs increases the accuracy and efficiency of the photogrammetric process. Highway design mapping often required careful preflight planning for optimal target placement. To reduce the possibility of pre-marked points being moved or prior to the aerial mission, it is important to either paint them on a hard surface or schedule the field panelling operation as close as possible to the anticipated flight. Targets should be located where shadows will not adversely affect the visibility of the panel.

13.16.1 Number of Photographs

The number of photographs needed to cover an area will be required not only to work out the cost involved, but also to estimate the amount of film required, and the points at which the film magazines should be changed. If possible, the magazines should be changed in the turns.

Consider an area of 200 km × 100 km to be flown at an average scale of 1/10,000. At this scale, the area is 20 m × 10 m. The photography has a format size of 230 mm × 230 mm of which 60% is overlapped.

Thus, the new ground covered at this scale is 40% of 230 mm = 92 mm.

Therefore, the number of photographs per strip = 20,000 ÷ 92 mm = 218 + 2 at each end to ensure complete coverage.

Hence the total number of photographs = 218 + 4 = 222.

The number of strips, assuming a 30% lateral overlap = 10,000 ÷ (70% of 230 mm) = 63 strips.

Therefore, the total number of photographs = 222 × 63 = 13,986.

13.16.2 Interpretation of Photos

Photo-interpretation refers to the accurate identification of the feature seen in photographs. Object seen in the photographs are often not easy to recognize, and it may take good skill on the part of the interpreter to identify the object and judge their significance. It is difficult to identify object in vertical photographs than in tilted photographs owing to the familiarity of view in oblique photographs. Colour photographs are easier to interpret than black and white photographs due to tonal variations. A stereoscopic pair of photographs is easier to interpret due to the depth available in the photographs when seen through stereoscope.

13.17 MOSAICS

A mosaic is an assembly of two or more air photos to form one continuous picture of the terrain. An aerial mosaic is an arrangement of aerial photographs, the edges of which have been cut and matched to form a continuous photographic representation of a portion of the topography.

After aerial photography, the photographic film is processed and each negative of a flight line is numbered consecutively. The flight number, along with the roll number of the negative, the time, the serial number of each photo and the day in which the photograph was taken will be noted. This information is usually shown in one corner of the photograph print, and is useful to construct the mosaic by arranging the photographs in order and matching the terrain features shown on each, so that all numbered information is visible. These mosaics, termed as index mosaics, are useful for the identification of photographs required to cover a particular area.

Based on the method of compilation, mosaics are classified into:

1. Controlled mosaics.
2. Semi-controlled mosaic.
3. Uncontrolled mosaics.

Controlled mosaics are the most accurate form of mosaics. The photographs are rectified and scaled to remove the errors due to tilt displacement and scale differences. The errors due to relief displacement are restricted within the tolerable limits by the careful selection of the scale of aerial photographs. Proper scaling can be done with the help of sufficient control points from each photograph. The controlled mosaic is compiled according to good accuracy level requirement. In a controlled mosaic, the superimposing of a coordinate system and giving a graphical scale are fully justified.

The semi-controlled mosaic is a compilation of photographs without rectified photographs or without utilizing control for the position of each photograph. When the arrangements for rectifications are not available and for the reasons of speed or when the proper number of required control points is not available, semi-controlled mosaics are carried out. This is done with the help of less number of control points, or even with azimuth line or series of lines to present the flight strip from swinging off line as the photographs are mounted. The semi-controlled mosaics are used where the terrain under study is not flat and higher accuracy is required. This means that by some means or other, the relative position of the principal point of adjacent photographs have to be determined.

Uncontrolled mosaic is a compilation of photographs, without regard to any horizontal control positions. The photographs are oriented in position with respect to the corresponding images on the adjacent photographs. This method is used where less accuracy is required. If an uncontrolled mosaic is to be used for graphical presentation purposes or as base for mapping terrain information, it is necessary to feather the photograph to avoid shadow along the edge of the overlapping air photos as well as to improve the appearance of the mosaic. Feathering is accomplished by cutting through the emulsion with a razor-edge knife and pilling the outside of the photograph towards the photo centre; thus, living only the thin emulsion where the photograph joins. The overlapping photograph edges are matched to the terrain features on both the adjoining photos as accurately as possible. The best means of permanently attaching the adjacent photos is using a hot roller to apply a special adhesive wax to the underside of the overlapping photograph. The joints between the photos are then taped securely on the back of the masking tape.

Mosaics have a lot of advantages over maps prepared by ground survey maps.

Advantages of Mosaics

1. Mosaics can be produced more rapidly than a map, because the time required to carry out ground surveys for the production of a map is time consuming.
2. A mosaic shows a wealth of detail which a map can never equal.
3. Mosaics are less expensive, even if the costs of acquiring the air photos are included.
4. A mosaic can be combined more rapidly and economically than a map.
5. Certain terrain features can be more easily recognized and interpreted on a mosaic.
6. Details of the terrain are not omitted on a mosaic, whereas a map represents only selected items in detail.
7. For air photo interpretation purposes, subtle terrain characteristics such as tone, texture and vegetation must be visible. Hence the use of mosaic for these purposes is essential.

Limitations of Mosaics

1. Mosaics are not topographical maps and therefore they do not show elevations.
2. Lake of name of places and features.
3. The accuracy of the horizontal scale between two points on a mosaic is limited due to relief displacement.
4. The wealth of terrain detail turns out to be a disadvantage due to excessive detail and obscuring of some important features.

13.18 STEREOSCOPY

The human ability to see three dimensions is the result of having two eyes that record slightly different images of an object around him. These slightly different images are fused by the brain to give the impression of the third dimension, that is, depth. The prime motive of aerial photography is that it must be possible to view an optical model of the terrain in three dimensions. This is made possible if the geometrical condition is similar to natural binocular vision. At successive exposures, the aerial camera positions are analogous to the separation of the eyes in normal vision (see Fig. 13.10).

If the photographic sequence is arranged so that a portion of terrain between the two camera stations is recorded successive exposure, two slightly different views of the area of terrain, called the overlap, are recorded on the photographic film. If the pair of photographs are viewed under suitable

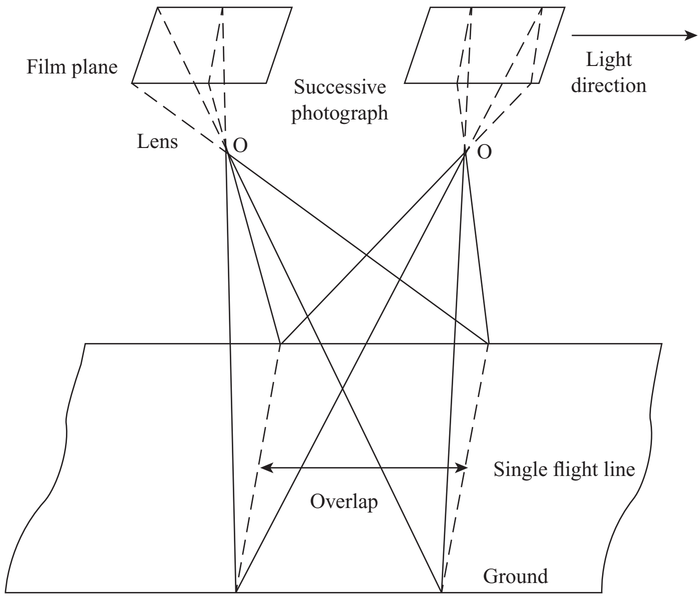


FIGURE 13.10 Conditions for aerial photography for stereoscopy

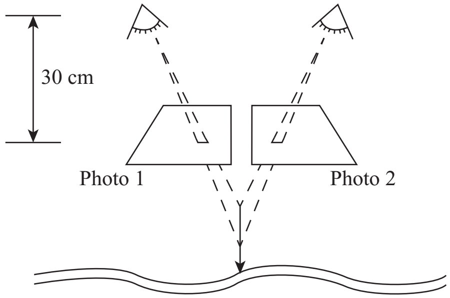


FIGURE 13.11 Stereo viewing principle

conditions, an optical modal of the area of overlap will be seen in relief. The condition to be satisfied is for each eye to have a slightly different view of the same area of terrain. This can be achieved if the left eye views only the left photograph and the right eye views only the right photograph (see Figs. 13.11 and 13.12).

The dimension of height not only enhances the interpretation of many features, but also allows the possibility of measuring heights within the stereoscopic model using only a minimal amount of ground derived information normally four height points for one overlap.

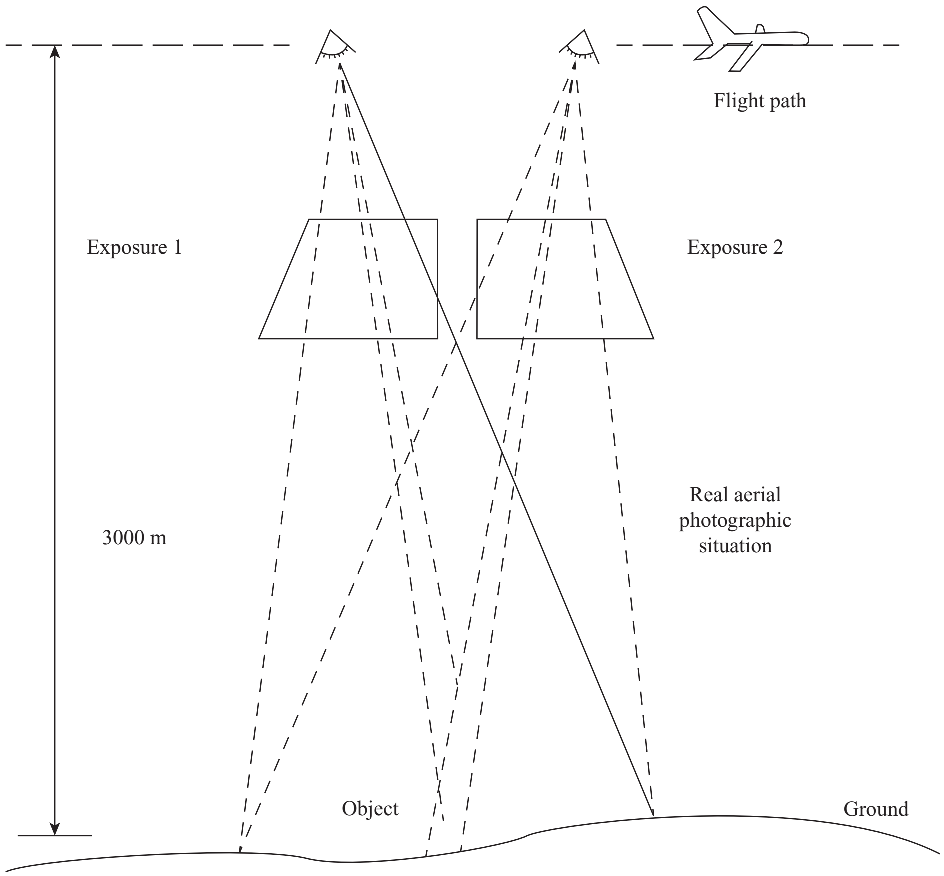


FIGURE 13.12 Visualizing the concept of stereo viewing in a real situation

13.19 LENS STEREOSCOPE AND MIRROR STEREOSCOPE

The simple and least expensive equipment for viewing the 3D image from aerial photography is the lens stereoscope. It consists of a pair of lenses mounted on a frame, which is supported over the photographs by a thin metal leg (see Fig. 13.13).

The lens separation many be fixed for the average eye separation, but more usually the separation may be varied to suit the eye base of the user. When the lens stereoscope is set up, the distance from the lens to the photograph will be equal to the focal length of the lenses, so that the photographic image will appear sharp in the focal plane of the lenses.

The user has to arrange the photographs below the lenses, so that the appropriate eye is viewing the same detail in the appropriate photograph at the same time.

It is necessary to place one photograph on top of the other, so that the separation of corresponding images is about the same as the lens separation. The limitation of a lens stereoscope is that only a limited portion of the photograph can be viewed at a time and the photographs have to be readjusted often if the whole overlap area is to be viewed stereoscopically.

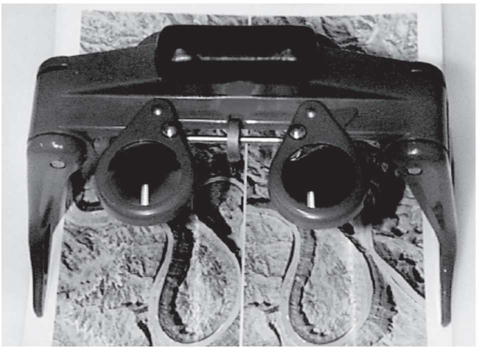


FIGURE 13.13 A lens stereoscope placed over a pair of images

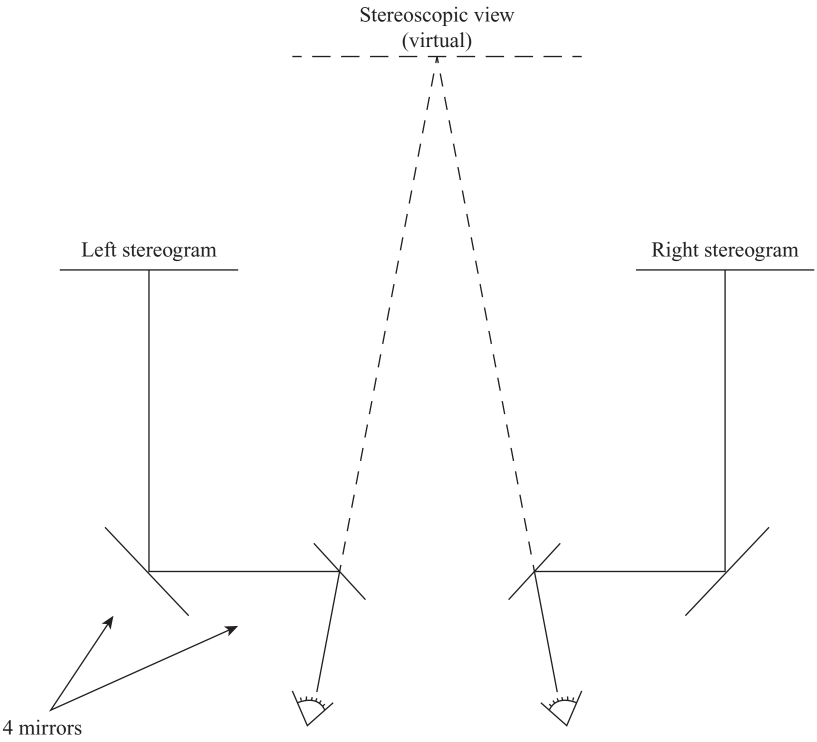


FIGURE 13.14 Schematic diagram of a mirror stereoscope

Almost all limitations of the lens stereoscope are overcome in the mirror stereoscope. This stereoscope consists of two pairs of parallel mirrors, one large and one small mirror in each pair arranged, so that the outer larger mirrors are inclined at 45 degrees to the horizontal, one over each photograph (see Figs. 13.14 and 13.15).

The image on each photograph is reflected from the mirror and that image is directed to the small mirror. The resultant image, which appears as two separate images, is seen through the lenses

FIGURE 13.15 A mirror stereoscope

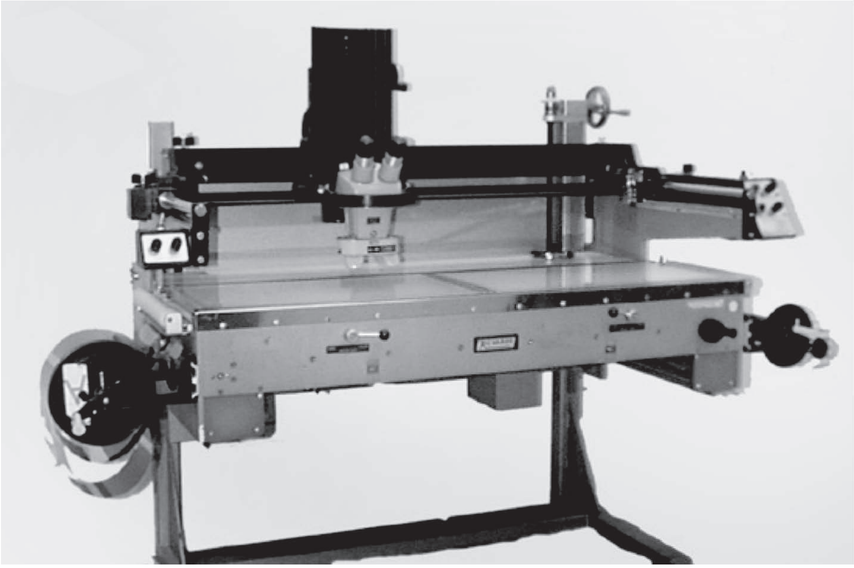


FIGURE 13.16 A stereoscope work station (zoom stereoscope)

placed vertically over the same mirror. The small viewing is seen lenses can be replaced by a binocular attachment to permit magnification of the images. A part from the increased magnification, the mirror stereoscope has an advantage that the optical paths from each photograph are physically separated. For most accurate large scale viewing, a zoom stereoscope is used (see Fig. 13.16).

13.20 PARALLAX

Parallax is the displacement along the flight line of the point on adjacent aerial photographs. Once it is correctly set up for viewing, the photographs can be taped to the table, and lifting the mirror stereoscope and readjusting it over the required section of the overlap achieve viewing of different portions

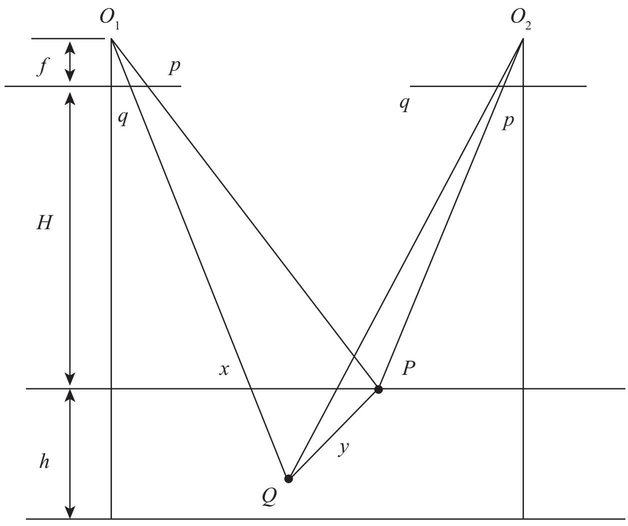


FIGURE 13.17 Parallax

of the overlap. In one design of the mirror stereoscope, the scanning of the overlap is achieved by moving the pair of photographs together, on a moveable base, below the large mirrors.

The parallax of a point can be measured from photographs, as the total movement of the image of the point between two exposure situations. For this two consecutive aerial photographs are taken, and the principal points of both exposures are marked in the two photographs (O_1 O_2). When joined, this will give the line of flight between the two exposures. If (x_1, y_1) and (x_2, y_2) are the coordinates in the two photographs, then the parallax P is given by $P = x_1 - x_2$. Please see Fig. 13.17.

If X and Y are the coordinates from plumb point,

$X = Bx_1/P, Y = By_1/p$ and $H - h = Bf/p,$ where B is the camera baseline length.

13.20.1 Parallax Bar and Measurement of Parallax

The parallax between points is measured using a parallax bar or a stereoscopic viewer. A parallax bar (see Fig 13.18) consists of a rigid bar with two glass reticules at its ends, each having an index point. One of them can be adjusted by moving it along the rigid bar and can be clamped in any position (fixed part). The other reticule can be moved very little using a micrometer screw. The glass reticules are provided with or all of the three types of stereo-marks, namely a dot,a circle or cross. The fixed part is provided with graduation in millimetres. The drum attached with the fixed parts of the instrument is provided with micrometre divisions having a lease count of 0.01 mm. The fixed part carries an inner rod on which the movable parts slides. The two reticule plates are inserted in metallic mount of movable and fixed parts. When two photographs have been adjusted for stereoscopic viewing, the photographs are fixed in position and the parallax bar is placed over them, with the glass reticules over the photographs. The index points are fused during viewing. The dots when fused for viewing can be placed over any point at its elevation, by moving the micrometre screw. Similarly, same procedure is followed with a different point, recording the micrometre reading in each case. The difference in micrometre reading gives the parallax difference between the points.

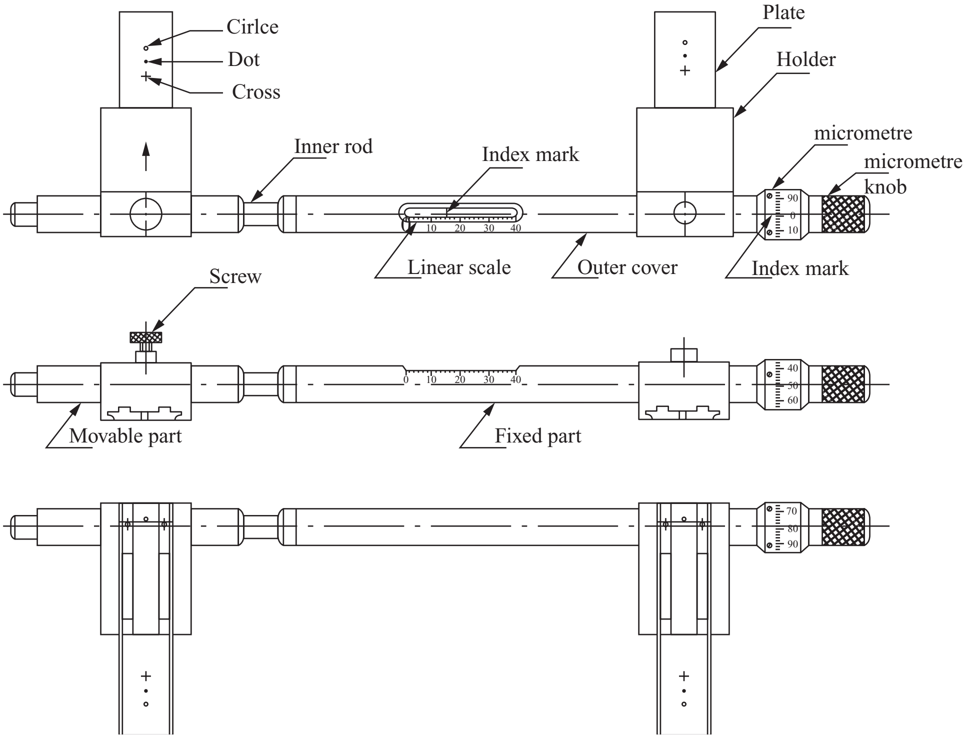


FIGURE 13.18 Parallax bar

Or in other words, take two consecutive aerial photographs, the photographs are adjusted under the stereoscope, and then parallax bar is so placed on the photographs that the reticule dots fuse together foaming a floating mark, which appears to move vertically or float by adjusting the micrometre. The floating mark can be adjusted at the level of a selected point say a_1 and the micrometre reading is noted. This reading will show the parallax at that point. Same procedure of floating mark is repeated for another selected point say a_2 and the micrometre reading is noted. The difference of reading taken at the points a_1 and a_2 will directly give the parallax between the points.

PROBLEM 1 Find the parallax difference between the top and bottom of a building having 30 m height, given that the mean base length in the photographs are 98.5 mm, the flying height during exposure is 1,500 m, and the focal length of the camera lens is 250 mm.

Solution:

Scale of the photograph = f/H = 0.250/1,500 = 1/6,000

Mean base length of photograph, b = 98.5 mm

Actual base length $B = bH/f = 98.5 \times 1500/250 = 591$ m

Parallax is calculated using $Bf/(H - h)$

Parallax at bottom = $591 \times 250/1{,}500 = 98.5$ mm

Parallax at top = 591 × 250/1500 – 30) = 100.51 mm.

Difference in parallax = 2.01 mm

13.21 AERIAL TRIANGULATION

Map-making basically aims at portraying the physically and man-made features in their correct relative position, both horizontally and vertically. This is done on an appropriate scale. The survey consists of providing a number of control points, and once a network of control points is available, the other details are surveyed to fit within the control points. Aerial triangulation, in its simple form, is making use of the elementary property of a triangle being fully determinable with one side and two angles from a known triangle, on a vertical photograph, for providing the control points. The aerial triangulation is worked out using radial triangulation.

13.22 RADIAL TRIANGULATION

The basic principle of the radial line method is that from a truly vertical photograph, the angles measured at the principal point in the plane of the photograph to the photographic images are horizontal angles to the corresponding points on the ground. Hence by using this principle, the aerial photographs can be used for measuring the horizontal angles and distances.

Extension of planimetric control between known control points is carried out by radial triangulation using the radial line principle described above. A minimum of nine control points is necessary for radial triangulation, and these control points have to be distributed as shown in Fig. 13.19.

Control points so obtained are called minor control points or pass points. Radial triangulation is a method of supplementing planimetric control. Radial triangulation can be carried out by a simple mechanical method such as the slotted template method or by using a simple instrument like the radial line plotter.

13.20.1 The Slotted Template Method

The slotted template method is used for the correction of relief and tilt displacement if relatively small number of points is concerned. The location of the principal point, the transferred principal and the minor control points are pricked through each photograph on to a sheet of stable plastic material (X-ray film), preferably a polyester-based or an aluminium-based one, and above the same size of the photograph. By means of a special punch, a circular hole is punched through the principal point on

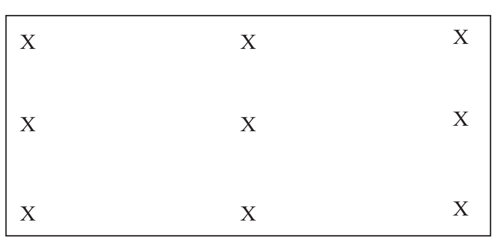


FIGURE 13.19 Control points for radial triangulation

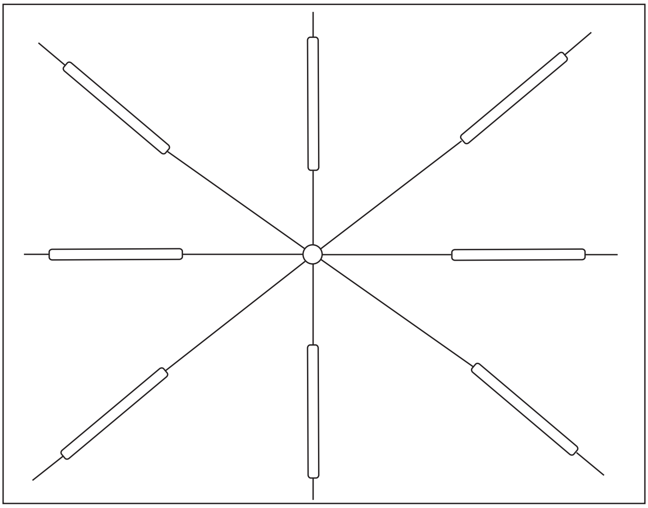


FIGURE 13.20 A slotted template

the stable plastic template, and with the help of a special slotted template cutter, the radial rays from the principal point through the other points are replaced by a precisely cut shot about 5 mm wide and 2–6 cm long as in Fig. 13.20.

After the preparation of a series of these templates for a strip of photographs, the templates are laid down one by one on a base sheet of a stable plastic drawing material that is marked with the known drawing points, which are fixed by pins to a base board. All slots (rays) that pass through the same point are held together by a stud, and for a correct lay down of the whole assembly, the studs for the control points should be fixed over their respective pins on the base sheet. The slots allow considerable free play in the movement of the whole assembly, so that on the expansion or contraction of the template assembly, the control point studs can eventually be located over their pins on the base board. When the control point pins have had the appropriate studs placed over them, the other studs represent the adjusted positions of the unknown points, and can be pricked through on the base board.

The slotted template method for extending planimetric control has been widely used by mapping agencies because of its simplicity and the low cost involved.

13.20.2 Radial Line Plotter

With the aid of an instrument known as a radial line plotter, rapid point-by-point corrections for relief and tilt displacements is possible. The design of a radial line plotter is based on the principle of the Arundel method. The points, which are fixed by intersection along the edges of a photograph, may be used as plan control points for the plotting of the details later. Extending plan control points in this way is known as the Arundel method.

The upper portion of the instrument (Fig. 13.21) consists of a mirror stereoscope, which permits stereo viewing of the aerial photographs, which are centered below the large mirrors on two circular metal photo-carriers. When the photographs are set for stereo viewing, a metal pin is inserted in a transparent Perspex cursor, and passes through the principal point of the photograph into a hole in the metal photocarriers. When centered on the principal point, the Perspex cursor is constrained to move

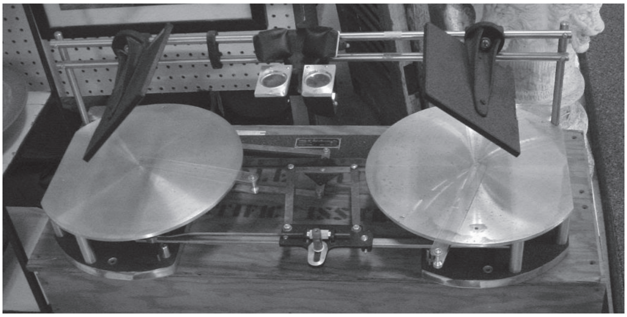


FIGURE 13.21 Radial line plotter

radially around that point. The radial index line is usually a fine, coloured thread or an engraved line on the Perspex cursor. This cursor is connected to the radial arm, which is located below the circular table.

The radial arm is linked to the plotting bar, which can move in the *x* and *y* directions, but is constrained by a linkage mechanism to keep parallel with the eye- base. When the plotting bar holding the pencil is moved, the radial index lines rotate in symmetry around the principal point of each photograph. With the continuous rotation of the radial index lines, an almost infinite number of intersection points can be made, but only the actual intersection point need be plotted by the pencil on the plotting bar.

It is necessary to place the control map sheet, or a sheet with four plan control points, on the table below the plotting bar. For most radial line plotters, there is a limited range of photograph-to-map scale differences within which the instrument must be used (a common limitation is between 2:3 reduction and 3:2 enlargement). There is a standard setting-up procedure, outlined in the instrument handbook, which ensures that when the radial index lines intersect on the photo-image of a control point, the plotting pencil is over the correct map position of that point.

If no existing map is available to act as a control base, plan control points may be obtained by the radial line method of control extension (Arundel method or slotted template) already described, or directly by ground surveying methods.

Once the photographs are correctly set up and linked to the control base, the plotting of the required detail features can be carried out point by point, and at each intersection, the relief displacement effect is corrected automatically.

For plotting the detail around the photo-base, where the intersection angle would be very oblique, it is necessary to decenter the Perspex cursors and the principal point. The errors made in this decentering procedure could be shown to be negligible for most radial line plotting work.

13.23 PHOTOGRAMMETRIC TECHNIQUES

Depending on the available material like the metric camera, stereopairs, shape of the recorded object, control information and the required results such as 2D or 3D, and accuracy, different photogrammetric techniques can be applied. Depending on the number of photographs, three main categories can be distinguished. They are:

1. Mapping from a single photograph.
2. Stereo photogrammetry.
3. Mapping from several photos.

13.23.1 Mapping from a Single Photograph

Mapping from a single photograph is useful for plane (2D) objects. Obliquely photographed plane objects show perspective deformations, which have to be rectified. For rectification, there exists a broad range of techniques. Some of them are very simple. However, there are some limitations. To get good results even with the simple techniques, the object should be plane (as, for example, a wall), and since only a single photograph is used, the mappings can only be done in 2D. The rectification can be neglected only if the object is flat and the picture is made from a vertical position towards the object. In this case, the photograph will have a unique scale factor, which can be determined if the length of at least one distance at the object is known. The common techniques such as the paper strip method, optical rectification, numerical rectification, mono plotting and digital rectification are described below.

Paper Strip Method This is the cheapest method, because only a ruler, a piece of paper with a straight edge and a pencil are required. Four points must be identified in the picture and in a map. From one point, lines have to be drawn to the others (on the image and the map) and to the required object point (on the image). Then the paper stripe is placed on the image and the intersections with the lines are marked. The strip is the placed on the map and adjusted such that the marks coincide again with the lines. After that, a line can be drawn on the map to the mark of the required object point. The whole process is repeated from another point, giving the object point on the map as an intersection of the two object lines.

Optical Rectification Optical rectification is done using photographic enlargeners. Again, at least four control points are required for optical rectification and not three on one line. The control points are plotted at a certain scale. The control point plot is rotated and displaced until two points match the corresponding object points from the projected image. After that the table has to be tilted by two rotations, until the projected negative fits to all control points. Then an exposure is made and developed.

Numerical Rectification Again, the object has to be plane and four control points are required. At the numerical rectification, the image coordinates of the desired object points are transformed into the desired coordinate system (which is again 2D). The result is the coordinates of the projected points.

Differential Rectification If the object is uneven, it has to be divided into smaller parts, which are plane. Each part can then be rectified with one of the techniques shown above. Of course, even objects may also be rectified piecewise, differentially. A prerequisite for differential rectification is the availability of a digital object model, that is, a dense raster of points on the object with known distances from a reference plane; in aerial photogrammetry, is called a digital terrain model (DTM).

Mono Plotting This technique is similar to numerical rectification, except that the coordinates here are transformed into a 3D coordinate system. First, the orientation elements that are the coordinates of the projection centre, and the three angles defining the view of the photograph are calculated by spatial resection. Second, using the calibration data of the camera, any ray that came from the archaeological feature through the lens onto the photograph, can be reconstructed and intersected with the digital terrain model.

Digital Rectification Digital rectification is a rather new technique. It is somehow similar to mono platting. But here, the scanned image is transformed pixel by pixel into the 3D real-world coordinate system. The result is an orthophoto, a rectified photograph that has a unique scale.

13.23.2 Stereo Photogrammetry

As the term implies, stereopairs are the basic requirements here. These can be produced using stereometric cameras. If only a single camera is available, two photographs can be made from different positions, trying to match the conditions of stereopair. They are made using special metric cameras that are built into an aircraft looking straight downwards. While taking the photographs, the aircraft files over a certain area in a way that the whole area is covered by overlapping photographs. The overlapping part of each stereo pair can be viewed in 3D, and consequently mapped in 3D using one of the following techniques:

Analogue Plotters The analogue method was mainly used until the 1970s. Two projectors, which have the same geometric properties as the used camera, project the negatives of the stereopair. Their positions then have to be exactly rotated into the same relationship towards each other as at the moment of exposure (relative orientation). After this step, the projected bundle of light rays from both photographs intersects with each other, forming a 3D optical model. Then, the scale of this model has to be related to its true dimensions, and the rotations and shifts in relation to the mapping coordinate system have to be determined. Therefore, at least three control points, which are not on one straight line, are required for absolute orientation.

The optical model is viewed by means of a stereoscope. The intersection of rays can then be measured point-by-point using a measuring mark. This consists of two marks, one on each photograph. When viewing the model, the two marks fuse into a single 3D mark, which can be moved and raised until the desired point of the 3D object is met. The movements of the mark are mechanically transmitted to a drawing device. In this way, maps are created.

Analytical Plotters The first analytical plotters were introduced in 1957. From 1970s, they become commonly available in the market. The idea is still the same as with analogue instruments. But here, a computer manages the relationship between the image-and real-world coordinates. The restitution of the stereopair is done within three steps.

After restoration of the inner orientation, where the computer may now also correct for the distortion of the film, both pictures are restively oriented. After this step, the pictures will be looked at in 3D. Then, the absolute orientation performed, where the 3D model is transferred to the real-world coordinate system. Therefore, at least three control points are required.

After the orientation, any detail can be measured out of the stereomodel in 3D. Like in the analogue instrument, the model and a corresponding measuring mark are seen in 3D. The movements of the mark are under the control of the user. The main difference with the former analogue plotting process is that the plotter does not plot directly onto the map, but onto the monitor screen or into the database of the computer.

The analytical plotter uses the computer to calculate the real-world coordinate, which can be stored as an ASCII file or transferred on-line into CAD-programmes. In this way, 3D drawings are created, which can be stored digitally, combined with other data and plotted later at any scale.

Digital Plotters (Soft Copy Photogrammetry) This latest generation of stereoplotting technique uses digital raster images (soft-copy medium) rather than aerial photographs (hard-copy medium) to perform the photogrammetric process. Aerial photographs, in the form of 23 cm × 23 cm photographs or continuous-roll films, are processed through high-resolution scanners to provide the digital images used in the process. The scanners convert the light transmitted through the photographic image into picture elements, called pixels, of fixed size, shape, spacing and brightness. The size of the pixel is important, with manufacturers claiming that a size of 7–10 μm (μm is micrometer, i.e. 10^{-6} m) is needed for this type of image processing. This scanning process could be bypassed if digital cameras were used in the first place.

Once the digital image files have been created, stereopaired images can be observed in three dimensions on a computer monitor by an operator wearing stereoglasses. The operator can, at this stage, perform the same functions available with the highly efficient analytical plotters. But because of the digital nature of the image files, the majority of the processes can be accomplished automatically. The five steps in aerial photography digital processing are:

1. Scanning of aerial photos (if film-based cameras are in use).
2. Aerotriangulation.
3. Elevation mapping (DEM/DTM).
4. Orthophoto (and mosaics) production.
5. Planimetric features mapping.

DEM refers to digital elevation model, and DTM refers to digital terrain model. DTM can be referred as a DEM with contour lines, which are needed to define the elevation surface properly. Software will arrange the huge data file into sub-files (called tiles) that can be manipulated more easily by the computers, at the appropriate time. But the computer processors should have a very large memory storage (cache memory) and fast speed. High refresh-rate monitors and various graphics accelerator boards will permit efficient processing of the huge data files. With the introduction of the network-based operating systems, users no longer have to invest huge sums in stereo plotters or workstation computers; instead, the principal part of the hardware process is a readily available, off-the-shelf computer that is also capable of performing a host of other functions ranging from CAD and GIS programmes to business software. The only photogrammetry-specific hardware needed are the scanners, stereo glasses, floating mark controller, three-dimensional mouse or pointer, and appropriate hard-copy plotter.

13.23.3 Mapping from Several Photographs

This kind of restitution, which can be done in 3D, has only become possible by analytical and digital photogrammetry. Since the required hardware and software is steadily getting cheaper, the application fields grow from day-to-day. Here, mostly more than two photographs are used. 3D objects are photographed from several positions. These are located around the object, where any object point should be visible on at least two or three photographs. The photographs can be taken with different cameras (even amateur cameras) and at different times (if the object does not move).

As mentioned above, only analytical or digital techniques can be used. In all methods, first a bundle adjustment has to be calculated. Using control points and triangulation points, the geometry of the whole block of photographs is reconstructed with high precision. Then the image coordinates of any desired object point measured in at least two photographs can be intersected. The results are the coordinates of the required points.

In this way, the whole 3D object is digitally reconstructed.

13.24 PHOTOGRAMMETRIC STEREOSCOPIC PLOTTING TECHNIQUES

Stereoplotters have been used for image rectification, that is, to extract planimetric and elevation data from stereopaired aerial photographs for the preparation of topographic maps. The photogrammetric process includes the following steps:

1. Establish ground control for aerial photos.
2. Obtain aerial photographs.
3. Orient adjacent photos so that the ground control matches.
4. Use aerotriangulation to reduce the number of ground points needed.
5. Generate a digital elevation model (DEM), also known as a digital terrain model (DTM).
6. Produce an orthophoto.
7. Collect data using photogrammetric techniques.

Steps 3 to 7 are accomplished using stereoplotting equipment and techniques. Essentially, stereoplotters incorporate two adjustable projectors that are used to duplicate the attitude of the camera at the time the film was exposed. Camera tilt and differences in flying height and flying speed can be noted and rectified. A floating mark can be made to rest on the ground surface when viewing stereoscopically, thus enabling the skilled operator to trace planimetric detail and deduce both elevations and contours.

In the past 50 years, aerial stereoplotting has undergone four distinct evolutions. The original stereoplotter (Kelsh Plotter) was a heavy and delicate mechanical device. Then came the analogue stereoplotter, and after that the analytical stereoplotter, which is an efficient technique that utilises computer-driven mathematical models of the terrain. The latest technique developed in the 1990s is soft-copy (digital) stereoplotting. Each new generation of stereoplotting reflects revolutionary improvements in the mechanical, optical and computer components of the system. Common features of the first three techniques were size, complexity, high capital costs, high operating costs of the equipment and the degree of skill required by the operator. Soft-copy photogrammetry utilises:

1. High-resolution scanners to digitise the aerial photos, and
2. Sophisticated algorithms to process the digital images on workstations (for example, Sun, Unix) and on personal computers with the Windows NT operating system.

The following three sub-sections cover the analytical stereoplotter, stereoplotting using photo prints, and soft-copy (digital) stereoplotting.

13.25 LIDAR

LIDAR is an acronym for Light Detection And Ranging. LIDAR uses the same principle as RADAR. The LIDAR instrument transmits light out to a target. The transmitted light interacts with and is changed by the target. The LIDAR beam, however, is very narrow, as it is basically a laser beam of visible or infrared radiation (very short wavelength). The wavelength for the LIDAR-transmitted energy is near/in the visible spectrum. Some of this light is reflected/scattered back to the instrument where it is analyzed. The change in the properties of the light enables some properties of the target to be determined. The time for the light to travel out to the target and back to the LIDAR is used to determine the range to the target.

There are three basic generic types of LIDAR:

1. Range finders
2. DIAL (Differential Absorption LIDAR)
3. Doppler LIDARS

Range finder LIDARs are the simplest LIDARs. They are used to measure the distance from the LIDAR instrument to a solid or hard target. Differential absorption LIDAR (DIAL) is used to measure chemical concentrations (such as ozone, water vapour and pollutants) in the atmosphere. A DIAL uses two different laser wavelengths, which are selected so that one of the wavelengths is absorbed by the molecule of interest, whilst the other wavelength is not. The difference in intensity of the two return signals can be used to deduce the concentration of the molecule being investigated.

Doppler LIDAR is used to measure the velocity of a target. When the light transmitted from the LIDAR hits a target moving towards or away from the LIDAR, the wavelength of the length reflected/scattered off the target will be changed slightly. This is known as a Doppler shift, hence it is called a Doppler LIDAR. If the target is moving away from the LIDAR, the return light will have a longer wavelength (sometimes referred to as a red shift); if moving towards the LIDAR, the return light will be at a shorted wavelength (blue shifted). The target can be either a hard target or an atmospheric target—the atmosphere contains many microscopic dust and aerosol particles which are carried by the wind. These are the target of interest to us, as they are small and light enough to move at the true wind velocity, and thus enable a remote measurement of the wind velocity to be made.

Laser radar, optical radar, and LIDAR are all names used for radar systems utilizing electromagnetic radiation at optical frequencies. The radiation used by laser radars is at wavelengths that are 10,000–100,000 times shorter than those used by conventional radars. Radiation (photons) scattered by the target is (are) collected and processed to yield information about the target and/or the path to the target. Early conventional and laser radars observed only the intensity of the collected radiation and the time delay from transmission to collection. Modern laser radars also observe intensity and time delay (see Fig. 13.22). However, some, called coherent laser radars (CLRs) record information about the phase of the scattered radiation with respect to a local reference. Laser radars may be continuous wave (CW) or pulsed, focussed or collimated.

FIGURE 13.22 OPTECH 2050 LIDAR equipment operated by laser map

CW laser radars are used when the signal may be integrated over long time periods and/or when the target is nearby. They are convenient to use when measuring average properties of the path to the target. Focussing is mainly used with CW laser radars to permit them to make a more sensitive measurement over a smaller span of ranges.

Pulsed laser radars use much higher power levels during the laser pulse than can be maintained with a CW laser, producing higher signal-to-noise ratios for the collected radiation. Pulsed LIDARs are usually chosen for long-range remote sensing and when long signal integration is impractical.

Laser radars may also be bistatic or monostatic. Bistatic sensing required the maintenance of two locations: one for the transmitter and one for the receiver. Monostatic sensing needs only one location to operate. Monostatic sensing is preferred for most scientific measurements because:

- Alignment and pointing are more straightforward
- There is only one instrument
- One location to staff
- One location from which to transmit and receive data

The location of a measurement taken by a monostatic, pulsed, laser radar is determined by knowledge of the laser pointing direction, and of the time delay between the firing of the laser pulse and the detection of the signal.

13.26 APPLICATIONS OF LIDAR

The advancements made in geographical information systems (GISs) in the last decade have created a large demand for 3D digital data. The most affordable method of obtaining location and elevation data for ground terrain, man-made objects and vegetation is though remote sensing. LIDAR provides a cost-effective solution for obtaining elevation data, and digital imaging (running the spectrum from panchromatic through hyperspectral) provides solution for classification of surface objects. The ideal solution for capturing this data is a signal platform capable of obtaining LIDAR and digital imagery simultaneously.

LIDAR technologies play a crucial role in the development of high-resolution topographic maps and digital terrain models. Present LIDAR system essentially paint the surface with a near infrared laser beam, collecting a dense cluster of elevation points with accuracies of the order of 15 cm and greater. In maintaining such vertical accuracies, the user is left with reflectance images that contain the elevations of the background natural terrain, and everything from vegetation cover (trees and shrubs) to man-made features (roads, bridges, building footprints, utility structures, cars and trucks). LIDAR data processing provides a unique challenge in the identification, delineation and removal of cultural and vegetation surface features, for the purposes of generating highly accurate bare earth digital terrain surfaces.

The commercially available software is overwhelmed by the large volumes of data generated by LIDAR systems, and by its complex data processing requirements. Little attention has been paid to the extraction of digital elevations outside the context of a bare earth surface. Thus, a whole set of surface features that include cultural and vegetation cover elevations are being treated more as a nuisance than legitimate 3D digital elevation features (see Fig. 13.23). The LIDAR data processing and analysis is a function of the following factors:

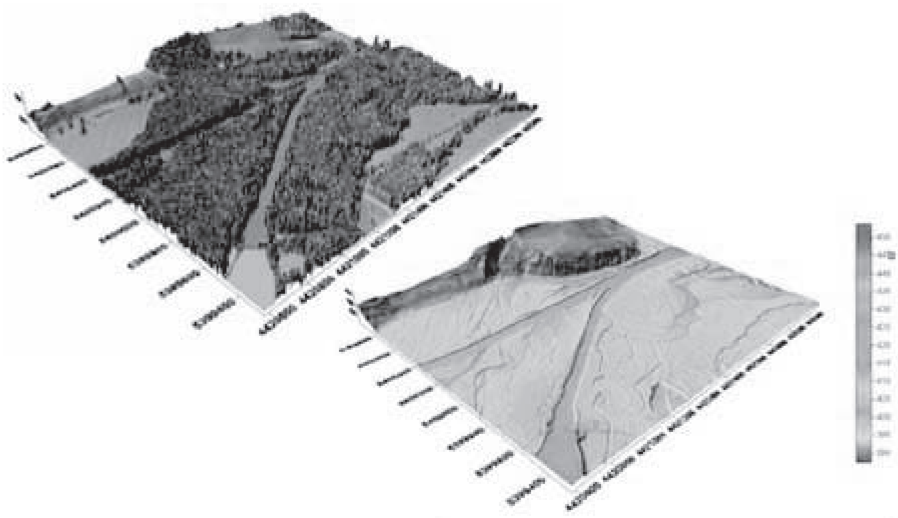


FIGURE 13.23 One-metre LIDAR digital surface model (DSM) and digital terrain model (DTM)

1. Terrain roughness
   a. Flat to mountainous terrains
2. Type of earth surface features
   a. Natural features (water bodies, vegetation and geomorphology)
   b. Cultural features (buildings, roads, utility structures and ditches)
3. Vegetation density
   a. Low- to high-density vegetation
4. Mixed terrain and surface features
   a. Heavily vegetated mountains
   b. Vegetated rolling terrain
   c. Urban areas
5. Spatial resolution of the LIDAR data
   a. Point spacing and resolution

13.27 HYPERSPECTRAL IMAGERY

Airbrone hyperspectral imaging is a new tool that can be used to map specific materials. Sunlight is reflected off the surface of materials and is received by the sensor in more than 100 different bands or channels. The combination of specific bands produces a unique signature for each material in the scene. These signatures are used to classify/identify the materials present at each location. It is, therefore, an excellent tool for environmental assessment, mineral mapping and exploration, vegetation communities/species and health studies, and general land management studies.

This imagery is especially powerful when combined with LIDAR points or the LIDAR generated surface. For example, the extraction of forest canopy heights can be accomplished using a combination of hyperspectral classification and LIDAR-based multiple return analysis techniques. The resultant hyperspectral vegetation classification could be used to generate polygons that exhibit the exact locations of forest canopy areas. It should be noted that these areas contain both forest canopy and ground elevation points.

13.28 ORTHOPHOTO

Manipulating the images on a photograph using mechanical, electronic, or computer techniques to eliminate the adverse effect of tip, tilt and relief displacement on scale produce an orthophoto.

Horizontal measurements made from an absolutely oriented stereoscopic model are very accurate, as they are made from orthogonal projections to form a map beneath the model. An orthophoto combines the accuracy of scaling, such as from a map with the detailed photographic representation of the mosaic. Orthophotos have been widely used for generating orthoimage maps (see Fig. 13.24), a preferable product for many GIS applications. The function is called, mono plotting in which vectors or features are delineated on top of orthophotos. The process does not require stereo viewing, and therefore shows an economic advantage. However, the accuracy of mono plotting is largely dependent on the underlying digital elevation model (DEM) used.

FIGURE 13.24 An orthoimage

REVIEW QUESTIONS

1. What is the general concept of photogrammetry, and what are the final results derived from photogrammetry?
2. Explain the various classifications of photogrammetry.
3. What are advantages of aerial photogrammetry?
4. Explain the different photographic devices used for aerial photography, giving the advantages and disadvantages of each.
5. Explain the various categories of aerial photographs.
6. What is a photographic scale? Derive an equation for the calculation of photographic scale with the help of a neat diagram.
7. Explain the interpretation of aerial photographs according to various factors.
8. How are flying heights calculated in aerial photography?
9. Explain the terms relief displacement and tilt displacement. How are relief displacement and tilt displacement calculated?
10. What is flight planning and what factors depend on the scale of photography?
11. To achieve good photogrammetric mapping, what are the points to be kept in mind?
12. Why is overlapping of photos required in aerial photography? What are the reasons for keeping overlap in aerial photography?
13. What is the effective coverage of a photograph and how it is calculated?
14. Derive an equation to calculate the number of photographs required in a strip; the net length and the net width covered by each photograph.
15. What are mosaics?
16. Explain the different classification of mosaics.
17. What are the limitations and advantages of mosaics?
18. What is stereoscopy? What are the conditions for aerial photography for stereo viewing?
19. Write short notes on:
    i. Lens and mirror stereoscope
    ii. Aerial triangulation
    iii. Parallax
20. How is radial triangulation carried out using the slotted template method?
21. Describe the working principle of a radial line plotter.
22. Write short notes on:
    i. Analogue plotters
    ii. Analytical plotters
    iii. Digital plotters
23. What is LIDAR? Define the three basic types of LIDAR.
24. What are the applications of LIDAR?
25. What is an orthophoto?

FUNDAMENTALS OF REMOTE SENSING 14

Chapter Outline

14.1 CONCEPT OF REMOTE SENSING

The advantage of collecting information from a distance was recognized years ago. Remote sensing is the most prominent technique of collecting information from a distance. The data collected from a distance is called remotely sensed data. Remote sensing is the science of acquiring information about the earth's surface without actually being in physical contact with the surface. It is also defined as the science or technology by which the characteristics of objects of interest can be identified, measured or analyzed without direct contact. This is done by sensing and recording reflected or emitted energy and processing, analyzing and applying that information. The science of remote sensing provides instruments, theory and methods by which objects and phenomena can be detected.

The technical term 'remote sensing' was first used in the United States in the 1960s. Technical terms like encompassed photogrammetry, photo-interpretation and photo-geology also were emerged with remote sensing. Since Landsat-1, the first earth observation satellite, was launched in 1972 remote sensing has become widely used.

The characteristics of an object can be determined using reflected or emitted electromagnetic radiation from the object. That is, each object has unique and different characteristics of reflection or emission, depending upon different environmental conditions. Remote sensing includes aerial photography and satellite imagery. Today, most natural resource mapping is done using remote sensing. Aerial photography has been used to produce virtually all topographic maps such as forest maps, geological maps, land use maps and soil maps. Aerial photography is used to prepare detailed city maps. Satellite-based systems can now measure phenomena that change continuously over time and cover large and inaccessible areas. Satellite-based systems are used to estimate chlorophyll levels near the sea surface, which is an indicator of the availability of the food on which commercial fish stock depend.

Remote sensing is a technology to identify and understand the object or the environmental condition, through the uniqueness of the reflection or emission. The remote sensing data will be processed automatically by computer and manually interpreted by humans, and finally utilized in agriculture, land use, forestry, geology, hydrology, oceanography, meteorology and environment.

14.2 PRINCIPLES OF REMOTE SENSING

All objects on the surface of the earth have spectral signatures. A spectral signature of an object or ground surface feature is a set of values for the reflectance or radiance of the feature, each value corresponding to the reflectance or radiance arranged over a different and well-defined wavelength interval. Spectral signature is the distinctive set of distinguishable characteristics. The response of ground surface materials to incident radiation is the reflectance, and the energy emitted by all objects as a function of their temperature and structure is the emittance. The reflectance and emittance determine the signatures. The knowledge of spectral signature is essential for exploiting the potential of the remote sensing technique. This knowledge enables one to identify and classify objects. It is also required for interpretation of all remotely sensed data, whether the interpretation is carried out visually or using digital techniques. Evaluation of the spectral signature implies a basic understanding of the interaction of electromagnetic radiation with various earth surface objects.

When radiation is incident on a surface, it is reflected, absorbed, scattered and transmitted. All the processes are strongly dependent on the wavelength of the incident radiation, as well as the atomic and molecular structure of the material. In view of these facts, one can identify the material constituting the object from a spectral plot, multiband photograph or any other record, which shows enough details of its spectral reflection, absorption, scattering or transmission properties.

The absorption of radiation by an object leads to thermal activity, which results in the emission of radiation at a different wavelength. The spectral emission from the object depends on the surface characteristics, as well as the molecular structure. Hence the spectral emission pattern can identify an object. Practically, the remote sensing technique is based on the observation of the reflectance of incident radiation and the emittance of radiation by the objects. Hence, spectral reflectance and emittance are the most important characteristic studies done by remote sensing techniques.

14.3 COMPONENTS OF REMOTE SENSING

The first requirement for remote sensing is to have an energy source to illuminate the target (unless the sensed energy is being emitted by the target). Radiant energy emitted from ground features is transmitted to the sensing instrument in the forms of waves. Remote sensing of land surface features is based on detection of radiant energy called electromagnetic radiation or Electromagnetic radiation, which is reflected or emitted from an object, is the usual source of remote sensing data. Most remote sensing instruments detect and record electromagnetic radiation. However, any media such as gravity or magnetic fields can be utilized in remote sensing.

Remote sensing instruments detect and record the energy photons in the band or bands of which the sensor is sensitive. The amount of energy deflected depends upon the inherent energy of the photons and the number of photons reaching the detector during the short-term interval of energy collection. The number of photons (energy intensity) reaching the detector varies according to the amount of energy emitted by the illumination source, the amount of energy absorbed by the atmosphere and the degree to which ground objects reflect and emit energy.

The range of possible electromagnetic radiation wavelengths and frequencies is termed as an electromagnetic spectrum. A remote sensing system primarily consists of the following stages:

1. A source of electromagnetic energy (from the sun, the self-emission).
2. Transmission of the energy from the source to the surface of the earth (scattering and absorption).
3. Interaction of electromagnetic radiation with the earth surface, reflection and reemission.
4. Transmission of energy from the surface to the remote sensor.
5. Sensor data output.
6. Processing and analyzing the sensor data output.

Hence the four basic components of a remote sensing system include a target, an energy source, a transmission path and a sensor (see Fig. 14.1).

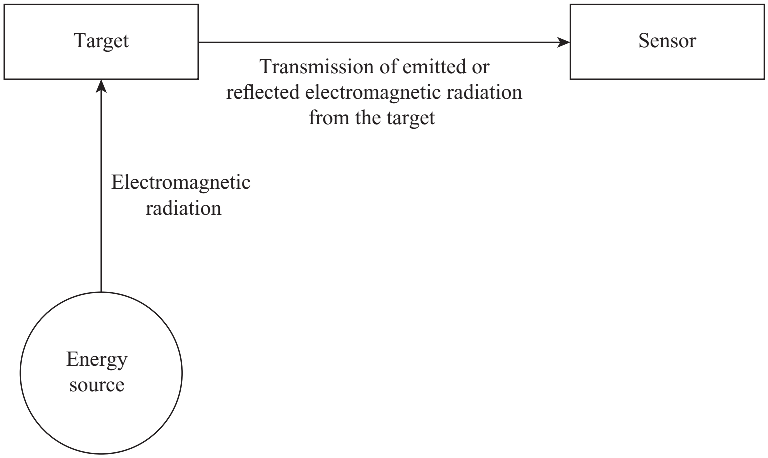


FIGURE 14.1 Components of a remote sensing system

The target is the object or material that is being studied. The components in the system work together to measure and record information about the target without actually coming into physical contact with it. There must also be an energy source, which illuminates or provides electromagnetic energy to the target. The energy interacts with the target, depending on the properties of the target and the radiation, and will act as a medium for transmitting information from the target to the sensor. The sensor is a remote device that will collect and record the electromagnetic radiation. Sensors can be used to measure energy that is given off (or emitted) by the target, reflected off the target or transmitted through the target.

Once the energy has been recorded, the resulting set of data must be transmitted to a receiving station, where the data are processed into a usable format, which is most often as an image. The image is then interpreted to extract information about the target. This interpretation can be done visually or electronically with the aid of computers and image processing software.

Weather satellite imaging of the earth is a familiar example of a remote sensing system. The target in such a system is the earth's surface, which gives off energy in the form of or heat energy. This energy travels through the atmosphere and space and reaches the sensor, which is mounted on a satellite platform. Varying levels of this energy are recorded, transmitted to ground stations on the earth and converted into images that depict differences in temperature across the planet's surface. In a similar manner, other weather satellite sensors measure the visible light energy from the sun as it is reflected off the earth's surface, transmitted through space to the satellite sensor, and recorded and sent to earth for processing.

Another familiar form of remote sensing that occurs on a relatively smaller scale is medical imaging technologies such as magnetic resonance imaging (MRI), sonograms, and X-ray imaging. These technologies all forms of energy to produce images of the human body internally. In each, various forms of energy are produced by a machine and directed at the target object. The sensors measure how this energy is absorbed, reflected or transmitted through the target object, and the results are compiled into an image. These technologies provide the obvious benefit in that they allow observation and measurement of the internal systems in the human body without potentially dangerous invasive surgery.

14.4 SEVEN ELEMENTS IN REMOTE SENSING

In remote sensing, the process involves an interaction between the incident radiation and the targets of interest. This is exemplified by the use of imaging systems where the following seven elements are involved. Note, however, that remote sensing also involves the sensing of emitted energy and the use of non-imaging sensors.

Energy Source or Illumination The first requirement for remote sensing is to have an energy source, which illuminates or provides electromagnetic energy to the target of interest.

Radiation and the Atmosphere As the energy travels from its source to the target, it will come in contact with and interact with the atmosphere it passes through. This interaction may take place a second time as the energy travels from the target to the sensor.

Interaction with the Target Once the energy makes its way to the target through the atmosphere, it interacts with the target depending on the properties of both the target and the radiation.

Recording of Energy by the Sensor After the energy has been scattered by, or emitted from the target, it requires a sensor (i.e., a remote sensor that is not in contact with the target) to collect and record the electromagnetic radiation.

Transmission, Reception and Processing The energy recorded by the sensor has to be transmitted, often in an electronic form, to a receiving and processing station where the data are processed into an image (hard copy and/or digital).

Interpretation and Analysis The processed image is interpreted, visually and/or digitally or electronically to extract information about the target, which was illuminated.

Application The final element of the remote sensing process is achieved by extracting the required information from the imagery about the target to better understand it, or to reveal some new information, or to assist in solving a particular problem.

The seven elements described above comprise the remote sensing process from the beginning to the end.

14.5 CHARACTERISTICS OF ELECTROMAGNETIC RADIATION

Electromagnetic radiation is a carrier of electromagnetic energy by transmitting the oscillation of the electromagnetic field through space or matter. The transmission of electromagnetic radiation is derived from the Maxwell equations. Electromagnetic radiation has the characteristics of both wave motion and particle motion.

Characteristics as Wave Motion Electromagnetic radiation can be considered as a transverse wave with an electric field and a magnetic field. A plane wave, for example, as shown in Fig. 14.2, has its electric field and magnetic field in the perpendicular plane to the transmission direction.

The two fields are located at right angles to each other. Here the wavelength, frequency and the velocity have the following relation:

where $c = \lambda v$

λ = is the wavelength (m)

v = the frequency (cycle per second, Hz); and

c = the speed of light (3×10^8 m/s)

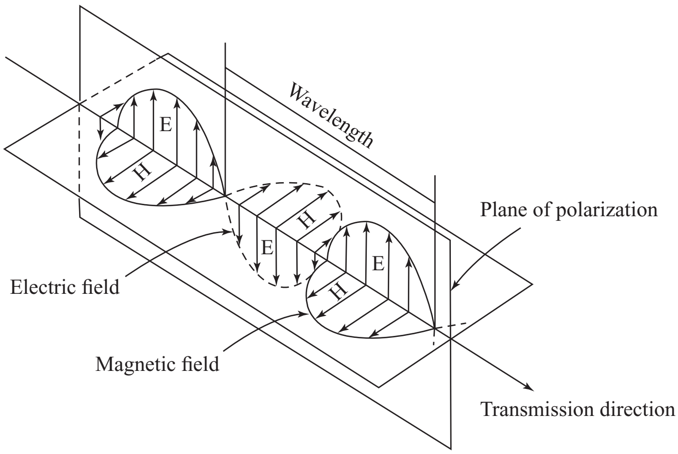


FIGURE 14.2 Electromagnetic radiation

Electromagnetic radiation is transmitted in a vacuum of free space with the velocity of light c (2.998×108 m/sec) and in the atmosphere with a reduced but similar velocity to that in a vacuum. The frequency n is expressed as a unit of hertz (Hz), which is the number of waves transmitted per second.

Characteristics as Particle Motion Electromagnetic radiation can be treated as a photon or a light quantum, and energy E can be expressed as

$$E = hv$$

where h is Plank's constant and v the frequency.

All electromagnetic radiation has fundamental properties and behaves in predictable ways according to the principles of wave theory. Electromagnetic radiation consists of an electric field (E), which varies in magnitude in a direction perpendicular to the direction in which the radiation is travelling, and a magnetic field (H) oriented at right angle to the electric field. Both these fields travel at the speed of light (c).

Two characteristics of electromagnetic radiation are important for understanding remote sensing as discussed earlier. These are the wavelength and frequency. The wavelength is the length of one wave cycle, which can be measured as the distance between successive wave crests (see Fig. 14.3). Wavelength is usually represented by the Greek letter lambda (λ). Wavelength is measured in metres (m) or some factor of metres such as nanometres (nm, 10^{-9} m), micrometres (μm, 10^{-6} m) or centimetres (cm, 10^{-2} m). Frequency refers to the number of cycles of a wave passing a fixed point per unit of time. Frequency is normally measured in hertz (Hz), equivalent to one cycle per second, and various multiples of hertz.

The photoelectric effect can be explained by considering the electromagnetic radiation as composed of particles. Electromagnetic radiation has four elements frequency (or wavelength), transmission direction, amplitude and plane of polarization. The amplitude is the magnitude of the oscillating electric field.

The square of the amplitude is proportional to the energy transmitted by electromagnetic radiation. The energy radiated from an object is called radiant energy. A plane, including the electric field, is called a plane of polarization. When the plane of polarization forms a uniform plane, it is called linear polarization. The four elements of electromagnetic radiation are related to different information content, as shown in Fig. 14.4. Frequency (or wavelength) corresponds to the colour of an object in the visible region, which is given by a unique characteristic curve relating the wavelength and the

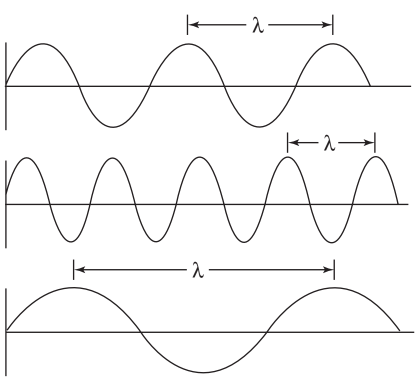


FIGURE 14.3 Wavelength of different waves

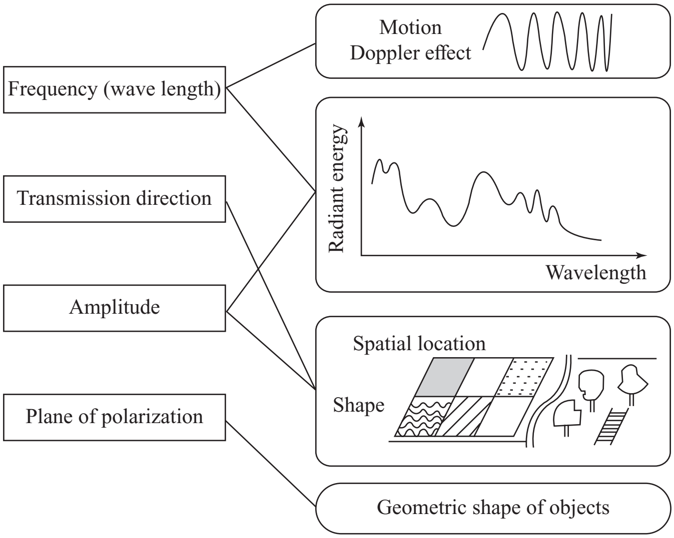


FIGURE 14.4 Information derived from elements of electromagnetic radiation

radiant energy. In the microwave region, information about objects is obtained using the Doppler shift effect in frequency that is generated by a relative motion between an object and a platform. The spatial location and shape of objects are given by the linearity of the transmission direction, as well as by the amplitude.

The plane of polarization is influenced by the geometric shape of objects in the case of reflection, or scattering in the microwave region. In the case of radar, horizontal polarization and vertical polarization have different responses on a radar image.

All matter reflects, absorbs, penetrates and emits electromagnetic radiation in a unique way. For example, the reason why a leaf looks green is that the chlorophyll in the leaf absorbs blue and red spectra and reflects the green spectrum (see Fig. 14.5). Radiation that is not absorbed or scattered in the atmosphere can reach and interact with the earth's surface. There are three forms of interaction that can take place when energy strikes, or is incident (I) upon the surface (see Fig. 14.5). These are

FIGURE 14.5 Three forms of interaction when energy strikes upon a surface

TABLE 14.1 Relation between characteristic state and electromagnetic radiations

Characteristic State	Energy (eV)	Associated Electromagnetic Wave
Nuclear transmissions and disintegrations	$10^7 \sim 10^5$	γ-ray
Ionization by inner electron removal	$10^4 \sim 10^2$	X-ray
Ionization of outer electron removal	$10^2 \sim 4$	Ultraviolet
Excitation of valence electrons	$4 \sim 1$	Visible
Molecular vibration, lattice vibration	$10 \sim 10^{-5}$	IR
Molecular rotations, electron spin resonance	$10^{-4} \sim 10^{-5}$	Microwave
Nuclear spin resonance	10^{-7}	Metre wave

Note: (unit) energy of 1 eV = $1.60219 \times 10^{-19} J$, wavelength of 1 eV = 1.23985 µm.

absorption (A), transmission (T), and reflection (R). The total incident energy will interact with the surface in one or more of these three ways. The proportions of each will depend on the wavelength of the energy and the material and condition of the feature.

As all matter is composed of atoms and molecules with a particular composition, the matter will emit or absorb electromagnetic radiation at a particular wavelength with respect to the inner state. The types of inner state are classified into several classes, such as ionization, excitation, molecular vibration and molecular rotation as shown in Table 14.1, which will radiate the associated electromagnetic radiation. For example, visible light is radiated by excitation of valence electrons, while IR is radiated by molecular vibration or lattice vibration.

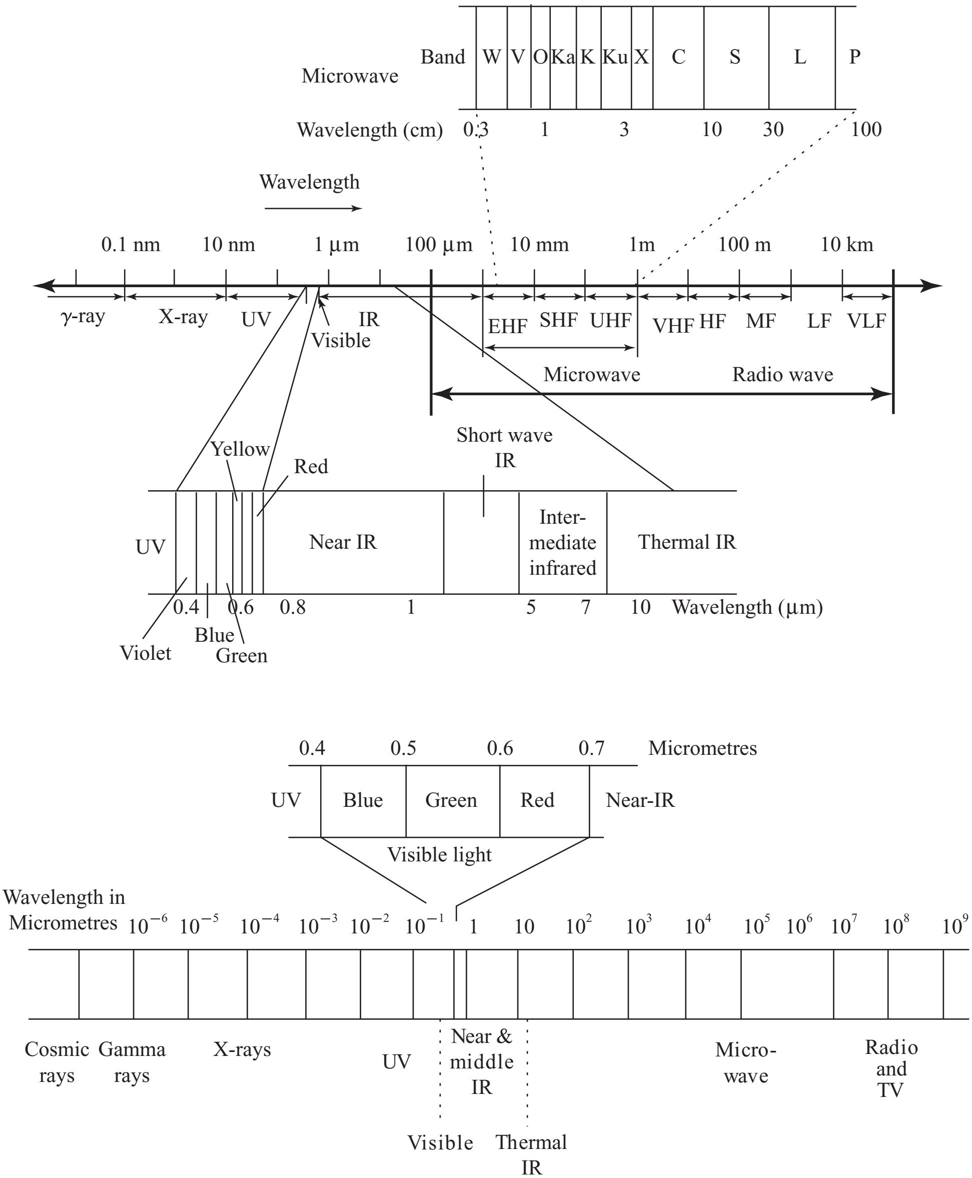
14.6 ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum ranges from the shorter wavelengths (including gamma rays and X-rays) to the longer wavelengths (including microwaves and broadcast radio waves). There are several regions of the electromagnetic spectrum which are useful for remote sensing. Wavelength regions of electromagnetic radiation have different names ranging from γ-ray, X-ray, ultraviolet (UV), visible light, IR to radio wave, in order from the shorter wavelengths. The shorter the wavelength, the more is the electromagnetic radiation characterized as particle motion with more linearity and directivity.

The electromagnetic radiation regions used in remote sensing are near UV (0.3–0.4 µm), visible light (0.4–0.7 µm), near short-wave and thermal IR (0.7–14 µm), and microwave (1 mm to 1 m).

Figure 14.6(a) and 14.6(b) shows the spectral bands used in remote sensing. The spectral range of near IR and short wave IR is sometimes called the reflective IR (0.7–3 µm) because the range is more influenced by solar reflection rather than the emission from the ground surface. In the thermal IR region, emission from the ground's surface dominates the radiant energy with little influence from solar reflection.

The light which our eyes or remote sensors can detect is a part of the visible spectrum. The visible portion is very small relative to the rest of the spectrum. There is a lot of radiation around us, which is invisible to our eyes, but can be detected by other remote sensing instruments. The visible wavelengths cover a range of approximately 0.4–0.7 µm. The longest visible wavelength is red and the shortest is violet. The common wavelength spectrum that human beings can perceive as particular



(a)

FIGURE 14.6 (a) Spectral bands used in RM and electromagnetic spectrum

colours from the visible portion of the spectrum are listed below. It is important to note that this is the only portion of the spectrum that is associated with the concept of colours.

1. Violet: 0.4–0.446 μm
2. Blue: 0.446–0.500 μm
3. Green: 0.500–0.578 μm

4. Yellow: 0.578–0.592 µm
5. Orange: 0.592–0.620 µm
6. Red: 0.620–0.7 µm

Blue, green and red are the primary colours or wavelengths of the visible spectrum. They are defined as such because no single primary colour can be created from the other two, but all other colours can be formed by combining blue, green and red in various proportions. Although sunlight is seen as a uniform or homogeneous colour, it is actually composed of various wavelengths of radiation in primarily the UV, visible and IR portions of the spectrum. The visible portion of this radiation can be shown in its component colours when sunlight is passed through a prism, which bends the light in differing amounts according to the wavelength.

The next portion of the spectrum that is of interest is the IR region, which covers the wavelength range of approximately 0.7–100 µm, which is more than 100 times as wide as the visible portion. The IR region can be divided into two categories based on their radiation properties, the reflected IR, and the emitted or thermal IR. Radiation in the reflected IR region is used for remote sensing purposes in ways very similar to radiation in the visible portion. The reflected IR covers wavelengths from approximately 0.7 to 3.0 µm. The thermal IR region is quite different from the visible and reflected IR portions, as this energy is essentially the radiation that is emitted from the earth's surface in the form of heat. The thermal IR covers wavelengths from approximately 3.0 to 100 µm. Visible light corresponds to the spectral colours. They are, in order from the longer wavelengths in the visible region, the so-called rainbow colours such as red, orange, yellow, green, blue, indigo and violet, and are located with respect to the wavelength.

Short wave IR has more recently been used for geological classification of rock types. Thermal IR is primarily used for temperature measurement, while microwave is utilized for radar and microwave radiometry. A special naming of *K* band, X band, C band and L band is given to the microwave region as shown in Fig. 14.6(a).

14.6.1 IR Region and Wein's Displacement Law

The non-visible IR spectral region lies between the visible light and the microwave portion of the electromagnetic spectrum. The IR region covers a wavelength range of 0.7–14 µm. This broad range of IR wavelengths is further subdivided into two smaller IR regions. Each of these regions exhibits very different characteristics.

The IR region closest to visible light contains two smaller bands labelled near the IR and short-wave IR with wavelengths ranging from 0.7 to 1.1 µm, and 1.1 to 3.0 µm, respectively. These IR regions exhibit many of the same optical characteristics as visible light. The sun is the primary source of IR radiation, which is reflected from an object. Cameras used to capture images in the visible light spectrum can capture images in the near IR region by using special IR film.

The other IR region with longer wavelengths ranging from 3.0 to 14.0 µm is composed of two smaller bands labelled mid-wave IR and long wave IR, with wavelengths ranging from 3.0 to 5.0 m, and 5.0 to 14.0 µm, respectively. Objects generate and emit thermal IR radiation, and thus these objects can be detected at night because they are not dependent on the reflected IR radiation and from the sun. Remote sensors operating in this IR wavelength range measure an object's temperature.

Electromagnetic radiation is a carrier of electromagnetic energy by transmitting the oscillation of the electromagnetic field through space or matter. Electromagnetic radiation is transmitted in a vacuum of free space with the velocity of light (~300,000 km/sec). In the atmosphere, it has a reduced but a similar velocity to that in a vacuum. It is important to note that the above division is entirely arbitrary. Nothing really changes from cosmic rays to radio waves, except that the wavelength and the spectrum are continuous. The properties of radiation undergo a gradual change with respect to increasing wavelength.

The quantity of radiation emitted by an object depends upon its temperature. Higher the temperature of an object, larger the quantity of radiation emitted. The lowest temperature attainable by an object is theoretically, 273°C or absolute zero, at which objects do not emit any radiation. If the energy emission by objects in different wavelengths at various temperatures is computed, the plot of energy versus wavelength is achieved. The hotter the objects, the more the radiation will result.

Wein's Displacement Law of Emission: It can be seen that although an object emits radiation in several wavelengths, there is always one wavelength in which emission is maximum. That is to say, the emission curves have a well-defined peak in every spectrum. This wavelength is known as the wavelength of maximum emission or I_{max}. The value of I_{max} is unique for every temperature. This implies that the temperature of an object and the wavelength at which most of the emission occurs are connected. When one looks at the various emission spectra, one notices that when the wavelength is maximum, emission I_{max} shifts to the left as the temperature of the object increases. One can then say that hotter an object, shorter the wavelength it emits with high-frequency radiation, while cooler objects emit low-frequency radiation.

Wein's displacement law states that the wavelength of maximum emission 'I_{max}' is displaced to the left as the temperature of the emitting object is raised, sums up this process.

Wein's displacement law can be mathematically stated as $I_{max} \times T = 3000$ (14.1)

where I_{max} is the wavelength of maximum emission in micrometres and T the absolute temperature in kelvin.

Using the Wein's displacement law, we can find the temperature of any object if we know its wavelength at which most of its emission occurs. This principle is used in an instrument called the precision radiation thermometer (PRT), which senses the temperature of an object by just looking at it. This is unlike other common thermometers, which has to physically come in contact with the object. PRT is probably the first remote sensing instrument in the true sense of the term. Most of the sun's emission is roughly at 480 nm (0.48 mm), a wavelength that gives us the sensation of blue colour. If one substitutes this value of I_{max} in the above equation, the sun's temperature at its surface works out to approximately 6,200°C.

In the case of the earth, its temperature varies between −50°C and +50°C. From the above equation, one can find that most of the earth's emission lies between 10 and 12 mm, which is known as the thermal IR radiation. Incidentally, the thermal IR does not encounter any atmospheric effects while passing through it. Hence, thermal IR red is used for remote sensing of earth's temperature. This logic is true for every other planet and star. One can sense their temperature by observing the wavelength in which they emit most of their radiation. It is equally important to note that the temperature sensed by this method is only the surface temperature of the object, whether it is the sun or the earth or any other celestial body. This method gives no clue to the temperature of its interiors.

Thus, what an object emits depends upon its temperature, for example, the earth is not so hot to emit light. But objects on earth can be heated to fairly high temperatures to emit light, UV rays or even X-rays. Radio waves have the longest wavelengths in the electromagnetic spectrum, while cosmic rays have the shortest wavelength. This means that the coolest objects in the universe would emit radio waves, while the hottest objects will emit cosmic rays. This visible light, which forms a small part of the electromagnetic spectrum, is itself composed of a series of wavelengths. They vary from as high as 700 nm (red) to as low as 400 nm (violet). The colour of a flame is an indication of its temperature. It is coolest when it looks red and hottest when it looks violet. This is the reason why a blue flame is hotter than a red flame.

14.7 TRANSMISSION PATH

Radiation may be defined as a process in which energy is transmitted across space, whether or not a material medium is present. Radiative transfer is fundamental to remote sensing in that energy transfer into and out of the earth–atmosphere–ocean system affects the way in which we can remotely sense properties of this system. Some of the effects of the atmosphere on remote sensing processes are illustrated schematically in Fig. 14.8.

Any or all of the waves can be used to carry information from one place to another. But information carrying capacity itself depends upon the frequency. The higher the frequency, the better the capacity. Thus, radio waves, which have the lowest frequency among the spectrum, can only carry sound. They cannot be used to carry television (TV) pictures. For TV transmission, higher-frequency waves like microwaves are required.

When short waves are used for remote sensing, they resolve the picture better because of their higher information carrying capacity. A microwave picture can resolve only up to 25- or 50-km wide features, while an IR picture can resolve nearly 1-km wide features. Hence, short waves are better for conceivable application in remote sensing. But shorter the waves, higher are the chances that they will be absorbed and scattered in the earth's atmosphere. As a result, short waves like light cannot pass through cloud, haze, mist and smoke.

IR have a higher potential to penetrate through cloud, haze and smoke. Hence, the information carried by short waves quickly degrades in the earth's atmosphere, and may never reach the satellite. Microwave and radio waves can pass through the thickest of fog or cloud because of their long wavelength. They travel in a straight line because they do not scatter in air like short waves. Hence, signals carried by microwaves do not degrade in the atmosphere. Cosmic rays, gamma rays and X-rays are too short to enter the earth's atmosphere. They are either absorbed or reflected from the top of the atmosphere itself. Hence, both remote sensing and communication on the earth is restricted only to four wave bands, namely, visible light, IR, microwave and radio waves. But radio waves are usually not used for remote sensing, because their resolution is very poor due to their very long wavelengths. They do not resolve objects of interest.

Light IR and microwaves are the only wavebands used for remote sensing. Light resolves the best, but the signal gets degraded in the earth's atmosphere. IR can pass though most atmospheric obstacles except clouds, while microwaves can pass through almost all atmospheric obstacles. Earth does not emit light. It merely reflects the sunlight. In case of IR, earth emits only one waveband called thermal IR, while the rest of the IR radiations are simply received from the sun, which the earth reflects. But the earth emits microwaves and that gives microwaves a special significance in remote sensing. The atmospheric effects do not affect the thermal IR and microwaves, except that the microwaves are scattered by raindrops. There are only two processes that hinder the smooth passage of radiation through a medium. One is absorption and the other is scattering. Shorter waves are more affected by absorption than the longer waves. There are a few wavelengths in the radiation spectrum that are simply not affected by the atmosphere. They behave as if the atmosphere does not exist. They are known as atmospheric windows.

14.7.1 Atmospheric Windows

Remote sensing is based on the measurement of reflected or emitted radiation from different bodies. Atmospheric scattering and absorption is not constant across the spectrum. Absorption in particular can be much greater in some parts of the spectrum when compared with other parts. In some regions of the spectrum termed atmospheric windows, the electromagnetic radiation passing through the atmosphere is less affected. Electromagnetic energy when travels through the atmosphere are partly absorbed by various molecules. The electromagnetic radiation is usually absorbed by ozone, water vapour and carbon dioxide present in the atmosphere. Atmospheric windows are that portion of the electromagnetic spectrum that can be penetrated through the atmosphere without any distortion or absorption. Only the wavelength regions of the electromagnetic spectrum outside the main absorption bands of the atmospheric gases can be used for remote sensing. An atmospheric window is a dynamic property of the atmosphere. But the spectral window is a static characteristic of the electromagnetic radiation absorption spectra of greenhouse gases, including water vapour. The atmospheric window provides the information on what actually happens in the atmosphere. Window radiation is radiation that actually passes through the atmospheric window. Non-window radiation is radiation that actually does not pass through the atmospheric window.

The IR atmospheric window is the property of the earth's atmosphere, taken as a whole at each place and time that lets some IR radiation from the land-sea surface pass directly to space without intermediate absorption and re-emission and without heating the atmosphere. Fig. 14.6(b) shows a schematic representation of atmospheric transmission in the 0–22 μm wavelength region.

Atmospheric window include four windows, one in the visible and reflected region between 0.4 and 2.0 μm (where the remote sensors and human eye operate) and other three windows in the thermal IR region. Strong absorption bands are found at longer wavelengths due to the moisture present in the atmosphere.

14.7.2 Scattering of Electromagnetic Radiation

Before radiation used for remote sensing reaches the earth's surface, it has to travel some distance through the earth's atmosphere. Particles and gases in the atmosphere can affect the incoming light and radiation. These effects are caused by the mechanisms known as scattering and absorption.

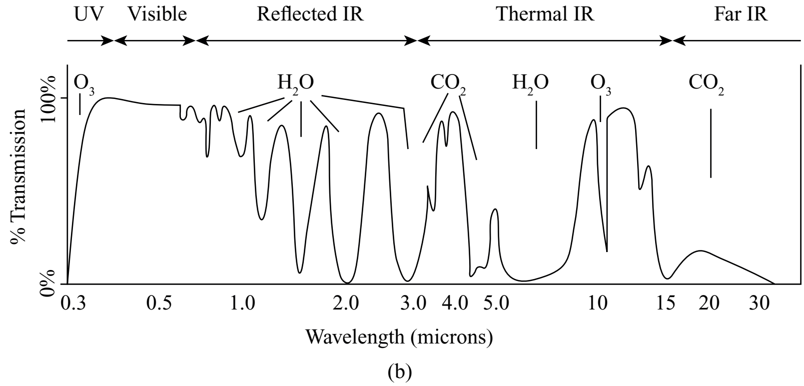


FIGURE 14.6 (b) Spectral bands used in RM and electromagnetic spectrum

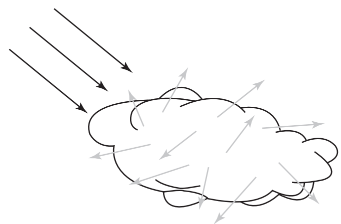


FIGURE 14.7 Scattering of waves

Rayleigh scattering: Rayleigh scattering is the dominant scattering mechanism in the upper atmosphere. The law of scattering of air molecules was discovered by Rayleigh in 1871. Rayleigh scattering occurs when the size of the particle responsible for the scattering is less than the wavelength of radiation. The fact that the sky appears blue during the day is because of this phenomenon. As sunlight passes through the atmosphere, the shorter wavelengths of the visible spectrum are scattered more than the other visible wavelengths. At sunrise and sunset, the light has to travel farther through the atmosphere than at midday, and the scattering of the shorter wavelengths is more complete; this leaves a greater proportion of the longer wavelengths to penetrate the atmosphere. Rayleigh scattering occurs when particles are very small compared with the wavelength of the radiation. The particles could be small specks of dust or nitrogen and oxygen molecules. Rayleigh scattering causes shorter wavelengths of energy to be scattered much more than longer wavelengths.

Mie scattering: Mie scattering occurs when the particles are just about the same size as the wavelength of the radiation. Dust, pollen, smoke and water vapour are common causes of Mie scattering, which tends to affect longer wavelengths rather than those affected by Rayleigh scattering. Mie scattering occurs mostly in the lower portions of the atmosphere where larger particles are more abundant, and dominates when cloud conditions are overcast (see Fig. 14.7).

This non-selective scattering occurs when the particles are much larger than the wavelength of the radiation. Water droplets and large dust particles can cause this type of scattering. Non-selective scattering gets its name from the fact that all wavelengths are scattered about equally. This type of scattering causes fog and clouds to appear white to our eyes because blue, green and red lights are all scattered approximately in equal quantities.

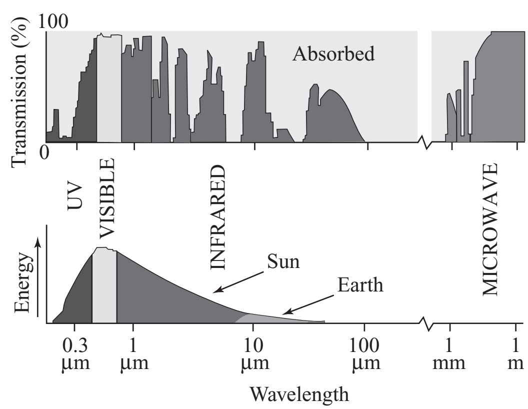


FIGURE 14.8 Atmospheric transmittance of electromagnetic energy

Those areas of the spectrum, which are not severely influenced by atmospheric absorption, and thus are useful to remote sensors, are called atmospheric windows. By comparing the characteristics of the two most common energy/radiation sources (the sun and the earth) with the atmospheric windows available to us, we can define those wavelengths that we can use most effectively for remote sensing. The visible portion of the spectrum corresponds to both an atmospheric window and the peak energy level of the sun. Note also that heat energy emitted by the earth corresponds to a small window around 10 μm in the thermal IR portion of the spectrum, while in a large window, wavelengths beyond 1 mm are associated.

14.8 PLATFORMS

The platform is the vehicle on which the instrument (or sensor) is carried. A remote sensing platform is designed with a relatively narrow set of purposes in mind. Many important decisions must be made when designing a remote sensing technology. The type of sensor and its capabilities must be defined. The platform on which the sensors will be mounted must be determined, and the means by which the remotely sensed data is received, transmitted and processed before delivery to its end user must be designed. All of these decisions are made based on knowledge of the target and the information about the target that is in demand, balanced by other factors such as cost, availability of resources and time constraints. The end result of this process is a tool that is specifically designed to perform a task or a set of related tasks that will assist researchers in better understanding the process that is under investigation.

A listing of platforms is given below:

Hand Held Some field instruments may be carried in the hand and pointed at the target in question.

Cherry Picker A platform similar to the one used by electrical workers when servicing lamplights, etc. It allows a larger region to be viewed with a hand held instrument.

Balloon It is generally tethered and suspended over a target region, and is useful for taking repeated measurement of a small area using either a hand held instrument or an imaging instrument, for example, a camera.

Radio Control Models Both fixed and rotary wing radio-control models have been used to carry remote sensing instruments. This type of platform has gained recent popularity with filmmakers wishing to emulate flight sequences through tight terrain. Non-piloted vehicles with pre-programmed flight paths that are controlled by GPS have also become available and provide a cheap and extended mission capability for remote sensing data acquisition.

Low Altitude Aircraft Ultralights may be employed up to 5,000 ft with light payloads. Conventional aircraft can extend this to over 15,000 ft. Jets and other pressurised aircraft can fly at higher altitudes.

High Altitude Aircraft Special planes like the U2 spy plane can fly at altitudes of 20 km in the upper atmosphere. At this altitude the aircraft is very stable and geometric distortions can be minimized. Operational expenses are high, however.

Low Altitude Spacecraft The space shuttle has flown many remote sensing missions at altitudes of around 250 km.

Medium Altitude Spacecraft Spot and Landsat orbit at approximately 700–900 km above the earth and can provide wide area coverage at high-geometric stability and high-spatial resolution.

High Altitude Spacecraft Geostationary satellite are placed at an altitude of 20,000 km or more above earth's surface to provide a broad coverage at a coarse spatial resolution.

FIGURE 14.9 Ground-based platform (sensor)

Together, these platforms introduce the concept of multilevel remote sensing. This wide range of platforms (and sensors) means that the right system can be selected for each project. It is when the wrong system is used that the results of remote sensing projects fail to meet expectations. In the past this had led to remote sensing receiving a poor reputation, while in reality, it was inexperience on behalf of the user that was to blame.

Platforms are broadly classified into three classes:

1. Ground-based platforms
2. Aerial platforms
3. Satellite platforms

14.8.1 Ground-Based Platforms

Remote sensing platforms that position the sensor at the earth's surface are called ground-based platforms. These systems are fixed to the earth and the sensors are often standard tools used to measure environmental conditions such as air temperature, wind characteristics, water salinity and earthquake intensity.

Ground-based sensors can be placed on tall structures such as towers, scaffolding or buildings to elevate the platform (see Fig. 14.9). Ground-based sensors are generally less expensive to operate and maintain than aircraft or satellite sensors, but they do not provide the aerial extent of the airborne platforms. Ground-based sensors are often used to record detailed information about the surface, which is compared with the information collected from aircraft or satellite sensors.

14.8.2 Aerial Platforms

Aerial platforms are mostly sensors mounted on fixed-wing aircraft, though other airborne platforms, such as balloons, rockets and helicopters, can be used. Aircraft are often used to collect very detailed images of the earth's surface and facilitate the collection of data virtually over any portion of the earth's surface at any time. Aerial systems elevate the sensor above the earth's surface to increase its aerial coverage. They also allow researchers to monitor very large areas of the surface, which would be impractical with ground-based sensors or impossible or dangerous to visit.

14.8.3 Satellite Platforms

In space, remote sensing is conducted from satellites. Satellites are launched for remote sensing, communication and telemetry (location and navigation) purposes. Satellites permit repetitive coverage of the earth's surface on a continuing basis because of their designed orbits. Cost is often a significant factor in choosing among the various platform options.

In the early 1960s, researchers started mounting sensors on satellites placed into orbit over the earth and ushered in a new era of environmental remote sensing that continues to grow at a rapid pace today. Satellites can be operated remotely from the ground and data from the satellite sensors must be transmitted to the surface. The communication technology in remote sensing are very complex and very expensive.

Environmental satellites have contributed greatly to our understanding of the earth's environment and continue to be used extensively for remote sensing research. For example, weather satellite technology, one of the first practical applications of satellite remote sensing, has vastly expanded our understanding of the earth's weather by providing a large scale view of our weather systems that was previously impossible. It was only after the advent of satellites that weather patterns such as hurricanes and mid-latitude cyclones were fully understood. Prior to satellites, any knowledge of these storms was collected through ground level observations, which unfortunately, did not provide the necessary information to adequately understand them. The contribution of satellites to our understanding of dangerous weather events has saved countless number of lives since the early 1960s.

14.9 TYPES OF REMOTE SENSING

Remote sensing can be classified into two categories based on the type of energy resources:

1. Passive remote sensing
2. Active remote sensing

Remote sensing can be classified into three categories based on wavelength region:

1. Visible and reflective IR remote sensing
2. Thermal IR remote sensing
3. Microwave remote sensing

The energy source used in the visible and reflective IR remote sensing is the sun. The sun radiates electromagnetic energy with a peak wavelength of 0.5 μm. Remote sensing data obtained in the visible and reflective IR regions mainly depends on the reflectance of objects on the ground surface. Therefore, information about objects can be obtained from the spectral reflectance. However, laser radar is exceptional because it does not use solar energy, but the laser energy of the sensor.

The source of radiant energy used in thermal IR remote sensing is the object itself, because any object with a normal temperature will emit electromagnetic radiation with a peak at about 10 μm, as illustrated in Fig 14.10. One can compare the difference of spectral radiance between the sun (a) and an object with normal earth temperature (about 300°K), as shown in Fig. 14.10. It should be noted that the figure neglects atmospheric absorption, though the spectral curve varies with respect to the reflectance, emittance and temperature of the object.

The curves of (a) and (b) cross at about 3.0 μm. Therefore, in the wavelength region shorter than 3.0 μm, spectral reflectance is mainly observed, while in the region longer than 3.0 μm, thermal

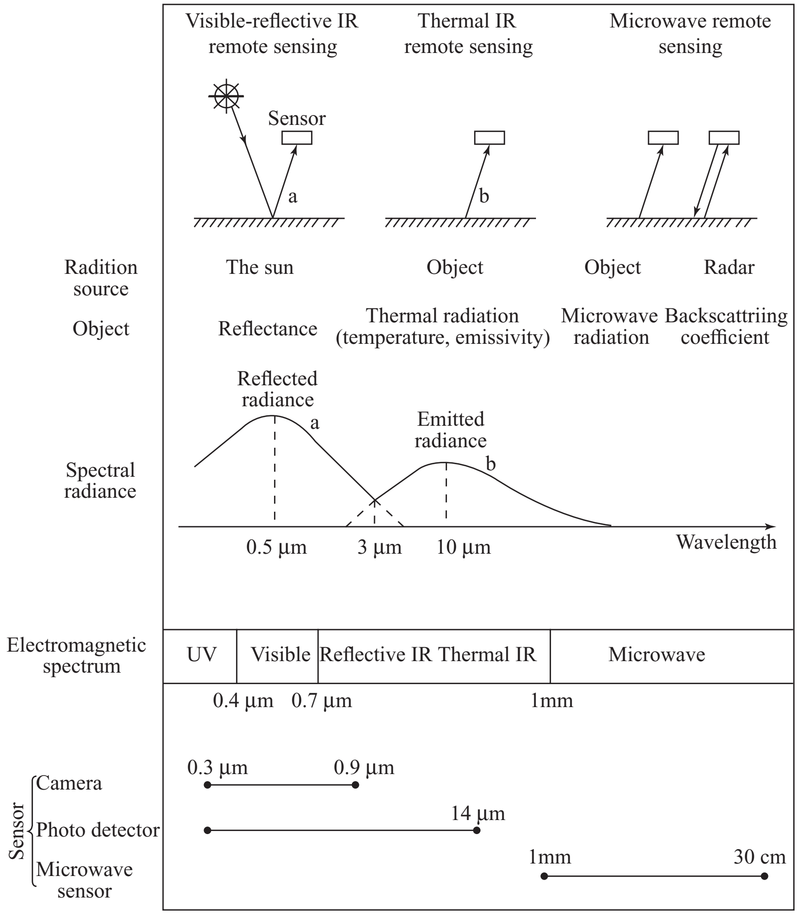


FIGURE 14.10 Different types of remote sensing

radiation is measured. In the microwave region, there are two types of microwave remote sensing, passive microwave remote sensing and active remote sensing. In passive microwave remote sensing, the microwave radiation emitted from an object is detected, while the back scattering coefficient is detected in active microwave remote sensing.

It should be noted that the two curves (a) and (b) in Fig. 14.10 show the black body's spectral radiances of the sun at a temperature of 6,000 K and an object with a temperature of 300 K, without atmospheric absorption.

14.10 PASSIVE REMOTE SENSING

Passive remote sensing makes use of sensors that detect the reflected or emitted electromagnetic radiation from natural sources. Many forms of remote sensing use passive detection, in which sensors measure levels of energy that are naturally emitted, reflected or transmitted by the target object. Passive sensors are those which are designed to detect naturally occurring energy. Most often, the source of radioactive energy is the sun. The sun's energy is either reflected, as it is for visible wavelengths, or absorbed and then re-emitted, as it is for thermal IR wavelengths.

Passive detection can only work when the naturally occurring energy is available. Detection of reflected solar energy, for example, can only proceed when the target is illuminated by the sun; thus, limiting visible light sensors on satellites from being used during a night-time pass. The amount of solar radiation present at polar latitudes is often insufficient for visible light sensors, limiting the use of passive detectors to lower latitudes. Clouds, dust, smoke and other particles in the atmosphere can block reflected energy from reaching a sensor.

The problems associated with passive sensing can be overcome while designing a remote sensing system. One common method is to use a sensor that is capable of detecting radiation in several different portions of the electromagnetic spectrum. For example, by using a combination of visible and thermal IR channels, weather satellites can provide imagery of the earth's cloud patterns during both day and night hours. A combination of visible channels and reflected IR channels can also be used to mathematically correct an image for atmospheric interference that is caused by energy interacting with it and being absorbed by particles in the atmosphere before it reaches a sensor.

14.10.1 Thematic Mapper

Thematic mapper (TM), the primary sensor on the Landsat remote sensing satellite, is a good example of a passive sensor. This sensor has seven bands or channels, each being sensitive to a different range of electromagnetic radiation. The sensors on the thematic mapper are sensitive to narrow portions of the visible and near IR portion of the spectrum, with one band sensitive to thermal IR. The selected range of wavelengths are specifically designed to detect differences in plant production, soil moisture, and mineral content in soils, providing a useful tool in assessing and monitoring land use practices. The sensors depend on available reflected solar energy, so the Landsat satellite is placed into an orbit that ensures that the satellite will pass overhead at the time when the amount of solar radiation is optimal for the sensor.

14.11 ACTIVE REMOTE SENSING

Active remote sensing makes use of sensors that detect reflected responses from objects that are irradiated from artificially generated energy sources, such as radar. Active remote sensing provides its own energy source for illumination of the target. These devices, known as active sensors, direct a burst of radiation at the target and use sensors to measure how the target interacts with the energy. Most often, the sensor detects the reflection of the energy, measuring the angle of reflection or the amount of time it took for the energy to return. Active sensors provide the capability to obtain measurements anytime, regardless of the time of the day or season. They can be used for examining energy types that are not sufficiently provided by the sun, such as microwaves, or to better control the way a target is illuminated in a better manner. However, active systems require the generation of a fairly large amount of energy to adequately illuminate targets. Most active and passive remote sensors employ a detecting or sensing system that scans the subject, a recording system that stores the information received, and an analysis or display system for giving the desired results. Sometimes, a combined analysis and display system is operated concurrently with the sensing system to aid in data gathering and to provide some preliminary information. In that case, recordings are made and later analyzed in the typical manner. The displays are two dimensional, usually composed of many scan lines from the sensor similar to aerial photographs or television pictures.

14.11.1 Doppler Radar

Doppler radar is an example of an active remote sensing technology. A Doppler radar device is a ground-based system that emits radio energy in a radial pattern as the transmitter rotates. A sensor measures the reflection or echoes of this energy off such atmospheric particles as dust, raindrops and even birds. These echoes, when plotted on a regional map, assist a meteorologist in determining the exact location of storm centres, measuring the speeds in the wind field of a storm and notifying the public of areas of potentially severe weather.

Another form of active collection is the atmospheric sounder, which uses various forms of energy, including lasers, microwaves and radar, to take measurements of the density of the atmosphere at certain altitudes, thus providing detailed data about a wide variety of phenomena that includes wind speeds, pollution levels and atmospheric composition. Sounders can be ground-based and measure from the ground up, or they can be mounted on an airborne or satellite platform and measure down through the atmosphere. Data from the sounding equipment can be used to construct three-dimensional models of the state of the atmosphere, and often form the basis of prediction models used to determine future weather patterns.

14.11.2 Precipitation Radar

Precipitation radar is the first space borne instrument designed to provide three-dimensional maps of storm structure. It is flown on the tropical rainfall measuring mission (TRMM) satellite. This is accomplished using narrow beam radar that is transmitted from the satellite through the atmosphere. When the radiation strikes the raindrops in the atmosphere, it is echoed back up to the satellite. The size and height of the raindrops is discerned from the pattern of the returned radar pulses. The measurements have yielded invaluable information on the intensity and distribution of the rain, on the rain type, on the storm depth and on the height at which snow melts into rain. The estimates of the heat released into the atmosphere at different heights based on these measurements can be used to improve models of the global atmospheric circulation.

14.12 THERMAL IR REMOTE SENSING

Any object is potentially a source of electromagnetic energy as a result of its temperature. When the temperature is sufficient to cause the object to emit energy in the form of heat, the wavelengths emitted are predominantly in the thermal IR portion of the spectrum (say, between approximately 3 and 25 μm). At about 3 μm wavelength, more than half the ambient energy is due to the reflectance of incident solar energy, but with increasing wavelength, the energy emitted from the surface progressively becomes the major portion of the ambient energy. As graphic emulsions are no longer suitable for sensing the ambient energy from the surface, other forms of detectors must be used.

Furthermore, due to the differential absorption characteristics of the atmosphere, only certain limited regions within the IR part of the spectrum (the 'windows') are suitable for thermal IR sensing of the earth's surface.

14.12.1 Stefan–Boltzmann Law and Temperature–Energy Relationships

The signal created as a result of recording the energy emitted by the earth's surface features is influenced both by the temperature of the feature and by its relative efficiency as an emitter of heat energy. The total power emitted per unit area is described by the Stefan–Boltzmann law.

$$M = \sigma\varepsilon(T)^4$$

where M is the power emitted per unit area (in W cm^{-2})

σ is the Stenfan–Boltzmann constant

ε is the emissivity of the body, and

T is the absolute temperature in Kelvin (K)

As remote sensing is mostly concerned with comparing the power emitted by different objects or targets, it is the relative power emitted which is important, and so the value of the constant σ may be disregarded.

The emissivity ε of a body is a measure of its performance as a radiator compared with the so-called 'black body', which has an emissivity of 1.0.

$$\text{Emissivity } (\varepsilon) = \frac{\text{Radiant emittance of an object at a given temperature}}{\text{Radiant emittance of a 'black body' at the same temperature}}$$

As a true 'black body' does not exist in reality, all objects have an emissivity of less than 1.0. However, as may be inferred from the above equation, two objects or surfaces at more or less the same temperature may record quite differently if they have significantly different emissivities (eg. ice within emissivity of 0.85). Likewise, two objects with similar emissivity values may produce different signals if they are at different temperature. The final appearance of a thermal IR image should therefore be seen, as being due to a combination of emissivity and temperature.

In general, the energy emitted at thermal IR wavelengths is largely due to the temperature of the object. However, the wavelength of peak emission is closely tied to the specific absolute temperature of the body, and if an object at 'T' different temperature is the primary interest, then the wavelength of peak emission will alter. This relationship is described by Wien's displacement law, which broadly states that as the temperature of the object or source rises, the peak emission shifts to progressively shorter wavelengths.

$$\lambda = \frac{2897}{T(\text{K})}$$

Some examples of the relationship between temperature and Wavelength of peak emission λ_{max} in µm are given in Table 14.2 for some common energy sources.

TABLE 14.2 Relationship between absolute temperature and wavelength of peak emission for some common energy sources

Object (at Surface)	Absolute Temperature (K)	Power at λ_{max} (Wcm^{-2})	Range of Wavelength (µm)		
				λ_{max}	
Sun	6,000 K	7,000	0.3	0.48	
Hot fire	1,160 K	10	0.8	2.5	3.2
Small fire	650 K	1	1.4	4.5	16
Air temperature					29
Earth's data surface	300 K	0.04	3.2	9.7	Infinite

14.13 DETECTORS

For remote sensing using the sun as the prime energy source, a sensing system operating at a wavelength of around 0.48 μm would be the most efficient. As it happens, this falls within the range of visible light (0.4 μm to 0.7 μm), and photographic film is therefore well suited as a detector for recording energy patterns having the sun as the prime source. If the other sources are considered, then photography is not well suited to their wavelengths of peak emission (see Table 14.2). For recording energy emitted by these other sources, different forms of sensors are used, which are more precisely responsive to the wavelengths of peak emission.

For thermal IR sensing, the sensor generally used is some form of crystal detector, which is maintained at a very low temperature. A thermal IR sensing system designed mainly for detecting emissions from surface fires (e.g., in forest fire surveillance) would operate best in the 3–5 μm waveband, using an Indium antimonite crystal detector. However, for recording thermal emissions from other earth surface features (at around 300 K temperature), a crystal sensitive to longer wavelengths (mercury cadmium telluride, 8–14 μm) would be more appropriate, because the peak emissivity for most terrain features is at around 10 μm wavelength (Table 14.3).

TABLE 14.3 Range of sensitivity of crystals commonly used in thermal IR imaging

Type of Crystal	Spectral Range in μm	Crystal Cooled to
Indium antimonide	3–5	–196°C (77 K)
Mercury doped germanium	5–14	–247°C (26 K)
Mercury cadmium telluride	8–14	–196°C (77 K)

The function of the crystal detector is to permit the recording of data about thermal emissive patterns of energy at the earth's surface by transforming the power 'caught' by the sensor at any instant into a signal, which can later be analyzed to allow inferences to be made about the pattern of thermal emissions from the surface. The detectors listed in Table 14.3 have a very rapid response time to the energy incident on them (less than a micro second) and are therefore suitable for use in an airborne imaging system, where the area of the ground being sensed is changing very rapidly.

14.14 THERMAL IR IMAGING

By contrast, thermal IR imaging involves the use of a 'point' detector (the supercooled crystal). To achieve coverage of an area of ground, a scanning mirror is made to rotate across the ground track of the aircraft, such that the thermal emissions from a series of small contiguous ground areas, along a line normal to the flight direction, is allowed to impinge on the crystal detector. The rapid response time of the detector ensures that each of these small ground segments will register as a separate signal on the magnetic tape record of the signals. As the aircraft moves forward, the scanning mirror will trace out a series of parallel narrow swaths on the ground, which eventually, will ensure recording of the thermal emissions for an area of ground below the aircraft, normally within a viewing angle between 90 and 120 degrees. The scan rate of the mirror can be adjusted according to the aircraft's altitude and speed to ensure that there are no gaps or overlaps between the swaths. Such a sensor is called a 'line-scanner', and although described here for thermal inferred sensing, the system is more widely applied as, for example, in a multispectral scanner system.

At any instant, the small angular field of view (FOV) will be receiving the energy emitted by a limited ground area (called the 'ground resolution element' (GRE) or instantaneous field of view (IFOV)).

The scan mirror directs the thermal radiation on to the crystal, and the temperature change induced in the crystal alters the electrical resistance of the detector. Those changes in electrical resistance, from one ground resolution element to the next, are recorded in analog form on magnetic tape. Playback of the magnetic tape can later be used to produce intensity modulations of a light spot on a cathode ray tube (CRT). In this way, the differences in energy emitted from the surface can be portrayed as grey tones on a photographic type of image. This image formation can also be carried out in real-time onboard the aircraft by positioning a film recorder in front of a CRT on which the intensity modulations of the light spot are displayed, a line at a time.

14.15 APPLICATIONS OF THERMAL IR IMAGING

Wherever features can be identified on the basis that their thermal emissions differ from their surroundings, there is a potential for the use of thermal IR imaging. The thermal imaging could be used successfully when the difference between the powers emitted by a feature and its surroundings are greater than the limiting 'power resolution' of the system. The feasibility of using thermal line scans for a particular application can usually be established in advance by means of preliminary calculation. An example of such calculations is necessary where it is required to detect and enumerate warm-blooded animals as part of a census of wildlife. If the animals are more readily counted at night, thermal line scan imaging may be the best method.

14.16 IMAGING WITH MICROWAVE RADAR (MICROWAVE REMOTE SENSING)

In the microwave region of the electromagnetic spectrum, the strength of naturally occurring emissions is very weak. Although these passive microwaves can be recorded using highly sensitive radiometers, the most common imaging system in this part of the spectrum uses active microwave energy in the form of a radar imaging system.

Radar was originally developed for detecting aircraft from the ground. Imaging radar reverses the procedure by having the radar system on the aircraft, with the artificially generated microwave energy directed towards the ground via an antenna.

Although microwaves are about 1,00,000 times longer than wavelengths of visible light, they are still short enough to resolve detail fine enough for most of the earth science applications. Microwaves also exhibit some of the surface penetration characteristics of the much longer radio wavelengths, which compensates to some extent for the perceptibly poorer spatial resolution when compared with aerial photography.

The spatial resolution of radar imagery in fact varies in both the along-track and the cross-track directions. Along-track or azimuthal resolution depends on the radar angular beam width and on the range to the ground, and the cross-track or range resolution depends on the radar pulse length and the ground incidence angle. These complications mean that it is no simple matter to use radar imagery for standard precision mapping tasks.

The most significant improvement in recent years has resulted from the development of synthetic aperture radar (SAR), whereby the azimuthal resolution is effectively made to be half the antenna length and to be independent of the range or altitude. This has facilitated the adoption of imaging radars at very great range on satellite platforms.

As an imaging radar system does not use the waveband of visible light, such a system can operate by day or night. Furthermore, microwaves are generally little affected by suspended particles in the atmosphere, such as water vapour, and to imaging radar has an ability to penetrate cloud wavelengths commonly used in imaging radar.

In practice, an imaging radar system operates by directing a stream of artificially generated microwave pulses towards the ground from an aircraft. The reflection of the microwave energy back to the aircraft depends on such details as feature orientation and slope, surface dielectric properties and surface roughness in relation to the wavelength. Furthermore, as the exact location of features within the radar beam is dependent on range discrimination (i.e., differences in slant distances from aircraft to ground features), it follows that the radar microwave energy should be directed to the side of the aircraft to avoid the range ambiguities which exist with features close to the vertical beneath the aircraft.

14.17 RADIOMETRY AND PHOTOMETRY

In remote sensing, electromagnetic energy reflected or emitted from objects is measured. The measurement is based on either radiometry or photometry, with different technical terms and physical units. Radiometry is used for physical measurement of a wide range of radiation from X-rays to radio waves, while photometry corresponds to the human perception of visible light based on the human eye's sensitivity.

14.18 BLACK BODY RADIATION

An object radiates unique spectral radiant flux depending on the temperature and emissivity of the object. This radiation is called thermal radiation because it mainly depends on temperature. Thermal radiation can be expressed in terms of black body theory. A black body is a matter which absorbs all electromagnetic energy incident upon it and does not reflect or transmit any energy. According to Kirchhoff 's law, the ratio of the radiated energy from an object in thermal static equilibrium to the absorbed energy is constant and dependent only on the wavelength and the temperature T. A black body shows the maximum radiation as compared with other matter. Therefore, a black body is called a perfect radiator. Black body radiation is defined as thermal radiation of a black body, and can be given by Plank's law as a function of temperature T and wavelength as shown in Fig. 14.11.

In remote sensing, a correction for emissivity should be made because normal observed objects are not black bodies. Emissivity can be defined by the following formula:

$$\text{Emissivity} = \frac{\text{Radiant energy of an object}}{\text{Radiant energy of a black body with the same temperature as the object}}$$

Emissivity ranges between 0 and 1 depending on the dielectric constant of the object, surface roughness, temperature, wavelength and look angle. Figure 14.12 shows the spectral emissivity and spectral radiant flux for three objects—a black body, a grey body and a selective radiator.

The temperature of the black body that radiates the same radiant energy as an observed object is called the brightness temperature of the object.

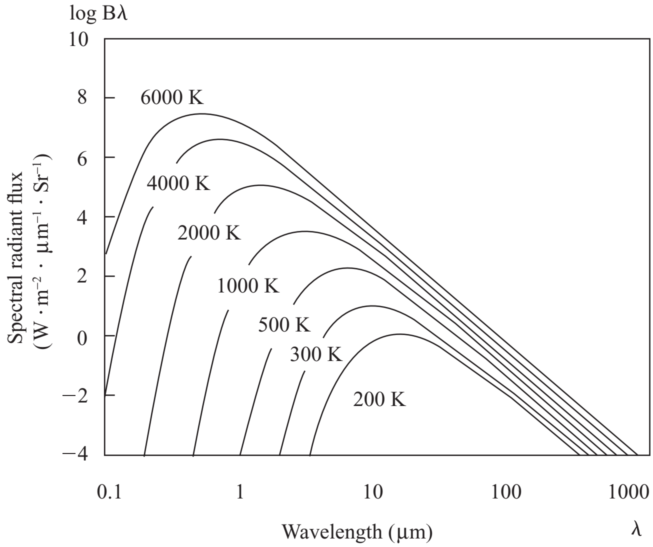


FIGURE 14.11 Plank's law of radiation

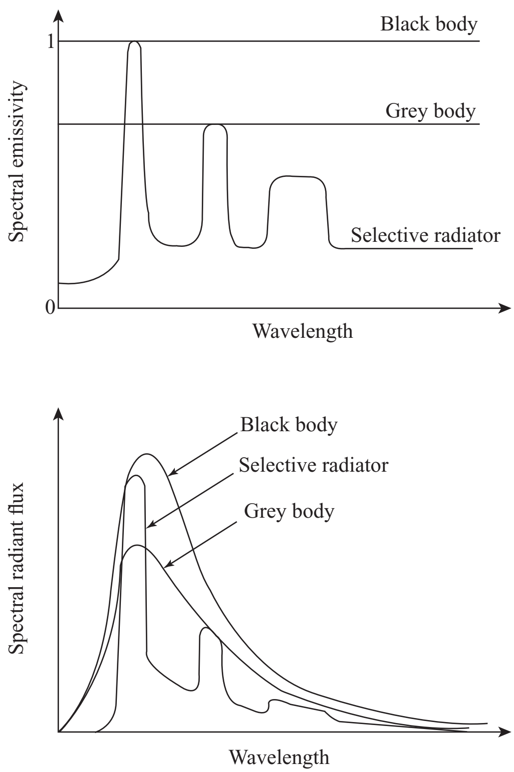


FIGURE 14.12 Radiators

14.19 REFLECTANCE

Reflectance is defined as the ratio of incident flux on a sample surface to reflected flux from the surface as shown in Fig. 14.13. Reflectance ranges from 0 to 1.

Reflectance was originally defined as a ratio of incident flux of white light to reflected flux in a hemispherical direction. The equipment to measure reflectance is known as a spectrometer. The reflectance factor is sometimes used as the ratio of reflected flux from a sample surface to reflected flux from a perfectly diffuse surface (see Fig. 14.13). Reflectance with respect to wavelength is called spectral reflectance (see Fig. 14.14). A basic assumption in remote sensing is that the spectral reflectance is unique and different from object to object.

Reflectance with a specified incident and reflected direction of electromagnetic radiation or light is called directional reflectance. The two directions of incidence and reflection can be straight, conical or hemispherical, making nine possible combinations.

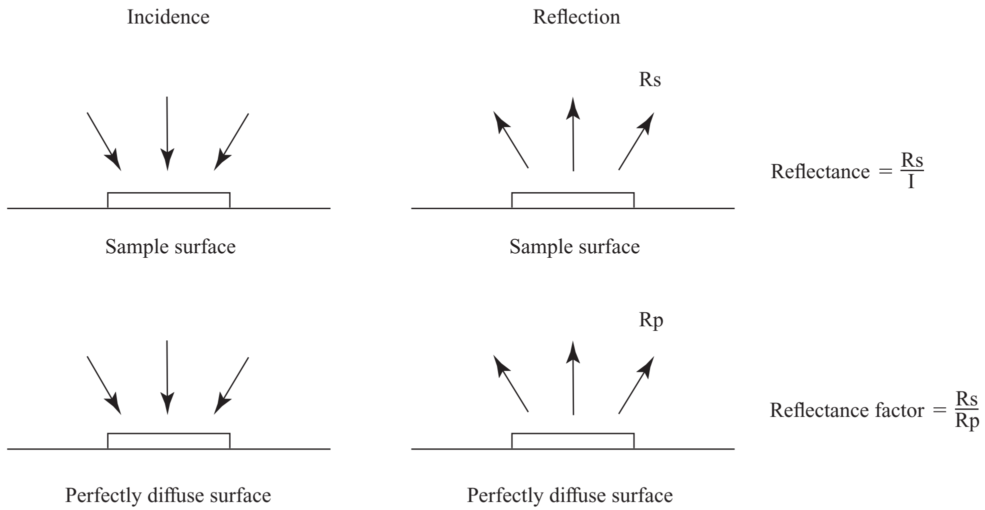


FIGURE 14.13 Reflectance and reflectance factor

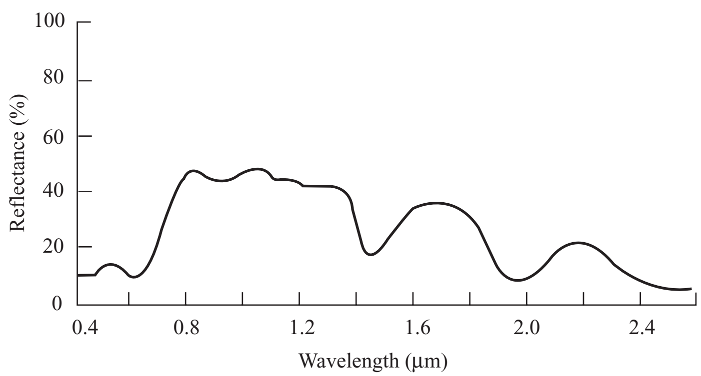


FIGURE 14.14 Spectral reflectance of vegetation

14.20 REMOTE SENSING SYSTEMS

These are two major categories in remote sensing systems. They are:

1. Framing system.
2. Scanning system.

Framing systems instantaneously acquire an image of a large area (or frame) on the terrain. Cameras and vidicons are common examples of such systems. A vidicon camera is a type of television camera that records the image on a photosensitive electronically charged surface called vidicon. As in a photographic camera, radiation is gathered by a lens system, passed through various filters and focussed on a flat target. Instead of a photosensitive emulsion, a vidicon target is coated with a transparent photoconductive material. The electrical conductivity of the target increases with the intensity of the illuminating radiation. An image is built up by scanning the target with an electron beam, which methodically sweeps the area of the target as a series of lines. A colour imaging system involves three vidicons with different filters.

A scanning system employs a single detector with a narrow field of view that is swept across the terrain in a series of parallel scan lines to produce an image.

14.21 SCANNER

Many electro-optical remote sensors acquire data using scanning systems, which employ a sensor with a narrow field of view that sweeps over the terrain with the motion of the platform to build up and produce a two-dimensional image of the surface. Scanners are different from cameras that are framing systems and acquire a near instantaneous snapshot of an area of the surface. Scanning systems can be used on both aircraft and satellite platforms, and have essentially the same operating principles.

A scanning system used to collect data over a variety of different wavelength ranges is called a multispectral scanner (MSS), and is the most commonly used scanning system. There are two main modes of scanning employed to acquire multispectral image data—across-track scanning, and along-track scanning.

14.21.1 Across–Track (Whiskbroom) Scanners

Using a rotating mirror, this system scans the terrain along scan lines that are perpendicular to the direction of motion of the sensor platform. This allows the scanner to repeatedly measure the energy from one side of the satellite platform to the other. As the platform moves forward over the earth, successive scans build up a two-dimensional image of the earth's surface.

A direct grating is used to separate thermal and non-thermal energy, and a prism is used to split the non-thermal energy into continuous UV, visible and near IR wavelength components. A bank of internal detectors, each sensitive to a specific range of wavelengths, detects and measures the energy for each spectral band and then, as an electrical signal, they are converted to digital data and recorded for subsequent computer processing. An array of electro-optical detectors is located on the focal plane. The system's IFOV (instantaneous field of view) and flight height determine the ground spatial resolution. The angular field of view is the sweep of the mirror, measured in degrees, used to record a scan line and determines the width of the imaged swath.

14.21.2 Along-Track (Push-Broom) Scanners

In this system, there is no rotating scanning mirror. Instead, a linear array of detectors located at the focal plane of the image scan along the flight track direction. These systems are also referred to as push-broom scanners, as the motion of the detector array is analogous to the bristles of a broom being pushed along a floor. Linear arrays consist of numerous charge-coupled devices (CCDs). Each individual detector measures the energy for a single ground resolution cell along any given scan line, and thus the size and the IFOV of the detectors determine the spatial resolution of the system.

A separate linear array is required to measure each spectral band or channel. For each scan line, the energy detected by each detector of each linear array is sampled electronically and recorded digitally. The linear array also uses the forward motion of the platform to record successive scan lines and build up a two-dimensional image, perpendicular to the flight direction.

14.22 MULTISPECTRAL SCANNER

A multispectral scanner (MSS) is an opto-mechanical device where the image is recorded in a number of spectral bands. MSS is a passive device, which records data in exact spatial and temporal registration. It essentially combines the capabilities of an imaging scanner and a spectro-radiometer in one instrument. The Space Application Center in Ahmedabad has developed an MSS, which was extensively used during the period from October 1977 to April 1978 for carrying out aerial survey flights all over the country. The SAC-MSS essentially consists of a scan mirror, collecting optics, dispersive system and the detector. The five spectral bands of the SAC-MSS have a direct correspondence to the four Landsat MSS bands and a thermal band.

14.23 ELECTRO-OPTICAL SENSORS

Electro-optical sensors use non-film detectors. In contrast, photographic cameras record radiation reflected from a ground scene directly onto film. Electro-optical detectors record the reflected and emitted radiation from a ground scene as analog electrical signals, which are converted into the image DN values.

14.24 SIGNATURE

The term signature is applied in remote sensing to any identifying feature that appears in the analysis or display process through which a desired subject can be positively identified against what may be a complex background or surroundings. For example, it may be necessary to identify a particular mineral, crop blight, or type of air or water pollution. Signature, as applied to imagery, usually refers to visual characteristics that identify the subject and separate it from other similar objects. However, other types of signatures may be much more complex, requiring spectral analysis or other techniques. The response of the material on the earth surface to incident radiation and the energy emitted by all objects as a function of their temperature and structure (emittance) essentially determines all signatures. The determination of signatures implies establishing a relationship between the spectral response pattern and the inherent characteristics of an object.

The knowledge of signatures is essential for exploiting the potential of remote sensing techniques. This knowledge enables us to identify and classify objects and is required for interpretation of all remotely sensed data, whether the interpretations are carried out visually or by computers. It also helps to specify the requirements for any remote sensing mission, like the optimal wavelength band to be used, and identify the type of sensor which will be best suited for the task. Signature data are also necessary for analyzing and designing sensor systems for specific applications.

Any set of observable characteristics, which directly or indirectly leads to the identification of an object or its condition, is termed as signature. The spectral signature of any object or its condition comprises a set of values for its reflectance and emittance in different spectral bands or regions. Spectral signatures are not completely deterministic. They are statistical in nature. There are four principal characteristics of electromagnetic radiation, which can be exploited for providing signatures, which characterize the nature and the condition of the object. They are:

1. Spectral resolutions,
2. Spatial resolutions,
3. Temporal resolutions,
4. Polarization resolutions.

14.25 RESOLUTION AND GRE

In the context of remote sensing, there are several *resolution*s such as image spatial, spectral, radiometric and temporal resolutions.

Image Resolution Image resolution defines the ability of a sensor to distinguish between spatial characteristics of objects on the earth's surface. Resolution can change due to sensor design, detector size, focal length, satellite altitude and time.

Spatial Resolution Spatial resolution indicates the physical size of the smallest feature or the closest separation of two features, which can be distinguished by the imaging system. Systems, which operate at shorter wavelengths, produce better spatial resolution than those operating at longer wavelengths. Spatial resolution in the reflectance and emittance are attributes of the shape, size and texture of objects. The camera-film system generally produces the best spatial resolution.

The GRE: GRE is related to the wavelength, the diameter of the energy collector and the sensor altitude and is given by the following expression:

$$\text{GRE} = \frac{1.2 \times \text{wavelength } (\lambda) \times \text{altitude}}{\text{collector diameter}}$$

Most remote sensing images are composed of a matrix of picture elements, or pixels, which are the smallest units of an image. Image pixels are normally square and represent a certain area on an image. It is important to distinguish between pixel size and spatial resolution—they are not interchangeable. If a sensor has a spatial resolution of 20 m and an image from that sensor is displayed at full resolution, each pixel represents an area of 20 m × 20 m on the ground. In this case, the pixel size and resolution are the same. However, it is possible to display an image with a pixel size different from the resolution. The spatial reflectance of a crop within a given wavelength band, will depend upon the size and shape of the plants, their spatial frequency, field size and the scene background. A photograph of a large cotton field taken from an aircraft at a high altitude will show an entirely different texture from the photograph taken from a low orbital altitude. The vertical photograph will also show different texture than an oblique photograph.

The ratio of distance on an image or map, to actual ground distance is referred to as scale. If you had a map with a scale of 1:100,000 an object of 1 cm length on the map would actually be an object 1,00,000 cm (1 km) long on the ground. Maps or images with small 'map-to-ground ratios' are referred to as small scale (e.g., 1:1,00,000), and those with larger ratios (e.g., 1:5,000) are called large scale.

Spectral Resolution Spectral refers to the number and width of the segments or bands of the electromagnetic spectrum, which are covered by a sensing system. Standard black and white photography, for example, covers the range of visible light (say, 400–700 nm wavelength) in one bank, and, therefore, has a relatively poor spectral resolution. Normal or true colour photography, which consists of a three-layer emulsion, effectively splits the same range of wavelengths into three spectral bands, such as blue sensitive (400–500 nm), green sensitive (500–600 nm) and red sensitive (600–700 nm).

Spectral resolution describes the ability of a sensor to define fine wavelength intervals. The finer the spectral resolution, the narrower the wavelengths range for a particular channel or band. Spectral response and spectral emissivity curves characterizes the reflectance and/or emittance of a feature or target over a variety of wavelengths (see Fig. 14.15). Different classes of features and details in an image can often be distinguished by comparing their responses over distinct wavelength ranges. Black and white film record wavelengths extending over much, or all of the visible portion of the electromagnetic spectrum. Its spectral resolution is fairly coarse, as the various wavelengths of the visible spectrum are not individually distinguished, and the overall reflectance in the entire visible portion is recorded. Colour film is also sensitive to the reflected energy over the visible portion of the spectrum, but has higher spectral resolution, as it is individually sensitive to the reflected energy at the blue, green and red wavelengths of the spectrum. Thus, it can represent features of various colours based on their reflectance in each of these distinct wavelength ranges.

By increasing the spectral resolution of a system, the potential of the system to discriminate between features is improved. Features, which may have rather similar reflectance over a broad band, may differ in detail if the spectral interval of sensing is narrowed. The use of several bands of the spectrum in conjunction is referred to as multispectral sensing. Advanced multispectral sensors called

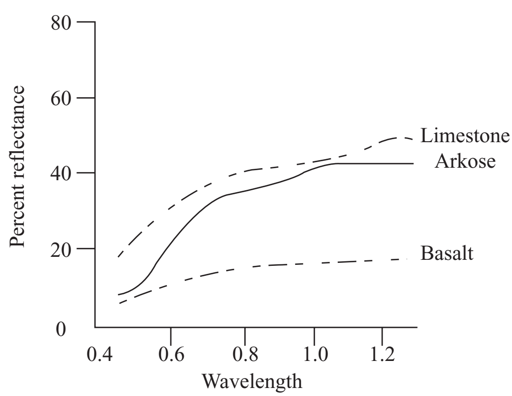


FIGURE 14.15 Spectral emissivity curve of various substances

hyperspectral sensors, detect hundreds of very narrow spectral bands throughout the visible, near IR and mid- IR portions of the electromagnetic spectrum. Their very high spectral resolution facilitates fine discrimination between different targets, based on their spectral response in each of the narrow bands.

Radiometric Resolution The radiometric characteristics describe the actual information content in an image. Every time an image is acquired on film or by a sensor, its sensitivity to the magnitude of the electromagnetic energy determines the radiometric resolution. The radiometric resolution of an imaging system describes its ability to discriminate very slight differences in energy. The finer the radiometric resolution of a sensor, the more sensitive it is for detecting small differences in reflected or emitted energy.

Imagery data are represented by positive digital numbers, which vary from 0 to (one less than) a selected power of 2. This range corresponds to the number of bits used for coding numbers in binary format. Each bit records an exponent of power 2 (e.g., 1 bit = $2^1 = 2$). The maximum number of brightness levels available depends on the number of bits used in representing the energy recorded. Thus, if a sensor used 8 bits to record the data, there would be $2^8 = 256$ digital values available, ranging from 0 to 255. However, if only 4 bits were used, then only $2^4 = 16$ values ranging from 0 to 15 would be available, and hence the radiometric resolution would be much less. Image data are generally displayed in a range of grey tones, with black representing a digital number of 0 and white representing the maximum value (e.g., 255 in 8-bit data). By comparing a 2-bit image with an 8-bit image, we can see that there is a large difference in the level of detail discernible, depending on their radiometric resolutions.

In the case of panchromatic photography, radiometric resolution refers to the number of different grey levels, which make up the tonal range of the photography. With non-photographic systems, radiometric resolution refers to the number of steps or intervals on the scale of signal values.

Polarization Resolution Polarization resolution can be introduced by an object in the radiation reflected or emitted by it, causing the electric vector of the electromagnetic radiation to lie preferentially along one direction. The resultant radiation is thus polarized and is characteristic of the object, and hence can help in distinguishing the object.

Temporal Resolution Temporal resolution is indicated by the time interval between successive overpasses of the sensor when the imaging is repeated. Hence temporal resolutions are the variations in reflectivity or emissivity with time. The revisit period refers to the length of time it takes for a satellite to complete one entire orbit cycle. The revisit period of a satellite sensor is usually several days. Therefore, the absolute temporal resolution of a remote sensing system to image the exact same area at the same viewing angle a second time, is equal to this period. The use of repeat coverage may be necessary when the phenomena of interest undergo significant changes with the passage of time. These changes may be significant in themselves, as for example, in monitoring the movement of potential fishing grounds via the detection of phytoplankton blooms, or the freezing and thawing of a commercially significant waterway. The changes may also form part of a pattern, which when taken together over a significant time-span, may improve the accuracy of identifying certain features, as for example, the identification of agricultural crops. The temporal resolution required will vary according to the application, and the user must establish what would be appropriate in each case. With airborne imaging, each repeat mission incurs more or less the same costs, and so the costs of introducing temporal resolution are high. In the case of satellite imaging, although the initial costs are high, there is automatic repeat coverage due to the orbital mode of operation. The temporal resolution, or repeat period, will depend upon the orbital characteristics of the satellite.

The time factor in imaging is important when:

1. Persistent clouds offer limited clear views of the earth's surface (often in the tropics)
2. Short-lived phenomena (floods, oil slicks, etc.) need to be imaged
3. Multi-temporal comparisons are required (e.g., the spread of a forest disease from 1 year to the next)
4. The changing appearance of a feature over time can be used to distinguish it from near-similar features (wheat/maize).

14.26 PIXEL SIZE AND SCALE

Most remote sensig images are composed of a matrix of picture elements, or *pixels,* which are the smallest units of an image. Image pixels are normally square and represent a certain area on an image. It is important to distinguish between pixel size and spatial resolution—they are not interchangeable. If a sensor has a spatial resolution of 20 m and an image from that sensor is displayed at full resolution, each pixel represents an area of 20 m × 20 m on the ground. In this case, the pixel size and resolution are the same. The ratio of distance on an image or map to the actual ground distance is referred to as scale. In a map with a scale of 1:1,00,000, an object of 1cm length on the map would actually be a 1,00,000 cm (1 km) long object on the ground. Maps or images with small map-to-ground ratios are referred to as small scale (e.g., 1:1,00,000), and those with larger ratios (e.g., 1:5,000) are called large scale.

14.27 SATELLITE ORBITAL CHARACTERISTICS, AND SWATHS

The remote sensing instruments can be placed on a variety of platforms to view and image targets. Although ground-based and aircraft platforms may be used, satellites provide a great deal of the remote sensing imagery commonly used today. Satellites have several unique characteristics, which make them particularly useful for remote sensing of the earth's surface.

The path followed by a satellite is referred to as its orbit. Satellite orbits are matched to the capability and objective of the sensors they carry. Orbit selection can vary in terms of altitude of the satellite and their orientation and rotation relative to the earth. Satellites at very high altitudes, which view the same portion of the earth's surface at all times are geostationary satellites and have geostationary orbits. These geostationary satellites, at altitudes of approximately 36,000 km from the earth's surface, revolve at speeds which match the rotation of the earth, and so they seem stationary relative to the earth's surface. This allows the satellites to observe and collect information continuously over specific areas. Weather and communications satellites commonly have these types of orbits. Due to their high altitude, some geostationary weather satellites can monitor weather and cloud patterns covering an entire hemisphere of the earth.

Many remote sensing platforms are designed to follow an orbit, basically north–south which, in conjunction with the earth's west–east rotation, allows them to cover most of the earth's surface over a certain period of time. These are near-polar orbits, so named for the inclination of the orbit relative to a line running between the North and South poles. Many of these satellite orbits are also sun-synchronous, such that they cover each area of the world at a constant local time of day called the local sun time. At any given latitude, the position of the sun in the sky as the satellite passes overhead will be the same within the same season. This ensures consistent illumination conditions when acquiring

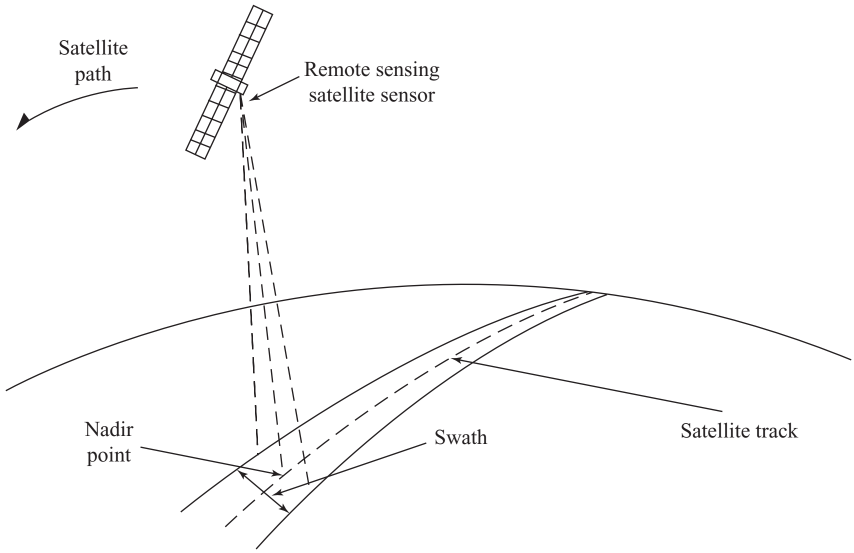


FIGURE 14.16 The swath

images in a specific season over successive years, or over a particular area over a series of days. This is an important factor for monitoring changes between images or for making mosaics of adjacent images together, as they do not have to be corrected for different illumination conditions.

Most of the remote sensing satellite platforms today are in near-polar orbits, which means that the satellite travels northwards on one side of the earth and then towards the southern pole on the second half of its orbit. These are called ascending and descending passes, respectively. If the orbit is also sun-synchronous, the ascending pass is most likely to be on the shadowed side of the earth, while the descending pass is on the sunlit side. Sensors recording reflected solar energy only image the surface on a descending pass, when solar illumination is available. Active sensors, which provide their own illumination, or passive sensors that record emitted (e.g., thermal) radiation can also image the surface on ascending passes.

As a satellite revolves around the earth, the sensor sees a certain portion of the earth's surface. The area imaged on the surface is referred to as the *swath* (see Fig. 14.16). Imaging swaths for space-borne sensors are generally between tens and hundreds of kilometres wide. As the satellite orbits the earth from pole to pole, its east–west position wouldn't change if the earth didn't rotate. However, as seen from the earth, it seems that the satellite is shifting westward because the earth is rotating from west to east beneath it. This apparent movement allows the satellite swath to cover a new area with each consecutive pass. The satellite's orbit and the rotation of the earth work together to allow complete coverage of the earth's surface, after it has completed one complete cycle of orbits. In near-polar orbits, areas at high latitudes will be imaged more frequently than the equatorial zone due to the increasing overlap in adjacent swaths as the orbit paths come closer together near the poles.

14.28 INSTANTANEOUS FIELD OF VIEW

For some remote sensing instruments, the distance between the target being imaged and the platform plays a large role in determining the information obtained and the total area imaged by the sensor.

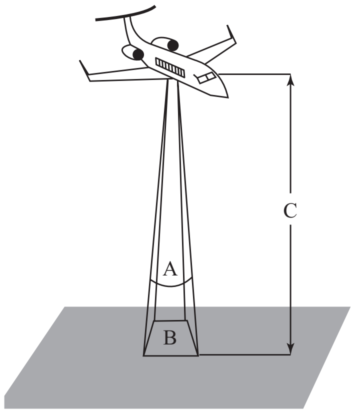


FIGURE 14.17 Instantaneous field of view

Sensors onboard platforms far away from their targets, typically view a larger area, but cannot provide much detail.

The detail discernible in an image is dependent on the spatial resolution of the sensor and refers to the size of the smallest possible feature that can be detected. Spatial resolution of passive sensors depends primarily on their *Instantaneous Field of View* (IFOV). The IFOV is the angular cone of visibility of the sensor (refer A, in Fig. 14.17), and determines the area on the earth's surface that is seen from a given altitude at one particular moment of time (refer B, in Fig. 14.17). The size of the area viewed is deter-mined by multiplying the IFOV by the distance from the ground to the sensor (refer C, in Fig. 14.17). This area on the ground is called the resolution cell, and determines a sensor's maximum spatial resolution. For a homogeneous feature to be detected, its size generally has to be equal to or larger than the resolution cell. If the feature is smaller than this, it cannot be detectable, as the average brightness of all features in that resolution cell will be recorded. However, smaller features may sometimes be detectable if their reflectance dominates within a particular resolution cell, allowing sub-pixel or resolution cell detection.

14.29 MAJOR SATELLITE PROGRAMMES

Some major satellite programmes delivering images used in agriculture today include the following:

1. Landsat 5 uses a thematic sensor (TM—thematic mapper), which operates in seven bands with a resolution of 30 m, except thermal IR, which has a resolution of 120 m. Space Imaging EOSAT of Thornton, Colorado, is the exclusive distributor of Landsat images.
2. Spot 1, 2, 3, and 4 use high-resolution visible (HRV) sensors that operate in four bands with a resolution of 10 m panchromatic and 20 m multispectral. Spot Image, headquartered in Toulouse, France, distributes spot images.
3. Indian Remote Sensing (IRS)-1C uses three sensors, namely, the LISS-III with 23 m resolution in four spectral bands, a panchromatic sensor with 5.8 m resolution, and a wide field sensor (WiFS) with 188 m resolution. Space Imaging EOSAT of Thornton, Colorado, distributes IRS images under an exclusive license from ANTRIX Corp. Ltd. of India, the commercial marketing company of the Indian Space Research Organization.

14.30 WEATHER MONITORING SATELLITE SENSORS

Weather monitoring satellites use sensors which have fairly coarse spatial resolution and provide large aerial coverage (see Fig. 14.18). Weather monitoring and forecasting was one of the first civilian applications of satellite remote sensing, dating back to the first true weather satellite, TIROS-1 (Television and Infrared Observation Satellite-1), launched in 1960 by the United States. Several other weather satellites were launched over the next 5 years, in near-polar orbits, providing repetitive coverage of global weather patterns. In 1966, NASA (the U.S. National Aeronautics and Space Administration) launched the geostationary Applications Technology Satellite (ATS-1), which provided hemispheric images of the earth's surface and cloud cover every half hour. For the first time, the development and movement of weather systems could be routinely monitored. Today, several countries, including India, operate weather or meteorological satellites to monitor weather conditions. Their temporal resolutions are generally quite high, providing frequent observations of the earth's surface, atmospheric moisture and cloud cover, which allows for near-continuous monitoring of global weather conditions, and hence, forecasting.

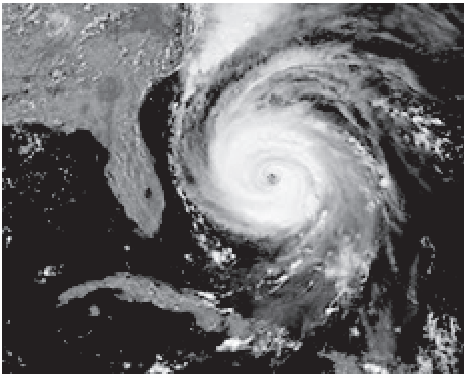


FIGURE 14.18 A weather monitoring satellite imagery

14.31 THE PRINCIPLE STEPS USED IN REMOTELY SENSED DATA ANALYSIS

The principle steps used to analyze all remotely sensed data are:

Definition of Information Needs The objective of using remotely sensed data is to generate information. Before the data acquisition or analysis begins, the information needed has to be defined. Only then can the techniques be identified that might best satisfy the requirement. This assessment should take into account such factors as the accuracy needed, how quickly it is needed, in what time period should the information have been collected (season), the cost to produce it, and the form (soft copy, paper map, statistics) needed.

Collection of Data Using Remote Sensing and Other Techniques Remotely sensed data are rarely used as the sole data source. Field observations and measurements, as well as existing information such as maps and reports are used together in the analysis. The data requirements must be defined, the available data assessed, and then the new data to be collected must be specified. The remote sensing data specifications should be planned and integrated with the other data collection activities.

Data Analysis There are three principal types of analyses applied to remotely sensed data. They are measurement, classification and estimation.

Measurement analyses use the values measured by the sensor to calculate environmental conditions like surface temperature, soil moisture, quantity of plant material or the condition of crops.

Classification analyses define regions that have the same characteristics. These results are commonly provided in the form of a map-like image, where regions with the same characteristics are shown colour-coded.

Estimation analyses are commonly applied to classification results. The objective of this type of analysis is to estimate the quantity of a material, such as the quantity of timber or the quantity of wheat for each management area.

Verification of Analysis Results To use information effectively, we need to know something about its accuracy. Information need not be 100% correct for it to be useful, so long as the expected level of accuracy is known and is taken into account when the information is used. The results of remote sensing analyses should be accompanied by a report on the quality of the data. This involves a test of the results produced in the previous analysis step, in order to verify that the results are of sufficient quality to be accepted for use. The accuracy of maps may be assessed using a formal accuracy test, such as the way topographic maps are tested by selecting sample points on the map and comparing them with independent measurements like field survey data. The amount of error is then calculated for each sample point, and if the overall errors are high, the map is rejected.

Reporting of Results to those who will use the Information Once the quality of the information has been assessed and found to be acceptable, the information can then be assembled into a suitable reporting format (paper, softcopy, report with diagrams). The format should convey the required information and take into account the way the data will be used. The map scale, projection, and units of measurement should be selected for compatibility with the format of other information with which it might be used.

Taking Action Based on the Information The objective of producing information is for decision-making. Even a decision not to take action is a decision. If information is being produced and not used, it is generally because there is no user, the information never gets to the users, or the information is not in a suitable format. As with any other service, when there is no clearly defined customer to be served and when there is no systematic assessment of the quality of the information produced, the information production tends to become self-perpetuating and the quality of the information tends to decline. Nothing promotes service and quality like having the producers of information answerable to their customers.

14.32 DATA RECEPTION, TRANSMISSION, AND PROCESSING

Data obtained during airborne remote sensing missions can be retrieved once the aircraft lands. It can then be processed and delivered to the end user. However, data acquired from satellite platforms need to be electronically transmitted to earth, because the satellite continues to stay in orbit during its operational lifetime. The technologies designed to accomplish this can also be used by an aerial platform if the data are urgently needed on the surface.

There are three main options for transmitting data acquired by satellites to the surface. The data can be directly transmitted to earth if a ground receiving station (GRS) is in the line of sight of the remote sensing satellite. If it is not so, the data can be recorded on board the satellite for transmission to a GRS at a later time. Data can also be relayed to the GRS through the tracking and data relay satellite system (TDRSS), which consists of a series of communications satellites in geo-synchronous orbit.

The data is transmitted from one satellite to another until it reaches the appropriate GRS. The data is received at the GRS in a raw digital format. It may then, if required, be processed to correct systematic, geometric and atmospheric distortions to the imagery, and be translated into a standardized format. The data is written to some form of storage medium such as tape, disk or CD. The data is typically archived at most receiving and processing stations.

Government agencies, as well as commercial companies responsible for each sensor's archives, manage the full libraries of data. For many sensors, it is possible to provide customers with quick-turnaround imagery when they need data as quickly as possible after it is collected. Near real-time processing systems are used to produce low-resolution imagery in hard copy or soft copy (digital) format within hours of data acquisition. Such imagery can then be faxed or transmitted digitally to end-users. One application of this type of fast data processing is to provide imagery to ships sailing in the Arctic, as it allows them to assess current ice conditions quickly in order to make navigation decisions about the easiest/safest routes through the ice. Real-time processing of imagery in airborne systems has been used, for example, to pass thermal IR imagery to forest fire fighters right at the scene. Low resolution quick-look imagery is used to preview archived imagery prior to purchase. The spatial and radiometric quality of these types of data products is degraded, but they are useful for ensuring.

14.33 INTERPRETATION AND ANALYSIS

To get the desired result from the remote sensed data, the meaningful information from the imagery must be extracted. Interpretation and analysis of remote sensing imagery involves the identification and measurement of various targets in an image in order to extract useful information about them.

Targets in remote sensing images may be any feature or object that can be observed in an image. Targets may be a point, line or area feature. This means that they can have any form, from a tower, a bridge or roadway, to a large water body or a field. The prime factor of a target is that it must be distinguishable and it must contrast with other features around it in the image. Interpretation and identification of targets in remote sensing imagery is done manually. In many cases, this is done using imagery displayed in a pictorial or photograph-type format, independent of what type of sensor was used to collect the data and how the data was collected. Remote sensing images can also be represented in a computer as arrays of pixels, with each pixel corresponding to a digital number, representing the brightness level of that pixel in the image. In this case, the data must be in a digital format. Examining digital imagery displayed on a computer screen may also perform visual interpretation. Both analog and digital imagery can be displayed as black and white or monochrome images, or as colour images by combining different channels or bands representing different wavelengths.

Digital processing can be used to enhance data as a prelude to visual interpretation. When remote sensing data is available in digital format, digital processing and analysis may be performed using a computer. Digital processing and analysis are carried out automatically to identify targets and extract information without intervention by a human interpreter. But digital processing and analysis is rarely carried out as a complete replacement for manual interpretation. Digital processing and analysis is more recent, that is, since the advent of digital recording of remote sensing data and the development of computers.

14.33.1 Manual and Digital Interpretation

Both manual and digital techniques for interpretation of remote sensing data have their respective advantages and disadvantages. Generally, manual interpretation requires less specialized equipment, while digital analysis requires specialized and expensive equipment. Manual interpretation is often limited to analyzing only a single channel of data or a single image at a time due to the difficulty in performing visual interpretation with multiple images. The computer environment is more amenable to handling complex images of several or many channels or from several dates. In this sense, digital analysis is useful for simultaneous analysis of many spectral bands and can process large data sets much faster than a human interpreter. Manual interpretation is a subjective process in which the results will vary with different interpreters. Digital analysis is based on the manipulation of digital numbers in a computer and is thus more objective, generally resulting in more consistent results. However, determining the validity and accuracy of the results from digital processing can be difficult. In most cases, a mix of visual and digital methods is usually employed when analyzing imagery.

14.34 ELEMENTS OF VISUAL INTERPRETATION

Targets may be defined in terms of the way they reflect or emit radiation. This radiation is measured and recorded by a sensor, and is ultimately depicted as an image product such as an air photo or a satellite image. Analysis of remote sensing imagery involves the identification of various targets in an image, and those targets may be environmental or artificial features, which consist of points, lines or areas. It is easy to get a good sense of depth when viewing a two-dimensional image viewed stereoscopically to simulate the third dimension of height. Viewing objects from directly above also provides a very different perspective.

Combining an unfamiliar perspective with a very different scale and lack of recognizable detail can make even the most familiar object unrecognizable in an image. Recognizing targets is the key to interpretation and information extraction. Observing the differences between targets and their backgrounds involves comparing different targets based on any, or all, of the visual elements of tone, shape, size, pattern, texture, shadow and association. Identifying targets in remotely sensed images based on these visual elements allows further interpretation and analysis.

Tone Tone refers to the relative brightness or colour of objects in an image. Generally, tone is the fundamental element for distinguishing between different targets or features. Variations in tone also allows the elements of shape, texture and pattern of objects to be distinguished (see Fig. 14.19).

Shape Shape refers to the general form, structure or outline of individual objects (see Fig. 14.20). Shape can be a very distinctive clue for interpretation. Straight edge shapes typically represent urban or agricultural targets, while natural features, such as forest edges, are generally more irregular in shape, except where man has created a road or clear cuts. Farms irrigated by rotating sprinkler systems would appear as circular shapes.

FIGURE 14.19 Variation of tone in an image

FIGURE 14.20 Identification of structures with respect to its shape from an image (Airport)

Size Size of objects in an image is a function of scale. It is important to assess the size of a target relative to other objects in a scene, as well as the absolute size, to aid in the interpretation of that target (see Fig. 14.21). A quick approximation of target size can make the direct interpretation to an appropriate result more quickly. For example, if an interpreter had to distinguish zones of land use, and had identified an area with a number of buildings in it, then the large buildings such as factories or warehouses would suggest commercial property, whereas small buildings would indicate residential use.

FIGURE 14.21 Size of the objects is assessed from the image (road, buildings and prominent structures)

FIGURE 14.22 Identification of pattern from an image

Pattern Pattern refers to the spatial arrangement of visibly discernible objects (see Fig. 14.22). Typically, an orderly repetition of similar tones and textures will produce a distinctive and ultimately recognizable pattern. Orchards with evenly spaced trees, and urban streets with regularly spaced houses are good examples of pattern.

Texture Texture refers to the arrangement and frequency of tonal variation in particular areas of an image (see Fig. 14.23). Rough textures would consist of a mottled tone where the grey levels change abruptly in a small area, whereas smooth textures would have very little tonal variation. Smooth textures are most often the result of uniform, even surfaces, such as fields, asphalt or grasslands. A target with a rough surface and irregular structure, such as a forest canopy, results in a rough textured appearance. Texture is one of the most important elements for distinguishing features in radar imagery.

Shadow Shadow is also helpful in interpretation, as it may provide an idea of the profile and relative height of a target or targets which may make identification easier (see Fig. 14.24). However, shadows can also reduce or eliminate interpretation in their area of influence, because targets within shadows are much less (or not at all) discernible from their surroundings. Shadow is also useful for enhancing or identifying topography and landforms, particularly in radar imagery.

Association Association takes into account the relationship between other recognizable objects or features in proximity to the target of interest. The identification of features that one would expect to associate with other features may provide information to facilitate identification. In the example

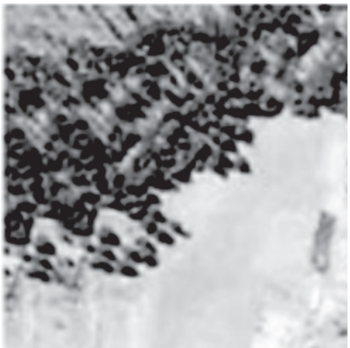


FIGURE 14.23 Texture variation in an image

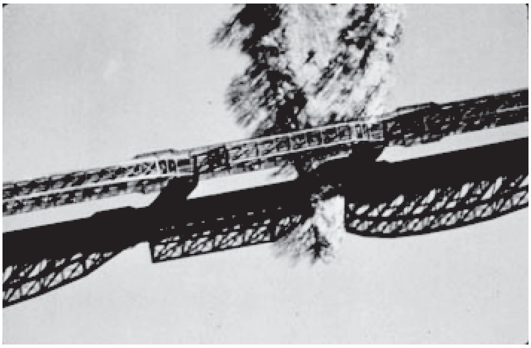


FIGURE 14.24 Shadowing in an image

FIGURE 14.25 Association in an image

given in Fig. 14.25, commercial properties may be associated with proximity to major transportation routes, whereas residential areas would be associated with schools, playgrounds and sports fields. In Fig. 14.25, a lake is associated with boats, a marina and adjacent recreational land.

14.35 DIGITAL IMAGE PROCESSING

Digital image processing involves numerous procedures including formatting and correcting of the data, digital enhancement to facilitate better visual interpretation, or even automated classification of targets and features entirely by computer. In order to process remote sensing imagery digitally, the data must be recorded and available in a digital form suitable for storage on a computer tape or disk. The main requirement for digital image processing is a computer system, known as an image analysis system, with the appropriate hardware and software to process the data. Many commercially available software systems have been developed specifically for remote sensing image processing and analysis. Most of the common image processing functions available in image analysis systems can be grouped into four categories. They are:

1. Preprocessing
2. Image enhancement
3. Image transformation
4. Image classification and analysis

Preprocessing functions involve operations that are normally required prior to the main data analysis and extraction of information, and are generally grouped as radiometric or geometric corrections. Radiometric corrections include correcting the data for sensor irregularities and unwanted sensor or atmospheric noise, and converting the data so that it accurately represents the reflected or emitted radiation measured by the sensor. Geometric corrections include correcting for geometric distortions due to sensor-earth geometry variations, and conversion of the data to real world coordinates like latitude and longitude on the earth's surface.

The objective of image enhancement is to improve the appearance of the imagery to assist in visual interpretation and analysis. Examples of enhancement functions include contrast stretching to increase the tonal distinction between various features in a scene, and spatial filtering to enhance or suppress specific spatial patterns in an image.

Image transformations usually involve combined processing of data from multiple spectral bands. Image transformations are operations similar in concept to those for image enhancement, and arithmetic operations are performed to combine and transform the original bands into new images, which better highlight certain features in the scene. These operations include various methods of spectral or band ratio, and a procedure called principal components analysis, which is used to represent the information in multi-channel imagery.

Image classification and analysis operations are used to digitally identify and classify pixels in the data. Classification is usually performed on multi-channel data sets, and this process assigns each pixel in an image to a particular class or theme, based on statistical characteristics of the pixel brightness values. There are a variety of approaches taken to perform digital classification.

14.35.1 Preprocessing

Preprocessing operations, sometimes referred to as image restoration and rectification, are intended to correct for sensor and platform-specific radiometric and geometric distortions of data. Radiometric corrections may be necessary due to variations in scene illumination and viewing geometry, atmospheric conditions, and sensor noise and response. Each of these will vary depending on the specific sensor and platform used to acquire the data and the conditions during data acquisition. Also, it may be desirable to convert and calibrate the data to known radiation or reflectance units to facilitate comparison between data.

Variations in illumination and viewing geometry between images can be corrected by modelling the geometric relationship and distance between the area of the earth's surface imaged, the sun and the sensor. This is often required, so as to be able to more readily compare images collected by different sensors at different dates or times, or to mosaic multiple images from a single sensor while maintaining uniform illumination conditions from scene to scene.

Scattering of radiation occurs as it passes through and interacts with the atmosphere. This scattering may reduce, or attenuate, some of the energy illuminating the surface. In addition, the atmosphere will further attenuate the signal propagating from the target to the sensor. Various methods of atmospheric correction can be applied, ranging from detailed modelling of the atmospheric conditions during data acquisition, to simple calculations based solely on the image data. The correction is applied by subtracting the minimum observed value, determined for each specific band, from all pixel values in each respective band. As scattering is wavelength dependent, the minimum values will vary from band to band. This method is based on the assumption that the reflectance from these features, if the atmosphere is clear, should be very small. If the observed values are greater than zero, then they are considered to have resulted from atmospheric scattering.

Noise in an image may be due to irregularities or errors that occur in the sensor response and data recording and transmission. Common forms of noise include systematic striping or banding and dropped lines. Both of these effects should be corrected before further enhancement or classification is performed.

For many quantitative applications of remote sensing data, it is necessary to convert the digital numbers to measurements in units, which represent the actual reflectance or emittance from the surface. This is done based on detailed knowledge of the sensor response and the way in which the analog signal (i.e., the reflected or emitted radiation) is converted to a digital number, called analog-to-digital conversion. By solving this relationship in the reverse direction, the absolute radiance can be calculated for each pixel, so that comparisons can be accurately made over time and between the different sensors.

All remote sensing imageries are inherently subject to geometric distortions. These distortions may be due to several factors, including: the perspective of the sensor optics; the motion of the scanning system; the motion of the platform; the platform altitude, attitude, and velocity; the terrain relief; and, the curvature and rotation of the earth. Geometric corrections are intended to compensate for these distortions, so that the geometric representation of the imagery will be as close as possible to the real world.

14.35.2 Image Enhancement

Image enhancements are used to make it easier for visual interpretation and understanding of imagery. The advantage of digital imagery is that it allows us to manipulate the digital pixel values in an image. Although radiometric corrections for illumination, atmospheric influences and sensor characteristics may be done prior to distribution of data to the user, the image may still not be optimized for visual interpretation. Remote sensing devices, particularly those operated from satellite platforms, must be designed to cope with levels of target/background energy, which are typical to all conditions likely to be encountered in routine use. With large variations in spectral response from a diverse range of targets (e.g., forest, deserts, snowfields and water), no generic radiometric correction could optimally account for and display the optimum brightness range and contrast for all targets. Thus, for each application and each image, a custom adjustment of the range and distribution of brightness values is usually necessary.

In raw imagery, the useful data often populates only a small portion of the available range of digital values. Contrast enhancement involves changing the original values, so that more of the available range is used, thereby increasing the contrast between targets and their backgrounds. The key to understanding contrast enhancements is to understand the concept of an image histogram. A histogram is a graphical representation of the brightness values that comprise an image. The brightness values are displayed along the *X*-axis of the graph. The frequency of occurrence of each of these values in the image is shown on the *Y*-axis. By manipulating the range of digital values in an image, graphically represented by its histogram, it is possible to apply various enhancements to the data.

14.35.3 Image Transformations (Multiimage Manipulation)

Image transformations typically involve the manipulation of multiple bands of data, whether from a single multispectral image, or from two or more images of the same area acquired at different times (i.e., multitemporal image data). Either way, image transformations generate new images from two or more sources, which highlight particular features or properties of interest, better than the original input images.

Basic image transformations apply simple arithmetic operations to the image data. Image subtraction is often used to identify changes that have occurred between images collected on different dates. Typically, two images, which have been geometrically registered, are used with the pixel (brightness) values in one image being subtracted from the pixel values in the other. Scaling the resultant image by adding a constant to the output values will result in a suitable difference image. Image division or spectral ratioing is one of the most common transforms applied to image data. Image ratioing serves to highlight subtle variations in the spectral responses of various surface covers.

14.35.4 Image Classification and Analysis

A human analyst attempting to classify features in an image uses the elements of visual interpretation to identify homogeneous groups of pixels, which represent various features or land cover classes of interest. Digital image classification uses the spectral information represented by the digital numbers in one or more spectral bands, and attempts to classify each individual pixel based on this spectral information. This type of classification is termed spectral pattern recognition. In either case, the objective is to assign all pixels in the image to particular classes or themes (e.g., water, coniferous forest, deciduous forest, corn, wheat). The resulting classified image is comprised of a mosaic of pixels, each of which belongs to a particular theme, and is essentially a thematic map of the original image.

Information classes are those categories of interest that the analyst is actually trying to identify in the imagery, such as different kinds of crops, different forest types or tree species and different geologic units or rock types. Spectral classes are groups of pixels that are uniform with respect to their brightness values in the different spectral channels of the data. The objective is to match the spectral classes in the data to the information classes of interest. There is rarely a simple one-to-one match between these two types of classes. Rather, unique spectral classes may appear, which do not necessarily correspond to any information class of particular use or interest to the analyst. Alternatively, a broad information class (e.g., forest) may contain a number of spectral sub-classes with unique spectral variations. Using the forest example, spectral sub-classes may be due to variations in age, species, and density, or perhaps as a result of shadowing or variations in scene illumination. It is the analyst's job to decide on the utility of the different spectral classes and their correspondence to useful information classes.

14.35.5 Data Integration and Analysis

In the early days of analog remote sensing, when the only remote sensing data source was aerial photography, the capability for integration of data from different sources was limited. Now, with most data available in digital format from a wide array of sensors, data integration is a common method used for interpretation and analysis. Data integration fundamentally involves the combining or merging of data from multiple sources in an effort to extract better and/or more information. This may include data that are multitemporal, multi-resolution, multi-sensor, or multi-data type in nature. Imagery collected at different times is integrated to identify areas of change. Multitemporal change detection can be achieved through simple methods such as these, or by other more complex approaches such as multiple classification comparisons or classifications using integrated multitemporal data sets. Multi-resolution data merging is useful for a variety of applications. The merging of data of a higher spatial resolution with the data of lower resolution can significantly sharpen the spatial detail in an image and enhance the discrimination of features.

Data from different sensors can also be merged, bringing in the concept of multisensor data fusion. An excellent example of this technique is the combination of multispectral optical data with radar imagery. These two diverse spectral representations of the surface can provide complementary information. The optical data provide detailed spectral information useful for discriminating between surface cover types, while the radar imagery highlights the structural detail in the image.

Applications of multi-sensor data integration generally require that the data be geometrically registered, either to each other or to a common geographic coordinate system or map base. This also allows other ancillary (supplementary) data sources to be integrated with the remote sensing data. For example, elevation data in digital form, called digital elevation or digital terrain models (DEM/DTM), may be combined with remote sensing data for a variety of purposes. DEM/DTM may be useful in image classification, as effects due to terrain and slope variability can be corrected, potentially increasing the accuracy of the resultant classification. DEM/DTM is also useful for generating three-dimensional perspective views by draping remote sensing imagery over the elevation data, enhancing visualization of the area imaged.

In a digital environment, where all the data sources are geometrically registered to a common geographic base, the potential for information extraction is extremely wide. This is the concept for analysis within a digital geographical information system (GIS) database. Any data source, which can be referenced spatially, can be used in this type of environment. A DEM/DTM is just one example of this kind of data. Other examples include digital maps of soil type, land cover classes, forest species, road networks, and many others, depending on the application. The results from a classification of a remote sensing data set in map format, could also be used in a GIS as another data source to update existing map data. In essence, by analyzing diverse data sets together, it is possible to extract better and more accurate information in a synergistic manner than by using a single data source alone. There are a myriad of potential applications and analyses possible for many applications. In the next and final chapter, we will look at examples of various applications of remote sensing data, mainly involving the integration of data from different sources.

14.36 REMOTE SENSING IN INDIA

Remote sensing technology has evolved in India at a fast pace. India is presently pressing ahead with an impressive national programme aimed at developing more and more earth observation satellites to meet the ever-increasing demands, which have been created with the use of this technology. The earth observation system, as an important space infrastructure, has now become an essential tool for today's government for achieving the goal of 'application of technology for national development'.

The dream of the Indian scientists to have an earth observation satellite of their own, and made by their own hands, took the solid shape of reality with the launch of Bhaskara-I on June 7, 1979 and the launch of Bhaskara-II in 1981. The Indian earth observation system became operational with the advent of Indian remote sensing satellites. The launch of the first operational remote sensing satellite, IRS-1A was on March 17, 1988.

14.36.1 Remote Sensing Satellites of India

IRS-1A

- Launched on: March 17, 1988
- Out of service since: 1995

TABLE 14.4 Technical specification of IRS 1A

Band	Wavelength (µm)	Bandwidth (µm)	Resolution (µm)	Swath Width (km)	Revisit Time (Days)
Band 1 (VIS)	0.46–0.52	–	72.5	148	22
Band 2 (VIS)	0.52–0.59	–	72.5	148	22
Band 3 (VIS)	0.62–0.68	–	72.5	148	22
Band 4 (NIR)	0.77–0.86	–	72.5	148	22

- Repeat Cycle: 22 days
- Orbit Height: 904 km
- Orbit Type: Sun synchronous

IRS-1B

- Launched on: August 29, 1991
- Out of service since: 1996
- Repeat cycle: 22 days
- Orbit height: 904 km
- Orbit type: Sun synchronous

(For technical specifications, see Tables 14.4 and 14.5.)

IRS-1C and IRS-1D IRS-1C and IRS-1D are identical and were launched in December 1995 and September 1997, respectively. They carry three cameras, panchromatic camera (PAN), linear imaging self-scanner (LISS-III) and wide field sensor (WiFS) as shown in Table 14.6.

IRS-P2

- Launched on: 1994
- Repeat cycle: 24 days
- Orbit height: 817 km
- Orbit type: Sun synchronous

TABLE 14.5 LISS II sensor bands

Band	Wavelength (µm)	Bandwidth (µm)	Resolution (m)	Swath Width (km)	Revisit Time (Days)
Band 1 (VIS)	0.46–0.52	–	36.25	74 (146.5*)	22
Band 2 (VIS)	0.52–0.59	–	36.25	74 (146.5)	22
Band 3 (VIS)	0.62–0.68	–	36.25	74 (146.5)	22
Band 4 (NIR)	0.77–0.86	–	36.25	74 (146.5)	22

* Two LISS II cameras combined swath.

TABLE 14.6 Pan, LISS-III, WiFS specifications

		LISS-III		
	PAN	VNIR	SWIR	WiFS
Spatial resolution (m)	5.8	23.5	70.5	188
Swath (km)	70	142	148	810
Spectral band (Microns)	0.5–0.75	0.52–0.59	1.55–1.7	0.62–0.68
		0.62–0.68		
		0.77–0.86		

Launched on October 15, 1994, the 5 m panchromatic data is especially useful for urban planning and mapping, the 25 m multispectral data is good for natural resource planning; and the 180 m wide-field data band has a 740-km swath and 5-day repeat coverage, which is excellent for large-area vegetation monitoring.

IRS-P3 The IRS-P3 satellite was launched using PSLV-D3 on March 21, 1996. IRS-P3 was put in a polar, sun synchronous orbit at an altitude of 817 km with equatorial crossing time of 10:30 AM in the descending node. IRS-P3 has an X-ray astronomy and two remote sensing payloads, namely WiFS and MOS. The mission caters to oceanography applications. IRS-P3 WiFS is similar to IRS-1C WiFS, but for the inclusion of an additional band in the middle IR (MIR) region. This sensor is primarily meant for vegetation dynamic studies, while MOS is meant for ocean related studies (for technical specifications, please refer Table 14.7 and Table 14.8).

IRS-P4 (OCEANSAT-1) India launched OCEANSAT-1 (IRS-P4) on May 26, 1999, which is the first Indian satellite dedicated fully for the study of oceans. It carries the ocean colour monitor

TABLE 14.7 LISS II sensor bands

Band	Wavelength (µm)	Bandwidth (µm)	Resolution (m)	Swath Width (km)	Revisit Time (Days)
Band 1 (VIS)	0.46–0.52	–	36.25	74 (146.5*)	22
Band 2 (VIS)	0.52–0.59	–	36.25	74 (146.5)	22
Band 3 (VIS)	0.62–0.68	–	36.25	74 (146.5)	22
Band 4 (NIR)	0.77–0.86	–	36.25	74 (146.5)	22

* Two LISS II cameras combined swath.

TABLE 14.8 IRS-P3 WiFS camera (sensor characteristics)

Sensor	WiFS
Resolution	188 m × 188 m (B3 & B4) 188 × 246 (B5)
Swath	770 km
Repetitivity	24 days
Spectral bands	0.62–0.68 µm (B3) 0.77–0.86 µm (B4) 1.55–1.69 µm (B5)

(OCM) and the multi-frequency scanning microwave radiometer (MSMR) as shown in Table 14.9 and Fig. 14.26, respectively. The satellite is helpful in the study of oceanographic phenomena such as sea temperature, sea surface height and rain over oceans, and would be useful in measuring various ocean parameters.

IRS-P6 (RESOURCESAT-1) The heaviest earth-observation spacecraft RESOURCESAT-1 was launched into an 817 km sun-synchronous polar orbit on October 17, 2003. RESOURCESAT-1 is the most advanced satellite bringing continuity to the current IRS-1C and IRS-1D programmes, and carries three sensors that deliver an array of spectral bands and resolutions ranging from 5.8 to 60 m. Data products derived from RESOURCESAT-1 can be used for advanced applications in vegetation dynamics, crop yield estimates and disaster management support. In addition, RESOURCESAT-1 has

TABLE 14.9 OCEANSAT-1: The specifications of OCM and MSMR payloads

OCM		MSMR
• Spectral Bands in nanometres		• Frequencies (GHz)
1	402–422	06.60
2	433–453	
3	480–500	10.65
4	500–520	
5	545–565	18
6	660–680	
7	745–785	21
8	845–885	
• Spatial resolution: 360 m × 236 m • FOV: +/– 43 degrees • Sweth: 1420 km • Radiometric quantisation: 12 bits		• Polarisation: V&H for all frequencies. • Spatial resolution: 120,80,40 and 40 km. • Sweth 1,360 km • Radiometric resolution: 12 bits
Along track steering +/– 20 degrees. In steps of 5 degrees. to avoid sun glint		

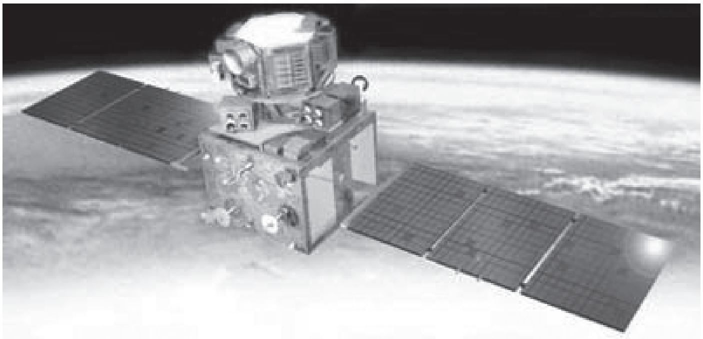


FIGURE 14.26 OCEANSAT-1 (IRS-P4)

120 gigabits of on-board memory that allows for out-of-contact imaging. It is scheduled to last for 5 years. RESOURCESAT. IRS-P6 or the RESOURCESAT-1 has several improved features over its predecessors. These include availability of 5.8 m spatial resolution in three bands from LISS-IV camera, improved LISS-III with MIR band information at 23.5 m resolution as other visible and NIR bands. In addition, the AWiFS provides data in the same spectral channels as LISS-III at about 56 m resolution with 10-bit radiometry, 5-day revisit and scene coverage of 740 km for regional studies. Unique to IRS-P6 is the availability of simultaneous multispectral data at three spatial resolutions from the same platform with scene coverage varying from 576 sq km to 19,600 sq km to 5,42,000 sq km.

OCEANSAT-2: The ISRO (Indian Space Research Organization) spacecraft OCEANSAT-2 is envisaged to provide service continuity for the operational users of Ocean Colour Monitor (OCM) data as well as to enhance the application potential in other areas. The main objectives of OCEANSAT-2 are to study surface winds and ocean surface strata, observation of chlorophyll concentrations, monitoring of phytoplankton blooms, study of atmospheric aerosols and suspended sediments in the water. It carries the following payloads:

- OCM
- Ku-band Pencil Beam scatterometer (OSCAT) developed by ISRO
- Radio Occultation Sounder for Atmosphere (ROSA) developed by the Italian Space Agency.

The OCEANSAT-2 spacecraft is envisaged to provide continuity of operational services of OCEANSAT-1 (IRS-P4) with enhanced application potential

Technology Experiment Satellite (TES) TES was launched on board in October 2001. TES carries a panchromatic camera with a spatial resolution of 1 m. Some of the technologies that are being demonstrated in TES are attitude and orbit control system using high torque reaction wheels; new reaction control system with optimized thrusters and a single propellant tank; light weight spacecraft structure; solid state recorder; X-band phased array antenna; improved satellite positioning system; miniaturized TTC and power system and, two-mirror-on-axis camera optics.

14.36.2 Data from IRS Satellites

Data from IRS satellites is received at the earth station of the National Remote Sensing Agency (NRSA) at Hyderabad. The data is processed after several stringent quality checks at various levels, and then supplied on user request on digital/photographic media.

14.36.3 NNRMS

Recognizing the need and importance of natural resources management in the country, the Government of India had set up the national natural resources management system (NNRMS). NNRMS is an integrated approach for management of natural resources, optimally utilizing the advantages of conventional systems and the information derived through remote sensing. The Department of Science established five regional remote sensing service centres (RRSSCs) in the country for speedy operation of remote sensing as an integral component of natural resources inventory, monitoring and management.

RRSSCs enable the use of remote sensing technology at a reasonable cost to derive necessary information on various aspects related to natural resources. The RRSSCs provide facilities for digital image analysis and GIS to the users, provide support service to execute national projects and much

TABLE 14.10 RESOURCESAT-1: Technical specifications

		LISS-IV Mono Mode	LISS-III MX Mode		AWiFS
Spatial resolution	Band 2 (green) Band 3 (red) Band 4 (NIR) Band 5 (SWIR)	5.8 m	5.8 m 5.8 m 5.8 m	23.5 m 23.5 m 23.5 m 23.5 m	60–70 m 60–70 m 60–70 m 60–70 m
Swath width	All Bands	70 km	23.9 km	140 km	700 km
Radiometric resolution, Quantisation	All Bands	10 bit 7 bit transmission	10 bit 7 bit transmission	7 bit (VNIR) 10 bit (SWIR)	10 bit
Spectral coverage	Band 2 (green) Band 3 (red) Band 4 (NIR) Band 5 (SWIR)	620–680 nm	520–590 nm 620–680 nm 770–860 nm	520–590 nm 620–680 nm 770–860 nm 1,550–1,700 nm	520–590 nm 620–680 nm 770–860 nm 1,550–1,700 nm
CCD arrays (number of arrays * no. of elements)	Band 2 (green) Band 3 (red) Band 4 (NIR) Band 5 (SWIR)	1 * 12,288	1 * 12,288 1 * 12,288 1 * 12,288	1 * 6,000 1 * 6,000 1 * 6,000 1 * 6,000	2 * 6,000 2 * 6,000 2 * 6,000 2 * 6,000

more. The Indian Institute of Remote Sensing at Dehradun, under the aegis of NRSA, provides a wide spectrum of training and educational courses in remote sensing applications.

14.36.4 Advanced Remote Sensing Satellites

IRS-P5 (CARTOSAT-1) IRS-P5 is launched by PSLV in 2005. The satellite is primarily intended for advanced cartographic applications. It will have two panchromatic cameras with a spatial resolution of 2.5 m and a swath of 30 km each. These cameras are mounted with a tilt of 26 and 5 degrees along the track with respect to the nadir to provide stereo pairs of images needed for the generation of digital terrain model (DTM)/digital elevation models (DEM) globally.

The data products will be used for cartographic applications, cadastral mapping and updating, land use and other GIS applications. The satellite is placed in a sun synchronous polar orbit of 617 km. It will have a revisit capability of 5 days, which can be realised by steering the spacecraft about roll axis by 26 degrees.

CARTOSAT-2: CARTOSAT-2 is an advanced remote sensing satellite with a panchromatic camera capable of providing scene-specific spot imageries for cartographic applications. It is placed in a sun synchronous polar orbit at an altitude of 630 km. It is having a revisit capability of 4 days. CARTOSAT-2 is the twelfth in the IRS satellite series and is an advanced remote sensing satellite capable of providing scene-specific spot imagery. CARTOSAT-2 was launched on January 10, 2007 and carries a single panchromatic camera onboard capable of providing better than 1-metre spatial resolution imagery, with a swath of 9.6 km. The satellite can be steered up to ±45 degrees along and ±26 degrees across the track to facilitate imaging of any area more frequently.

The data from the satellite is being used for cartographic applications at cadastral level, urban and rural infrastructure development and management, as well as applications in Land Information System (LIS).

Radar Imaging Satellite (RISAT-1) Radar imaging satellite (RISAT) mission envisages to support and augment the operational remote sensing programme by enhancing agricultural and disaster related applications. RISAT will have all-weather and day–night observation capability. It is slated for launch during 2006 with a mission life of 5 years. RISAT carries a C-band synthetic aperture radar (SAR) operating in multi-polarization and in multi modes (ScanSAR, Strip and Spot Modes). The satellite will provide spatial resolutions of 3–50 m with swaths varying from 10 to 240 km. The RISAT is expected to launch into a polar sun synchronous orbit of 609 km.

RISAT-2: RISAT-2 is a radar imaging satellite using an active SAR imager with all-weather capability to take images of the earth. This satellite will enhance ISRO's capability for Disaster Management Applications. RISAT-2 was launched as the sole passenger on an Indian PSLV-CA launch vehicle and is India's first satellite with SAR. It has a day–night, all-weather monitoring capability.

REVIEW QUESTIONS

1. What is meant by remote sensing? What is its importance?
2. Describe the principles of remote sensing.
3. With the help of a neat diagram, describe the components of remote sensing.
4. What are the seven elements in remote sensing?
5. Explain the characteristics of electromagnetic radiations.
6. What are the three forms of interaction when the electromagnetic energy strikes upon a surface?
7. Explain the spectral bands and its characteristics used in remote sensing.
8. Write short notes on the IR region of the electromagnetic spectrum.
9. Write short notes on:
   - i. Transmission path
   - ii. Atmospheric windows
   - iii. Rayleigh scattering
   - iv. Mie scattering
   - v. Different platforms used in remote sensing
   - vi. Passive remote sensing
   - vii. Active remote sensing
   - viii. Thermal IR remote sensing
   - ix. Sensors
   - x. Detectors
   - xi. Thematic mapper
10. Explain different types of remote sensing.

11. Explain thermal IR imaging and its importance in remote sensing.
12. What is blackbody radiation? Explain the term reflectance.
13. Describe briefly about microwave remote sensing.
14. What is the role of a scanner in remote sensing? Describe the different types of scanners used in remote sensing.
15. What is spectral signature? Explain its importance in remote sensing.
16. What are the four principal characteristics of electromagnetic radiation, which can be exploited for providing signatures? Explain each term with examples.
17. Explain the term swath and IFOV.
18. What are the principle steps used for analyzing remotely sensed data?
19. How is remote sensing data analyzed for getting desired results?
20. What are the various elements of visual interpretations?
21. Explain the term digital image processing.
22. Describe the common image processing functions available in an image analysis system.
23. Write short notes on data interpretation and analysis in remote sensing.
24. Write a short note on the Indian Remote Sensing programme.

BASICS OF GLOBAL POSITIONING SYSTEM 15

Chapter Outline

15.1 INTRODUCTION

Global Positioning System (GPS) is satellite-based radio navigation and positioning system built and run by the US DoD (US Department of Defence). Initially the system was developed for defence purpose, but later it was made available to the civilians. GPS is an all weather, global, continuous, navigation positioning and timing system. GPS is a passive or an one-way ranging system. The user can only receive the satellite signals. GPS is the shortened form of NAVSTAR GPS. NAVSTAR GPS is an acronym for NAVigation System with Time And Ranging Global Positioning System.

Traditionally, celestial bodies, such as sun, moon and stars, were observed for both navigation and positioning. Observing celestial bodies with the help of a navigational sextant or astronomical theodolite, and with the aid of a nautical almanac, one can easily define the geographic coordinates of a point anywhere on the surface of the earth. But one cannot define a position precisely and easily using the conventional methods. Hence to overcome the difficulties, in early seventies, a new project was devised by the US DoD; the GPS. The system is found successful in sudden calculation of positional value of any point on the surface of earth, accurately at any time, and in all weather conditions. To a soldier in a field, position accuracy means about 15 m. To a ship in coastal waters, accuracy means 5 m. To a land surveyor, accuracy means 1 cm or less. GPS can be used to achieve all of these accuracies in all of these applications, the difference being the type of GPS receiver used and the technique employed.

To define any point (in two dimensions) on the surface of earth, we require two parameters, the latitude and longitude, known as geographical coordinates. To define the three-dimensional position of a point, on the surface of the earth, we require a third parameter known as height from a reference datum, in addition to latitude and longitude.

15.2 OVERVIEW OF GPS

The GPS consists of a constellation of 24 operational satellites that revolve around the globe once every 12 hours, to provide worldwide position, time and velocity information. GPS makes it possible to precisely identify locations on the earth by measuring distance from the satellites. First, GPS satellite was launched in 1978 and a constellation GPS satellites for initial operation capability was achieved in 1993 and US DoD made official announcement about initial operational capability of the system in December 8, 1993. The GPS system was officially declared to have achieved full operational capability on July 17, 1995, ensuring the availability of 24 fully operational, non-experimental, GPS satellites worldwide. In order to ensure continuous worldwide coverage, GPS satellites are well placed in six orbital planes, that is, four satellites in an orbit, and with this satellite constellation geometry, four to ten GPS satellites will be visible anywhere in this world, when an elevation angle of 10 degrees is considered.

Even though GPS was originally designed for military purpose, the US DoD had made clear that the civilians could also use GPS for various scientific or civilian applications. The first major civilian applications to emerge were Marine Navigation and Surveying.

The advantages of GPS systems are:

- It is a worldwide, all weather system and continuously available 24 hours a day.
- GPS is designed to achieve relative high positioning accuracies from a few metres down to millimetre level.
- GPS is a positioning system with no user charges and uses relatively low cost hardware and software.
- Signal availability is guaranteed to users anywhere on the globe, whether it is in the air, on the ground or at sea.
- Capability of determining velocity and time.
- Able to provide service to unlimited number of users.
- Redundancy provisions to ensure the survivability of the system.
- Real-time positioning, velocity and time determination capability with good accuracy.
- Suitable for all platforms such as land, ship, aircraft and space.
- Able to handle a wide range of dynamics.
- Three-dimensional position information is provided by the system (in x, y and z coordinates).
- Passive positioning system or one way ranging system.

15.3 GPS SEGMENTS

GPS is a satellite-based system that uses a constellation of 24 satellites to give a user an accurate position on earth. The full GPS consists of three distinct segments:

1. The Space Segment: The space segment consists of all the GPS satellites orbiting around the earth.
2. The Control Segment: The control segment consists of all the control and monitoring stations, positioned at various location of earth to control and monitor all the GPS satellites.
3. The User Segment: Anyone, who uses the GPS signal for various applications, comes under this segment.

15.3.1 The Space Segment

The space segment is also known as satellite segment. The space segment is designed to have 24 fully operational satellites well placed at an altitude of 20,200 km above earth's surface. The orbits are nearly circular and equally spaced about the equator at 60 degrees separation with an inclination relative to the equator of 55 degrees. The orbital radius (distance from the centre of mass of the earth to the satellites) is approximately 26,600 km. The satellite constellation provides a 24-hour global user navigation and time determination capacity. The space segment is designed in such a way that there will be a minimum of four satellites visible above a 15-degree elevation angle from any point on the earth's surface at any one time.

The GPS satellite orbits are nearly circular, with eccentricity less than 0.02, a semi-major axis of 26,560 km, that is, an altitude of 20,200 km. Orbits in this height are referred to as medium earth orbit. The satellites are revolving around the earth with a velocity of 3.9 km/sec and a nominal period of 12-hour sidereal time (11 hour 58 minutes 2 seconds solar time), repeating the geometry each sidereal day. The satellites are well arranged on six orbital planes so that each orbit is having at least four slots. There are spare satellite slot in each orbital plane.

The following points have been taken into account for the design consideration of satellite segment.

1. Higher the satellites, the longer it is visible above horizon.
2. The higher a satellite, the better the coverage area due to longer fly-over passes and extended visibility of the satellite across large areas of the earth.
3. The higher a satellite, the less the rate of change of distance and lower the Doppler frequency of a transmitted signal.
4. The greater the angle of inclination, the more northerly the track of the sub-satellite point across the surface of earth.
5. No satellite can be seen simultaneously from all locations on the earth.
6. Depending on the positioning principles being employed, there may be requirement for observations to be made to more than one satellite simultaneously from more than one ground station.

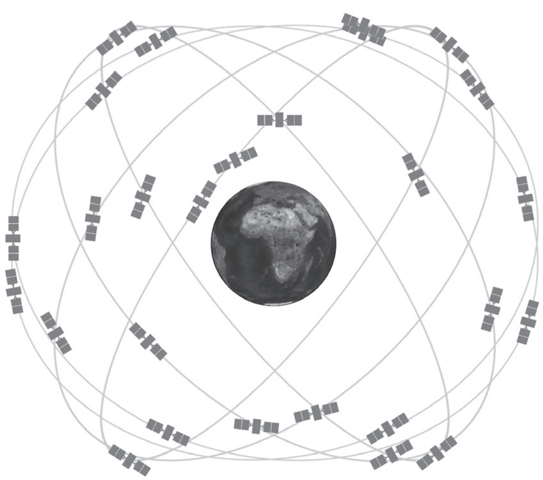


FIGURE 15.1 The space segment (the constellation of GPS satellites)

15.3.2 The Control Segment

The control segment comprises three physical components, the master control station, monitoring stations and the ground antennas. It consists of the ground facilities carrying out the task of satellite tracking, orbit computations, data telemetry and supervision necessary for the daily control of space segment. The control segment consists of one master control station, five monitoring stations and 4 ground antennas distributed amongst five locations near the earth's equator, as shown in Fig. 15.2. The control segment has the responsibility for maintaining the satellites and their proper functioning. This includes maintaining the satellites in their proper positions and monitoring satellite sub-system health and status. The control segment tracks the GPS satellites, updates their orbiting position and calibrates and synchronises the satellite clocks. A further important function is to determine the orbit of each satellite and predict its path for the following 24 hours. There may be a possibility of satellites travelling slightly out of orbits. Hence the ground monitor stations keep track of the satellite orbits, altitude, location and speed. The ground stations send the orbit data to the master control stations, which in turn send corrected data of the satellite. The corrected data is called Ephemeris data, valid for 6 hours, and transmitted in coded form to the GPS receivers. This information is uploaded to each satellite and subsequently broadcast from it. This enables the GPS receiver to know where each satellite can be expected to be found.

There are five ground facility stations; Hawaii, Colorado Spring, Ascension Island, Diego Garcia and Kwajalein. The satellite signals are read at Ascension, Diego Garcia and Kwajalein. All are owned and operated by US DoD. All five stations are monitoring stations and are equipped with tracking systems to track satellites. The tracking data is then sent to the Master Control Station in Colorado Springs where they are processed to determine any errors in each satellite. The information is then sent back to the four monitor stations equipped with ground antennas and uploaded to the satellites.

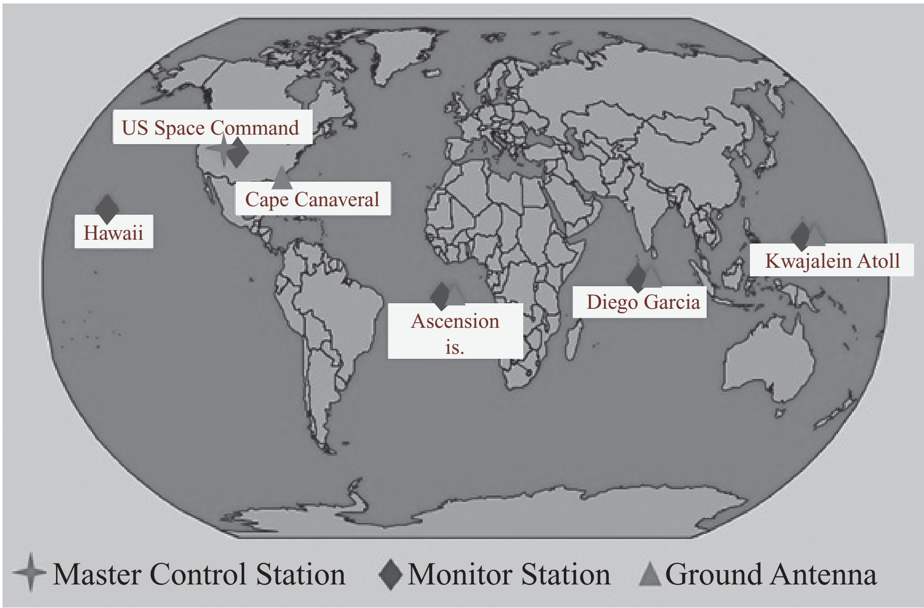


FIGURE 15.2 Control segment of GPS

15.3.3 The User Segment

The user segment comprises anyone using a GPS receiver to receive the GPS signal for navigation or to determine their position and time. It includes the entire spectrum of applications, equipment and computational techniques that are available to the users. Typical applications within the user segment are marine navigation, surveying, aerial navigation, recreational activities, machine control and GIS application. GPS has a lot of applications on land, sea and in the air. GPS can be used anywhere on the surface of the earth, except where it is impossible to receive direct signal from the satellite, such as underside of bridge, caves, enclosed parking areas, very thick forest, other sub-terrain locations and underwater.

Land-based application of GPS is more diverse. The scientific community uses GPS for precision timing applications. A surveyors and GIS practitioners use GPS for various applications in geomatics. GPS offers incredible cost savings by reducing setup time at the survey site and also gives amazing accuracy. The important aspect of GPS-based surveys is that it does not depend on line of sight between the survey points. Basic survey units are now offering 1 m accuracy. More expensive GPS receivers provide millimetre accuracy by employing special methods such as Differential Global Positioning System (DGPS), Real Time Kinematic DGPS (RTK DGPS), etc.

Each day new applications are being identified, each with its unique requirements with respect to accuracy of the results, reliability, operational constraints, user hardware, data processing algorithms, latency of the GPS results, etc. A general classification of various types of GPS receivers for various GPS applications are listed below.

1. Navigations Receivers: Used for land air and sea (marine) navigation and tracking including enroute as well as precise navigation, collision avoidance, cargo monitoring, vehicle tracking, search and rescue operation, etc. This kind of applications require modest accuracy and the user hardware is comparatively of low cost, and the integrity and speed with which the results are needed is generally high.
2. Surveying Receivers: Used for surveying and mapping on land, at sea and from air, and the applications include geophysical and resource mapping, GIS data capture, surveying and positioning, and general engineering applications. The applications are of relatively high accuracy, for positioning in both the static and moving receiver mode, and generally require specialized receivers and software.
3. Military Receivers: The military grade GPS systems are developed according to military specifications and they are capable of tracking encrypted codes such as precision code (P-code).
4. GPS Receivers for Recreational Activities: Used for recreational purposes on land, sea and in the air. The requirement is low-cost equipment that are easy to use.
5. Receivers for Scientific Applications: Used for scientific applications such as time transfer, altitude determination, spacecraft applications, atmospheric studies, crustal deformation studies, geodetic studies, etc. Such applications require special type of costly receivers.

15.4 SATELLITE RANGING

The GPS is a one-way ranging system. GPS is known as an one way or passive system in the sense that only the satellites transmit signals and the users simply receive them and the user cannot communicate to the satellite. The fundamental observable is the signal travel time between the satellite antenna and the receiver antenna. The signal travel time is scaled into range measurements using the signal propagation velocity. The GPS positions are based on measuring the distance from the satellites to the GPS receiver on the earth. These distances to each satellite are determined by the GPS receiver itself. The basic principle behind the determination of position is trilateration or resection. That is, if the distance to three points relative to a single point (point occupied by the instrument) is known, then the position occupied by the instrument can be defined relative to those three points and determine your own position relative to those three points. From the distance to one satellite it is known that the received must of at some point on the surface of an imaginary sphere of radius equal to that distance with origin at the satellite. By intersecting three imaginary spheres the receiver position can be determined accurately. To calculate the distance to each satellite, Sir Isaac Newton's equation can be used.

$$\text{Distance} = \text{Velocity} \times \text{Time}$$

Satellite ranging is done with the aid of radio signals that are broadcast from the GPS satellites to the GPS receivers in the microwave part of the electromagnetic spectrum. GPS receiver calculates the distance from the receiver to the satellite using the above equation, where velocity is the velocity of the radio signal which is equal to the velocity of light, that is, 290,000 km/sec, and time is the time taken by the radio signal to travel from the satellite to the receiver. This time calculation is difficult, because it requires the precise time at which the radio signal leaves the satellite and the time it reaches the receiver. The solution for this is to have a detailed study of radio signals and the codes and other information carried by that signal. All GPS satellites are carrying very accurate atomic clocks on board, which operates at a fundamental frequency of 10.23 MHz. This fundamental frequency of 10.23 MHz is used to generate the signals that are broadcasted from the satellite (see Figs. 15.3 and 15.4).

The GPS satellites communicate all the information to the receivers using codes. GPS satellite broadcasts two carrier waves and the codes are modulated within the carrier waves. Hence the carrier wave takes the codes (coded information in the modulated form) from the satellites to the receiver. The two GPS carrier waves are radio waves known as L1 and L2 carrier wave in the L-band (L-band width is 390–1,550 MHz), and the two carrier waves are derived from the fundamental frequency of 10.23 MHz (fundamental frequency of atomic clocks provided with the GPS satellites). The carrier waves travel to the earth at the speed of light. These high-frequency radio transmissions from the satellite travel in straight lines and have a very low power. The power of radio signal transmission from the satellite is as low as 50 W. Hence for receiving this low-power radio transmission (which is transmitted from 20,200 km above earth's surface), the antenna of the GPS receiver must have a direct view of satellite.

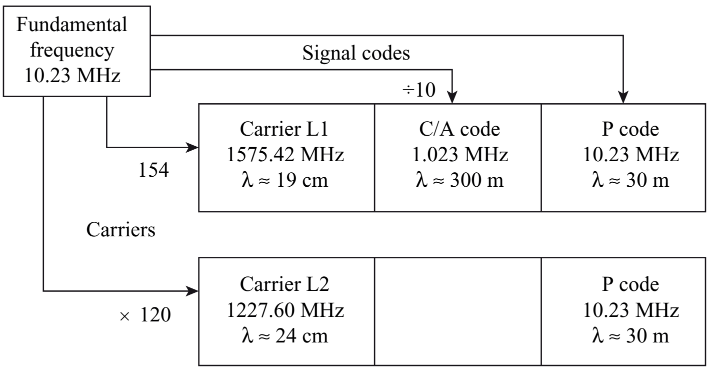


FIGURE 15.3 GPS signal structure

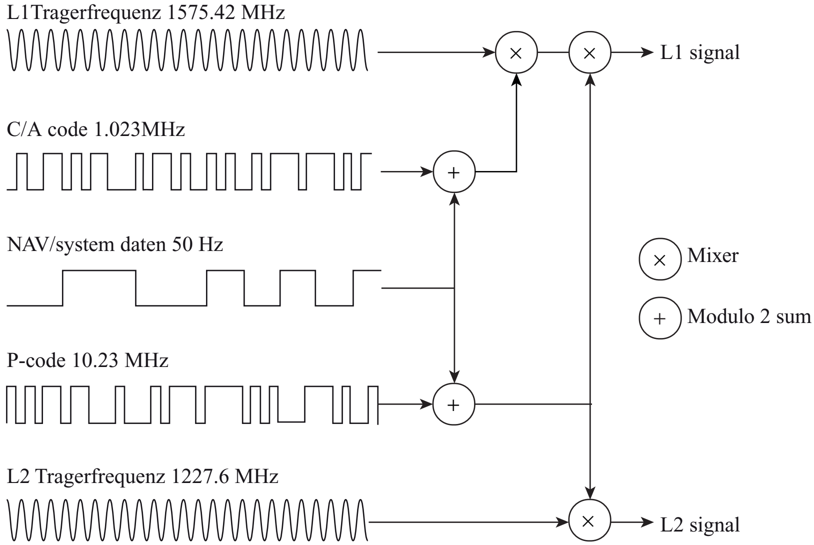


FIGURE 15.4 GPS signal structure showing the wavelength (λ) of all signals

L1 carrier wave is broadcasted at 1,575.42 MHz (fundamental frequency of 10.23 MHz ×154)
L2 carrier wave is broadcasted at 1,227.60 MHz (fundamental frequency of 10.23 MHz × 120)

MHz is the acronym of megahertz, the unit of radio frequency, which is million cycles per second. L1 carrier has two codes modulated upon it, namely the Coarse Acquisition code usually termed as C/A code, and the P code. The C/A code is modulated at 1.023 MHz, and the P-code is modulated at 10.23 MHz. L2 carrier has only the P-code modulated at 10.23 MHz upon it. The principal purpose of these codes is the calculation of travel time of radio signals from the satellite to the GPS receiver. This travel time is called the time of arrival. The travel time or time of arrival multiplied by the velocity of radio signals will give the satellite range, which is the distance from the satellite to the receiver.

The navigation message having a very low frequency of 50 bits/sec. It is a 1,500 bit sequence and, therefore, takes 30 sec to transmit.

The navigation message includes information on the broadcast ephemeris (satellite orbital parameters), satellite clock corrections, almanac data (a crude ephemeris for all satellites), ionosphere information and satellite health status. A navigation message is modulated within the L1 and L2 carriers. GPS receivers use different codes to distinguish between satellites. These codes can also be used as a basis for making pseudo-range measurements which enable the calculation of position.

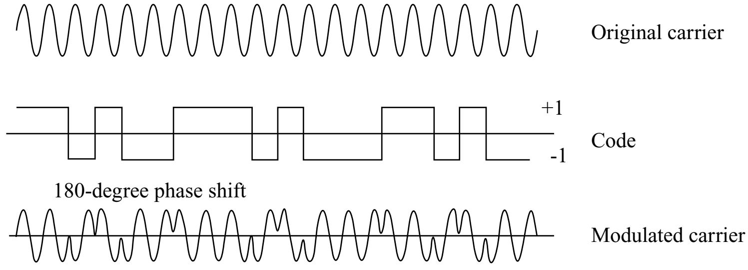


FIGURE 15.5 Modulated carrier wave

15.5 PSEUDO-RANGE AND PSEUDO-RANDOM CODE

The measurement of apparent signal propagation time from GPS satellite to the GPS receiver antenna, scaled into distance by speed of light, is known as pseudo-range, and the process of measurement is called pseudo-ranging. The apparent propagation time is the difference between the time of signal transmission (or emission) and the time of signal reception (at the antenna of the GPS receiver). When the position of a satellite is right overhead of an observer, the travel time of signal would be about 0.06 sec. This time difference gives the range measurements, but is called a pseudo-range because at the time of measurement the receiver clock is not precisely synchronized to the satellite clock. Pseudo-ranges differ from the actual range due to the influences of satellite orbit errors, user clock error, ionospheric delays, tropospheric delays, etc.

GPS codes are binary strings of zeros and ones. The three basic codes in GPS are the C/A code, the P-code and the navigation code (navigation message). These codes modulated within the carrier waves are known as pseudo-random noise (PRN) code. The pseudo-random code is actually a sequence of very precise time marks that permit the ground receivers to compare and compute the time of transmission between the satellite and ground station. The C/A code pulse interval is approximately 300 m in range and the more accurate P-code pulse interval is 30 m. The very complex pattern of pseudo-random code ensures that the receiver does not accidently synchronize with some other signal. Each satellite has its own unique pseudo-random codes. This complexity guarantee that the receiver will not accidently picks up a signal from another satellites. All the satellites can use the same frequency without signal jamming due to the complex nature of individual PRN code of each satellite. The complexity of pseudo-random codes makes it possible to use information theory to amplify the GPS signals and this is why the GPS receivers do not require big satellite dished to receive the GPS signals. The main features of all three signal types used in GPS observation namely the carrier, code and data signals are given in Table 15.1.

TABLE 15.1 Main features of GPS satellite signals

Atomic Clock Fundamental Frequency	10.23 MHz
L1 carrier signal	154 × 10.23MHz
L1 frequency	1575.42 MHz
L1 wavelength	19.05 cm
L2 carrier signal	120 × 10.23 MHz
L2 frequency	1227.60 MHz
L2 wavelength	24.45 cm
P-code frequency (chipping rate)	10.23 MHz (Mbps)
P-code wavelength	29.31 m
P-code period	266 days, 7 days/satellite
C/A code frequency (chipping rate)	1.023 MHz (Mbps)
C/A code wavelength	293.1 m
C/A code period	1 msec
Data signal frequency	50 bits/sec (bps)
Data signal cycle length	30 sec

The receiver can distinguish the L1 and L2 signals from different satellites because the GPS technology uses a code division multiple access (CDMA) spread-spectrum technique where the low-bit-rate message data are encoded with a high-rate PRN sequence that is different for each satellite. The receiver identifies the PRN codes for each satellite and uses it to reconstruct the actual message data. The message data is transmitted from the satellite at 50 bits/sec. Two distinct CDMA encodings are used: the C/A code (the so-called Gold Code) at 1.023 million chips per second, and the precise (P) code at 10.23 million chips per second. The C/A code is public and used by civilian GPS receivers, while the P-code can be encrypted as a so-called P(Y) code which is only available to military equipment with a proper decryption key. Both the C/A and P(Y) codes impart the precise time-of-day to the user. These codes can also be used as a basis for making ***pseudo-range*** measurements, which enables calculation of position.

15.6 GPS BROADCAST MESSAGE AND EPHEMERIS DATA

All the GPS satellites periodically broadcasts data related to time correction, satellite and system status, position of satellites in the form of ephemeris data. There are two types of ephemeris data, the broadcast ephemeris and precise ephemeris. The broadcast ephemerides are actually predicted satellite positions broadcasted with the navigation message transmitted from the satellites in real time. A receiver capable of acquiring either C/A code or P-code can acquire the ephemeris in real-time. The precise ephemerides are based on actual tracking data that is post processed to obtain more accurate satellite positions. This ephemeris is available at a later date and is more accurate than the broadcast ephemeris because they are based on the actual tracking date.

The data pertain to the position of the satellite in space at any given time stored in the memory of GPS receiver is called satellite almanac data. This data are received by the receiver from the satellites. When a GPS receiver is not turned on for a long time, the almanac data gets outdated. So after a long time when a GPS receiver is switched on, it will try to get latest almanac data from the satellite and will take more time for downloading the latest almanac data. Hence the position calculation will be a bit delayed due to this process. This condition of the receiver is called a cold start. A receiver is considered warm when the almanac data stored within the receiver is updated and accurate. While purchasing a new GPS receiver, the cold and warm start time should be noted, as the tie taken by the GPS unit to lock on to the satellite signals and calculate a position is important. The **almanac** consists of coarse orbit and status information for each satellite in the constellation, an ionospheric model and information to relate GPS-derived time to Coordinated Universal Time (***Coordinated Universal Time* (*UTC*)** *is a time standard based on International Atomic Time* (*TAI*) *with leap seconds added at irregular intervals to compensate for the earth's slowing rotation. Leap seconds are used to allow UTC to closely track UT1, which is mean solar time at the Royal Observatory, Greenwich*). A new part of the almanac is received for the last 12 sec in each 30-sec frame. Each frame contains 1/25th of the almanac, so 12.5 min are required to receive the entire almanac from a single satellite. The almanac serves several purposes.

i. To assist in the acquisition of satellites at power-up by allowing the receiver to generate a list of visible satellites based on stored position and time, while an ephemeris from each satellite is needed to compute position fixes using that satellite.

ii. For relating time derived from the GPS (called GPS time) to the international time standard of UTC.

The almanac allows a single-frequency receiver to correct ionospheric errors by making use of a global ionospheric model. The corrections are not as accurate as augmentation systems like wide area augmentation systems or dual-frequency receivers. However, it will provide better accuracy as ionospheric error is the largest error source as far as single-frequency GPS receiver is concerned. In the past lack of an almanac in a new receiver would cause long delays before providing a valid position. This was due to the slow process in searching of each satellite. Advances in hardware technology had made the acquisition process much faster, and hence lack of updated almanac is no longer an issue. Each satellite transmits not only its own *ephemeris*, but also transmits an *almanac* for all other satellites is the most important point to remember while acquiring satellite signals.

15.7 TIME CALCULATION

The GPS is a one-way-ranging system in which a clock reading at the transmitter antenna is compared with a clock reading at the receiver antenna. But the satellite clock and the receiver clock will not be in perfect synchronization. The observed signal travel time thus contain a systematic synchronization error which is known as time bias. Biased ranges are called pseudo-ranges. Hence, the basic observation principle of GPS can be regarded as the determination of pseudo-ranges.

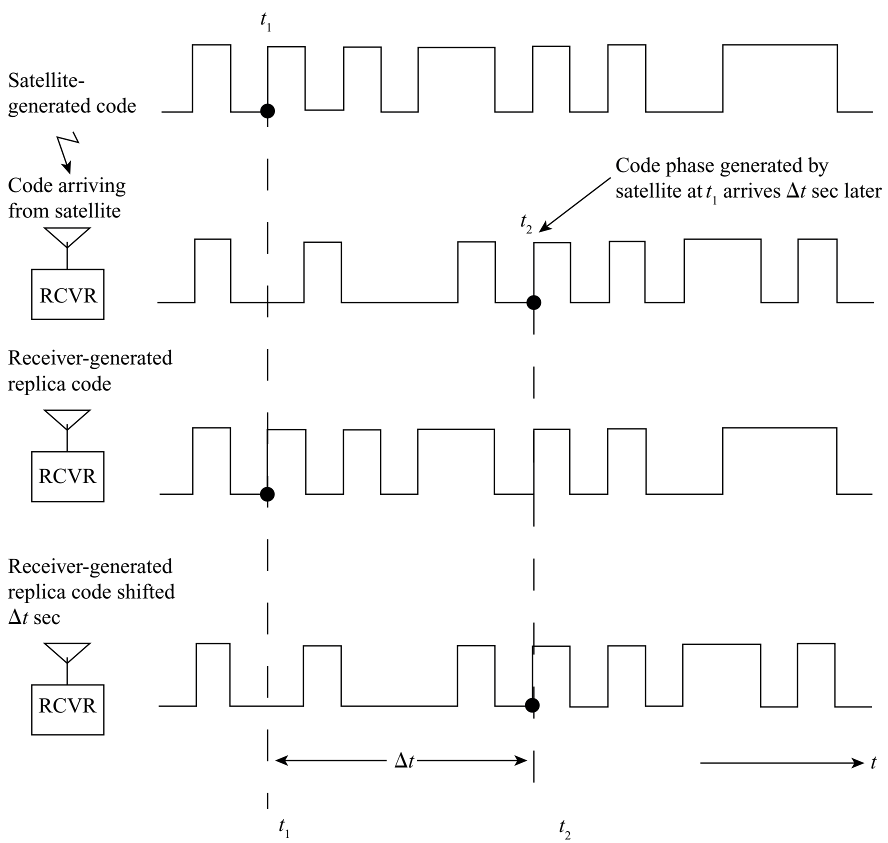


FIGURE 15.6 Code correlation

The GPS receiver requires the time taken by the radio signal to travel from the GPS satellite to the receiver for ranging, and the time is determined using the coded signals that the satellite transmits. The transmitted code is called pseudo-random code because it looks like a noise signal. When a pseudo-random code generated by a satellite reaches the GPS receiver, it generates the replica code and tries to match it with the satellite's code. The receiver then compares the two codes to determine how much delay (or shift) is required in its code to match the satellite code. This delay time (see Fig. 1.3, the time shift Δt) , which when multiplied by the speed of light will give the satellite range or distance of satellite from the receiver. The satellite signal has two codes modulated upon it, the C/A code and the P-code. The C/A code is based upon the time given by an accurate atomic clock. It is a digital code that appears to be random but is repeated 1,000 times in a second. The receiver also contains a clock that is used to generate a matching or replica of C/A code generated by the satellite. The GPS receiver is able to match or correlate the incoming satellite code to the receiver-generated code.

This is how the time taken for the radio signal to travel from the satellite to the GPS receiver is calculated. The measurement of travel time of radio signal from satellite to receiver is very important in position calculation. Therefore, ordinary quartz clocks cannot be used because they are not precise. When the clock is having an error of one by thousand of a second, then the radio signal which is travelling at the speed of light will create an error of 350 km. The satellite time is very precise because it is generated by the atomic clocks on board of the satellite. But the most crucial point in satellite ranging is that the satellite and GPS receiver on earth must be able to synchronize their pseudo-random codes to match the system work. When the GPS receivers are provided with atomic clocks to get timing accuracy, then the GPS receivers would become very expensive due to the high cost of atomic clocks. Hence the designers of GPS system come up with a good solution to achieve high accuracy in the clock of the receivers. The accuracy was achieved by making an extra satellite measurement. When three perfect measurements can locate a point in three-dimensional space, then four imperfect measurements can also do the same thing. This concept is one of the key elements in GPS technology.

When the receiver's clock is perfectly synchronized with satellite clock, then all the satellite ranges would intersect at a single point (which is not practical). With imperfect clock, a fourth measurement, measured by the receiver as a crosscheck, will not intersect with the previous three. So the receiver knows that there is a discrepancy in its measurements and is not perfectly synchronized with the satellite clock which follows universal time. As any offset from universal time will affect measurements, the receiver looks for a single correction factor that it can subtract from all its timing measurements that would cause them all to intersect at a single point. That correction brings the receiver's clock back into synchronization with universal time, hence the receiver gets the atomic accuracy time. When this correction is applied to all its measurements, the receiver gets a precise position. Hence with the pseudo-random code and the perfect synchronized universal time, the receiver gets everything it needs to measure the distance to a satellite in space. But for a triangulation to work, the GPS receiver not only needs to know distances, but also where exactly the satellites are. The receiver can get location of all the satellites at the time of ranging from the almanac data (navigational message) down loaded from the satellite.

15.8 POSITION CALCULATION

There are four unknowns to be determined for the calculation of a position. They are the position coordinates (X, Y, Z) and the precise time of travel of the signal. Observation of four satellites produces four equations, enabling these unknowns to be determined. When a GPS receiver picks up a

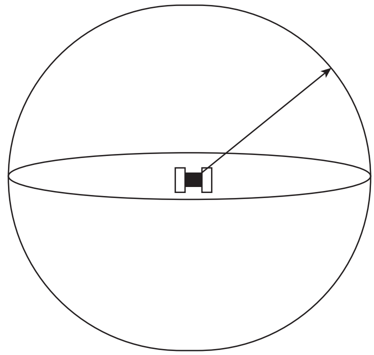


FIGURE 15.7 One satellite in an imaginary sphere with satellite at its centre

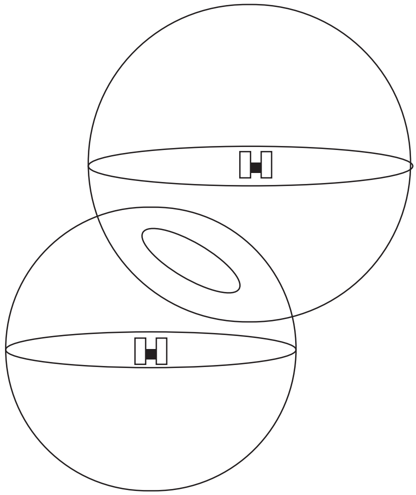


FIGURE 15.8 Intersection of two imaginary spheres with satellites at its centres is a circle

signal from a single satellite, the position of the GPS receiver will be on an imaginary sphere that has satellite at its centre with radius equal to the distance of the satellite from earth (see Fig. 15.7). The GPS receiver then starts picking up signals from two satellites. This helps to narrow down its position. The position of the GPS receiver will now be somewhere on a circle formed due to the intersection of two spheres (see Fig. 15.8).

When the receiver picks up signals from a third satellite, the position of the receiver will be on the point of intersection of the three circles, which is two points, one being the actual position of the receiver and the other being an imaginary point. Here the GPS receiver got a position but there is no check on its accuracy (see Fig. 15.9)

With four satellites the receiver will get a precise position and it will be able to compute the elevation of the position with reference to a selected datum. (See Fig. 15.10.)

Suppose a GPS receiver picks up signals from a satellite, which is about 19,000 km away, the position of the GPS receiver would be somewhere on an imaginary sphere that has the satellite at the centre with a radius of 19,000 km. When the GPS receiver picks up signals from a second satellite placed at a distance of 19,500 km, the second sphere would intersect the first sphere to create a

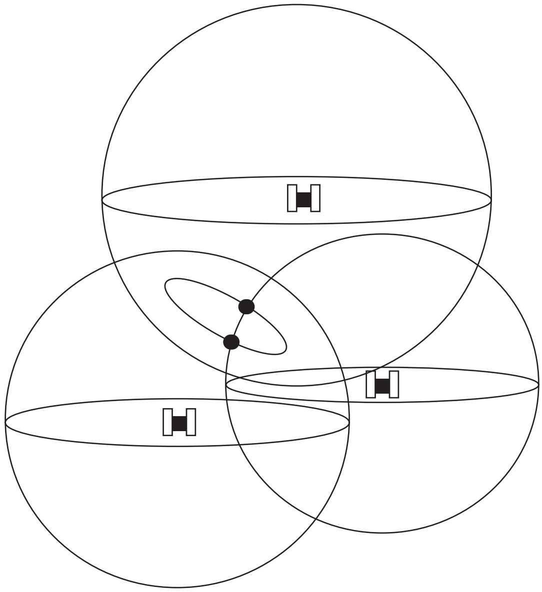


FIGURE 15.9 Intersection of three spheres is two points

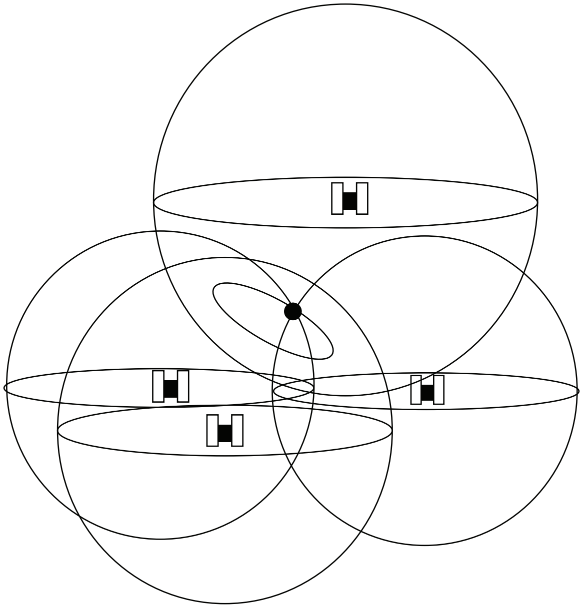


FIGURE 15.10 Intersection of four spheres is a single point

common circle. When the GPS receiver gets signals from a third satellite, which is at a distance of 20,000 km (say) the intersection of three spheres will create two common points. If a fourth satellite is available, the fourth sphere will intersect the first three spheres at one common point, giving the precise position of the receiver.

15.9 POSITIONING SERVICES

To provide services to civil users while ensuring that the national security interests of the US DoD, two types of GPS services are provided: the Standard Positioning Service (SPS) and the Precise Positioning Service (PPS).

15.9.1 SPS

The SPS is designed to provide a less accurate positioning capability than PPS for civil and all other users throughout the world The SPS uses the less precise C/A code pseudo-ranges for position calculation and for real-time GPS navigation. The SPS is freely available to all users worldwide. There are no restrictions on SPS usage.

15.9.2 PPS

The PPS is originally designed for defence purposes and can be tracked with GPA equipment equipped with PPS receivers. In PPS pseudo-ranges are obtained using the higher pulse rate P-code on both frequencies (L1 and L2), thereby giving higher accuracy. Real-time three-dimensional accuracies at sub-metre level (below 10 m horizontal) can only be achieved with PPS. The P-code is encrypted to prevent unauthorized civil or foreign use and requires a special key to obtain the accuracy offered by PPS. The PPS is specified to provide a predictable accuracy of at least 2 drms, 95% in the horizontal plane. The distance root mean square (or drms) is a common measure used for navigation. Twice the drms value, or 2 drms, is the radius of a circle that contains at least 95% of all possible fixes that can be obtained with a system at any point. PPS is primarily intended for military purposes. Civilian use is permitted only upon special US DoD approval. Access to the PPS position accuracies is controlled through two cryptographic feature denoted as antispoofing (AS). The antispoofing is a mechanism intended to defeat deception jamming. Deception jamming is a technique in which an adversary would replicate one or more satellite ranging codes, navigation data signals and carrier frequencies. The PPS was declared fully operational in the year 1995, when entire GPS satellite constellation was made fully operational.

15.10 CURRENT GPS SATELLITE CONSTELLATION

The current GPS constellation (as of July 2001) contains five Block II, 18 Block IIA and six Block IIR satellites (see Table 15.2). This makes the total number of GPS satellites in the constellation to be 31, which exceeds the original design concept of 24-satellite constellation by seven spare satellites.

15.11 GPS ERRORS AND THEIR CORRECTIONS

There are various errors resulting from various sources which affects satellite ranging and hence affects GPS performance. **Errors in range measurements create a range of uncertainty around the GPS position**. The sum of all systematic errors or biases contributing to the measurement error is referred to as range bias. The observed GPS range, without removal of biases, is referred to as a biased range or pseudo-range. GPSA errors can arise from inaccuracies in estimate of satellite position and satellite clock and electronic inaccuracies, tropospheric and ionospheric effect along signal

TABLE 15.2 Current constellation of GPS satellites as on March 1, 2017 (31 satellites)

Satellite Type	Pseudo-Random Number (PRN)	Satellite Vehicle Number (SVN)	Date of Launch	Type of Oscillator (Atomic Clock)	Orbital Plane and Slot
IIR-2	13	43	Jul 23, 1997	Rubidium	F6
IIR-3	11	46	Oct 7, 1999	Rubidium	D5
IIR-4	20	51	May 11, 2000	Rubidium	B6
IIR-5	28	44	Jul 16, 2000	Rubidium	B3
IIR-6	14	41	Nov 10, 2000	Rubidium	F1
IIR-7	18	54	Jan 30, 2001	Rubidium	E4
IIR-8	16	56	Jan 29, 2003	Rubidium	B1
IIR-9	21	45	Mar 31, 2003	Rubidium	D3
IIR-10	22	47	Dec 21, 2003	Rubidium	E2
IIR-11	19	59	Mar 20, 2004	Rubidium	C3
IIR-12	23	60	Jun 23, 2004	Rubidium	F4
IIR-13	02	61	Nov 6, 2004	Rubidium	D1
IIR-14M	17	53	Sep 26, 2005	Rubidium	C4
IIR-15M	31	52	Sep 25, 2006	Rubidium	A2
IIR-16M	12	58	Nov 17, 2006	Rubidium	B4
IIR-17M	15	55	Oct 17, 2007	Rubidium	F2
IIR-18M	29	57	Dec 20, 2007	Rubidium	C1
IIR-19M	07	48	Mar 15, 2008	Rubidium	A4
IIR-20M	04	49	Mar 24, 2009	Rubidium	
IIR-21M	05	50	Aug 17, 2009	Rubidium	E3
IIF-1	25	62	May 28, 2010	Rubidium	B2
IIF-2	01	63	Jul 16, 2011	Rubidium	D2
IIF-3	24	65	Oct 4, 2012	Cesium	A1
IIF-4	27	66	May 15, 2013	Rubidium	C2
IIF-5	30	64	Feb 21, 2014	Rubidium	A3
IIF-6	06	67	May 17, 2014	Rubidium	D4
IIF-7	09	68	Aug 2, 2014	Rubidium	F3
IIF-8	03	69	Oct 29, 2014	Rubidium	E1
IIF-9	26	71	Mar 25, 2015	Rubidium	B5
IIF-10	08	72	Jul 15, 2015	Cesium	C5
IIF-11	10	73	Oct 31, 2015	Rubidium	E6

propagation path, , atmospheric absorption, receiver noise generated through signal processing errors, signal multipath effect etc..,

Errors that are induced by Department of Defence for the intentional degradation of signal are the selective availability (S/A) and AS. In addition to these major errors, GPS also contains random observation errors, such as unexplainable and unpredictable time variation. These errors are impossible to model and correct. Many of the errors are either eliminated or significantly minimized when GPS is used in differential mode. This is because these errors are common to two or more receivers during simultaneous observing sessions and hence get cancelled.

The GPS errors are associated with absolute GPS positioning mode are:

i. Ephemeris errors and orbit perturbations
ii. Clock stability
iii. Ionospheric delays
iv. Troposhperic delays
v. Multi-path
vi. Satellite and receiver clock errors
vii. Selective availability
viii. Anti spoofing
ix. Receiver noise.

15.11.1 Ephemeris Errors and Orbit Perturbations

Satellite ephemeris errors are errors in the prediction of a satellite position which may then be transmitted to the user in the satellite data message. Ephemeris errors are satellite dependent and very difficult to correct and compensate while modelling the orbit of a satellite because many forces acting on the predicted orbit of a satellite are difficult to measure directly. Ephemeris errors produce equal error shifts in the calculated absolute point positions.

15.11.2 Clock Stability

GPS depends on accurate time measurements. GPS satellites carry rubidium and caesium time standards that are usually accurate to 1 part in 10^{12} and 1 part in 10^{13}, respectively, while most receiver clocks are accurate by a quartz standard accuracy of 1 part in 10^8. A time offset is the difference between the time recorded by the satellite clock and that recorded by the receiver. Range errors observed by the user as a result of the time offset between the satellite and receiver clock have a linear relationship and can be approximated by the following relation.

$$R_E = T_0 \times c$$

where R_E is the user equivalent range error (UERE); T_0 the time offset c the speed of light.

Example for the calculation of user equivalent range error:

When time offset $T_0 = 1.5$ μsec (1.5×10^{-6})

c = 29,97,92,458 m/sec

$R_E = 1.5 \times 10^{-6} \times 299792458 = 450$ m = user equivalent range error

In general unpredictable transient situations that produce high-order deviations in clock time can be ignored over short periods of time. The predictable time drift of the satellite clocks is closely monitored by the ground control stations. Through this the ground control stations are able to determine second-order polynomials, which accurately model the time drift. The second-order polynomial is included in the broadcast message in an effort to keep this drift to within 1 msec. The time synchronization between the GPS satellite clocks is kept within 20 nsec and the synchronization of GPS standard time to the Universal Time Coordinated (UTC) to within 100 nsec through the broadcast clock correction as determined by the ground control stations. Random time drifts are unpredictable, thereby making them impossible to model.

GPS receiver clock errors can be modelled in a manner similar to GPS satellite clock errors. In addition to modelling the satellite clock errors and in an effort to remove them, an additional satellite should be observed during operation to obtain an extra clock offset parameter along with the required coordinate parameters. This procedure is based on the assumption that the clock bias is independent at each measurement epoch. Rigorous estimation of the clock terms is more important for point positioning than for differential positioning. Many of the clock terms cancel when the position equations are formed from the observations during a differential survey session.

15.11.3 Ionospheric Delays

GPS signals are electromagnetic signals and as such are nonlinearly dispersed and refracted when transmitted through a highly charged environment like the ionosphere. Dispersion and refraction of the GPS signal is referred to as the ionospheric range effect because it results in an error in the GPS range value. Also as the satellite signal passes through the ionosphere, it is slowed down because the medium is denser compared with outer space.

These delays can introduce an error in the range calculation as the velocity of the radio signals from the satellite is affected. The ionospheric range effects are frequency dependent and are not constant. There are several factors that influence the amount of delays caused by the ionosphere.

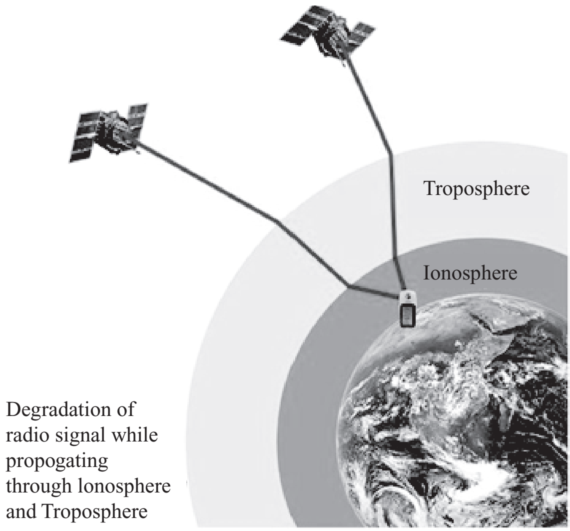


FIGURE 15.11 Degradation of radio signal while propagating through ionosphere and troposphere

15.11.3.1 Satellite Elevation and Satellite Geometry

Signals from low-elevation satellites will be affected more than the signals from high-elevation satellites. This is due to the increased distance that the signal has to travel through the atmosphere in the case of lower elevation satellites. Hence a signal originating from a satellite near the observer's horizon must pass through a longer distance of the ionosphere to reach the receiver than a signal from a satellite near the observer's zenith. The longer the signal is in the ionosphere, the greater the ionospheric effect.

15.11.3.2 Density of Ionosphere

The long travel of the GPS signal through the virtual vacuum of space changes as it passes through the earth's atmosphere. The atmosphere changes the apparent speed of the signal through refraction and diffraction and hence the direction of the signal alters the direction of the signal. This causes an apparent delay in the radio signal transmitted from the satellite to the receiver. The ionosphere contains ionized plasma comprised of negatively charged electrons which remain free for long periods before being captured by positive ions. It extends from about 50 to 1,000 km above the earth's surface and is the first part of the atmosphere that the signal came across on its propagation towards earth. The magnitude of these delays is determined by the state of the ionosphere at the time at which the signal passes through it. Hence it's important to note that its density and stratification of charged ions in ionosphere varies. The sun plays a key role in the creation and variation of these aspects as the daytime ionosphere is rather different from the ionosphere at night.

At night, there are very little ionospheric influences and during day time the light and heat from the sun increases the effect of the ionosphere and slows down the signal. The electron density of the ionosphere is also increased with solar cycles (sunspot activity). During the daylight hours in the midlatitudes the ionospheric delay may be as much as 5 times greater that it is at night. Ionospheric delay is usually least between midnight and early morning and maximum at about local noon. It is also nearly four times greater in November, when the earth is nearing its *perihelion*, its closest approach to the sun, than it is in July near the earth's *aphelion*, its farther point from the sun. The effects of the ionosphere on the GPS signal usually reach their peak in March, at the time of vernal equinox.

Sunspot activity peaks approximately once in 11 years. In addition to this, solar flares can also randomly occur and also have an effect on ionosphere. The ionospheric errors may be as large as 150 m when the satellite is near the observer's horizon, the vernal equinox is near and the sunspot activity is at its maximum. However, it is usually not that much severe. It varies with magnetic activity, location, time of day and the direction of observation.

15.11.3.3 Resolution of Ionospheric Delay

The ionospheric delay can be divided into two distinct categories, the phase delay and the group delay. All the modulations on the carrier wave, the P-code, the C/A code, and the navigation message, appear to be slowed while passing through ionosphere. They are affected by the Group delay. But the carrier wave itself appears to speed up in the ionosphere. It is affected by the phase delay.

Resolution of ionospheric delays can be accomplished by using a dual frequency GPS receiver (a receiver capable of tracking both L1 and L2 frequency measurements). During a period of uninterrupted observation of the L1 and L2 signals, these signals can be continuously counted and differenced. The resultant difference reflects the variable effects of the ionospheric delay on the GPS signal. The time delay for a higher frequency carrier wave is less than it is for a lower frequency wave due to the dispersive nature of the ionosphere. That means that L1 carrier wave which is at 1575.42 MHz is not affected as much as L2 carrier wave which is 1227.60 MHz. A single frequency GPS receiver is able to track only the L1 signal where as a dual frequency GPS receiver is able to track both L1 and L2 signals. This fact provides one of the greatest advantages of a dual frequency GPS receiver over the single frequency receivers. By tracking both carrier waves (L1 and L2 signals) a dual frequency receiver has the facility of modelling and removing not all, but a significant portion of the ionospheric bias. The frequency dependence of the ionospheric effect is described by the following expression.

$$v = \frac{40.3}{cf^2}.\mathrm{TEC}$$

where v is the ionospheric delay,

c the speed of light in metres per second,

f the frequency of the signal in Hz, and

TEC the quantity of free electrons per cubic metre.

As per the above equation the time delay is inversely proportional to the square of the frequency. In other words, the higher the frequency, the less is the delay. Hence the dual-frequency receiver is capable to discriminate the effect on L1 from that on L2. Single frequency GPS receivers used for absolute and differential positioning mode typically depend on an ionospheric models that model the effects of the ionosphere. Recent studies have shown that ionospheric delay removal can be achieved using single frequency receiver as well

15.11.4 Troposphere Delays

The troposphere is that part of the atmosphere closest to the earth. It extends from the surface of earth to about 9 km over the poles and 16 km over the equator. When GPS signal is concerned the troposphere is taken as a combined part of stratosphere and taken to be at a height of 50 km above the surface of earth. The tropospheric delay adds a slight distance to the range the receiver measures between itself and the satellite. The troposphere is an electrically neutral layer of the earth's atmosphere. Hence it is neither ionized nor dispersive. Therefore, the GPS signals are refracted while in troposphere. Like the ionosphere, the density of the troposphere also governs the severity of its effect on the GPS signal. The delay caused by the troposphere is maximum when the satellite is close to horizon and minimum when the satellite is at zenith, that is, directly above the GPS receiver.

The troposphere and the ionosphere are by no means alike in their effect on the satellite's signal. While the troposphere is refractive its refraction of a GPS satellite's signal is not related to its frequency. The refraction is the amount of delay in the arrival of a GPS satellite's signal. It can also be conceptualized as a distance added to the range the receiver measures between itself and the satellite. The troposphere is part of the electrically neutral layer of the earth's atmosphere meaning it is not ionized. The troposphere is also non-dispersive for frequencies below 30 GHz or so. Therefore L1, L2 and L5 are equally refracted. This means that the range between a receiver and a satellite will be shown to be a bit longer than it actually is.

When a satellite at horizon, the delay of the signal caused by the troposphere is maximized, and when a satellite is at zenith the tropospheric delay is minimized. The GPS signal that travels the shortest path through the troposphere will be the least delayed by it. So, even though the delay at an elevation angle of 90 degrees at sea level will only be about 2.4 m, it can increase to about 9.3 m at 75 degrees and up to 20 m at 10 degrees. There is less tropospheric delay at higher altitudes.

Modelling the troposphere is a technique used to reduce the bias in GPS data processing, and it can be up to 95% effective. However, the residual 5% can be quite difficult to remove. Refraction in the troposphere has a dry component and a wet component. The dry component which contributes most of the delay, perhaps 80% to 90%, is closely correlated to the atmospheric pressure. The dry component can be more easily estimated than the wet component and the dry component contributes the most—80–90% of the attenuation.

Even though the wet component of the troposphere is nearer to the earth's surface, measurements of temperature and humidity are not strong indicators of conditions on the path between the receiver and the satellite. While instruments that can provide some idea of the conditions along the line between the satellite and the receiver are somewhat more helpful in modelling the tropospheric effect the high cost of sending water vapour radiometers and radiosondes aloft generally restricts their use to only the most high-precision GPS work. In most cases this aspect must remain in the purview of mathematical modelling, such calculations include a hydrostatic model with corrections and a horizontal gradient component. It is important to recognize that the index of tropospheric refraction decreases as height increases.

The troposhperic conditions causing refraction of the GPS signal can be modelled by measuring the dry and wet components. The dry component is best approximated by the following equation.

$$D_c = (2.27 \times 0.001)\, P_0$$

where D_c is the dry term range contribution in zenith direction in metres and P_0 the surface pressure in millimetres.

For example, if the atmospheric pressure is 760 mm of Hg.

$D_c = (2.27 \times 0.001)\ 760 \sim 1.7$ m, is the dry term ranges error contribution in the zenith direction.

The tropospheric delay appears to add a slight distance to the range the receiver measures between itself and the satellite. The troposphere and the ionosphere are by no means alike in their effect on the satellite signal. The troposphere is non-dispersive for frequencies below 30 GHz. That is the refraction of a GPS satellite signal is not related to its frequency in the troposphere. Modelling the troposphere is one among the technique used to reduce the bias in GPS data processing and it can be up to 95% effective.

15.11.5 Signal Multipath

Signal Multipath is an error affecting positioning that occurs when the signal arrives at the receiver from more than one path due to reflection of signals. This occurs when the GPS receiver is positioned close to large reflecting surfaces such as a lake, a big rock, a building or other manmade structures. In the case of reflected signal, the satellite signal does not travel directly to the antenna but hits the nearby object and is reflected into the receiver's antenna creating a false measurement. This increases the travel time of the signal, thereby causing errors. Multipath effect depends on the antenna closeness to the reflecting structures, and it is important when the signal comes from the satellite with low elevation.

Multipath error is different for different frequencies. It affects the phase measurements, as well as the code measurements. The interference by multipath is generated when a signal arrives at

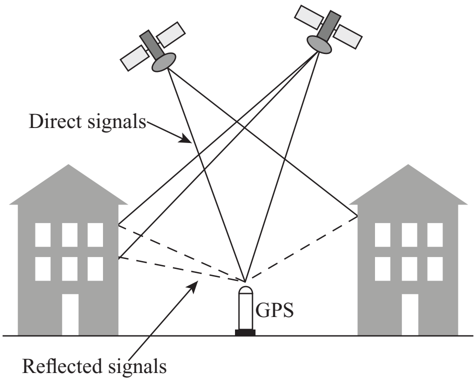


FIGURE 15.12 Signal multipath

FIGURE 15.13 Choke ring antenna. (Courtesy of M/S.Trimble.)

the antenna as shown in Fig. 15.12. Now most modern receivers use sophisticated signal rejection technique to minimize signal multipath problem

Good receivers use sophisticated signal rejection technique to minimize this problem. For example by using a choke ring antenna that incorporate a ground plane the low elevation signal reaching the antenna can be prevented. A choke ring antenna is a unidirectional antenna and it consists of a number of concentric metallic cylinders fences (metallic rings) around a central antenna. The concentric circles along with antenna in its centre are often enclosed in a protective cover (dome cover). Choke ring are used to reject multipath signals coming from a source. As the path that a signal takes from a transmitter to receiver is used to measure the distance between the two, choke ring antenna makes it highly recommended for Precise GPS observations required for surveying and other applications where millimetre precision is required. Rejections of Multipath signals are necessary for only high accuracy measurements.

15.11.6 Satellite and Receiver Clock Errors

GPS receivers use ordinary quartz crystal oscillators, to keep the cost of GPS receivers at an economical level. These oscillators are small and consume less power. In absolute positioning, the receiver clock offset has to be estimated as an unknown parameter in the navigation solution which estimates the receiver position and receiver clock at the same time. The receiver clock offset can be estimated within 1 μs or better. The limit in the accuracy results from the effects of measurement noise and the atmospheric impact on the GPS signal. A standard model ionospheric and tropospheric correction has to be applied or in the case where dual Frequency GPS receivers are used the ionospheric free combination can be employed.

In the single frequency navigation solution the clock error is dominant over atmospheric residuals and noise, when raw observations are used. Even though the clocks in the satellite are very accurate to about 3 nsec, they do sometimes drift slightly and cause small errors, affecting the accuracy of the position. The satellite clocks are independent of each other. The clocks are made up of rubidium and caesium oscillators. These oscillators are stable unless frequent tweaking does not disturb them and adjustment is kept to a minimum. While GPS time itself is designated to be keeping within one microsecond, the satellite clocks can be allowed to drift up to a millisecond from GPS time. The control segment of GPS system continuously monitors the satellite clocks and corrects the drift if any.

15.11.7 Selective Availability

S/A was an intentional degradation of public GPS signals implemented for national security reasons. It is a process applied by the US DoD to the GPS signal and is intended to deny civilian and hostile foreign powers from getting full accuracy of GPS. It is done by subjecting the satellite clocks to a process known as dithering, which alters their time slightly. In addition, the ephemeris (or path that the satellite follow), which is broadcasted from satellite, is slightly changed altered from what it is in reality.

Selective availability is the biggest single source of inaccuracy in the GPS system. Military receivers use a decryption key to remove the selective availability errors and so they are much more accurate. The selective availability was turned off by the Department of Defense on May 2, 2000, and is currently not active.

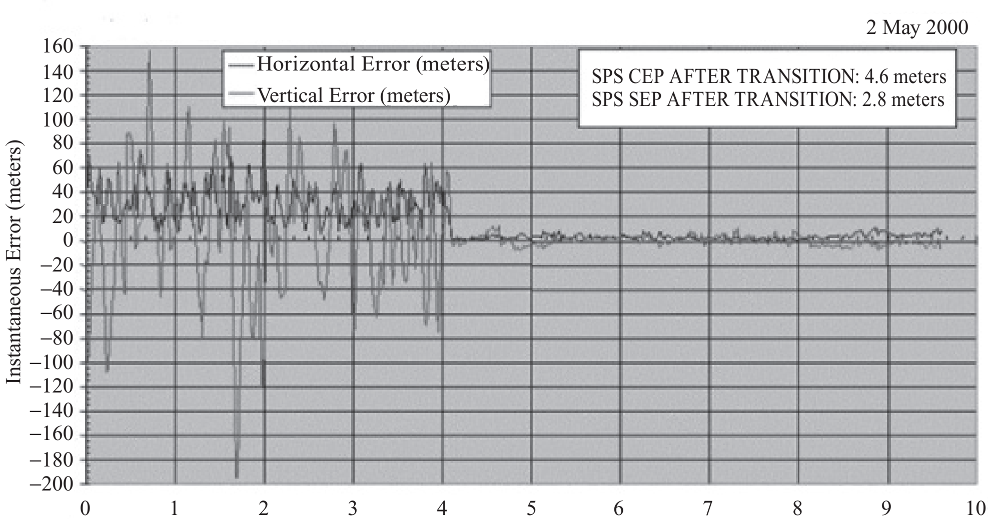


FIGURE 15.14 GPS positioning accuracy before and after switching off of S/A

15.11.8 Anti-Spoofing (A-S)

Anti-spoofing (AS) of the GPS system is designed for an anti potential spoofer (or jammer). A spoofer generates a signal that is similar to the GPS signal and attempts to cause the receiver to track the wrong signal. When the AS mode of operation is activated, the P-code will be replaced with a secure Y code which is available only to military grade receivers, and other receivers become a single L1 frequency receiver. Anti-Spoofing is made to deny the civilian access to the P-code part of the GPS signal, thereby forcing the user to use the C/A code only, which has selective availability applied to it. Anti-Spoofing encrypts the P-code into a signal called the Y-Code. Only users with military GPS receivers can de-crypt the Y-code. Military receivers are more accurate because they do not use the Coarse/Acquisition code to calculate the time taken by the signal to reach the receiver.

15.11.9 Receiver Noise

GPS receivers are not perfect device. They cannot measure the GPS observables with infinite precision. They are always subjected to some level of noise that contaminates the observations. How precisely a GPS receiver can measure the pseudo-range and carrier phase largely depends on how much noise accompanies the signals in the receiver's tracking loops. The more the noise, the worse the performance of the receiver is. This noise either comes from the receiver electronics itself or is picked up by the receiver's antenna. The effects of receiver noise on carrier phase measurements are small when compared to their effects on pseudo-range measurements. Generally, the receiver noise error is about 1% of the wavelength of the signal involved. In other words in code solutions the size of the error is related to chip width. For example, the receiver noise error in a C/A code solution can be around 3 m which is about an order of magnitude more than it is in a P-code solution, about 3 cm. And in carrier phase solutions the receiver noise error contributes millimetres to the overall error.

Receiver noise includes a variety of errors associated with the ability of the GPS receiver to measure a finite time difference. These include signal processing, clock/signal synchronization and correlation methods, receiver resolution, signal-to-noise ratio, etc. Receiver noise is unavoidable and is a relatively small contributor to the GPS error budget. Signal multipath error and receiver noise are uncorrelated, that is, that are not related to the length of the baseline between GPS receivers.

15.12 USER EQUIVALENT RANGE ERROR

The ranging errors described above are principal contributors to the overall GPS error. UERE is the summary of the total error budget affecting a pseudo-range summarized above. User Equivalent Range Error refers to the error of a component in the distance from receiver to a satellite. These UERE errors are given as ± errors thereby implying that they are unbiased or zero mean errors. User equivalent range error used in satellite is the square root of the sum of the squares of the individual biases.cd` These errors can be removed or at least effectively suppressed by developing models of their functional relationships in various parameters that can be used as a corrective supplement for the basic GPS information.

Dilution of Precision

Dilution of Precision (DOP) is a term used to measure the strength of satellite geometry or the strength of a satellite configuration and is related to the spacing and position of GPS satellites in the

TABLE 15.3 Limits of DOP that should be considered while GPS observation

DOP Value Calculated by the Receiver	Normal Rating	Explanation on the Rating
<1	Ideal and best	Best possible confidence level to be used for applications which require highest precision
1–2	Excellent	At excellent confidence level, positional measurements are considered accurate and the accuracy level is good enough to meet most of the sensitive applications
2–5	Good	This level is appropriate for making positional measurements reliable for navigation and for large scale mapping
5–10	Moderate	Positional Measurements may be used but a more open view of sky is recommended
10–20	Fair	Represents a low confidential level. Position information should be discarded or shall be used for a very rough estimate of position.
>20	Poor	Position Measurements are in accurate as much as 300m and should not be adopted for any purpose.

sky. The arrangement of satellites in the sky also affects the accuracy of GPS positioning. The ideal arrangement (of the minimum four satellites) is one satellite directly overhead, three others equally spaced near the horizon. The relative satellite-receiver geometry is important in determining the accurate positions. The precision of multiple satellites in view of a receiver combine according to the relative position of the satellites to determine the level of precision in each dimension of the receiver measurement. When satellites are clustered close together in the sky, the geometry is said to be weak and the DOP value is high, and when the satellites are well arranged apart, the geometry becomes strong and the DOP value will be low. Hence it should be noted that the DOP factor multiplies the uncertainty associated with UERE. Let us say that the DOP for an ideal arrangement of the satellite constellation gives a lesser value (say DOP value of 1), which will not magnify UERE., then a DOP value of 2 can doubles the uncertainty associated with UERE. The DOP values for various types of observations are shown in Table 15.3.

GPS receivers report several components of DOP, including horizontal dilution of precision (HDOP) and vertical dilution of precision (VDOP). The combination of these two components of the three-dimensional position is called position dilution of precision (PDOP). A key element of GPS mission planning is to identify the time of day when PDOP is minimized. Since satellite orbits are known, PDOP can be predicted for a given time and location. Various software products allow you to determine when conditions are best for GPS work.

15.12.1 Geometric Dilution of Precision

Geometric dilution of precision (GDOP) is a quantity which is used to determine the information content due to satellite geometry and results in a measure of the overall geometrical strength to the navigation solution. It provides a method of quantitatively determining whether particular satellite geometry is good or bad. To achieve high accuracy in GPS positioning systems, both accurate measurements and good geometric relationship between the satellite geometry and the receiver are required. GDOP is widely used as a criterion for selecting measurement units, as it represents the geometric effect on the relationship between measurement error and positioning determination error. In the calculation of GDOP value, the maximum volume method does not necessarily guarantee the selection of the optimal four measurement units with minimum GDOP. The general object of the GPS satellite selection algorithm is to minimize the GDOP to improve the position accuracy. The smaller value of GDOP is calculated, the better geometric configuration we will have. The redundant measurements will bring large amount of computation and may not provide significantly improved location accuracy. When enough measurements are available, the optimal measurements selected with the minimum GDOP can prevent the poor geometry effects and have the potential of obtaining greater location accuracy.

The main form of DOP used for absolute GPS positioning is the GDOP. As mentioned earlier, it is a measure of accuracy in a three-dimensional position and time. The relationship between final positional accuracy, actual range error and GDOP can be expressed as

$$\sigma_a = \sigma_R \times \text{GDOP}$$

where σ_a is the final positional accuracy and σ_R the actual range error (UERE).

$$\text{GDOP} = [(\sigma^2_E + \sigma^2_N + \sigma^2_U + (c \,.\, \delta_T)^2)^{1/2}]/\sigma_R$$

where σ_E is the standard deviation in east value, in metres,

σ_N the standard deviation in north value, in metres,

σ_U the standard deviation in up direction, in metres,

c the speed of light = 29,97,92,458 m/sec,

δ_T the standard deviation in time, s, and

σ_R the overall standard deviation in range, in meters, usually in the range of 6 m for P-code usage and 12 m for C/A code usage.

15.12.2 Positional Dilution of Precision

PDOP can be interpreted as the reciprocal of the volume V of a tetrahedron that is formed from the satellite and user positions or PDOP = $1/V$. The best geometric situation exists when the volume is maximized, and hence PDOP is minimized.

$$\text{PDOP} = \{[\sigma^2_E + \sigma^2_N + \sigma^2_U]^{1/2}\}/\sigma_R$$

where, all the symbols have the same meaning as above. PDOP is the combination of both the horizontal and vertical components of position error caused by satellite geometry.

PDOP value between 2 and 4 is considered as excellent, 4 and 6 is considered as good, 6 and 8 is considered as fair, 8 and 10 as poor, and 10 and 12 as marginal and above 12 is not advised to use. PDOP represent position recovery at any instant in time and is not representative of a whole session time. PDOP error is generally given in unit of metres of error per 1 m error in the pseudo-range measurements. PDOP values in the range 4–5 m/m is considered very good for pseudo-ranging and PDOP values greater than 10 m/m is considered as poor. High values of PDOP and GDOP over a period of time are associated with satellites in a constellation of poor geometry. Higher the values of PDOP or GDOP, makes the solution poorer for that instant of time. PDOP values are generally developed from satellite ephemeris prior to the survey so that they can be used to determine the adequacy of a

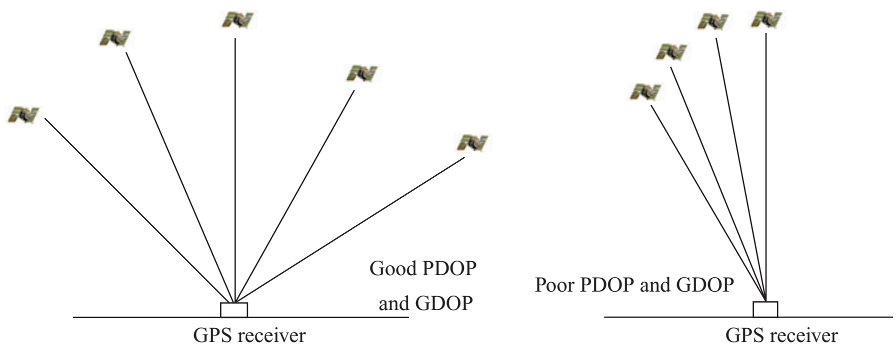


FIGURE 15.15 Good GDOP and PDOP and bad GDOP and PDOP

particular survey schedule. PDOP represents position recovery at any instant in time and is not representative of a whole session of time.

15.12.3 Horizontal Dilution of Precision

HDOP is important in evaluating GPS surveys intended for establishing horizontal control HDOP relates to the horizontal position measurements suggested by GPS data. HDOP is basically the error determined from the final variance-covariance matrix divided by the standard error of the range measurements.

$$\text{Mathematically HDOP} = [(\sigma^2{}_E + \sigma^2{}_N)^{1/2}]/\sigma_R$$

HDOP will be high when GPS receiver is observing a partial satellite constellation.

15.12.4 Vertical Dilution of Precision VDOP

VDOP tends to be larger at higher altitudes (NORTH OR SOUTH). The reason is that the inclination of the GPS orbits is 55 degrees, and at higher altitudes there are no satellites overhead. A low VDOP require a satellite neat zenith and good coverage in azimuth with satellites that are close to the horizon. Actually VDOP would be lower if a receiver could track satellites below the horizon, as would be case with space borne receivers.

$$\text{VDOP} = \sigma_U/\sigma_R$$

GDOP is a combination of all Dilution of Precisions Hence it is the most useful DOP. Some receivers are capable of calculating PDOP and HDOP. The best way to minimize GDOP is to observe as many as satellites possible.

15.13 PSEUDO-RANGE OBSERVATION EQUATION

GPS is primarily a navigation system. The fundamental navigation principle is based on the measurement of so called pseudo-ranges between the used and four satellites. The GPS receiver computes its three-dimensional coordinates and its clock offset from four or more simultaneous pseudo-range measurements or biased ranges between the receiver and each of the satellites being tracked. The pseudo-range is calculated by cross-correlating the pseudo-random noise code received from the satellite and with a replica generated by the receiver. The basic pseudo-range equation is given by

$$p = \rho + \mathrm{d}\rho + c(\mathrm{d}t - \mathrm{d}T) + d_{\mathrm{ion}} + d_{\mathrm{trop}} + \varepsilon_{\mathrm{mp}} + \varepsilon_{\mathrm{p}}$$

where p is the pseudo-range measurement,

ρ the true range between receiver's antenna and the satellite's antenna at the time of signal transmission,

$\mathrm{d}\rho$ satellite orbital errors,

c the speed of light,

$\mathrm{d}t$ the satellite clock offset from GPS time,

dT the receiver clock offset from GPS time,

d_{ion} the ionospheric propagation delay on pseudo-range,

d_{trop} the tropospheric propagation delay on pseudo-range,

$\varepsilon_{\mathrm{mp}}$ the multipath on pseudo-range,

ε_{p} the receiver noise, and

c the the speed of light.

Clock offsets are only one of the errors in pseudo-ranges. Note that the pseudo-range, p, and the true range, ρ, cannot be made equivalent, without consideration of clock offsets, atmospheric effects and other biases that are inevitably present.

The above pseudo-range observable equation can also be written as

$$P = \rho + c(\mathrm{d}t - \mathrm{d}T) + d_{\mathrm{ion}} + d_{\mathrm{trop}} + \varepsilon$$

where ε represents the combined effect of multipath and receiver measurement noise. Pseudo-range measurements can be made using either the C/A code or P-code. The P-code generally provides higher observations because 300 m resolution of the C/A code is 10 times lower than the P-code. However, recent technological advancement in receiver technologies have resulted in higher precision C/A code measurements such as narrow code correlation etc. some of the ranging errors can be corrected, such as the tropospheric delay can be corrected using a tropospheric delay model, and ionospheric delays can be corrected using a dual frequency GPS receiver. The errors will cause inaccuracy in the calculation of user position. It should be noted that the user clock error cannot be corrected through received information. Thus it will remain as an unknown.

Assuming the receiver accounts for the satellite clock offset (using navigation message) and atmospheric delays (from models programmed into the firmware), the pseudo-range observable equation can further be simplified as follows

$$\mathrm{P_c} = \rho + c \cdot \mathrm{d}T + e_{\mathrm{c}}$$

where e_{c} represents the original measurement noise model errors and any un-modelled effects such as S/A, etc. There are n numbers of such equations that a receiver must solve using the n-simultaneous measurements. The parameter ρ is a non-linear function of the receiver and satellite coordinates.

15.14 CARRIER PHASE OBSERVATION EQUATION

The carrier phase measurement is a measure of the range between a satellite and receiver expressed in units of cycles of the carrier frequency. This measurement can be done with very high precision. It is very easy to measure the fraction of a carrier wave, but there will be an error in measuring the number of full carrier cycles. For example just imagine a measuring tape extending from the satellite to the receiver that has numbered markers every 1 mm. But the numbering scheme of the tape returns to zero with every wavelength (~20 cm for GPS L1). This allows measuring the range very precisely, but with an ambiguity in the number of whole carrier cycles. Hence in carrier phase measurement, the whole number of cycles between satellite and receiver is not measurable. The basic carrier phase observable is a measure of the difference between the carrier signal generated by the receiver's internal oscillator and the carrier signal from the satellite. The Carrier phase observable is the number of full carrier cycles and the fractional cycle, between the antennas of the satellite and the receiver. The main problem in carrier phase tracking is that the GPS receiver has a way of distinguishing one carrier cycle from another. The best it can do is to measure the fractional phase and keep track of changes to the phase. Hence the initial phase is undermined or ambiguous, by an integer number of cycles. To use the carrier phase as an observable for positioning this unknown number of cycles or the phase ambiguity must be estimated.

$$\phi = \rho + \mathrm{d}\rho + c(\mathrm{d}t - \mathrm{d}T) + \lambda N - d_{\mathrm{ion}} + d_{\mathrm{trop}} + d_{\phi\mathrm{mp}} + \in_{\phi}$$

where ϕ is the carrier phase observation in unit of metres,

$d_{\phi\mathrm{mp}}$ the multipath on the carrier phase,

$\in_{\phi}$ the carrier phase observation noise,

λ the wave length of carrier phase,

N the integer ambiguity,

ρ the satellite position error,

$\mathrm{d}\rho$ the effect of ephemeris error,

$\mathrm{d}t$ the satellite clock error with respect to GPS time,

$\mathrm{d}T$ the receiver clock error with respect to GPS time,

d_{ion} the ionospheric delay,

d_{trop} the tropospheric delay, and

c the velocity of light.

15.15 MASK ANGLE

Mask angle is the minimum acceptable satellite elevation above the horizon to avoid blockage of line-of-sight. Mask angle is an elevation mask filter to ignore satellites that are below a certain angle on the horizon. The angle is modifiable by the user of a GPS receiver, depending on the type of receiver used. For example, if the elevation mask on a GPS receiver is set to 15 degrees, the GPS receiver would be unable to use the satellite that are rising or setting <15 degrees elevational cone around the position of GPS receiver. Signals from low-elevations satellites are generally influenced to a great extent by noise sources. As a general rule it is best to observe satellites that are >15 degrees above the horizon for surveying using GPS. This is known as mask angle (see Fig. 15.15).

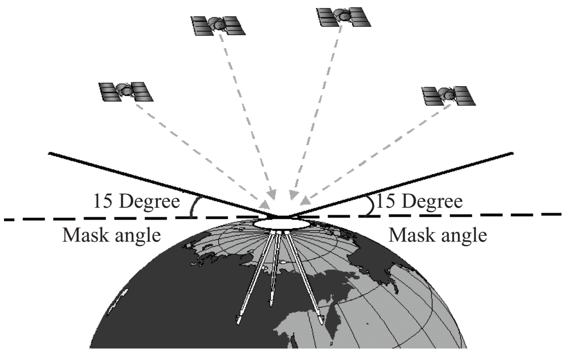


FIGURE 15.16 Mask angle

A typical mask angle is between 10 and 15 degrees. Ranging errors can be minimized by using a mask angle between 10 and 15 degrees, depending upon the terrain, visibility and geographical conditions.

REVIEW QUESTIONS

1. What are the basic concepts of Global Positioning System?
2. Define GPS and narrate its advantages.
3. A number of stringent conditions have to be met by the GPS by considering the design. Explain them.
4. What technology is used for the measurement of time in GPS receivers? Explain.
5. What are the segments of GPS. Define each one of them.
6. What are the basic functions of GPS satellites?
7. How many satellites define a full satellite segment and give reasons for that.
8. Write short notes on
   i. The control segment of GPS
   ii. The user segment of GPS
   iii. The satellite segment of GPS
9. What are the basic concepts of ranging a satellite?
10. Why it is necessary to calculate the range between a GPS satellite and a GPS receiver?
11. What is the nature of radio signals transmitted from a GPS satellite?
12. How is C/A code assisted ranging is done? What is the importance of pseudo-random code?
13. Define SPS, PPS and ephemeris data.
14. Write short notes on
    i. Ephemeris data
    ii. Almanac data

iii. Selective availability
iv. Anti-spoofing
v. Code correlation
vi. Almanac data

15. What is signal multipath? How one can overcome signal multipath while calculating a positional value
16. What is DOP? Explain GDOP, PDOP, HDOP and VDOP.
17. When will we get good GDOP? What is mask angle?

SURVEYING USING GLOBAL POSITIONING SYSTEM 16

Chapter Outline

16.1 INTRODUCTION

Global Positioning System (GPS) technology is having numerous advantages over the traditional surveying methods. Generally a good surveyor gives more importance to the accuracies and speed of work. With the aid of GPS, very high geodetic accuracy can be achieved within short time by applying suitable methodology. Efficient utilization of a GPS receiver in surveying field depends on the efficiency and dedication of the surveyor. Proper planning and preparation are essential ingredients for a successful survey using GPS. A surveyor must be well versed with the capabilities, limitations of the GPS receiver and the methodology to be followed for getting good results, while using it.

The advantages of GPS over traditional methods of surveying are

1. Inter-visibility between points is not required.
2. Can be used at any time, day or night, and in all weather conditions.
3. Geodetic accuracy can easily be achieved.
4. Large area can be surveyed in small duration.
5. GPS surveys are less labour intensive and hence cost effective.
6. More work can be completed by utilizing less time and manpower.
7. Limited or less calculation and tabulation works are required.
8. Network-independent site selection can be used.
9. Three-dimensional (3D) coordinates of points are obtained.
10. Very few skilled persons are required to execute the work.
11. GPS can compute a position (latitude, longitude and height) directly, without the need to measure angles and distances between intermediate points.
12. Survey control can be established anywhere with a clear view of the sky to get uninterrupted signals from the GPS satellites.

The disadvantages of GPS-based surveying methods are

1. High initial cost (high cost of survey grade GPS receiver and related software).
2. GPS antenna must have a clear view to the sky for getting uninterrupted satellite signals.
3. Satellite signals can get blocks by high-rise buildings, trees, over bridges, etc.
4. GPS antenna must get signals from at least four satellites to compute a positional value.
5. Difficult to use in dense forests and town centres.
6. Coordinates derived from GPS must be transformed for conventional survey applications.
7. Highly skilled persons with a good knowledge about GPS-based surveying are necessary to accomplish the projects.
8. Good computer skills and good experience in operating electronic equipment are required for attaining error free GPS results.
9. Software skills such as CAD are required for the surveyors for mapping.

GPS eliminates the need for establishing control before a survey. GPS can establish control as and when needed and establish points at strategic locations to start and close conventional traverses The following values could easily be achieved directly from the field or after post-processing the data.

- Latitude, longitude, height above a datum and X, Y, Z in Cartesian coordinates.
- X, Y, Z coordinates in local datum.
- Forward and back bearing (azimuth) of the baseline.
- Geodetic distance or station to station distance and slope distance of baselines.
- Vertical angle between points.
- Bearing of base line with respect to True North.
- Calculation of ellipsoidal height of points.

Local coordinates can be directly computed from the latitudes and longitudes derived from GPS. GPS is able to determine the geodetic azimuth between two points directly thereby eliminating the need for converting an astronomic azimuth to geodetic azimuth by applying Laplace correction equation. As ellipsoidal heights are known for the computed points, slope distances can be reduced to the ellipsoid accurately. Even though the baseline components such as distances and azimuths are accurate, the accuracy of coordinates of new points is dependent on the quality of known points included in the survey.

16.2 DIFFERENCE BETWEEN GPS NAVIGATION AND GPS SURVEYING

Basic differences between GPS navigation and GPS surveying are as follows:

1. GPS navigation is based on pseudo-ranges in which biases are dealt with the exception of the clock errors. In GPS surveying almost all biases are eliminated to get a good solution.
2. Navigation type GPS receivers are low-cost code-correlating type instruments that measures only pseudo-range, while GPS receivers are expensive phase measuring instruments having built-in complex type software to support best solution.
3. GPS navigation is employed for the safe passage of a ship/boat or an aircraft from the port of departure to the point of arrival. GPS surveying is associated with the traditional function of establishing geodetic control, supporting engineering construction, hydrographic/land surveying and mapping.

16.3 CHARACTERISTICS OF GPS SURVEYING AND GPS NAVIGATION

The characteristics of GPS navigation are:

1. Positional values will have relatively low accuracy.
2. The points being measured are generally in motion.
3. Instant collection of positional data and real-time solution.
4. Measurements are made on pseudo-random codes.
5. Usually associated with navigation of ships and aircrafts.

The characteristics of GPS surveying are:

1. Computed positional values will be accurate.
2. The points being measured are static or stationary.

3. GPS data is collected according to observation session.
4. Measurements are made on carrier wave and require special type of instrument and complex software.
5. Mostly associated with traditional surveying and mapping functions.

16.4 ACCURACY REQUIREMENTS IN GPS SURVEYING

Accuracy requirement for various applications in surveying can be grouped into

Category A (scientific application):	Better than 1 ppm (1 part in million or 1 part in 10^7)
Category B (geodetic surveying):	1–10 ppm
Category C (general surveying):	lower than 10 ppm

Category A surveys include those surveys undertaken for precise point positioning, deformation analysis and geodynamic applications. Category B surveys include geodetic surveys undertaken for an establishment, densification and maintenance of control networks to support mapping. Category C surveys include lower accuracy surveys, undertaken for urban, cadastral, geophysical surveys GIS and other general mapping applications. This classification is arbitrary and does not follow any order of survey as classified by survey authorities.

16.5 ABSOLUTE AND RELATIVE POSITIONING

The most common GPS positioning technique is absolute positioning. Commercial hand held GPS receivers provide absolute positioning. Two kinds of absolute positioning accuracy can be obtained from GPS, they are standard positing service and precise positioning service. Standard positioning service is authorized to civil users by the system provider (US Department of Defence, US DoD). This service includes the Course Acquisition Code and navigational message on L1 signal. L2 signal and precision code (P-code) on L1 are not the part of standard positioning service.

Absolute positioning involves the use of only a single passive receiver at one station location to collect data from multiple satellites to determine the station's location. It is not accurate for precision surveys. Absolute position is widely used by military grade receivers and commercial hand held navigational receivers for real-time navigation and location determination. The accuracies obtained by GPS absolute positioning are dependent on the authorization given by the US DoD to access satellite signals. A standard positioning service can provide positional accuracies of 20–25 m without selective availability The precise positioning service (PPS) user capable of tracking P-code, can achieve a positional accuracy of 5–10 m with a single-frequency receiver. Sub-metre accuracies can be obtained from absolute measurements, when special type receivers and post-processing techniques are used.

16.6 ABSOLUTE POSITIONING WITH THE CARRIER PHASE

In absolute positioning with the carrier phase, positional information is gathered using a GPS receiver which is capable of tracking both C/A code and carrier phase. Making use of broadcast ephemeris, the user is able to use pseudo-range values in real-time for the determination of absolute point positions with an accuracy of 3 m in the best condition and 20 m in the worst. By using post-processed ephemeris data, the user can achieve sub-metre accuracy in absolute point positioning mode in best conditions.

In absolute point positioning with the carrier phase, positional information is gathered using a GPS receiver, which is capable of tracking both the C/A code and carrier phase. By using broadcast ephemerides, the user is able to use pseudo-range values in real time to determine absolute point positions with an accuracy of 3 m in the best of conditions and 20 m in the worst. By using post-processed ephemerides, the user can expect absolute point positions with sub-metre accuracy in the best of conditions and 15 m in the worst.

16.7 PSEUDO-RANGING

For GPS navigation solution using the C/A code, only an approximate range, or pseudo-ranges to concerned satellites are measured. To determine precise position using GPS, the known range to the satellite and the position of those satellites must be known. In pseudo-ranging, the GPS measures an approximate distance between the antenna and the satellite by correlating satellite-transmitted code with a reference code created by the receiver, without applying any corrections for errors in synchronization between the clock of the transmitter and that of the receiver. When the velocity of the radio signal transmitted is multiplied by the time taken by the signal to travel from the satellite to the GPS receiver, will give the distance travelled by the signal.

The accuracy of the positioned point is a function of the range measurement accuracy and the geometry of the satellites, as reduced to spherical intersections with the earth's surface. A description of the geometrical magnification of uncertainty in a GPS-determined point position is dilution of precision (DOP). Repeated and redundant range observations will generally improve range accuracy. However, the dilution of precision remains the same. In a static mode (i.e., the GPS antenna stays stationary), range measurements to each satellite may be continuously re-measured over varying orbital locations of the satellites. The varying satellite orbits cause varying positional intersection geometry. In addition, simultaneous range observations to numerous satellites can be adjusted using weighting techniques based on the elevation and pseudo-range measurement reliability.

Four pseudo-range observations are required to resolve a GPS position in three dimensions. Only three pseudo-range observations are needed for a two-dimensional (2D) location. More pseudo-ranges are required to resolve the clock biases contained in both satellite and ground-based receiver. Thus, in solving for the X–Y–Z coordinates of a point, a fourth unknown (i.e., the clock bias) must also be included in the solution. The solution of four pseudo-ranges will give a solution of a point in 3D. The solution of the 3D position of a point is simply the solution of four pseudo-ranges observation equation containing four unknowns namely the U, V, W and the clock biases Δt. A pseudo-range equation is equal to the true range from the satellite to the user ρ', plus delays due to satellite and receiver clock biases and other effects as shown in Fig. 16.1a and 16.1b.

The pseudo-range observation equation R equals

$$R = \rho^{t} + c\Delta t + d \tag{5.1}$$

where R is the observed pseudo-range, ρ^{t} the true range to satellite (unknown), Δt the clock biases (receiver and satellite), d the propagation delays due to atmospheric conditions and c the velocity of radio signals.

These are usually estimated from models. The true range ρ is equal to the 3D coordinate difference between the satellite and user.

$$\rho^2 = (X - U)^2 + (Y - V)^2 + (Z + i)^2$$

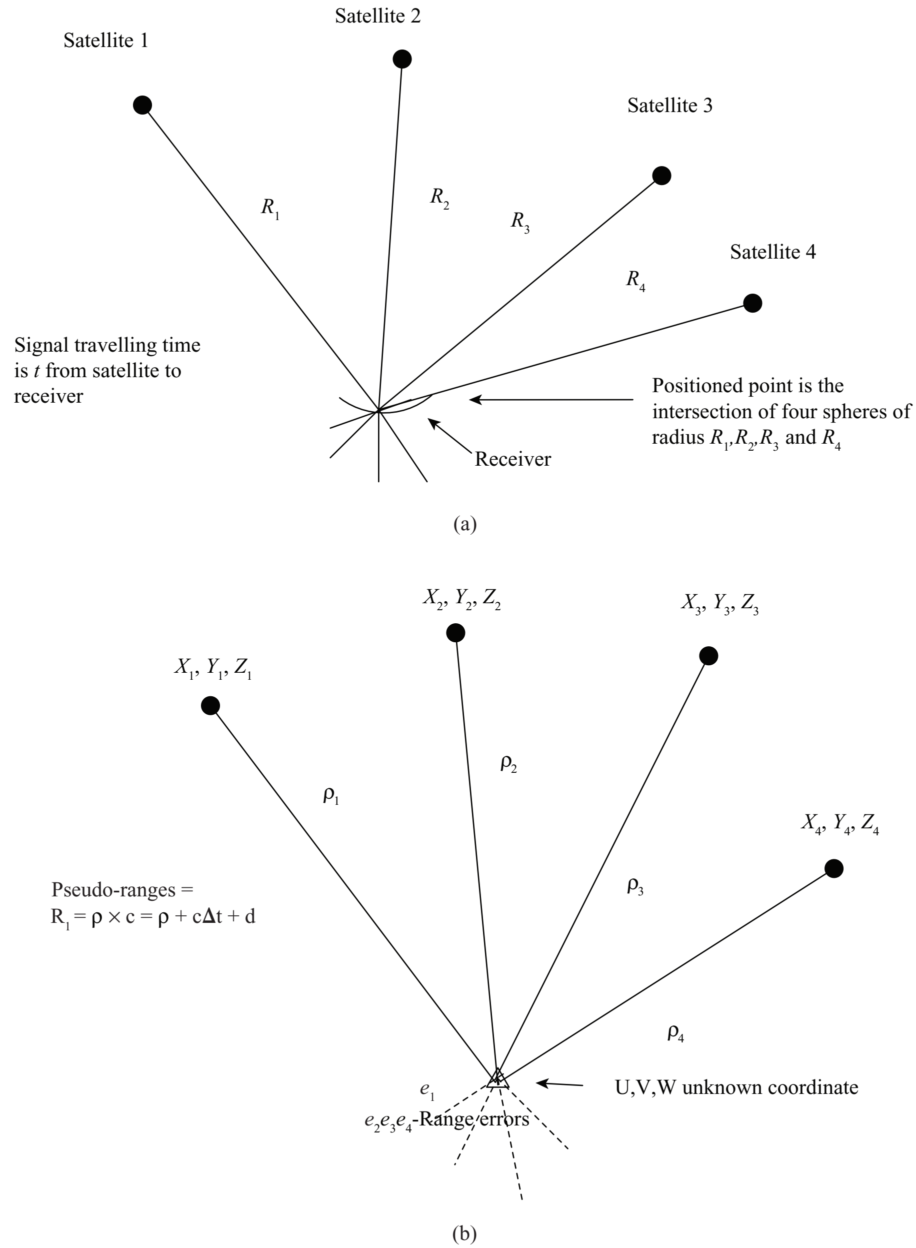


FIGURE 16.1 (a) General positioning concept. (b) Pseudo-ranging concept

where *X, Y, Z* is the known satellite coordinates from ephemeris data; *U, V, W* the unknown coordinates of user (station point), which is to be determined

When four pseudo-ranges are observed we get four equations as below

$$(R_1 - c\,\Delta t - d_1)^2 = (X_1 - U)^2 + (Y_1 - V)^2 + (Z_1 - W)^2$$

$$(R_2 - c\ \Delta t - d_2)^2 = (X_2 - U)^2 + (Y_2 - V)^2 + (Z_2 - W)^2$$
$$(R_3 - c\ \Delta t - d_3)^2 = (X_3 - U)^2 + (Y_3 - V)^2 + (Z_3 - W)^2$$
$$(R_4 - c\ \Delta t - d_4)^2 = (X_4 - U)^2 + (Y_4 - V)^2 + (Z_4 - W)^2$$

In these equations, the only unknowns are U, V, W (the coordinate of GPS receiver) and Δt (clock biases). Hence solving the four simultaneous equations will give the user's 3D position coordinates. Adding more pseudo-range observations will give redundancy to the solution. If seven satellites are simultaneously observed, seven equations are formed with the same four unknown and the solution will give better results.

This kind of absolute positioning solution is highly dependent on the accuracy of the known coordinates of each satellites (i.e., X, Y, and Z), the accuracy with which the atmospheric delay d can be estimated though modelling, and the accuracy of the resolution of the actual time measurement process performed within the GPS receiver such as clock synchronization, signal processing, signal noise, etc. In measurement process, repeated and long-term observations from a single point will enhance the overall positional reliability and confidence level.

The accuracies needed for most of the control project will not be provided by absolute positioning due to the existing and induced errors. Hence to eliminate these errors and to obtain higher accuracies, GPS is used in a relative positioning or differential positioning mode. The terms differential is having the same meaning as relative. The term differential is used whenever the technique of positioning one receiver with respect to another is employed. GPS receiver determines the travel time of a signal from a satellite by comparing the pseudo-random generated within the

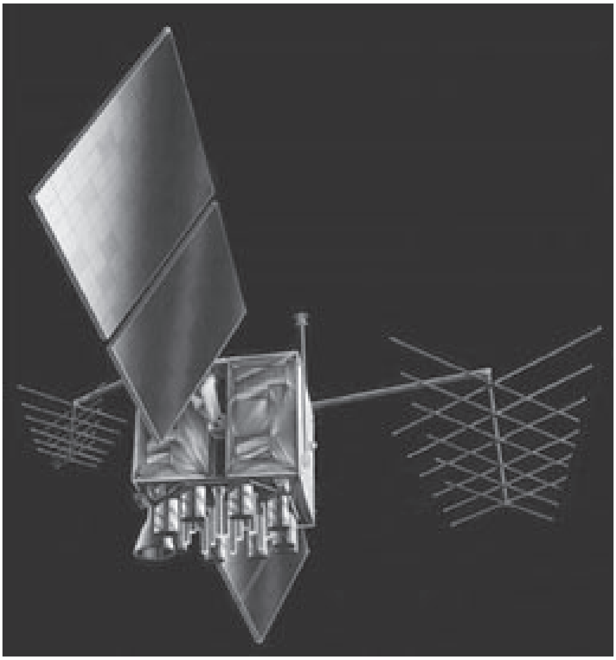


FIGURE 16.2 A GPS satellite

receiver, with an identical code within the signal from the satellite. The receiver correlates the code with the satellite's code. The amount it has to move the code during code correlation is equal to the signal's travel time. The problem is that the bits (or cycles) of the pseudo-random code are so wide that even after it gets perfectly synchronized, there will be still plenty of errors. A GPS receiver compares the pseudo-random codes that have a cycle width of almost a microsecond. At the speed of light a microsecond is almost 300 m of error. But the new narrow code correlating facility in new generation receivers are capable of correlating signals perfectly, so that the signals will be perfectly in phase. They can provide positioning accuracy of 3-6m. Fig. 16.2 shows a typical GPS satellites.

16.8 DIFFERENTIAL POSITIONING

Differential or relative positioning requires at least two receivers set up at two different stations. One receiver is placed at a well-defined position or known position called reference station or base station and the second receiver is placed at an unknown station, whose value have to be determined. The second receiver is called a rover receiver or navigator. Both receivers are set to collect the satellite data simultaneously to determine the coordinate differences (see Fig. 16.3). This method will position the two stations relative to each other and can provide the accuracies required for land surveying and for other application which may require precise positioning accuracies.

The process involves the measurement of the differences in ranges between the satellite and two or more ground observing points. Both the reference and rover receivers acquire data simultaneously for later data computation known as post-processing or, in real-time, the reference receiver transmits the difference data to the rover receiver for real-time computation of position. The range measurement is performed by a phase difference comparison, using either the carrier phase or code phase. The basic principle of differential positioning is that the absolute positioning errors at the two receiver points will be approximately the same for a given instant of time. The resultant accuracy of these coordinate differences is at the metre level for code phase observations and at centimetre level for carrier phase observations.

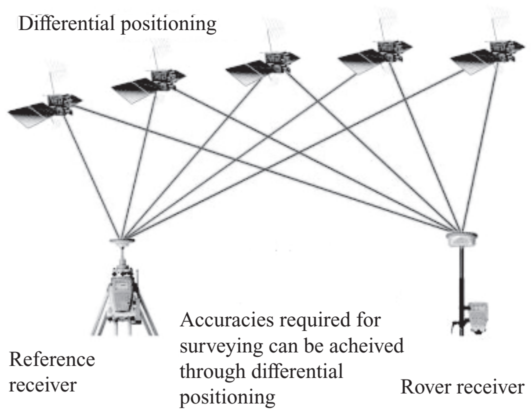


FIGURE 16.3 Differential positioning

16.9 DIFFERENTIAL PSEUDO-RANGE POSITIONING (DIFFERENTIAL CODE-BASED POSITIONING)

Differential pseudo-ranges processes results in absolute coordinates of the user on the earth's surface. Errors in range are directly reflected in resultant coordinate errors. Differential positioning is not so concerned with the absolute position of the user but with the relative difference between two user positions, which are simultaneously observing the same satellites. As errors in the satellite position and atmospheric delay estimates are effectively the same (i.e., highly correlated) at both receiving stations, they cancel each other to a large extent.

For example, if the true pseudo-range distance from a known control point to a satellite is 100 m and the observed or measured pseudo-range distance was 92 m, then the pseudo-range error or correction is 8 m for that particular satellite. A pseudo-range correction (PRC) can be generated for each satellite being observed. If a second receiver is observing at least four of the same satellites and is within a reasonable distance (300 km) it can use these PRCs to obtain a relative position to the known control point as the errors will be similar. Thus, the relative distance between the two stations is relatively accurate (i.e., within 0.5–5 m) regardless of the poor absolute coordinates. In effect, the GPS observed baseline vectors are not different from azimuth/distance observations. Figure 16.4 shows the basic concept of code-based differential positioning

Code pseudo-range tracking has primary application to real-time navigation systems where accuracies at the 0.5–5-m level are tolerable. Considering these tolerances, engineering survey applications of code pseudo-oranges tracking GPS are limited, with two exceptions being hydrographic survey and dredge positioning. The absolute GPS coordinates will not coincide with the user's local project datum coordinates. As differential survey methods are concerned only with relative coordinate differences, disparities with a global reference system used by the NAVSTAR GPS are not significant for survey purposes. Therefore, GPS coordinate differences can be applied to any type of local project reference datum (i.e., NAD 27, NAD 83, EVEREST or any local project grid reference system).

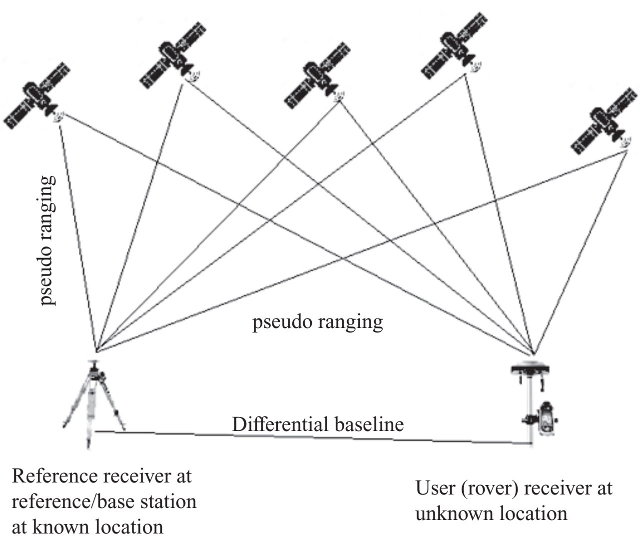


FIGURE 16.4 General concept of code-based differential positioning

16.10 DIFFERENTIAL POSITIONING (CARRIER PHASE TRACKING)

Carrier phase tracking of GPS signals has resulted in a revolution in land surveying. A line of sight along the ground is no longer necessary for precise positioning. Positions can be measured up to 30 km from reference point without intermediate points. This use of GPS requires specially equipped carrier tracking receivers.

The L1 and/or L2 carrier signals are used in carrier phase surveying. L1 carrier cycles have a wavelength of 19 cm. If tracked and measured, these carrier signals can provide ranging measurements with relative accuracies of millimetres under special circumstances. Tracking carrier phase signals provides no time of transmission information. The carrier signals, while modulated with time tagged binary codes, carry no time-tags that distinguish one cycle from another. The measurements used in carrier phase tracking are differences in carrier phase cycles and fractions of cycles over time. At least two receivers must track carrier signals at the same time. Ionospheric delay differences at the two receivers must be small enough to ensure that carrier phase cycles are properly accounted for (see Fig. 16.5). This usually requires that the two receivers be within about 30 km distance each other. Carrier phase is tracked at both receivers and the changes in tracked phase are recorded over time in both receivers. All carrier-phase tracking is differential, requiring both a reference and remote receiver tracking carrier phases at the same time. Unless the reference and remote receivers use L1–L2 differences to measure the ionospheric delay, they must be close enough to insure that the ionospheric delay difference is less than a carrier wavelength. Using L1–L2 ionospheric measurements and long measurement averaging periods, relative positions of fixed sites can be determined over baselines of hundreds of kilometres. Phase difference changes in the two receivers are reduced using software to differences in three position dimensions between the reference station and the remote receiver. High-accuracy-range difference measurements with sub-centimetre accuracy are possible.

Differential positioning uses carrier phase tracking. The process becomes somewhat more complex when the carrier signals are tracked such that the range changes are measured by phase resolution. In carrier phase tracking, an ambiguity factor is added which must be resolved to obtain a derived range. Carrier phase tracking provides for a more accurate range resolution due to the short wavelength (~19 cm for L1 and 24 cm for L2) and the ability of a receiver to resolve the carrier phase

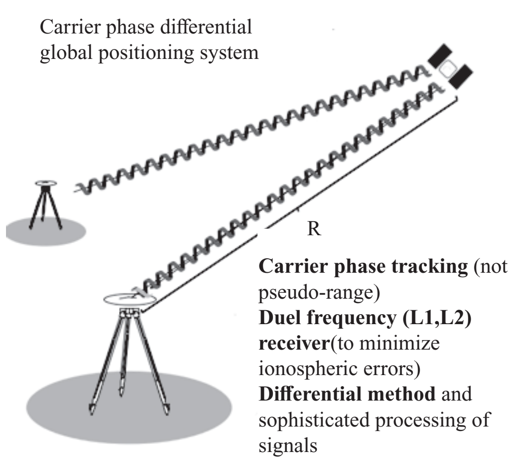


FIGURE 16.5 Carrier phase tracking

down to about 2 mm. This method, therefore, has primary application to engineering, topographic, and geodetic surveying, and may be employed with either static or kinematic methods. There are several techniques, which use the carrier phase to determine a station's position.

16.11 AMBIGUITY RESOLUTION

Cycle ambiguity (or integer ambiguity) is the unknown number of whole carrier wavelengths between the satellite and the receiver. Successful ambiguity resolution is required for successful baseline formulations. In general, static surveying can resolve instrumental error and ambiguity through long-term averaging and simple geometrical principles, resulting in solutions to a linear equation that produces a good positional value. Ambiguity resolution can also be achieved through a combination of pseudo-range and carrier-beat measurements, which are made possible by the Pseudo Random Noise modulation code. Continuous 30 min observation is required to resolve the ambiguities in a static survey. Numerous physical and mathematical techniques have been adopted to resolve the carrier phase ambiguities. The physical methods include observations over known length baselines. A good method is to set the base and rover receivers up over a known datum (best if it is WGS84) and to collect the data for at least 30 sec. Initialization can be done over very short baselines (such as using an initialization bar). Most of the modern GPS systems can automatically resolve ambiguities mathematically on the fly technique used for Real Time Kinematic applications.

GPS Baseline

The GPS receivers used for surveying are more complex and expensive than those used for navigation. Survey grade GPS receivers' use the dual frequencies (L1 and L2) broadcast by the GPS satellites. Survey grade receivers usually have a separate high-quality antenna. A GPS baseline uses two survey-quality GPS receivers, with one at each end of the line to be measured. They collect data from the same GPS satellites at the same time. The duration of these simultaneous observations varies with the length of the line and the accuracy needed, but is typically an hour or more. When the data from both points is later combined, the difference in position (latitude, longitude and height) between the two points is calculated with the help of processing software. Many of the uncertainties of GPS positioning are minimized in these calculations because the distortions in the observations are similar at each end of the baseline and cancel out.

The accuracy obtained from this method depends on the duration of the observations, but is typically about 1 ppm (1 mm/km) so a difference in position can be measured over 30 km with an uncertainty of about 30 mm, or about 100 mm over 100 km. The GPS satellites are placed in a very high orbit (~20,000 km) and the distance between the GPS baseline can be hundreds of metres, or even kilometres apart, but the two GPS receivers placed at two stations will still observe the same satellites.

Although a single baseline from a known position is enough to give the position at the other end of the baseline, additional GPS baselines to other points are often measured to give a check on the results and an estimate of the uncertainty of the calculated position.

16.12 GENERAL FIELD SURVEY PROCEDURES FOR SURVEYING USING GPS

General field survey procedure that should be performed at each occupied point is as noted below. This procedure should be followed for static, kinematic and real-time and post-processed data collection.

1. *Antenna Setup*: All tribrachs for mounting antennas should be calibrated and adjusted. Both optical plummet and standard plumb bobs must be used to minimize centering errors as this error represents a major source of error in survey works. The reference line marked on the antenna should be oriented towards north direction with the help of a magnetic compass. All tripods, range poles and fixed height tripods should be checked and calibrated regularly.
2. *Receiver Setup*: All GPS receivers have to be set up in accordance with the manufacturer's instructions prior to any observation. Base station antennas should be mounted on a tripod and kinematic rover receivers and its antenna should be mounted on range poles. For kinematic observation in real-time (RTK method), a radio modem for the transmission of real-time corrections have to be setup along with the base station and rover receiver. Modern GPS receiver are having separate data controller to record, coordinate and process all GPS data and they are having the facility to collect the data from the receivers through Bluetooth, wireless or wired connection.
3. *Height of Instrument* (*HI*): This refers to the correct vertical height of the GPS antenna above the reference monument (or datum). The height measurement is made to the fixed point on the antenna mounting device from which the calibrated distance to the antenna phase centre can be added (which will be given by the OEM (Original Equipment Manufacturer) itself). HI measurement should be done both before and after each observation session. The standard reference point for each antenna is established prior to the commencement of observations so that all observations are made with respect to the measurements of the same point. HI measurements should be made to the nearest millimetre in metric units. It should be noted whether the HI is measured vertical or diagonal. All GPS receiver/antenna will have specific antenna height measuring guidance with their instrument operating manual.
4. *Field Observation and Recording Procedure*: Field recording log sheets, log forms or full-text input data collectors have to be completed for each station and session. For that any acceptable recording media can be used. The following typical data can be included on the field logging records.

   a. Date, weather condition.
   b. Station designation.
   c. Station file number.
   d. Time of starting and ending of observation session.
   e. Receiver, antenna, data recording unit, and tribrach make, model and serial number.
   f. Antenna height, vertical or diagonal measures.
   g. Space vehicle designations of satellite observed during the session.
   h. Sketch of station location.
   i. Approximate geodetic location and elevation.
   j. Problems encountered.

5. *Field Calibration and Initializations*: For kinematic surveys, it is necessary to calibrate the base station to a known local coordinate point and reference datum. An initialization process may also be required for some types of kinematic surveys. Check the OEM's manual and follow the specific techniques for calibrating RTK surveys to a local datum. The calibration procedure should be noted on log records.

6. *Field Processing and Verification*: GPS data processing and verification have to be done in the field itself, as far as possible, to identify any problems that can be corrected before returning from the field.
7. *Survey Session*: Survey session refers to a single period of observations. Sessions and station designations are denoted by the original manufacturer of the receiver. The station and session designation should be correlated with entries on the log forms so that there will not be any doubt during baseline processing. Some GPS software requires Julian dates for efficient software operation, and in that case, the Julian date of observation session should also be noted. In addition for best processing, the following points should also be noted.
   a. Satellite visibility for each stations.
   b. Name of persons designated to each station.
   c. Site reconnaissance data for stations to be occupied.
   d. Project sketch.
   e. Explicit instruction on which each session to begin and end and follow-up sessions.
   f. Providing observers with data logging sheets for each occupied station.
8. *Log Forms of GPS Station*: GPS data logging sheets.

16.13 ABSOLUTE POINT POSITIONING

Absolute point positioning is usually employed where differential positioning becomes impractical and a new reference point is needed. There are two techniques used for point positioning in absolute mode:

1. *Long term averaging* (*self-survey mode*): A receiver is setup over a station to store positions over a long time. The length of observation time varies based on the accuracy requirement. The longer the period of data acquisition, the better the average position be. The observation time can range between 1 and 48 hours. This technique can also be done when the receiver averages the position in real time itself. Positions can be stored at 15, 30 and 60 sec intervals depending on the storage capacity and length of observation. A 24-hour observation period is used to calculate an absolute point position to sub-metre accuracy.
2. The process of differencing between signals can be performed in a post-processed mode. There are software that can perform this operation.

16.13.1 Navigation Receivers

General vehicle and vessel navigation system use inexpensive single-frequency GPS receivers. Operation of these receivers is simple and will be explained in operating manuals provided with the devise. Some receivers can log code and carrier phase data for post-processing adjustments to a reference station such as CORS (Continuously Operating Reference Stations).

16.13.2 Mapping Grade GPS Receivers

Mapping grade GPS receivers are capable to provide absolute positioning, real-time differential positioning, post processed carrier based differential positioning and can correlate the position values with

CAD map features. The geo referenced features can be exported into any GIS platform. Mapping grade receivers include software intended for mapping applications.

16.14 DIFFERENT METHODS USED IN GPS SURVEYING (DIFFERENTIAL POSITIONING BY CARRIER PHASE TRACKING)

General methodology used in GPS surveying is the differential positioning by carrier phase tracking. There are several measuring techniques that can be used by most of GPS survey receivers. The surveyor should choose the appropriate technique for the application.

There are varieties of differential GPS surveying techniques, and some of the most common methods include

1. Static
2. Rapid Static
3. Stop-and-Go Kinematic
4. True Kinematic
5. Pseudo-Kinematic
6. Kinematic On-the-Fly (OTF)
7. Real-Time Kinematic (RTK)
8. Real-Time DGPS (code)

Some of the above methods are identical, or performed similarly, with minor differences depending on the type of GPS receiver used. The major difference between static and kinematic baseline measurements involves the method by which the carrier wave integer cycle ambiguities are resolved. Most carrier phase surveying technique requires post-processing except RTK method. Post-processing of observed satellite data involves the differencing of signal phase measurements recorded by the receiver. The differencing process reduces biases in the receiver and satellite oscillator.

16.14.1 GPS Antenna for Absolute and Relative Measurements

Defining positions using GPS can be done with millimetre accuracy. For positioning, GPS receivers determines the distance between the electrical phase centre of its antenna and the phase centre of radio signal transmitting antenna fitted in the GPS satellite by performing a pseudo-range or a carrier phase observation. The phase centre of the antenna is not a stable point. For each GPS receiver, the phase centre of the antenna changes with respect to the elevation angle and azimuth of the incoming satellite signal. Different antenna calibration models are given in general by a set of antenna offsets with respect to the mechanical reference point and a list of variation values that help a user to refer to the ground points under determination. Knowing Phase Centre Variations is especially important in case of different antenna types, which are used at the end points of a baseline. Mixing antennas usually occurs in regional and permanent GPS networks like national networks. A non-calibrated antenna will certainly introduce errors that combined with other error sources and will result in significant erroneous point estimations. Even if identical antenna types are used the effects of Phase Centre Variation values do not cancel out for long GPS baselines. This is due to the effect of earth's curvature that causes elevation differences and therefore common satellites are seen at different elevation angles by the end points of a baseline. In order to overcome the above problems various calibration models have been generated and used namely the relative and more recently the absolute antenna calibration models

16.15 IMPORTANT POINTS FOR A GPS SURVEY SOLUTION

1. The fundamental unit of a GPS solution is a 3D baseline vector joining the antennas of two GPS receivers that have been tracking simultaneously the same satellites. GPS software to carry out the solution task is usually provided by the instrument manufacturer.
2. One end of the baseline is held fixed and its coordinates are known, and the other station's coordinates are determined relative to it.
3. Solutions may be obtained from ambiguity-free or ambiguity-fixed double-differenced data solutions, with different resultant accuracies and reliabilities.
4. All results are obtained in the quasi-WGS84 reference system, but relative to a fixed station (the WGS84 coordinates of one end of the baseline are known).
5. All results refer to the antenna phase centres, and the height of antenna and any offsets must be applied to reduce the coordinates to the ground marks. The quality of the baseline vector solution is dependent on
   a. Length of the observing session
   b. The number of satellites tracked by the receivers
   c. The quality of the data (multipath and cycle slips, single or dual-frequency data, presence of noise and other biases)
   d. The type of baseline solution: triple-difference, double-difference, etc.
   e. The software used to reduce the data
6. If during an observation session more than two receivers were deployed, independent baselines need to be processed, either in a single combined solution or separately.
7. If a network needs to be surveyed over a number of sessions, a combination of separate baseline solutions is needed in a subsequent network adjustment step.
8. This network solution may then be constrained and transformed into the local geodetic datum if sufficient geodetic control stations are also surveyed as well.

GPS survey results should be quality controlled at all stages of survey and data processing.

16.16 STATIC SURVEYING METHOD

Static GPS surveying is the most common method used for establishing project network control. This was the primary method of surveying using GPS and still it continues to be the primary technique. Static surveying is the most widely used differential technique for control and geodetic surveying. It involves 1–2 hours observation time depending on number of visible satellites, in order to resolve the integer ambiguities between the satellite and the receiver. Relative static positioning involves several stationary receivers collecting data simultaneously from at least four satellites during an observation session that usually last 30 min to 2 hours. In this method at least two receivers are used to collect carrier-phase data in stationary or static mode for a long duration of time. Post-processing software analyses all data from the receivers and obtains the differential position between the two receivers. This method is used for long lines, geodetic networks, tectonic plate studies, etc. This method offers high accuracy over long distances from 1 to 0.1 cm over 10 km.

In this method, two GPS receivers are used to measure a GPS-baseline distance. The line between a pair of GPS receivers from which simultaneous GPS data have been collected and processed is a vector referred to a baseline. The station coordinate differences are calculated in terms of a 3D coordinate system that uses X, Y and Z values based on the WGS-84 ellipsoid. These coordinate differences are then subsequently shifted to the project's coordinate system. GPS receiver pairs are set up.

GPS receivers are set up over stations of either known or unknown location. One receiver is positioned over a point whose coordinates are known (reference station or base station) and that receiver is known as reference receiver or base receiver. The second one is positioned over another point whose coordinates are unknown and the receiver is known as rover receiver. Both GPS receivers (the reference receiver and rover receiver) must receive signals from the same four or more satellites for a period of time that can range from a few minutes to several hours, depending on the conditions of observation and the precision required.

Requirements for static method are:

1. A clear unobstructed sky above the station to be occupied.
2. Direct signals from four or more satellites.
3. More than one receiver.
4. Enough power supply to complete the observations.
5. Enough memory space to store all the data for post-processing.

The determination of the session's duration depends upon several particulars such as the length of the baseline and the geometry of the satellites, etc. The larger the constellation of satellites, the better the available geometry, the lower the positioning dilution of precision (PDOP) and the shorter the length of the session needed to achieve the required accuracy. A static receiver gathers a huge amount of data during an observation session of 1–2 hours. This much data information is required to refine the receiver's measurements down to millimetre.

16.16.1 Equipment for Instrument Station for Static Surveying

The following list for instrument stations may be taken as a guide for static surveying:

1. GPS receiver, antenna with its cable and connectors.
2. Antenna, tripod, tribrachs or adaptors for mounting antenna.
3. External batteries (including spares), and battery charger.
4. Data storage facilities such as memory cards, and possibly a personal computer for logging or downloading data.
5. Compass and clinometer for determination of the azimuth and elevation of possible obstructions to satellite signals.
6. Pocket tape, 30 m tape, plumb bobs, surveying umbrella and supports, etc.
7. Thermometer, and barometer for meteorological observations (if insisted upon).
8. Field-books, maps, access details, observation schedule, skyplots, instructions, etc.
9. Useful ancillary equipment like camera, watch, two-way radio communications equipment.

10. Surveyor's toolkit like spanners for trig stations, construction of masts (to raise antenna above trees), clearing undergrowth, etc.
11. Vehicle for transportation.

The following equipment have to be provided for the field office:

1. Portable computer with suitable software for the downloading, checking, pre-processing and perhaps baseline processing of GPS data collected by individual field parties.
2. List of station coordinates topographic maps, observation schedules, access diagrams, client instructions, useful contact addresses and telephone numbers.
3. Data storage facilities such as SD cards, pen drives for the archiving and storage of tracking data.
4. Cables and ancillary equipment for downloading data from GPS receivers.
5. Two-way radio communication equipment to ensure contact between field parties and head office.
6. Spare GPS receivers, cables, batteries, and other field equipment that may not be needed by field parties.
7. Vehicle for transportation.

GPS receivers are complex and expensive instruments. Keep the antenna, receiver, battery, battery charger and cables and connectors together in a kit. Special care should be taken to keep the connections clean and to keep cables properly.

16.16.2 Static Survey Methodology

In static survey method one receiver is placed on a point whose coordinates are known accurately in WGS84 datum. This is known as the reference receiver. The other receiver placed on the other end of the baseline is known as the rover. Data are then recorded at both stations simultaneously. It is important that data must be recorded simultaneously at the same rate by the entire receivers.

The receivers have to collect the data for certain period of time as stated earlier. This time is influenced by:

1. The number of satellite observed and the satellite geometry.
2. The length of the baseline.
3. The precision of the result required.

In general, 30 min to 2 hours is a good occupation time for baselines of 1–30 km. The static survey methodology is as follows:

General: GPS receiver pairs are to be setup over station of either known or unknown location. One of the receivers has to be positioned over a point of known coordinates and the second receiver should be positioned over another point whose coordinates are unknown (see Fig. 16.6) Both receivers must get signals from at least four satellites for a period of time that ranges from a couple of minutes to several hours depending on the condition of observation and precision required.

Satellite Visibility Requirement: The station that are selected for static survey observations should have an unobstructed view of the sky of at least 15 degrees or greater above the horizon during the observation session (or observation window). An observation window is the period of time when observable satellites are in the sky and the survey can be successfully completed.

Common Satellite Observation: It is critical for a static survey baseline solution that the receivers simultaneously observe the same satellites during the same time interval.

Data Processing: After the observation session, the logged data from both the receivers are downloaded in a computer for processing. The processing is done to calculate the 3D baseline vector components between the two observed points. From the vector components, local or geodetic coordinates are computed and adjusted.

Survey Configuration: Static baselines may be extended from existing control using any of the control densification methods such as traverse, networking, etc.

Receiver Operation and Data Reduction: GPS receiver operation and baseline data processing requirement are manufacturer dependent. The surveyor must go through the OEM's operational manual before performing a static survey.

It is very important to introduce redundancy into the network that is being measured. This involves measuring points at least twice and creates safety checks against problems that would otherwise go undetected. The output of the work can be increased with the addition of extra rover receivers. Good coordination is required between the survey crews to maximize the potential of having three receivers.

Many receivers offer the user facility to pre-program parameters at the occupied station. Some are so automatic that they do not require operator interaction at all once they are programmed and set on-site. The selection of satellite to track, start and stop times, mask elevation angle, assignment of data file names, reference position, bandwidth and sampling rate are some options useful in the static mode, as well as in other GPS surveying methods. Satellite selection is unnecessary when using a receiver with sufficient independent channels to track all satellites above the receiver's horizon without much difficulty. But it is useful when the need arises to eliminate data from a satellite that is unhealthy, before the commencement of observations.

The receiver does not need long sessions to make the fine distinctions between millimetres. It needs long sessions to solve the integer number of cycles between itself and the satellites known as cycle-ambiguity problem. In fact, it is the unique handling of this difficulty that allows the kinematic method (another GPS surveying method) to achieve high accuracy with very shorter occupation times. In the static application, the receiver must resolve the phase ambiguity with each occupation and hence the observation sessions are long.

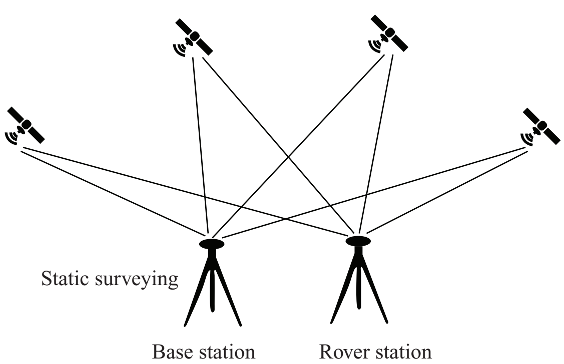


FIGURE 16.6 Static surveying

Single-frequency receivers are best applied to relatively short baselines say under 20 km and as the dual-frequency receivers have the capability to eliminate the effects of ionospheric refraction and hence can handle longer baseline.

The stations that are selected for survey must have unobstructed view of the sky for at least 15 degree or greater above horizon during the observation window. The observation window is the period of time when observable satellites are in the sky and the survey can be done successfully.

After the observation session has been completed, the logged signal data from both receivers are downloaded and processed in a computer using an application program to calculate 3D baseline vector components between the two observed points. From these vector distances, local or geodetic coordinates may be computed and adjusted. This process is called post-processing. Specific receiver operation and baseline data post-processing requirements are manufacturer dependent. The surveyor is strongly advised to consult and study the manufacture's operation manual thoroughly along with the baseline data reduction examples provided in the manual. The accuracy of GPS static surveys will usually exceed 1 ppm. Comparing all GPS processing methods static is the most accurate and can be used for any order survey.

16.16.3 Static Survey Field Procedures

Procedures for the operation of GPS receivers such as warm-up times for the receiver oscillator, operations as cold and warm starts, minimum power operation, data storage capacity, etc. are to be followed according to the manufacturer's instructions in the operator's manual of the GPS equipment.

1. Antenna Setup and Height Measurement

The following points should be noted for antenna setup:

a. Normally a Survey type GPS antenna bears a direction indicator that should be oriented in the same direction at all sites using a compass (usually in the North direction). This ensures that any antenna centre offset (as measured from the mechanical centre to the electrical phase centre) will propagate into the baseline solution (ground mark to ground mark) in a systematic manner.

b. The same antenna, receiver and cabling should be maintained together in a kit.

c. Due of the high precision of GPS surveys, the centering of the antennas is important. If centering is poor, the accuracy of the overall survey will suffer. Hence plumb bobs should be avoided and tribrachs with built-in optical plummets should be used.

d. The antenna assembly should therefore be mounted on a standard survey tribrach with an optical plummet fitted on a good quality survey tripod.

e. Setting up on a pillar (station mark) is reasonably effortless and is preferred.

f. If the receiver is to remain on-site for two or more observing sessions, the antenna should be re-positioned each time.

g. Care must be taken with antenna height measurement.

2. Observation session synchronization

For the synchronization of observation session the following points should be noted.

a. All receivers must track satellites simultaneously, during the same time period.

b. All receivers must track the same constellation of satellites.
c. All receivers must record data for the same epochs to within a few microseconds, hence they should start at a well-defined instant of time and they should track at the same data rate (e.g., start on the minute and make observations every 15 sec: 0, 15, 30, 45, 0, etc.).

Modern GPS receivers will independently synchronise to GPS time, and as long as they are programmed for the same data observation rate, then the receivers will be automatically synchronized to the required accuracy.

3. Field Log Sheets

A field log has to be maintained, in which information concerning the site being occupied and the data collection process is entered. Such a sheet contains the following information:

1. Date and time, field crew details, etc.
2. Station name and number (site codes, etc.).
3. Session number, or other campaign indicator.
4. Serial numbers of receiver, antenna, data logger, memory card, etc.
5. Start and end time of observations (actual and planned).
6. Satellites observed during session (actual and planned).
7. Antenna height and eccentric station offsets (if used).
8. Weather and meteorological observations if requested (such as temperature, pressure and relative humidity).
9. Receiver operation parameters such as data recording rate, type of observations being made, elevation mask angle imposed, data format used, etc.
10. Receiver problems or tracking problems that were noticed.
11. Sketch of the site showing all marks, possible obstructions, etc.

16.16.4 General Checklist for Onsite Procedures

The following is a list of some GPS on-site field procedures:

1. GPS receiver initialization procedures.
2. Set-up and orientation of antenna.
3. Correct cable connection of antenna to receiver, receiver to battery, etc.
4. Double checking of centring and antenna height measurement.
5. Receiver start-up procedure, for example, entry of site number, height of antenna, etc.
6. Start time of tracking.
7. Survey of eccentric station.
8. Temperature, pressure and humidity measurements (if required).
9. Monitoring receiver operation and data recording.
10. Field log entries.
11. Photographs of point occupancy.

12. Procedures at completion of session, for example communication, data transfer to P.C.
13. Instructions in event of receiver problems, contingency plans, etc.

16.16.5 General Check List for Monitoring the GPS Receiver While Surveying

1. Battery status
2. Memory capacity left
3. Satellites being tracked
4. Real-time navigation position solution
5. Satellite health
6. Date and time (UTC or local)
7. Elevation and azimuth of satellites (compare with predictions or sky plot)
8. Signal-to-noise ratios
9. Antenna connection indicator
10. Tracking channel status

16.16.6 Applications of Static Method of Survey

1. Geodetic control over large areas
2. National and continental networks
3. Monitoring tectonic movements
4. High accurate survey networks

16.17 RAPID STATIC OR FAST STATIC METHOD

Rapid static or fast static surveying is a form of static surveying technique. The rover receiver spends only a short time on each unknown points, loss of locks is allowed while the rover receiver is transferred between points, and accuracies are similar to those of static survey methods. The concept of rapid static is to measure baselines and determine positions to centimetre level with short observation time of about 5–20 min. The observation time is dependent on the length of the baseline and number of visible satellites. Loss of lock to satellites when moving from one station to the next can also occur as each baseline processed is independent of the other. In rapid static surveys, a reference point is chosen and one or more rovers operate with respect to it. Observed rapid static data are post-processed. Rapid static surveys are normally performed over small project areas. The rapid static technique does not require the use of dual frequency (L1 and L2) GPS receivers but it is good to use dual frequency receivers to get precise results. Rapid static surveying is similar to the static method, but consists of a shortened site occupation time.

The reduction in the observation time results from faster ambiguity resolution, which is achieved either by combining pseudo-range measurement technology with carrier-phase measurements or making use of carrier phase measurements only. The reference receiver is usually set up at a known point. If known point is not available, it can be setup anywhere within the network by continuous observation for a period of 24 hours over a station and finding out its value. After setting up the reference station, the rover receiver moves to each stations to be observed. The observation time that the rover receivers require to observe at each stations/points depends on the length of baseline from the reference point and the geometry of the satellites. The data collected are post-processed back at the office. Checks should be carried out to ensure that no gross errors exist in the measurements.

More than one rover receiver can be used in rapid static surveying and if possible all rovers should operate at each occupied point simultaneously. This allows data from each station to be used as either reference or rover during the post-processing and is the most efficient way to work, but most difficult to synchronize. Locks on to the satellite's signals are not necessary. Success in Rapid Static method depends on the use of receivers that can combine code and carrier phase measurements from both GPS frequencies.

Rapid static can provide the user with nearly the same accuracy available from 1- to 2-hour session of static positioning with observations of 5–20 min. This is because it uses a technique called *wide laning*, which is based on the linear combination of the measured phases from both GPS frequencies, L1 and L2. Carrier phase measurements can be made on L1 and L2 separately. When they are combined to distinct signals result, one is called a narrow lane, which has a short wavelength of 10.7 cm and the other is known as a wide lane having 86.2 cm wavelength. The frequency of the wide lane is 347.82 MHz, which is 3 times slower than the original carriers. Furthermore, it is 86.2 cm wavelength is about 4 times longer than the wavelength of L1 (19 cm) and L2 (24.4 cm). These changes greatly increase the spacing of the phase ambiguity, thereby making its resolution much easier. For rapid static surveys, the receivers used must be capable of dual-frequency tracking.

16.17.1 Reoccupation Mode in Rapid Static Survey

This method is also similar to rapid static method. Here a reference point is chosen and the rover receiver visits each of the points starting from one end to the other end giving an enclosed polygon. And the reoccupation starts from the same starting point by repeating the same procedure once again at least after 1 hour of previous observation.

Advantages

1. Satellite constellation less critical than with rapid static
2. Results with four satellites per site occupation
3. Ideal method when conditions are not suitable for rapid static

Disadvantages

1. Must wait at least one hour before re-observing
2. Each point has to be visited twice for ascertaining accuracy

Applications

1. When the constellation is very weak for rapid static.
2. When it is difficult to get clear sky on a particular side of observations.

16.18 THE STOP-AND-GO TECHNIQUE IN KINEMATIC METHOD

Differential GPS surveying method, known as stop and go is typically used for accurate topographic mapping and construction control points. Stop-and-go technique is similar to static surveying method in that each method requires at least two receivers recording observations simultaneously. But for stop-and-go technique require an initial calibration process, prior to conducting the survey. A major difference between static and stop-and-go surveying is in the amount of time required for a rover receiver to stay fixed over a point of unknown station. In stop-and-go method, the reference receiver remains fixed on a known position. The rover receiver collects observation statically on a point of unknown position for few minutes, and moves to the subsequent unknown points to collect data for a short period of time. During surveying at least four or five common satellites need to be continuously tracked by both the receivers. Once the rover receiver has occupied all the required points, the observed data are downloaded and post-processed using a good post-processing software to calculate baseline vector/coordinates difference between the known control point and the points occupied by the rover receiver during the survey session. The main advantage of this kind of GPS survey over static survey is the reduced occupation time required over the unknown points. The term kinematic is applied to GPS surveying methods where the rover receivers are in continuous motion. But for relative positioning the more typical arrangement is a stop-and-go technique. The main difference with rapid static and stop-and-go kinematic survey is that loss of lock of satellite is not allowed in stop-and-go kinematic surveys.

Stop-and-go surveying is performed similar to a conventional electronic total station radial survey. After setting the reference station, the rover receiver traverses between unknown points continuously maintaining satellite signal lock. The method is sometimes referred to as semi-kinematic surveying, in which the carrier phase ambiguities are resolved before the actual survey starts. In this mode the surveyors can accurately determine the differential position of the remote sites with observation periods as minimum as a few seconds.

During stop-and-go survey the rover receiver must maintain lock on at least four satellites during the period of survey. The reference station must also be observing the same four satellites. Loss of lock occurs when the receiver is unable to continuously record satellite signals or the transmitted satellite signal is disrupted and the receiver is not able to record it. If satellite lock is lost, the roving receiver must re-observe the last fixed point surveyed before loss of lock. The operator must closely monitor the GPS receiver when performing the stop-and-go survey to ensure loss of lock does not occur.

Survey site condition and route between points to be surveyed are critical. All observing points must have a clear view of satellites having a mask angle of 15 degrees or greater. The routes between rover receiver occupation points must also be clear of obstructions so that the satellite signal will get interrupted. Remote points should be occupied 2 or 3 times to provide redundancy between observations.

16.18.1 Antenna Swap Calibration Procedure

The antenna swap calibration procedures are required for stop-and-go kinematic surveys. The antenna swap initialization procedure requires two active receivers kept nearby points and collecting data simultaneously and both points should maintain an unobstructed view of horizon. A minimum of five satellites and constant lock to satellites are required. To perform an antenna swap, one receiver antenna is placed over a point of known control and the second a distance of 10–100 m away from the first receiver. The receivers at each station must collect data for about 2–5 min. The receivers' antennas are made swap, that is, the antenna at the known point is changed to the unknown point and antenna at the unknown point is changed to known point. Satellite data are again collected for 2–5 min. The receivers are then swapped back to their original location. This completes one antenna swap calibration. If satellite lock is lost during the process, the calibration procedure must be repeated. Although an antennae swap procedure is used to initialise a stop-and-go kinematic survey, the same technique can also be used to determine a precise baseline and azimuth between two points. While this technique is viable without a control point, it is obviously an advantage if it begins at a known position. One receiver occupies a control point while the other receiver occupies the first point of the kinematic survey. This is helpful if the two points are conveniently close together. If it is shorter than the length of their antenna cable, the procedure is most convenient.

Following a few minutes of observation the receivers are switched over. Then the reference receiver is moved from the control point to the starting point, and the rover receiver is moved from the starting point to the control point. Again after a few minutes of observation they are switched over back to their original positions. With typical tribrach-mounted antenna, or receivers with built-in antennas, the antenna swap procedure is fast and easy because the tribraches may stay on their tripods and simply be locked and unlocked to accommodate the instruments. After taking few minutes of data in their final position the vector between the two points will be determined to millimetre accuracy. With the receivers both initialized, the kinematic survey may begin.

Applications of Stop-and-go Kinematic Survey

1. For detailed and engineering surveys in open areas
2. Used where points are close together.

Advantages of Stop-and-go Kinematic Survey

1. This method is fast and economical
2. This is the fastest way to survey detail points.

Disadvantages of Stop-and-go Kinematic Survey

1. New static or rapid static fix is needed if complete loss of satellite lock
2. Must maintain phase lock to at least four satellites for a successful survey

16.19 KINEMATIC SURVEYING METHOD (TRUE KINEMATIC)

Kinematic surveying using differential carrier phase tracking is similar to static carrier phase methods. It requires two receivers recording observation simultaneously. The reference receiver is fixed on a known control station and the roving receiver collects data on a constantly moving platform such as a vehicle, aircraft, or on the backpack of surveyor itself, as shown in figure. Kinematic positioning is faster than static methods. Kinematic surveying uses differential carrier phase tracking similar to the previous types of differential carrier phase GPS surveying because it requires two receivers for recording observations simultaneously. Kinematic surveying is often referred to as dynamic surveying. As in stop-and-go surveying, the reference receiver remains fixed on a known control point while the roving receiver collects data on a constantly moving platform. The observation data are later post-processed to calculate the relative vector/coordinate differences to the roving receiver. In kinematic surveying, before the rover receiver starts collection of data, a period of static initialization or antenna swap is required, as described in antenna swap calibration procedure in Section 16.18.1. The period of initialization depends on the geometry of satellites. Once initialization is completed, the rover receiver can move point to point as long as satellite locks are maintained. If a loss of satellite lock occurs, a new period of static initialization will be required.

A kinematic survey requires two GPS receivers. One receiver is set over a known point called reference station and the other is used as a rover (i.e., moved from point to point or along a path) (see Fig. 16.7). Before the rover receiver can rove, a period of static initialization or antenna swap must be performed. This is the same as measuring a rapid static point and enables the post-processing software to resolve the *ambiguity* when back in the office. The reference and rover are switched on and remain absolutely stationary for 5–20 min, collecting data. The actual time depends upon the baseline length from the reference and the number of satellites observed.

When compared with other relative positioning methods, the advantage of true kinematic surveying is that in very short sessions of the kinematic or real-time kinematic method can produce the largest number of positions in a small duration of time. But the remarkable point is that this technique can do with only slight degradation in the accuracy of the work. In the kinematic method the receiver resolves the phase ambiguity once, and only once, at the beginning of the project. Then by keeping a continuous lock on the satellite's signals, it maintains that solution throughout the work. The kinematic technique needs initialization. The receivers can occupy each end of a baseline between two control points and because the distance between the points is known, the phase ambiguity is resolved in a few minutes. Perhaps the best way to establish this baseline is with static GPS techniques.

The kinematic data processing techniques are similar to those used in static surveying. When processing kinematic GPS data, the user must ensure that satellite lock was maintained on four or more satellites and that cycle slips are adequately resolved in the data recorded.

Carrier phase differential kinematic survey errors are correlated between observations received at the reference and rover receivers, as in differential static surveys. Experimental test results indicate kinematic surveys can produce results with centimetre accuracy. Test results of an experimental true kinematic GPS survey under ideal conditions had verified that the true kinematic GPS surveying could achieve centimetre-level accuracy over distances up to 30 km.

Application of Kinematic Method

1. Measuring trajectory of moving objects
2. In hydrographic surveys
3. In road surveys

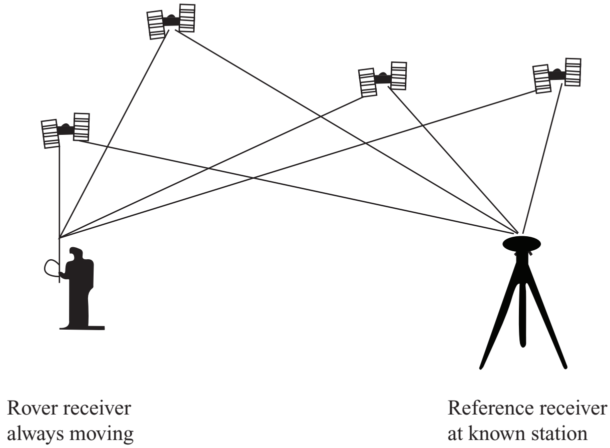


FIGURE 16.7 True kinematic surveying

4. Photogrammetry with ground control
5. Collection of data for the preparation of highly accurate topographic maps

Advantages of Kinematic Method

1. Fast and economical
2. Continuous measurements

Disadvantages of Kinematic Method

1. New static or rapid static fix needed if complete loss of satellite lock.
2. Occupies stations should be free of overhead obstructions.
3. The route between stations must be clear.

16.20 PSEUDO-KINEMATIC GPS SURVEY

Pseudo-kinematic GPS surveying is similar to stop-and-go techniques except that loss of satellite lock is tolerated when the receiver is transported between occupation sites. That is the roving receiver can be turned off during movement between occupation sites even though this is not recommended. This feature provides the surveyor with a more favourable positioning technique because obstructions, such as bridge overpasses, tall buildings and overhanging vegetation, are common. Loss of lock that may result due to these obstructions is more tolerable when pseudo-kinematic techniques are employed.

The pseudo-kinematic techniques require that one receiver be placed over a known control station. A rover receiver occupies each unknown station for 5 min, approximately 1 hour after the initial station occupation. The same rover receiver must reoccupy each unknown station. The pseudo-kinematic technique requires that at least four of the same satellites are observed between initial station occupation and the requisite reoccupation. For example, the rover receiver occupies Station A for the first 5 min and tracks satellites 3, 6, 11, 12, 13; then 1 hour later, during the second occupation of Station A, the rover receiver tracks satellites 3, 6, 8, 9, 19. In this example, only satellites 3 and 6 are common to the two sets, so the data cannot be processed because four common satellites were not tracked for the initial station occupation and the requisite reoccupation.

Prior mission planning is essential in conducting a successful pseudo-kinematic survey. It is in the determination of whether or not common satellite coverage will be present for the desired period of the survey. In addition, during the period of observation the base receiver must continuously occupy a known control station.

Pseudo-kinematic survey satellite data records and resultant baseline processing methods are similar to those performed for static GPS surveys. As the pseudo-kinematic technique requires each station to be occupied for 5 min and then reoccupy for 5 min approximately an hour later, this technique is not suitable when control stations are widely spaced and transportation between stations within the allotted time is impractical. Pseudo-kinematic survey accuracies are similar to kinematic survey accuracies of a few centimetres.

This method is less productive than kinematic but more productive than the static method. In pseudo-kinematic work it is necessary to occupy the unknown stations twice. The time between the two occupations allows the satellites of the second session to reach a configuration different enough from that of the first for resolution of the phase ambiguity. As the production advantage of the procedure over the static mode is diminished by the reoccupation requirements, the application of this method is limited over short baselines, preferably under 10 km, where the access to the station is quick and easy. Pseudo-kinematic and stop-and-go techniques can be considered as the ideal GPS measurement technique for large scale surveying purposes. Pseudo-kinematic technique can be used advantageously in areas where there is a fear of signal shading due to vegetation, built-up areas, as there is no requirement for the rover receiver to maintain its lock to the satellite during movement. But in open areas, stop-and-go technique may prove useful.

16.21 KINEMATIC ON-THE-FLY (OTF)

Kinematic On-the-Fly or OTF surveying is similar to kinematic differential GPS surveying, as it requires two receivers recording observations simultaneously and allows the rover receiver to be moving. Unlike the kinematic surveying, OTF surveying technique use dual frequency L1/L2 GPS observations and can handle loss of satellite lock. As this method uses the L2 frequency, the GPS receiver must be capable of tracking the L2 frequency during antispoofing. There are several techniques used to obtain L2 during the anti-spoofing. This includes the squaring and cross-correlation methods. In OTF method successful ambiguity resolutions are required for baseline formulations. The OTF technology allows the rover receiver to initialize and resolve the ambiguity integers without a period of static initialization. With OTF if loss of satellite lock occurs, initialization can occur while on motion. The integers can be resolved at the rover within 10–30 sec, depending upon distance from the reference station. OTF uses the L2 frequency transmitted by the GPS satellites for the ambiguity resolution. After the integers are resolved, only the L1 C/A code is used to compute the position. In this method one of the GPS receivers is set over a known point, and the other is placed on a moving or mobile platform. If the survey is performed in real time, a data link and processor are required and the method is known as Real-Time Kinematic Surveying (RTK Method).

16.22 REAL-TIME KINEMATIC SURVEYING (RTK)

In the early 1990s, RTK technology (Real-Time Kinematic) was born and the GPS industry has not looked back. RTK allows the user to obtain centimetre-level positioning in real-time. That is the point when using GPS for staking became possible, and GPS for topographic surveys became very efficient. RTK is the fundamental technology that makes machine control possible. The basic concept behind RTK is that a base station receiver set on a known point somewhere around the project site. The base station receiver sends correction data to the rover receiver. The correction data is typically sent via Ultra High Frequency or spread spectrum radios that are built specifically for wireless data transfer. The corrections from the base station receiver can be sent to an unlimited number of rovers. Real-time positions on the rover receiver are calculated as fast as 20 times per second or as little as once per second. RTK is a method that can offer positional accuracy in real time very near to static carrier phase positioning. Unlike DGPS that is differential GPS on pseudo-ranges, RTK is a differential GPS method that uses carrier phase observations corrected in real-time and, therefore, depends upon the fixing of the integer cycle ambiguity.

RTK systems resolve the carrier phase ambiguity. The method requires dual frequency GPS receivers capable of making both carrier phase and precise pseudo-range measurements. Here a search area is defined in the volume of the possible solutions, but that group is narrowed down quite a bit by using pseudo-ranges. If the number of integer combinations to be tested is greatly reduced with precise pseudo-range, the search can be very fast. The possible solutions in that volume are tested

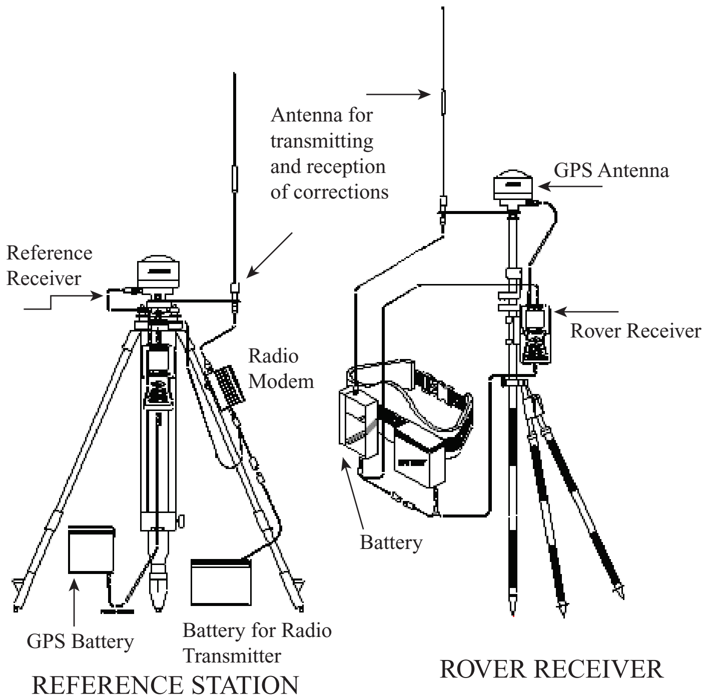


FIGURE 16.8 A full RTK DGPS system

statically, according to a minimal variance criterion, and the best one is found. The process may take less than 10 sec under the best circumstances such as the receivers are tracking a large constellation of satellites, the PDOP is small, the receivers are dual frequency, there is no multipath, and the receiver noise is low. Observation on L1 and L2 are combined into a wide lane, which has an ambiguity of about 86 cm, and the integer ambiguity is solved in a first pass. This information is used to determine the kinematic solution on L1. Therefore, it is good idea to restrict RTK to situations where there is good correlation of atmospheric biases at both ends of the baseline, or in other words RTK is best when the distance between the base and rovers is less than 20 km.

RYK surveying method requires a radio link between the receiver at the base station and the receiver at the rover, and they both must be tuned to the same frequency. The usual configuration operates at 4,800 or 9,600 baud rate. A full RTK system is shown in Fig. 16.8.

16.23 REAL-TIME DIFFERENTIAL GPS CODE PHASE HORIZONTAL POSITIONING GPS

The code phase differential GPS system is a functional GPS (DGPS) survey system for marine positioning such as positioning facility for hydrographic observations and positioning of dredgers. It also has small applications for topographic, small-scale mapping surveys and input to GIS database. Although greater positional accuracies can be obtained with use of P-code, the implementation of anti-spoofing will limit its use. Real-time code phase DGPS is a method that improves GPS pseudo-range accuracy.

This is also known as real time (code) DGPS surveying. Real-time code phase Differential GPS involves the use of two receivers; one is stationary and kept at a known station and the other roving around making position measurements. The stationary receiver is known as reference station, base station or reference receiver. The second receiver that is roving is known as a rover receiver, mobile receiver or navigator. The reference receiver antenna is mounted on a previously measured point with known coordinates. The reference station should be placed on a known survey station in an area having an unobstructed view of sky. It consists of a GPS receiver, GPS antenna, processor and a communication link (radio-link).

When reference receiver is switched on it will start to track satellites. The reference station measures the timing and ranging information broadcast by the satellites and computes range corrections and transmit it to the user equipment It can also calculate it's position from the received signals from the satellites. The actual co-ordinates of the known station over which the reference receiver antenna mounted are fed manually into the reference receiver. Hence the reference receiver will have two positional values, the computed value by tracking the satellite signals and the actual value of the position fed manually into it. The reference receiver will estimate very precisely the ranges to the various satellites with the help of the actual value of the position, which is fed into it. The reference receiver works out the difference between the computed and measured value of the ranges to satellites. These differences are known as PRCs. As the reference receiver has no way of knowing which of the many available satellites a roving receiver might be using to calculate its position, the reference receiver quickly runs through all the visible satellites and computes errors of all the visible satellites.

As the reference receiver is attached to a radio data link, it will be able to transmit these corrections to the rover receiver through that radio link. The rover receiver is also connected with a radio data link and it will be able to receive these corrections transmitted by the reference receiver while taking its positional values.

The rover receiver calculates ranges to the satellites and then applies the corrections received from the reference receiver to the corresponding satellite ranges. This enables the rover receiver to calculate its position more accurately. Using this technique all of the error sources are minimized, hence a much more accurate positional solution is achieved. It is worthwhile that multiple rover receivers can receive corrections from one single reference receiver.

In addition, it takes a little time for the base station to calculate these errors and it takes a little time for it to transmit them through a radio modem. The reference station's data is put in to a definite format. The data makes its way from the reference station to the rover over the data link. Then the rover receives this transmitted data from the reference station and then it decodes the data and applies it through its software. The time it takes for all of this to happen is called *Latency* of the communication between the reference and the rover. It may be as little as a quarter of a second or as long as a couple of second. As the base station's corrections are only accurate for the moment they were created, the base station must send a range rate correction along with them. Using this rate correction, the rover is able to give correction to match the moment it made that same observation.

The biases calculated by the base station can reduce the range of common errors, or correlated errors with its differential corrections. If results are required in real-time, the base station transmits these differential corrections via a radio link to the rovers. Hence the rovers can calculate earth-centred-earth-fixed co-ordinates using these corrections. One large consideration is the radio link. There are many types of radio link that will transmit over different ranges and frequencies. The performance of the radio link depends on a range of factors including

1. Frequency of the radio
2. Power of the radio

3. Type and 'gain' of radio antenna
4. Antenna position
5. Distance from the reference station

A radio transmitter (or data transmitter) consists of a radio modulator, an amplifier and an antenna. The modulator converts the correction data into a radio signal. The amplifier increases the signal's power, which determines how far information can travel. The information is transmitted through an antenna. The main requirement of the communication link is that the transmission be at a minimum rate of 300 bits per second.

The type of radio transmission system is dependent on the user's requirement. All communication links necessitate a reserved frequency for operation to avoid interference with other activities in the area. No transmission can occur over a frequency until the frequency has been officially authorized for use in transmitting data. This applies to all government agencies also. The Telecom Department of Government of India handles allocating a frequency. The base station transmitter may be very high frequency (VHF), UHF, or a spread spectrum to have sufficient bandwidth to handle the data. Communication links between UHF and VHF are viable systems for the broadcast of DGPS corrections. UHF and VHF can extend out some 20–50 km, depending on local conditions. The disadvantages of UHF and VHF links are their limited range to line of sight and the effects of signal shadowing due to obstructions like island, structures, buildings, etc., and signal multipath. DGPS is the best when the distances between the base station and the rover are less than 20 km.

Networks of GPS receivers and powerful radio transmitters have been established, transmitting on a maritime only safety frequency. These are known as Beacon transmitters. The users of this service just need a rover receiver that can receive the beacon signal. These types of receivers are called beacon GPS receivers. Such beacon transmitters have been set up around the coast of many countries. Other devices such as mobile phones can also be used for transmission of data. Developed countries and other international agencies have established reference stations all over the world especially around popular harbours and waterways. These stations often transmit on the radio beacons that are already in place for radio direction finding (usually in 300 kHz range). It is best to place the transmitter antenna as

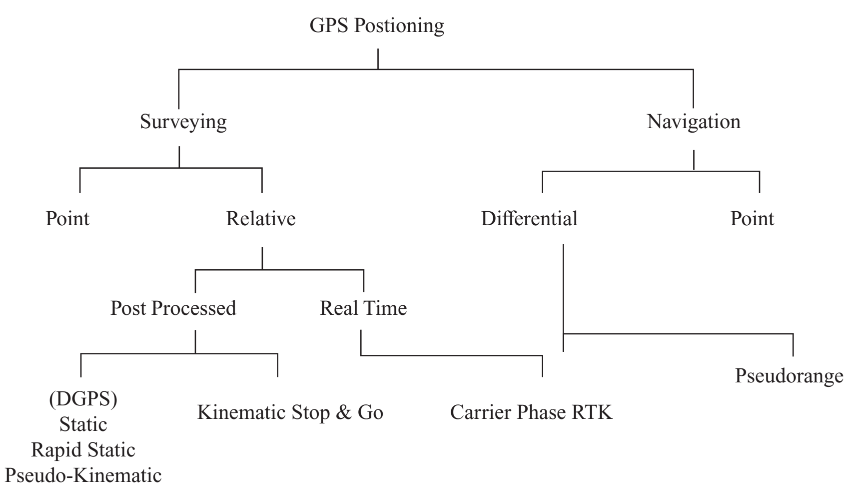


FIGURE 16.9 Flow chart of GPS observations

high as possible for maximum coverage. It is best if the base station occupies a control station that has no overhead obstruction to avoid multipath, etc. Figure 16.9 shows a flow chart of GPS observations. Anyone in that area can receive these corrections and improve the accuracy of their GPS measurements. Most ships already have radio beacons capable of automatic tuning to the beacons, adding DGPS quite easy. The new GPS receivers are designed to accept corrections, and some are even equipped with built-in-radio receivers. For hydrographic survey work usually code-based DGPS is employed.

16.24 OFFICE PROCEDURES AFTER DATA COLLECTION

The survey data should be systematically catalogued and archived between observation sessions, or at the end of the working day at the very latest. Many problems can be identified at this stage. The following are some typical field office procedures:

1. Data handling tasks, for example, transfer of data from receiver to computer.
2. Data verification and backup.
3. Preliminary computation of baselines.
4. Preliminary quality control procedures, such as the inspection of repeated baselines, loop closures, and evaluation of minimally constrained network.
5. Command and control of survey parties for repeated observation sessions.
6. Supervise calibration and testing of field equipment.
7. Preparation of campaign report, and maintain ensure reporting to head office.

Without safe data downloading from the GPS receiver, the survey work should never be considered complete. For efficient office work the following points should be noted

- Go through the OEM's operator's manual and download the data.
- Follow the procedures as stipulated in the operator's manual.
- Most GPS receivers have many hours of internal memory, so daily download is a reasonable routine.
- Delete files from receiver memory when data download procedure has been verified.
- Download the data to the computer and make backup copies.
- Store backup copies separately.
- Cross-reference booking sheets to data files.

It is best to re-observe a session than to spend a lot of effort in the office trying to track down problems. This option can only be exercised if the field parties are still in the area and have not moved on to a different area, or have returned to their head office. This is therefore another good reason for insisting on some field office processing. A common problem that may be identified at an early stage is incorrectly measured antenna heights. Confusion often results when antenna heights are extracted from field log sheets, and compared with heights entered into the receiver message file by the field party.

16.25 POST-PROCESSING OF DIFFERENTIAL GPS DATA

GPS baseline solutions are usually generated through iterative process. Observation values are compared with computed values and an improved set of positions occupied is obtained using least square minimizing procedures and equations modelling potential error sources. Observed baseline data are also evaluated over a loop or network of baselines to ascertain the reliability of the individual baselines. The flow of the process used in reducing GPS baselines is as shown below.

1. Create a new project
2. Download baseline data from receivers or data collectors
3. Download precise ephemeris data
4. Make changes and edits to raw baseline data
5. Process all baselines
6. Review, inspect, and evaluate adequacy of baseline reduction results
7. Make changes and rejects
8. Reprocess baselines and re-evaluate results
9. Designate independent and trivial baselines
10. Review loop closure and adjust baseline network

The ability to determine positions using GPS is dependent on the effectiveness of the user to determine the range or distance of the satellite from the receiver located on the earth's surface. There are two general techniques to determine this range, viz., differential code pseudo-ranging and differential carrier phase measurements. Differential carrier phase reductions provide centimetre level accuracy and are most commonly used for GPS-based surveying applications. Post-processed differential code phase reduction will give accuracies ranging from 0.25 to 5 m and are not intended for precise control surveys. Baseline processing time is dependent on the required accuracy, processing software, computer hardware speeds, data quality and amount of data collected. Special care should be taken while processing baselines with observations from different GPS receiver manufacturers. It is also important to ensure that observables being used for the formulation of the baseline are of a common format called RINEX (Receiver-Independent Data Exchange Format).

16.26 DIFFERENTIAL REDUCTION TECHNIQUE

Differential reduction technique involves the analysis of Doppler frequency shifts that occur between the moving satellites and ground-based receivers. Integration of Doppler frequency offsets along with interferometric processing and differencing techniques provide a resultant baseline vector between two ground-based stations or velocity measurements on a moving receiver. Differencing and interferometric analysis techniques may be performed on both carrier L1 and L2, the frequency differences and on the code-phase observations. Floating and fixed baseline solutions are computed from these interferometric differencing techniques. These process are simple for static observations, but they get complicated in the case of real-time (OTF) integer ambiguity resolution. Numerous GPS data reduction software are available in the market.

Carrier beat phase observable is the phase of the signal remaining after the internal oscillated frequency generated in the receiver is differenced from the incoming carrier signal of the satellite. The carrier phase observable can be calculated from the incoming signal or from observations recorded during a survey process. By differencing the signal over a period of epoch time, one can count the number of wavelengths that cycle through the receiver during given duration of time. The unknown number of cycles between the satellite and receiver antenna is known as integer cycle ambiguity. There is one integer ambiguity value per each satellite and receiver pair as long as the receiver maintains continuous phase lock during the observation period. The value found by measuring the number of cycles going through a receiver during a specific time, can be used to develop a time measurement for transmission of the signal. Once again the time of transmission of the signal can be multiplied by the speed of the light yield an approximation of the range between the satellite and the receiver. The biases for carrier phase measurement and pseudo-ranges are almost the same. A good range between the satellite and receiver can be formulated when the biases are taken into account during derivation of the approximate range between the satellite and the receiver.

The accuracy achievable by pseudo-ranging and carrier phase measurements in both absolute and relative positioning surveys can be improved through processing that incorporates differencing of the mathematical models of the observables. Basically there are three broad processing techniques that incorporates differencing, they are:

1. Single differencing
2. Double differencing
3. Triple differencing

Differencing solutions generally proceed in the following order, differencing between receivers takes place first, between satellites comes second and between epochs comes third.

16.26.1 Single Differencing between Receivers

Single differencing between receivers is the process of single differencing the mathematical model for pseudo-range or carrier phase observable measurements between receivers and will eliminate satellite clock errors and large amount of satellite orbit and atmospheric delays (see Fig. 16.10).

16.26.2 Single Differencing between Satellites

Single differencing the mathematical models for pseudo-range code or carrier phase observable measurements between satellites eliminates receiver clock errors. Single differencing between satellites can be done at each individual receiver during observations as a precursor to double differencing and in order to eliminate receiver clock errors. Figure 16.10 shows single differencing between satellites and receiver.

16.26.3 Single Differencing between Epochs

Single differencing the mathematical models between epochs takes advantage of Doppler shift or apparent change in the frequency of the satellite signals by the relative motion of the transmitter and receiver. Single differencing between epochs is generally done in effort to eliminate cycle ambiguities

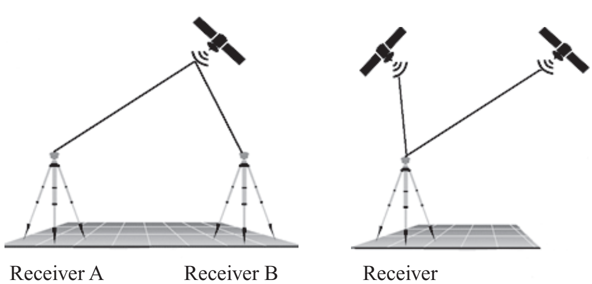


FIGURE 16.10 Single differences are computed between the two receivers A and B and single differencing between two satellites

There are three forms of single differencing techniques between epochs. They are Intermittently Integrated Doppler (IID), Consecutive Doppler Counts (CDC), and Continuously Integrated Doppler (CID). IID uses a technique where Doppler count is recorded for a small portion of the observation period, Doppler count is reset to zero, and then at a later time the Doppler count is restarted during the observation period. CDC uses a technique where Doppler current is recorded for a small portion of observation period, reset to zero, and then restarted immediately and continued throughout the observation period.

16.26.4 Double Differencing

Double differencing is differencing of two single differences. There are two general double differencing processing techniques, viz., receiver-time and receiver-satellite. Double difference processing technique eliminates clock errors.

16.26.4.1 Receiver-Time Double Differencing

This technique uses a change from one epoch to the next in the between receiver single differences for the same satellite. This technique eliminates satellite dependant integer cycle ambiguities and simplifies editing of cycle slips.

16.26.4.2 Receiver-Satellite Double Differencing

There are two different techniques that can be used to compute a receiver–satellite double difference. One technique involves using two single differences between receivers (see Fig. 16.11).

This technique also uses a pair of receivers, recording different satellite observations during a survey session, and then differencing the observations between two satellites. The second technique involves using two single differences between satellites. This technique also uses a pair of satellites.

16.26.5 Triple Differencing

There is a single type of triple differencing procedure, namely, the receiver satellite-time (epoch). All errors eliminated during single and double differencing processing are also eliminated during triple differencing (see Fig. 16.12). When used in conjunction with carrier beat phase measurements, triple differencing eliminates initial cycle ambiguity. During triple differencing, the data are also automatically edited by the software to detect any data that cannot be solved, so that the unresolved data are ignored during the triple difference solution. This feature is advantageous to the user because of the reduction in the editing of data required. But degradation of the solution may occur if too much of the data is eliminated during triple differencing.

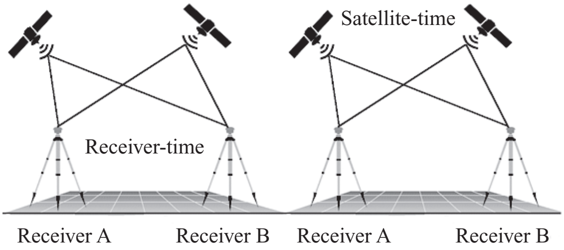


FIGURE 16.11 Double differencing between receivers and satellites

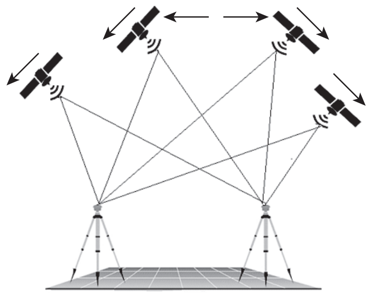


FIGURE 16.12 Triple differencing

16.27 BASELINE SOLUTION BY CYCLE AMBIGUITY RECOVERY

The resultant baseline vector solution produced when differenced carrier phase observations resolve the cycle ambiguity is called a fixed solution. The exact cycle ambiguity does not need to be known to produce a solution. If a range of cycle ambiguities are known, then a float solution can be formulated from the range of cycle ambiguities. A floating baseline solution is a least-square fir that may be accurate to only a few integer wavelengths. It is always desirable to formulate a fixed solution. However, when the cycle ambiguities cannot be resolved, which sometimes occurs when a baseline distance is greater than 75 km in length, a float solution may be the best solution. Differences between floating and fixed solutions can be calculated over all the epochs observed.

16.28 BASELINE PROCESSING

Current baseline processing software available in the market are fully automatic and user friendly. The common baseline processing steps are:

1. Downloading GPS data: The first step in baseline processing is transferring the observation data from the GPS data collector device to a personal computer for processing.
2. Pre-processing: Once observation data have been downloaded pre-processing of data can be completed. It depends on the type of GPS data collected (e.g., static, rapid static, kinematic). Pre-processing consists of smoothing and editing of the data and ephemeris determination. Activities done during smoothing and edition include determination and elimination of cycle slips, editing gaps in information and checking station names and antenna heights.
3. Ephemeris data: Retrieval of post-processed ephemerides may be required depending on the solution and type of survey being conducted. Code receivers do not require post-processed ephemerides during the survey.
4. Baseline solutions: Carrier phase baseline processing is fairly automatic process on almost all commercial software packages. Groups of baselines are processed in a defined or selected order. After an initial code solution is performed, a triple difference, then double-difference, solution is performed. If the integer ambiguities are successfully resolved, then a fixed solution can result. If all observed baselines are processed, any dependant baselines should be removed so that they will not be used in subsequent network adjustments.

16.29 STANDARD GPS DATA FORMAT

Each GPS manufacturers are having their own formats for storing GPS measurements and it will be difficult to combine data from different receivers. To overcome these limitations, a number of research groups have developed standard formats for various user needs and the most widely used standard formats, are RINEX, RTCM SC 104, and NMEA 0183.

16.29.1 RINEX Format

RINEX or Receiver-Independent Exchange Format is a data interchange format for raw satellite navigation system data. RINEX format allows the user to post-process the received data to produce a more accurate result. The final output of a navigation receiver is usually its position, speed or other related physical quantities. The calculations of these quantities are based on a series of measurements from one or more satellite constellations. Although receivers calculate positions in real time, in many cases it is interesting to store intermediate measures for later use. RINEX is the standard format that allows the management and disposal of the measures generated by a receiver, as well as their off-line processing by a multitude of applications, whatever the manufacturer of both the receiver and the computer application.

16.29.2 The RTCM SC-104 Message Format

There is an accepted format for the communication between Base Station and the Rovers in DGPS system. Real-time DGPS operations require the estimation of the pseudo-range corrections at the reference receiver, which is then transmitted to the rover receiver through a communication link. To ensure efficiency of operations, the pseudo-range corrections are sent in an industry standard format known as the RTCM SC-104. This Message Format is designed primarily for marine navigation. In 1985 *the Radio Technical Commission for Maritime Services, RTCM, Study Committee, SC-104* created this standard format for DGPS correction. This is the commonly used standard for the format of transmission of GPS data and is commonly called RTCM format. This format is commonly used all over the world. Correspondingly for aviation the standard is RTCA.

16.29.3. NMEA Format

NMEA is an acronym for the National Marine Electronics Association. NMEA existed well before GPS was invented. NMEA is an accepted standard data format supported by all GPS manufacturers. The purpose of NMEA is to give equipment users the ability to mix and match hardware and software. NMEA-formatted GPS data also makes life easier for software developers to write software for a wide variety of GPS receivers instead of having to write a custom interface for each GPS receiver. Most of the GPS and surveying software accepts NMEA-formatted data from any GPS receiver and graphically displays it. NMEA data can be transmitted through different types of communication interfaces such as RS-232, USB, Bluetooth, Wi-Fi, UHF, and many others.

16.30 ACCURACY OF GPS HEIGHT DIFFERENCES

The height (h) component of GPS measurements is the weakest plane. This is due to the orbital geometry of the X–Y–Z position determination. Thus, GPS ellipsoidal height differences are usually less accurate than the horizontal components. Currently, GPS-derived elevation differences will not meet third-order standards as would be obtained using conventional spirit levels. Accordingly, GPS-derived elevations must be used with caution.

16.31 TOPOGRAPHIC MAPPING WITH GPS

GPS positioning, whether operated in an absolute or differential positioning mode, can provide heights (*Z* coordinate or height differences) of surveyed points. The height *h* or height difference d*h* obtained from GPS is in terms of height above or below the WGS 84 ellipsoid. These ellipsoid heights are not the same as orthometric heights, or elevations, which would be obtained from conventional differential or spirit levelling. This distinction between ellipsoid heights and orthometric elevations is critical to many engineering and construction projects; thus, users of GPS must exercise extreme caution in applying GPS height determinations to survey projects, which are based on conventional orthometric elevations.

GPS uses WGS 84 as the optimal mathematical model best describing the shape of the true earth at sea level based on an ellipsoid of revolution. The WGS 84 ellipsoid adheres very well to the shape of the earth in terms of horizontal coordinates but differs somewhat with the established mean sea level definition of orthometric height. The difference between ellipsoidal heights, as derived by GPS, and conventional levelled (orthometric) heights is required over an entire project area to adjust GPS heights to orthometric elevations. Many agencies developed geoid-modelling software like GEOID90, GEOID91, and GEOID93 to be used to convert ellipsoidal heights to approximate orthometric elevations. These values should be used with extreme caution.

Static or kinematic GPS survey techniques can be used effectively on a regional basis for the densification of low-accuracy vertical control for topographic mapping purposes. Existing benchmark data (orthometric heights) and corresponding GPS-derived ellipsoidal values for at least three stations in a small project area can be used in tandem in a minimally constrained adjustment program to reasonably model the geoid. More than three correlated stations are required for larger areas to ensure proper modelling of the geoidal undulations in the area. The model from the benchmark data and corresponding GPS data can then be used to derive the unknown orthometric heights of the remaining stations occupied during the GPS observation period.

16.32 CYCLE SLIP

The error sources encountered in the position determination using differential GPS positioning techniques are explained earlier. In addition to these error sources, the user must ensure that the receiver maintains lock on at least three satellites for 2D positioning and four satellites for 3D positioning. When loss of lock occurs, a cycle slip, that is, a discontinuity of an integer number of cycles in the measured carrier beat phase as recorded by the receiver may occur. In GPS absolute surveying, if lock is not maintained, positional results will not be formulated. In GPS static surveying, if lock is not maintained, positional results may be degraded, resulting in incorrect formulations. Sometimes, in GPS static surveying, if the observation period is long enough, post-processing software may be able to average out loss of lock and cycle slips over the duration of the observation period and formulate positional results that are adequate. If not, reoccupation of the stations may be required. In all differential-surveying techniques, if loss of lock does occur on some of the satellites, data processing can continue easily if a minimum of four satellites has been tracked. Generally, the more satellites tracked by the receiver, the more insensitive the receiver is to loss of lock. In general, cycle slips can be repaired.

16.33 LATENCY

In differential positioning it takes some time for the base station to calculate corrections and it takes some time for it to put the data into packets in the correct format and transmit them. Then the data makes its way from the base station to the rover over the data link. It is decoded and must go through the rover's software. The time this takes is called the latency of the communication between the base station and the rover. It can be as little as a quarter of a second or as long as a couple of seconds. And because the base stations corrections are only accurate for the moment they were created, the base station must send a range rate correction along with them. Using this rate correction, the rover can back date the correction to match the moment it made that same observation.

16.34 GPS AUGMENTATION

A GPS augmentation is any system that aids GPS by providing accuracy, integrity, availability, or any other improvement to positioning, navigation, and timing that is not inherently part of GPS itself. Augmentation of a satellite based navigation and positioning system is a method of improving the system's accuracy, reliability and availability, through the integration of external information into the calculation process. GPS augmentations can be grouped into two: Satellite-Based Augmentation Systems (SBASs) and Ground-Based Augmentation Systems (GBASs).

16.34.1 Ground-Based Augmentation System

Ground-Based Augmentation System (GBAS) uses augmentation through the use of a network of terrestrial radio transmission stations, which transmits the satellite corrections continuously. GBASs are commonly composed of one or more accurately surveyed ground stations, which take measurements concerning the Global Navigational Satellite System (GNSS), and transmit the GNSS satellite differential correction information directly to the end user from the ground stations. This system avoids the constraints associated with SBASs. Usually GBAS is localized, supporting receivers within a small geographical area, and transmitting the differential corrections in the VHF band. Various GBASs include Local Area Augmentation Systems (LAASs) , Radio transmission towers of Real-Time RTK DGPS, such as Continuously Operating Reference stations (CORS), etc.

16.34.2 Satellite-Based Augmentation System

Satellite-based augmentation system (SBAS) is a system that supports wide-area or regional augmentation through the use of additional satellite-broadcast messages. SBASs use geosynchronous satellite systems that provide services for improving the accuracy, integrity and availability of basic GNSS signals. SBASs are commonly composed of multiple ground stations, located at accurately surveyed points. The ground stations take measurements of one or more of the GNSS satellites, the satellite signals, or other environmental factors which may impact the signal received by the users. Using these measurements, information messages are created and sent to one or more satellites for broadcast to the end users. The SBAS concept is based on GNSS measurements by accurately located reference stations. The GNSS errors are then transferred to a computing centre, which calculate differential corrections and integrity messages which are then broadcasted over the continent using geostationary satellites as an augmentation or overlay of the original GNSS message. SBAS messages are broadcast via geostationary satellites to cover vast areas. The advantages of SBAS are:

- Accuracy is enhanced through the transmission of wide-area corrections for GNSS range errors.
- Integrity is enhanced by the SBAS network quickly detecting satellite signal errors and sending alerts to receivers that they should not track the failed satellite.
- Signal availability can be improved if the SBAS transmits ranging signals from its satellites.

SBASs include reference stations, master stations, uplink stations and geosynchronous satellites, SBAS reference stations, which are geographically distributed throughout the SBAS service area, receive GNSS signals and forward them to the master station. As the locations of the reference stations are accurately known, the master station can accurately calculate wide-area corrections. Corrections are uplinked to the SBAS satellite then broadcast to GNSS receivers throughout the SBAS coverage area. User equipment receives the corrections and applies them to range calculations.

Examples for SBAS are:

1. The Wide Area Augmentation System (WAAS) operated by the US Federal Aviation Administration.
2. The European Geostationary Navigation Overlay System (EGNOS), owned and operated European Union's GSA.
3. The Multi-Functional Satellite Augmentation System (MSAS) of Japan.
4. The GPS-Aided Geo Augmented Navigation (GAGAN) system of India.
5. The commercial StarFire Navigation System, Starfix DGPS system, OmniSTAR system, etc.

16.35 WIDE AREA AUGMENTATION SYSTEM

The US Federal Aviation Administration (FAA) has developed the WAAS to provide GPS corrections and a certified level of integrity to the aviation industry to enable aircraft to conduct precision approaches to airports. A Wide Area Master Station (WMS) receives GPS data from Wide Area Reference Stations (WRS) located throughout the United States. The WMS calculates differential corrections then uplinks these to two WAAS geostationary satellites for broadcast across the United States. Separate corrections are calculated for ionospheric delay, satellite timing, and satellite orbits, which allows error corrections to be processed separately, if appropriate, by the user application. WAAS broadcasts correction data on the same frequency as GPS, which allows for the use of the same receiver and antenna equipment as that used for GPS. To receive correction data, user equipment must have line of sight to one of the WAAS satellites. Two master stations, located on either coast, collect data from the reference stations and create a GPS correction message. This correction accounts for GPS satellite orbit and clock drift plus signal delays caused by the atmosphere and ionosphere. The corrected differential message is then broadcast through one of two geostationary satellites, or satellites with a fixed position over the equator. The information is compatible with the basic GPS signal structure, which means any WAAS-enabled GPS receiver can read the signal.

16.36 EUROPEAN GEOSTATIONARY NAVIGATION OVERLAY SERVICE

The European Space Agency has developed the European Geostationary Navigation Overlay Service (EGNOS), an augmentation system that improves the accuracy of positions derived from GPS signals and alerts users about the reliability of the GPS signals. Three EGNOS satellites cover European Union member nations and several other countries in Europe. EGNOS transmits differential correction data for public use. EGNOS satellites have also been placed over the Eastern Atlantic Ocean, the Indian Ocean Region (IOR-10), and the African mid-continent.

16.37 MTSAT SATELLITE-BASED AUGMENTATION NAVIGATION SYSTEM

MTSAT Satellite-Based Augmentation Navigation System (MSAS) is an SBAS that provides augmentation services to Japan. It uses two Multi-Functional Transport Satellites (MTSAT) and a network of ground stations to augment GPS signals in Japan.

16.38 GPS-AIDED GEO AUGMENTED NAVIGATION SYSTEM

GPS-Aided GEO Augmented Navigation System (GAGAN) is an SBAS that supports flight navigation over Indian airspace. The system is based on three geostationary satellites, 15 reference stations placed at various locations throughout India, three uplink stations and two control centres. GAGAN is similar to other SBAS systems, such as WAAS, EGNOS and MSAS.

16.39 GLOBAL NAVIGATION SATELLITE SYSTEM

Global Navigation Satellite System (GNSS) is the standard term for satellite navigation systems that provide autonomous global geo-spatial positioning. A GNSS allows receivers to determine their location in terms of longitude, latitude and altitude to within a few metres using the time signals transmitted from satellites. Receivers on the ground with a fixed position can also be used to calculate the precise time as a reference for scientific experiments.

As of January 2010, the United States NAVSTAR Global Positioning System (GPS) and the Russian GLONASS are the fully operational Global Navigation Satellite Systems. The European Union's GALILEO positioning system is a GNSS and the system is nearing completion. China is expanding their regional satellite navigation system (Beidou) into a global satellite navigation system named Compass. India's IRNSS, a regional navigational satellite, and its operation have begun since 2016.

16.40 GNSS CLASSIFICATION

GNSS, which provide enhanced accuracy and integrity monitoring usable for civil navigation are classified as follows:

1. GNSS-1 is the first generation system and is the combination of existing satellite navigation systems (GPS and GLONASS), with SBASs or GBASs. In United States, the satellite-based augmented component is known as WAAS, in Europe it is known as European Geostationary Navigation Overlay Service (EGNOS) and in Japan it is the Multi-Functional Satellite Augmentation System (MSAS). Ground-based augmentation is provided by systems like the Local Area Augmentation System (LAAS).
2. GNSS-2 is the second generation systems that independently provide a full civilian satellite navigation system, such as the European Galileo positioning system. These systems will provide the accuracy and integrity monitoring necessary for civil navigation. Galileo system consists of L1 and L2 frequencies for civil use and L5 for system integrity. Development in GPS is also in progress to provide with civil use L2 and L5 frequencies, making it also a GNSS-2 system.
3. Present core satellite navigation systems are GPS, Galileo and GLONASS.
4. Global SBASs such as Omnistar and StarFire.
5. Regional SBASs including WAAS (United States), EGNOS (European Union), MSAS (Japan) and GAGAN (India).
6. Regional Satellite Navigation Systems such a QZSS (Japan), IRNSS (India) and Beidou (China).
7. Continental scale GBAS, for example, the Australian GRAS and the US Department of Transportation National Differential GPS (DGPS) service.
8. Regional scale GBAS such as CORS networks.
9. Local GBAS typified by a single GPS reference station operating Real-Time Kinematic (RTK) corrections.

16.41 EARLY GROUND-BASED POSITIONING SYSTEMS

Early ground-based positioning systems such as DECCA, LORAN and Omega systems, used terrestrial long wave radio transmitters instead of satellites. These positioning systems broadcast a radio pulse from a known master location, followed by repeated pulses from a number of slave stations. The delay between the reception and sending of the signal at the slaves was carefully controlled, allowing the receivers to compare the delay between reception and the delay between sending. From this the distance to each of the slaves could be determined, providing a positional value or fix.

The first operational satellite navigation system was TRANSIT. It was developed in 1960s to provide more accurate position for ships and submarines. Transit's operation was based on the Doppler effect. Six transit satellites provided worldwide coverage with an accuracy of 200 m. The satellites travel on well-known paths and broadcast their signals on a well-known frequency. The received frequency will differ slightly from the broadcast frequency because of the movement of the satellite with respect to the receiver (Doppler effect). By monitoring this frequency shift over a short time interval, the receiver can determine its location to one side or the other of the satellite, and several such measurements combined with a precise knowledge of the satellite's orbit can fix a particular position. The broadcast from the orbiting satellite includes its precise orbital data. In order to ensure accuracy, the US Naval Observatory (USNO) continuously observed precisely the orbits of these satellites. As a satellite's orbit deviated, the USNO would send the updated information to the satellite. Subsequent broadcasts from an updated satellite would contain the most recent accurate information about its orbit. TRANSIT continued to perform reliably for over 25 years (till 1996).

Modern systems are more direct. The satellite broadcasts a signal that contains the position of the satellite and the precise time the signal was transmitted. The position of the satellite is transmitted in a data message that is superimposed on a code that serves as a timing reference. The satellite uses an atomic clock to maintain synchronization of all the satellites in the constellation. The receiver compares the time of broadcast encoded in the transmission with the time of reception measured by an internal clock, thereby measuring the time-of-flight to the satellite. Several such measurements can be made at the same time to different satellites, allowing a continual fix to be generated in real time.

Each distance measurement, regardless of the system being used, places the receiver on a spherical shell at the measured distance from the broadcaster. By taking several such measurements and then looking for a point where they meet, a fix is generated. However, in the case of fast-moving receivers, the position of the signal moves as signals are received from several satellites. In addition, the radio signals slow slightly as they pass through the ionosphere, and this slowing varies with the receiver's angle to the satellite, because that changes the distance through the ionosphere. The basic computation thus attempts to find the shortest directed line tangent to four oblate spherical shells centred on four satellites. Satellite navigation receivers reduce errors by using combinations of signals from multiple satellites and multiple corelators, and then using techniques such as Kalman filtering to combine the noisy, partial and constantly changing data into a single estimate for position, time and velocity.

16.42 NEED FOR GNSS

The original motivation for satellite navigation was for military applications. But the GNSS systems have a wide variety of uses, such as:

1. Navigation, ranging from personal hand-held devices to devices fitted to cars, trucks, ships and aircraft.
2. Time transfer and synchronization.
3. Location-based services.
4. Surveying.
5. Geographic Information System.
6. Search and rescue.
7. Geophysical sciences.
8. Tracking devices used in wildlife management.
9. Asset tracking, as in tracking fleet management.
10. Road pricing.
11. Location-based media services.

It should also be noted that the service providers of satellite navigational system have the ability to degrade or eliminate satellite navigation services over any territory that they desire.

TABLE 16.1 Showing measuring modes and the time duration for measurements and its accuracy

Type	Observation Time	Accuracy
Single point with a single receiver	1. 30–60 sec 2. 6–4 h	10–50 m with SA Off 1–3 m
Static	1–4 h	5 mm + 1 ppm
Rapid static	5 min	5 mm + 1 ppm
Reoccupation	5 min	1–2 cm + 1 ppm
Stop and go	2 epochs	1–3 cm + 1 ppm
Kinematic	1 sec	1–3 cm + 1 ppm
Code differential real-time (DGPS)	1 sec	30–50 cm
Phase differential real time (DGPS)	1 sec	1–2 cm +2 ppm

TABLE 16.2 Showing observation time for static survey

Length of Baseline	Number of Satellites	GDOP	Time of Observation	Accuracy
10–50 km	4–6	<6	60–90 min	5 mm + 1 ppm
50–100 km	4–6	<6	90–150 min	5 mm + 1 ppm
Above 100 km	4–6	<6	120–150 min	5 mm + 1 ppm

TABLE 16.3 Showing observation time for rapid static survey

Length of Base Line	Number of Satellites	GDOP	Time of Observation	Accuracy
0–5 km	4–6	<6	5–10 min	1 cm + 1 ppm
5–10 km	4–6	<6	10–15 min	1 cm + 1 ppm
Above 10 km	4–6	<6	Minimum 20 min	1 cm + 1 ppm

TABLE 16.4 Showing data recording intervals for different modes

Mode	Data Recording Interval
Static	15–30 sec
Rapid static	5 sec
Kinematic	1–5 sec

REVIEW QUESTIONS

1. Enumerate the advantages of GPS surveying over traditional surveying methods.
2. What are the limitations of GPS surveying?

3. Explain the fundamental difference between GPS navigation and GPS surveying.
4. What are the characteristics of GPS surveying and navigation?
5. Explain different modes of positioning techniques used using GPS.
6. Discuss about the general field survey procedures involved in GPS surveying.
7. Explain the importance of making measurements to four or more GPS satellites for the determination of a position using GPS.
8. What is absolute positioning? Narrate the different methods used for absolute positioning.
9. What are the differences between absolute positioning and differential positioning?
10. How does a GPS receiver calculate a positional value using pseudo-ranges?
11. Explain about pseudo-ranges and pseudo-ranging equation.
12. What is the concept of Differential Global Positioning System?
13. Explain the difference between code phase and carrier phase differential positioning?
14. Explain about the various methods used in GPS surveying.
15. Explain about real-time differential code-phase horizontal positioning.
16. What are the advantages and disadvantages of RTK DGPS method?
17. Explain the methodology adopted for static surveys.
18. When you are asked to observe four control points on the ground using static survey, how will you map the control points following static survey methodology to get millimetre accuracy, if the accurate positional value of one control station is provided?
19. Explain the methodology of rapid static survey using GPS.
20. Why rapid static surveys are preferred to static surveys?
21. Differentiate between static and rapid static surveys.
22. Differentiate stop-and-go kinematic method and true kinematic method.
23. What is antenna swap and why it is performed?
24. Discuss about the application of DGPS, briefly with examples.
25. Explain briefly about the methodology of real-time kinematic method.
26. Discuss the concept of GNSS.
27. Write short notes on
    i. RTCM—104 format
    ii. RINEX format
    iii. Antenna Swap and height measurements static survey field procedure
    iv. Kinematic survey field procedure
    v. Office procedure involved after data collection in GPS surveying
    vi. RTK GPS surveying
28. What are the requirements of software in GPS surveying?
29. Explain about differential reduction technique.
30. Explain about single differencing and double differencing.

31. Explain about GNSS classification and the need for GNSS.
32. What is augmentation and explain about satellite-based augmentation.
33. Briefly explain about ground-based augmentation.
34. Write short notes on:
    i. Latency
    ii. Cycle slip
35. A 150-sq km terrain is to be surveyed using GPS for making contour map with 50 cm contour interval. The final map must show permanent structures such as buildings, bridges, roads, culvers, transmission line, communication towers and geographical features like streams, rivers, ponds, agricultural lands, canals, drains, hills, etc. The shore lines of all water bodies and boundary lines of paddy fields, cultivated land, forest area, etc., should be clearly mapped. The final map should be given in digital format also. The area has been evaluated to perform successful GPS surveys. The accuracy parameters should be as below:
    i. Vertical control points: Accuracy must be within 3 cm for contouring, and less than 2 cm for control points.
    ii. For horizontal control points: within 1 cm.
    iii. Permanent structures and boundary lines: within 5 cm.
    iv. River, streams and pond boundaries: within 10 cm.
36. Explain about the detailed procedure to be followed for the execution of the above survey using GPS. What are the equipment, human resources and other inventories required to complete the work? What are the methodologies to be followed for establishing control points and what method will be adopted for taking various points for getting precise contour interval.

This page is intentionally left blank

Appendix A

BASIC GEODETIC ASPECTS

Geodesy: Geodesy is the branch of science which determine the size and shape of the earth, the precise positions and elevations of points, lengths and direction of lines on the earth's surface and the terrestrial gravity and its variations. It is the science of accurate measurements and studying the three fundamental properties of the earth, namely, the geometric shape of earth, its orientation in space, and its gravity field. The three main branches of Geodesy are Geometric Geodesy, Gravimetric or Physical Geodesy and Satellite Geodesy. Basic knowledge of Geodesy is required for global positioning system (GPS) surveying and geographic information system (GIS).

Geoid: The Geoid is a surface along which the gravity potential is equal everywhere and to which the direction of gravity is always perpendicular. It is a surface formed by joining gravity points on the surface of earth. The Geoid can be defined as the surface of the earth's gravity field, which approximates mean sea level. The surface of Geoid is always perpendicular to the direction of gravity pull. The mass of the earth is not uniform at all points, and the magnitude of gravity varies. Hence the shape of the Geoid is irregular.

The Ellipsoid: The earth is flattened slightly at the poles and bulged near the equator. The geometrical figure used to represent earth in geodesy is an ellipsoid, which nearly approximates the shape of earth. The ellipsoid is the figure, which would be obtained by rotating an ellipse about its minor axis. The ellipsoid has no physical surface but is a mathematically defined surface.

Geodetic Coordinate: Geodetic coordinate components consist of latitude (ϕ), longitude (λ) and ellipsoid height (h). The geodetic latitude of a point is the angle from the equator plane to the vertical direction of a line normal to the reference ellipsoid. The geodetic longitude of a point is the angle between a reference plane and a plane passing through the point, both planes being perpendicular to the equatorial plane. The geodetic height at a point is the distance from the reference ellipsoid to the point, in a direction normal to the ellipsoid (see Fig. A.1).

Earth Centred Earth Fixed *X*, *Y* and *Z* (ECEF Coordinate System): Earth centred, earth-fixed X, Y and Z is a geographic **coordinate system** and Cartesian **coordinate system**. It is a three-dimensional position with respect to the centre of mass of the reference ellipsoid. The Z-axis point towards North pole and X-axis is defined by the intersection of the plane defined by the prime meridian and the

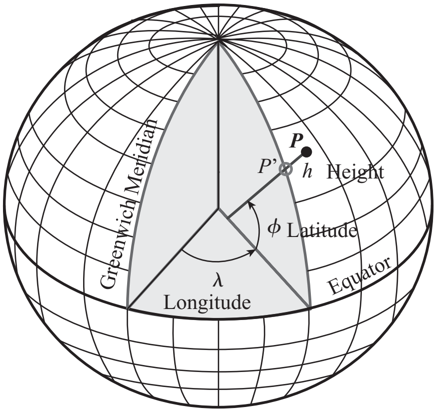


FIGURE A.1 Geodetic coordinates

equatorial plane. The Y-axis completes a right-handed orthogonal system by 90-degree plane, east of the X-axis and its intersection with the equator. In this three-dimensional right-handed coordinate system the X-coordinate is a distance from the Y–Z plane measured parallel to the X-axis. It is always positive from the zero meridian to 90° W longitude and from the zero meridian to 90° E longitude. In the remaining 180° the X-coordinate is negative. The Y-coordinate is a perpendicular distance from the plane of the zero meridian (see Figs. A.2 and A.3).

Geodetic Datum: A geodetic datum is a set of reference points, used to locate places on the earth. The results of a GPS survey require the coordinates of points to be given in relation to a previously defined geodetic datum. This datum may be the one previously established by the GPS or a conventional local geodetic datum, usually comprising of two distinct systems, one for the horizontal position and the other for the height component.

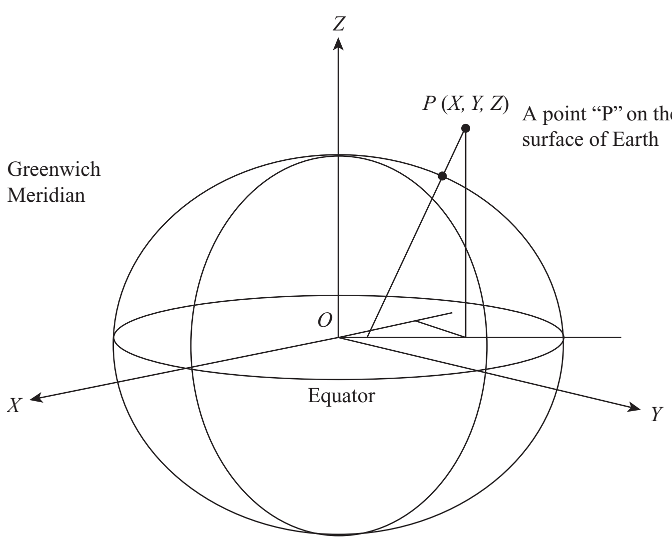


FIGURE A.2 Earth centred, earth fixed *XYZ* (ECEF coordinates)

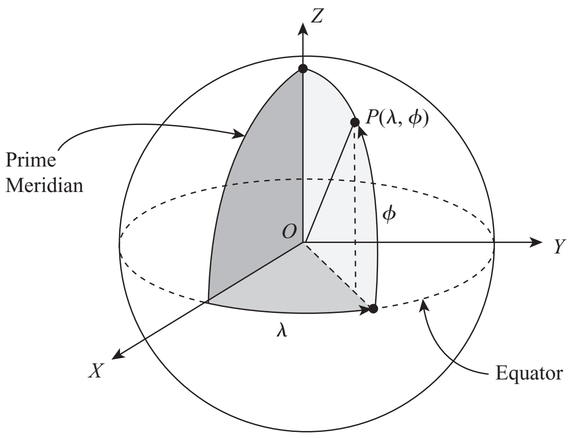


FIGURE A.3 ECEF reference ellipsoid

Horizontal Datum: The horizontal datum is the model used to measure positions on the earth. A specific point on the earth can have substantially different coordinates, depending on the datum used to make the measurement. There are hundreds of local horizontal data around the world, usually referenced to some convenient local reference point. Contemporary data, based on increasingly accurate measurements of the shape of the earth, are intended to cover larger areas, for example, WGS84, Everest and PZ90.

Vertical Datum: A vertical datum is used as a reference point for elevations of surfaces and features on the earth including terrain, water levels and man-made structures. Vertical data are either tidal, which are based on sea levels, gravimetric, which are based on a geoid or geodetic, which are based on the same ellipsoid models of the earth used for computing horizontal data.

GPS Datum: The datum used by the GPS for defining a point on the surface of earth is known as GPS datum. The commonly space-based data are founded on the geocentric (earth centred) three-dimensional Cartesian system defined by a global ensemble of US military tracking station established since the start of the space age. The standard datum is the Word Geodetic System 1984 (WGS84).

WGS84 provides an ellipsoidal model of the earth's surface in which cross-section of the earth parallel to the equatorial plane are circular. The equatorial cross-section of the earth has a radius of 6,378,137 m, which is the mean equatorial radius of the earth. In WGS84, cross-section of the earth normal to the equatorial plane is ellipsoidal. The semi-major axis (a) of WGS84 has the same value as the mean equatorial radius and equal to 6,378,137 m and the semi-minor axis (b) is taken as 6,356,752.3142 m.

Geodetic Datum Elements: The ellipsoidal model bulging at the equator and flattened at the poles has been used as a representation of the general shape of the earth's surface. It is called an oblate spheroid. Several reference ellipsoids have been established for various regions of the planet. They were precisely defined by their semi-major axis and flattening. The relationship between these parameters are expressed in the formula.

$$f = \frac{a-b}{a}$$

where f is the flattening, a thesemi-major axis and b the semi-minor axis.

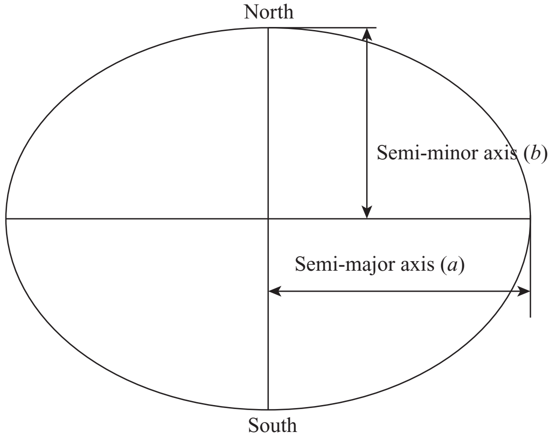


FIGURE A.4 The semi-major and semi-minor axes

A reference ellipsoid is defined by a series of parameters, which include a semi-major axis (a) and a semi-minor axis (b) (see Fig. A.4), its first eccentricity (e) and its second eccentricity (e'). The ellipsoid flattening (f) may be required depending upon the formulation.

$$e = \sqrt{\frac{a^2 - b^2}{a^2}} \qquad e' = \sqrt{\frac{a^2 - b^2}{b^2}}$$

WGS84 parameters $a = 6{,}378{,}137$ m,

$b = a(1 - f) = 6{,}356{,}752.31424518$ m

$f = 1/298.257223563$.

Everest Ellipsoid: In India the reference ellipsoid used is the ***'Everest' (India 1956).*** The shape Everest ellipsoid is completely defined by a semi-major axis of 6,377,301.243 m, semi-minor axis of 6,356,100.228 m and a flattening of 1/300.8017. *The* first Surveyor General of India *Sir George Everest*, devised the Everest spheroid in 1830. In this datum minor axis is assumed to be parallel to or coinciding with the mean polar axis or rotation of earth. *Kalyanpur* in Central India was taken as its origin, which is at the intersection of two primary series—the greater arc meridional series connecting *Cape-comerin* and *Mussorie,* and the longitudinal series connecting *Calcutta* and *Karachi.*

Everest (Indian) 1956 parameters: $a = 6{,}377{,}301.243$ m,

$b = 6{,}356{,}100.228$ m

$f = 1/300.8017$.

The Local Coordinates: Local coordinates are coordinates used in a particular country's maps based on a local ellipsoid, and designed to match the geoid in that area (see Fig. A.5). The ellipsoids used in most local coordinate systems throughout the world were first defined many years ago before the advent of space techniques. This ellipsoid tends to fit the area of interest, but could not apply to

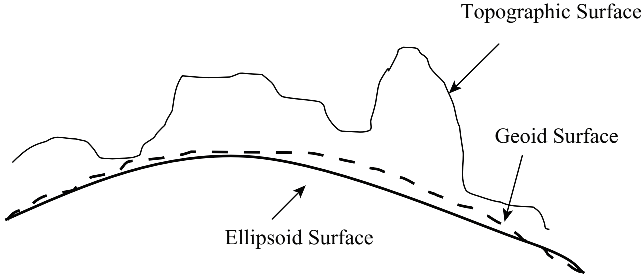


FIGURE A.5 The ellipsoid, geoid and topographical surface of a place

other areas of the earth. Hence, each country defined a mapping system based on a local ellipsoid. When using GPS, the coordinate of the calculated positions is based on the WGS84 ellipsoid. But coordinate systems used in different areas were usually a local coordinate system and therefore the GPS coordinates have to be transformed into this local system.

GPS and Height: The accuracy of GPS height measurements depends on several factors but the most crucial one is the imperfection of the earth's shape. The GPS uses height (h) above the reference ellipsoid that approximates the earth's surface. The orthometric height (H) is the height above an imaginary surface called the geoid, which is determined by the earth's gravity and approximated by Mean Sea Level. The signed difference between the two heights, that is, the difference between the ellipsoid and geoid is known as the geoid height (N). Figure A.6 shows the relationships between the different models and explains the reasons why the two hardly ever match spatially.

Scale Factor: The scale factors are ratios that can be used as multipliers to convert ellipsoidal lengths, also known as geodetic distances, to grid distances (lengths on the map projection surface). Or, in other words, scale *factor is a* multiplier to change *geodetic* distances based on the earth model (ellipsoid) to the grid plane.

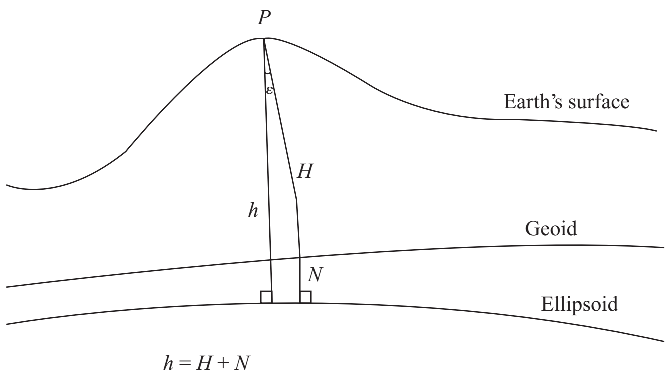


FIGURE A.6 Relation between orthometric height and ellipsoidal height

SAMPLE EQUIPMENT PROCEDURE OF VARIOUS EQUIPMENT

Appendix B

B.1 TOPCON: SURVEY PROCEDURE

B.1.1 Setting Up the Total Station

Levelling the Instrument

- Centre the instrument above the maker point using the optical plumb bob, moving two legs of the tripod to maneuver the total station.
- Press the tripod feet into the ground.
- Centre the bull's eye level, first using tripod leg adjustments, and then using the three knuckle screws for finer adjustment. After centering the bull's eye, centre the line level, first using two knuckle screws concurrently; then turn total station 90 degrees, and use the third knuckle screw to finally centre the line level.

Preparing the Instrument

- Connect the data collector with the 18" coiled cable to the total station.
- Turn on the total station (with data collector off).
- Break the plane by moving the scope up and down.
- Measure height of the instrument with tape measure, to centre the mark on the side of the scope.

B.1.2 Preparing the Data Collector

Creating a New Data File

- Press F2 'Collect Data'.
- Press F1 'Job File' Type in job file name followed by ENTER.
- Press F2 'Job Prompt' Select FIELDM followed by ENTER.
- Press F3.
- Enter JOB ID.
- Enter JOB #.

- Enter PARTY CHIEF.
- Enter INSTRUMENT ID—goes to next screen.

Adding to Existing Data File

- Press F2 ‘Collect Data’.
- Press F1 ‘Job File’ Use arrows to select file. Press ENTER.
- Press F3.
- Press F1—goes to next screen.

Obtaining First Back Sight for Total Station

- Enter OCC. PT# (occupied point, normally numbered 1).
- Enter OCC. ID (occupied point ID, descriptive label for marker point).
- Enter INST. (instrument height) in metres.
- Enter OCC.NORTH. Press ENTER to accept default, or type in new north coordinate.
- Enter OCC.EAST.
- Enter OCC.ELEV.
- Press ENTER.
- Enter back sighting point, normally point number 0.
- Enter azimuth of back sighting point.
- Aim total station at back sighting point and press ENTER of data collector.
- Press F1 for angles. The azimuth on the total station LCD should show the entered azimuth.
- Recheck the back sighting point in scope, and press ENTER on data collector.
- Angles in both horizontal and vertical direction are shown in the LCD.

Obtaining Sighting with the Total Station

- Next point # will be shown in the LCD, it will increment up, press ENTER to accept next point or use digit keys to change point #.
- Enter *ID* (target point description).
- Enter *target height* (in metres).
- Collect next point appears, enter *COARSE* (F2).
- Continue acquiring (Y/N question). Yes continues, no repeats.
- Collect *TOPO POINTS* (F3).
- Enter *OCC.PT#* (what point is the total station at).
- Enter *OCC.ID* (same as before).
- Enter *INST.HEIGHT* (in metres).
- Instrument will enter auto mode; if the target height does not change, then use F1 to acquire target; if rod height or data label is changed, use F2 to obtain a new topo point.

Closing Down the Total Station

- Hit F0 to exit to top menu.
- Enter Y to query ‘Exit <Y/N>’.
- Enter Y to query ‘Generate NEZ <Y/N>’.
- At top menu, turn total station off.

Moving the Instrument

- Move the station to a new point.
- Set up as described in Sections 1.1 and 1.2.
- Back sight on previously shot point.

Downloading Data

- Log on to a DEC station with FC-4 connected to /dev/tty0 (at the rear) via RS-232 cable.
- Turn on FC-4.
- Enter ‘File Manager’ (F3).
- Select ‘Data Transfer’ (F2).
- Select ‘Send A File’ (F1).
- Select ‘Send FC-4 Data’ (F1).
- At prompt type pcomm (to start Pro Comm programme).
- Type <ctrl>A D select TOPCON from menu.
- Type <ctrl>A N select ASCII (#6).
- Type in FILENAME then press RETURN.
- Use FC-4 arrow keys to select data file then press ENTER.
- When finished transferring file, press ENTER to send a new file.
- Repeat:

  Type <ctrl>A N select ASCII (#6)

  Type in FILENAME then press RETURN

  Use FC-4 arrow keys to select data file then press ENTER until all files are transferred.
- Exit out of Pro Comm. Type <ctrl>A X.
- If problems arise in transferring data, be sure to check that the FC-4 and ProComm have the same settings ‘N (parity not checked) 8 (data bits) 1 (stop bit)’ and the same baud rate. Baud rates of 9600 and higher have been known to have problems in transferring all the data in a file.

Preparing Data Files for MATLAB

- Using an xxxx.n file from the Topcon (NEZ files).
- Type the following at the Unix prompt:

  tr ‘012\’ ‘ <xxxx.n >xxxx.txt

  In other words: tr (a Unix command) (one space) ‘(single quote) 012\(ASCII code for a carriage return)’ ‘(a single quote, followed by a space and another single quote)’ (then a space followed by another single quote) < filename ('less than' character and input file name) > filename ('greater than' character followed by the output file name).

The above converts the Topconto to a single string.

- Use NOTEPAD (dxnotepad) or other file editor to insert a carriage return after each data label in the file.

  This converts the file to five columns containing:
  | Point Number | Northings | Eastings | Elevation | Data Label |

B.2 LEICA TC800 TOTAL STATION FIELD PROCEDURES

Field Trip Preparations

- Insure that the unit was properly stored from last use.
- All units should be clean and free of debris.
- Keep plugs clean and dry. Blow out any dirt lodged in the plugs of the connecting cables.
- Always use the original transport case for each item.
- Insure that both batteries are properly charged. The GKL-22 charger should be able to fully charge an NiCd battery within 14 hours. The charging procedure starts automatically whenever a battery is connected to the charger and a red control lamp indicates a fully charged battery.
- The battery chargers are intended for indoor use only. Use only in a dry place, never outdoors. Short-circuiting the battery terminals can cause fire.
- Finally, make sure whether you are using the TC605L or the TC800. Except for some minor differences, these machines are essentially the same in appearance and operation. The text is written for the TC605L. Modifications for the TC800 will be shown in green.

B.2.1 Getting to Know the Equipment

On-Site Set-Up

- Place the tripod firmly into the soil and raise legs to a comfortable height. The telescope of the total station should be approximately at eye height or slightly lower. Make allowance for

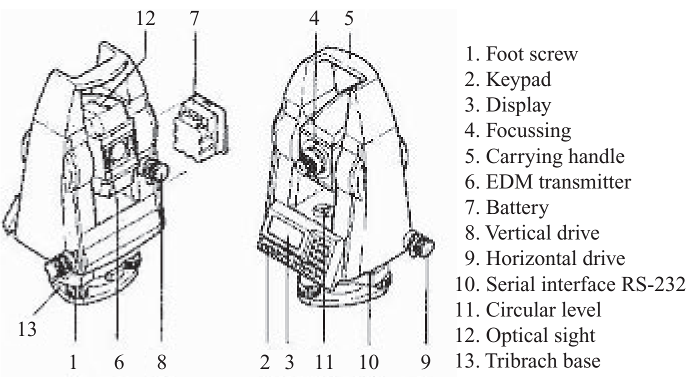
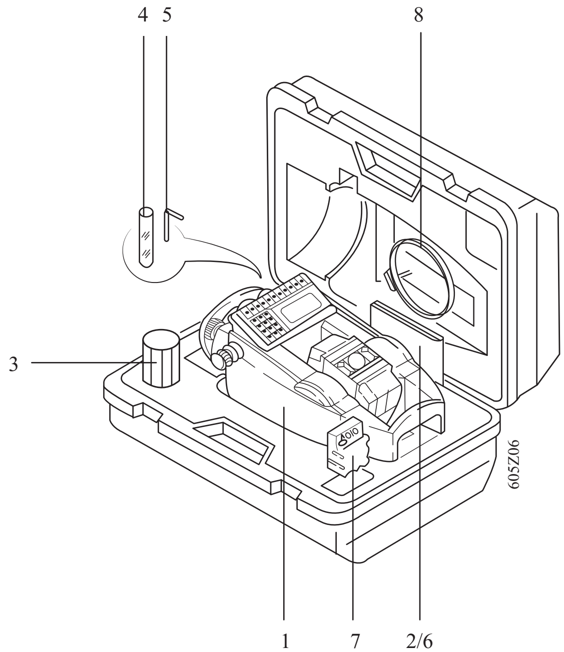


FIGURE B.1 Total station and its parts (TC605 model of Leica)



1. Instrument
2. User's manual
3. Protective cover
4. Set of pins
5. Allen key
6. Instrument
7. User's manual
8. Protective cover
9. Set of pins
10. Allen key
11. Diskette with TCTOOLS
12. Spare battery
13. Data cable (RS232)

FIGURE B.2 Transport case of total station

the instrument height once it is attached. Take into account that multiple people will be using the total station through the day and adjust the station height to the eye-height of the shortest group member. The tripod cannot be moved during a measurement session, so get it right the first time!

- Use leg height adjustments to visually level the top surface of the tripod. Check with the bull's-eye level of your Brunton compass.
- Remove tribrach (13) from the instrument by rotating the black lever on the tribrach counter-clockwise until freed from the tripod. Mount the tribrach on the tripod with a large aluminium screw. Insure you do not strip the threads.
- Verify that the three foot-screws (1) are in a neutral position (near the centre). This will allow for the greatest range of adjustments.
- Using the three foot-screws (1), level the tribrach (13) using the bull's-eye level on the tribrach.
- Mount the total station to the tribrach base (13) and secure by rotating the latch clockwise.
- Turn on the total station.
- Press Menu.
- Go to Level.
- Release the horizontal drive gross adjustment screw (9).
- Rotate the total station so that the eyepiece of the telescope is between two of the three foot-screws. Using only these two foot-screws, level the instrument as indicated on the total station display. (Off level and level images will be inserted.)

- Rotate the total station one third of a turn and repeat using the next pair of foot-screws. Repeat again for the third and final combination of two foot-screws. Then repeat all three combinations in the same order until perfectly level.
- Press the ESC key. For the TC800, press CONT.
- Measure and record the instrument height using the small indentation on the right side of the total station. Set the prism rod to match the instrument height.
- Have one person walk out at least 10 m due north; the farther the better. Using a compass properly, shoot a bearing from the prism rod to the total station. Adjust the prism rod position until the bearing is due south. Rotate the total station and sight the prism. The total station should now be facing due north. It is critical during this step that the prism rod is exactly vertical and that the cross hair in the total station optical sight is exactly in the centre of the prism.
- Press the Menu key, go to set horizontal. Set the horizontal angle to zero and press ENTER. Finally, press ENTER once again to return to the measurement screen. For the TC800, press MENU, SET, Horizontal, CONT, CONT.
- The total station is now ready to take measurements.

Collecting Data

- Begin with the benchmark identified for you.
- Each team member should become proficient with each of the three responsibilities of operating the total station. The rod holder handles the prism rod; the note taker records the point number and identifies the location of the point in words and on a sketch; the surveyor operates the total station by sighting the prism in the cross hairs and digitally records the data.
- Insure that the prism is pointing towards the total station.
- On the total station, release the gross adjustment levers on both the horizontal (9) and vertical (8) drives. Site the white arrow on the top of the telescope (the optic sight (12)) towards the prism rod. Lock the gross adjustment levers.
- Now site through the telescope at the prism. Focus the eyepiece using the large focussing ring (4) around the ocular. Use the vertical (8) and horizontal (9) drives for fine adjustments to align the cross hair into the centre of the prism.
- Press the distance key. Listen as the unit makes a measurement. If the conditions are such that a measurement is possible, you should hear a ‘beep’ (indication of a measurement has been requested), some mechanical sounds like the shutter of a slow moving camera (this is the laser activating and emitting the laser pulse), and finally another ‘beep’ when the measurement is successful. If you hear the mechanical shutter noise repeat several times, the total station was unable to detect the returning laser beam. Check that the prism is pointed towards the telescope, move any obstacles and realign the telescopic sight to the centre of the prism and try again.
- When an acceptable reading has been taken, the unit will ‘beep’ and display the horizontal angle, the vertical angle, the horizontal distance, and the vertical elevation difference from the instrument location. Inspect these values to make sure they are reasonable and record these values in a notebook or data log.
- Press the REC key to record this measurement. The reading number will briefly flash on the screen. Record this and consult frequently with the rod holder and note taker to verify that this number is the same as that being recorded in the field notes.

- Continue taking measurements.

Disassembly

- Turn off power to the station by pressing the DSP and 'light' button simultaneously. (For the TC800, simply press the OFF button.) All data will be saved automatically.
- Insure the horizontal and vertical drives are in the unlocked position. Remove the station and tribrach by unscrewing the large metal hand screw in the centre of the tripod. Place carefully into the case. Replace yellow cap and attach with hand screw.
- Remove tripod legs from the ground; shorten the telescoping leg using the hand screw near the centre of the leg. Be sure to retighten the screw when the leg is in the shortest position. Bring the legs together and bind with Velcro strap.

B.2.2 Field Procedure of Sokkia Total Station Model SET6F

Setup

1. Place the tripod approximately over a known point locking legs at a convenient height, so that the machine will be at or lower than eye level and the legs are at equal distances from each other. Eyeball the head of the tripod, so it is as close to level as possible.
   - Be sure the legs of the tripod are firmly planted into the ground.
   - For smooth surfaces (such as concrete, asphalt or tile), use a folding metal tripod footing to secure the legs.
2. Remove the instrument carefully from the casing with both hands. Place on top (supporting with top handle) of the tripod and tighten the centering screw below the platform into the instrument, aligning the three corners of the machine and the platform. Use sight (a) on the backside of the LCD display (b) to centre the instrument over the exact known point to be surveyed.

Power and Preparation

1. Attach one of the batteries (e) to the side of the instrument with the clamp side up. Press any one of the five buttons below the display to turn on the machine. It shall beep and the display

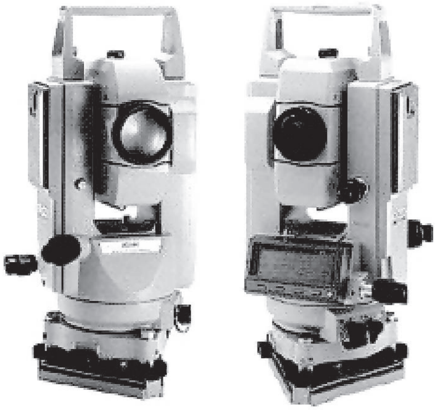


FIGURE B.3 Sokkia total station model SET6F should indicate the instrument is not level and must be levelled and indexed (precisely level internal components).

- To switch the power off, hold ESC button and press indicated button that corresponds to OFF on the display.
- If the battery is at a low level, the following will be displayed, '**Battery is low!**'—switch batteries and charge the drained one using the jack provided.
- Prior to storing the instrument for its next use, check the status of both provided batteries. If either is only *entirely* drained, charge overnight using the given equipment.

2. Locate the horizontal level bubble above the LCD display. Rotate the instrument by loosening the horizontal clamp (h) and align the display with any two of the levelling screws (c). Tighten or loosen the left screw so that the bubble is in the centre. Rotate the instrument clockwise to the next two screws, and again use the left one to centre the bubble. Rotate to the final two pair of screws and centre the bubble. Check stationary levelling bubble (i) to see if it is in the centre. If not, repeat the previous levelling process.

   - If the error message '**Tilt out of range**' is displayed, it is indicating the instrument is off-level. Relieve the instrument.

3. To index the vertical circles, loosen the vertical clamp (f), and manually rotate the telescope either way (g) twice. The beep should be heard and the zenith angle (**Za** vertical angle) will appear on the LCD display.
4. Loosen the horizontal clamp (h) and rotate the instrument clockwise twice to index the horizontal circles. The beep is heard again and the horizontal angle (**HAR**) is displayed.

   - Vertical and horizontal indexing have now been completed.

5. Note the menus displayed. Each option shown on the home page (reached by pressing ESC) opens a section, which contains several (up to 3) pages. To scroll through these pages to reach other options, press the button left of the yellow ESC button that reads → PX.
6. Set the target and instrument height by pressing **Ht** in **S-O** mode. Measure the target height by reading the measurement on the reflector pole at the clamp (set at any arbitrary height suitable for the job). Measure the instrument height by taping the distance from the black point on the side of the instrument (level with centre of the telescope) to the known point on the ground.

   - Be sure to note the units used (currently default set at feet and decimal fractions of feet; see manual to change to metric units) and height of the instrument and target in the field book.
   - When using two reflecting poles, be sure to set each at the same height.

Angle Measurement

1. Sight the first point (focus with eye piece and align centre hairs with the centre of reflector) using the horizontal clamp and the fine motion screw. Set the angle to zero by pressing **0SET** in **THEO** mode. Sight the second target and read the **HAR** on the display.

   - If you wish to read the angle by rotating the instrument to the left, press **R/L** in the **THEO** mode (display will read **HAL** or **HAR** for left or right, respectively).

2. For higher accuracy, the average of a number of readings can be taken using repetition. Sight the first target and press **REP** in the **THEO** mode. Press **BS** (back sight), then sight the second target. Press **FS** (fore sight) and the angle between the two will be displayed. Sight the first target again, press **BS**, and site the second target again and press **FS**. The average of the two readings will be displayed. Repeat up to 10 times for higher accuracy.
3. The slope of the line being shot can be displayed as a percentage by pressing **ZA%** in the **THEO** mode. This is read as **VA** and gives the percentage grade of the line. Press it again to return to the **ZA** reading.

   - **VA%** will be displayed when the parameter is set to 'Horizontal 0', instead of 'Zenith 0', but performs the same function.

Distance and Angle Measurement (Most Useful and Suggested Method)

1. Sight the target and select for slope, horizontal or height (**SHV**) measurement. Press **Sdist** to start the measurement and **STOP** to end. The distance, vertical, and horizontal angle are displayed. Press **SHV** to view the other measurements (horizontal distance or height difference).
2. To measure the horizontal distance several times and display the average, sight the target and press **Hdist** in the **THEO** mode. Three measurements are taken and the average (**H-A**) is displayed after a few seconds.

*The most recently taken data can be recalled and displayed by pressing **RCL** in the **EDM** mode.

Coordinate Measurement (Not Suggested for Use, See Manual for Use)

1. In order to begin the coordinate measure, set the initial coordinates of the station. This is done by pressing the **S-O** button in the main menu. Then press the **Stn-P** button on the second page of the **S-O** menu. Choose the **Input** button, then set the initial coordinates and press **ENTER**.
2. Sight the target and press **COORD** in the **S-O** mode, then press **STOP** to end the measurement. The coordinates of the target are given with respect to the initial starting position (0,0,0) and designated direction to be north.

Measuring the Distance between Two Points

1. Sight the first position and press **Sdist**, **Hdist** or **Vdist** in **EDM** mode to start the measurement. Stop the measurement by pressing the **STOP** and sight the next point. Press **MLM** on the same page to start the measurement, then press **STOP** to stop the measurement. The slope, horizontal and height difference between the two points is displayed. This can be repeated as many times as necessary.
2. The slope may be read as a percentage by pressing **S%** in the same mode after the missing line measurement has finished. This displays the per cent grade between the two points.

Distance Setting-Out Measurement

1. To find the direction and distance of a point, set out a wanted distance from the instrument station, sight the reference direction and press **0SET** in the **THEO** mode to set the **HAR** at 0. Turn the theodolite until the required angle is displayed and the horizontal movement is locked.
2. Press ESC to go to the basic mode and go to the **S-O** mode. Go to **S-O_D** for the data and input the desired distance to set out. Set the reflecting prism in the sighting line and press **SO_Hd** to

start the distance measurement. The difference between the desired distance and the measured distance is displayed on the first line.

3. Move the reflecting prism towards or away from the instrument until **H** distance becomes 0 m to determine the point at the desired distance.
   - If there is negative (−) data: Move the prism away from the instrument.
   - If there is positive (+) data: Move the prism towards the instrument. Press **STOP** to end the measurement.

Coordinates Setting-Out Measurement

1. Set the station coordinates and initial azimuth angle. Press **S-O_P** in the **S-O** mode and input the desired coordinates for N and E and press **YES** to store the data. Press **SO_HA** in the **S-O** mode to start the angle measurement. The setting-out horizontal angle, **dHA** is displayed. Use the horizontal clamp and fine motion screw to turn the theodolite until **dHA** reads 0°00′00″ and lock the clamp.
2. Sight the reflecting prism on the sighting line and press **SO_HD** and move the reflecting prism until **H** reads 0 m.Basics of Total Station.

SOKKIA TOTAL STATION CX SERIES, FIELD PROCEDURE

Appendix C

Chapter Outline

C.1. OBSERVATION PROGRAM MENU

C.1.1. Topography Observation

1. Job Select

   Press '★' key, and then 'Entry menu'. Select where you store the date.

```
Entry menu
 JOB select
 JOB detail
 Topography
 Coordinate
 S-O
```

→

```
JOB selection
 : JOB1
Coord search JOB
 : JOB1

LIST
```

2. Job Detail

JOB name and/or scale factor can be changed, if necessary. Press ‘ENT’ to complete inputting.

```
Entry menu
 JOB select
 JOB detail
 Topography
 Coordinate
 S-O
```

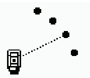
→

```
JOB details
JOB name                    A
              JOB1
SCALE  :  1.00000000
                        OK
```

3. Topography Observation

Select ‘Topography’ and press ‘ENT’.

```
Entry menu
 JOB select
 JOB detail
 Topography
 Coordinate
 S-O
```



↓

```
PTTRN   :
No of SETs : 1      - > 1)
No of Obs  : 1      - > 2)
F1/F2 Obs  : No     - > 3)
PresnterPt : No     - > 4)
PTTRN                  OK
```

Page 1

```
BS Obs-Dist  :  No  -> 5)&6)



PTTRN              OK
```

Page2

1. Number of distance sets
2. Number of distance readings
3. RL observation
4. Pre-entered point registration
5. Backsight distance measurement
6. Backsight distance check

F 4 : OK

↓

```
N0  :          0.000
Y0  :          0.000
Z0  :        < NULL >
Pt .            100
Inst .h        0.000m
LOAD      RESEC   OK
```

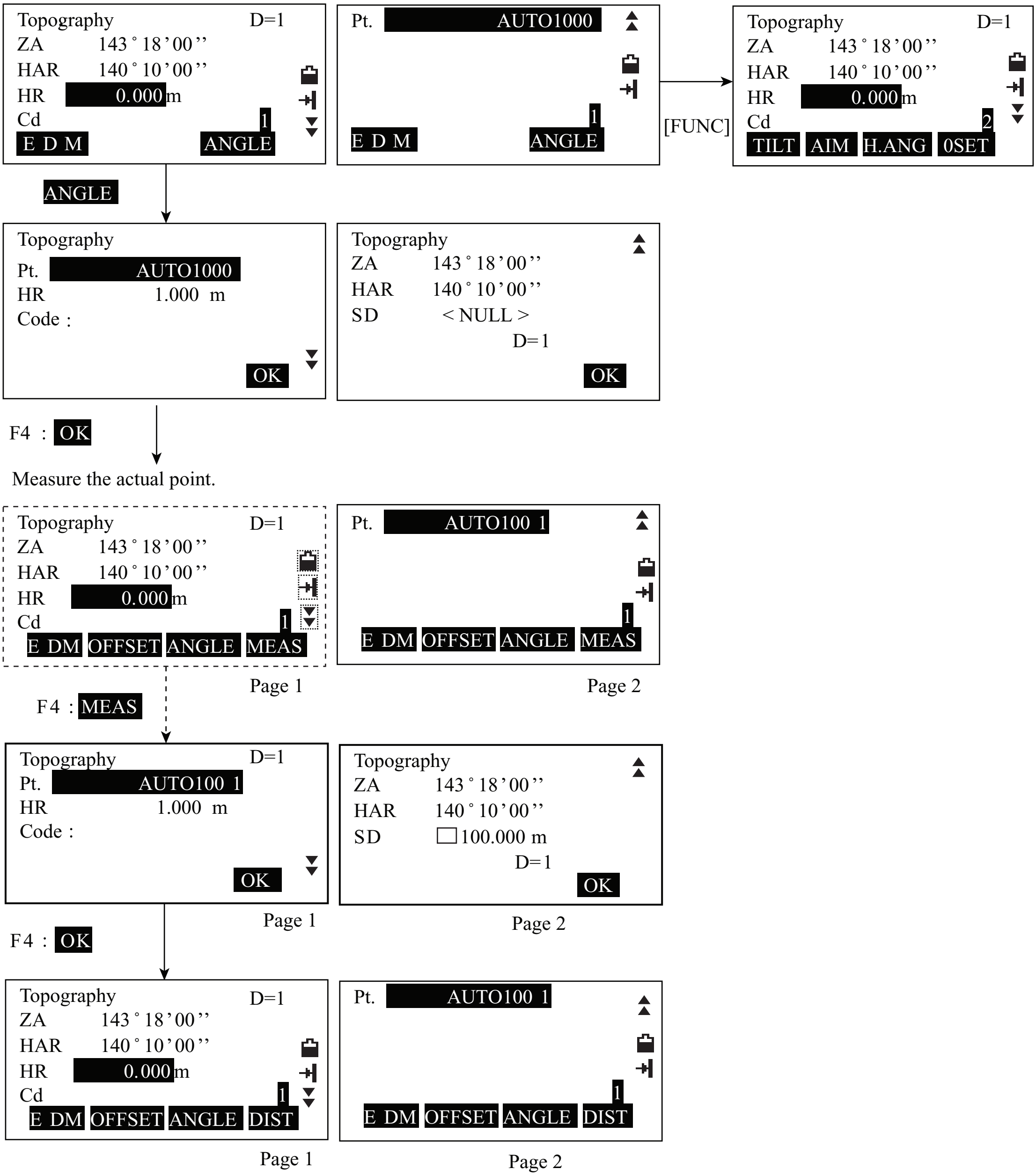
Enter the instrument station data.

Page 1

Page2

F 4 : OK

↓


Topography D=1
ZA 143°18'00"
HAR 140°10'00"
HR 0.000 m
Cd 1
E D M ANGLE
Pt. AUTO1000
1
E D M ANGLE
[FUNC]
Topography D=1
ZA 143°18'00"
HAR 140°10'00"
HR 0.000 m
Cd 2
TILT AIM H.ANG 0SET
ANGLE
Topography
Pt. AUTO1000
HR 1.000 m
Code :
OK
Topography
ZA 143°18'00"
HAR 140°10'00"
SD < NULL >
D=1
OK
F4 : OK
Measure the actual point.
Topography D=1
ZA 143°18'00"
HAR 140°10'00"
HR 0.000 m
Cd 1
E DM OFFSET ANGLE MEAS
Page 1
Pt. AUTO100 1
1
E DM OFFSET ANGLE MEAS
Page 2
F4 : MEAS
Topography D=1
Pt. AUTO100 1
HR 1.000 m
Code :
OK
Page 1
Topography
ZA 143°18'00"
HAR 140°10'00"
SD 100.000 m
D=1
OK
Page 2
F4 : OK
Topography D=1
ZA 143°18'00"
HAR 140°10'00"
HR 0.000 m
Cd 1
E DM OFFSET ANGLE DIST
Page 1
Pt. AUTO100 1
1
E DM OFFSET ANGLE DIST
Page 2

END

C.1.2. Setting-Out (Coordinates)

1. Press '★' key, and then 'Entry menu'. Select where you store coordinates.

```
Entry menu
  JOB select
  JOB detail
  Topography
  Coordinate
  S-O
```

```
JOB selection
  : JOB 1
Coord search JOB
  : JOB1

LIST
```

2. Press '★' key, and then 'Entry menu' and select 'S-O'.

```
Entry menu
  JOB select
  JOB detail
  Topography
  Coordinate
  S-O
```

3. Select 'Occ.Orien.' to set the instrument station date and azimuth angle of the backsight point.

```
S-O
  Occ.Orien.
  S-O data
  Observation
  EDM
```

4. First, input coordinates of station date directly, or when you wish to read in the registered coordinate date, press 'LOAD' and select Station Number. Second, press 'BS NEZ' to calculate azimuth angle from backsight coordinates. Finally, press 'OK' to set the input values.

```
N0 :          0.000
E0 :          0.000
Z0 :          <Null>
PT               T1
HI       0.000m
LOAD  BS AZ  BS NEZ  RESEC
```

```
Backsight
NBS:        10.000
EBS:         0.000
ZBS:        <Null>

LOAD            OK
```

5. Set backsight and press 'YES'. If you would like to check out distance from instrument station to there, press 'MEAS' to measure the actual distance and check it out. If it is okay press 'OK'.

```
Backsight
Take BS
ZA          79° 00' 00"
HA -R      143° 18' 00"
Azmth      143° 18' 00"
REC  MEAS  NO  YES
```

```
BS Hdist check

calc HD     10.000m
obs HD       0.500m
dHD         -9.500m
REC  HT         OK
```

```
S-O
 Occ.Orien.
 S-O data
 Observation
 EDM
```

6. Select 'S-O data'. <S-O Coord> is displayed. Enter the coordinates of the setting-out point and press 'OK'. When you wish to read in the registered coordinate date, press 'LOAD' and select Point Number. Press 'ENT' to set the data.

```
S-O Coord
N0:        10.000m
E0:        10.000m
Z0:        <Null>
HR          0.000m
LOAD  DISP      OK
```

7. The difference in the distance an angle calculated with the set instrument station and the target point is displayed. Rotate the top of the instrument until 'dHA' is 0° and place the target on the sight line.

```
S-O△ HD      -12.704m
dHA        45° 00' 00"
HD            1.439m
ZA         79° 00' 00"
HA-R       00° 00' 00"
   REC  DISP  ⇦ ⇨  MEAS
```

```
⇧ Back       -12.704m
 ⇨ R       45° 00' 00"
HD            1.439m
ZA         79° 00' 00"
HA-R       00° 00' 00"
   REC  DISP  ⇦ ⇨  MEAS
```

- By pressing "'⇐ ⇒'" an arrow pointing to the left or right displays which direction the target should be moved.

8. Move the prism forward and backward until the setting-out distance is 0 m.

```
⇧ Back       -12.704m
 ⇨ R       00° 00' 52"
HD            1.439m
ZA         78° 44' 35"
HA-R       44° 59' 08"
   REC  DISP  ⇦ ⇨  MEAS
```

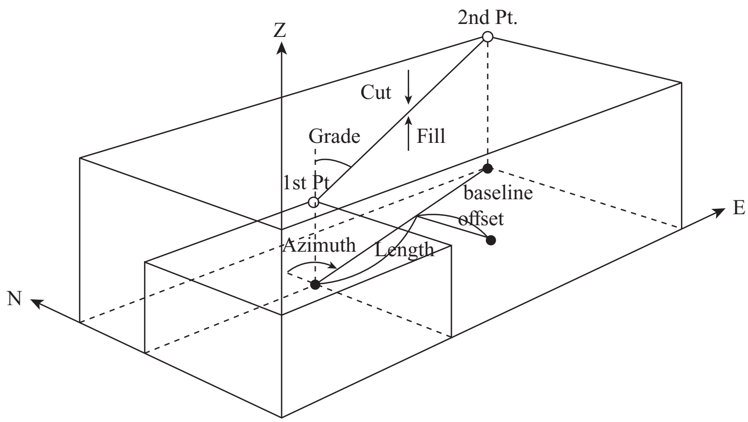
- When the target is within measurement range, all four arrows are displayed.

9. Press 'ESC' to return step 6.

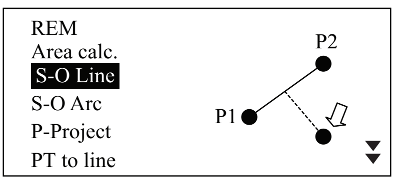
END

C.1.3. Setting-Out (Line)

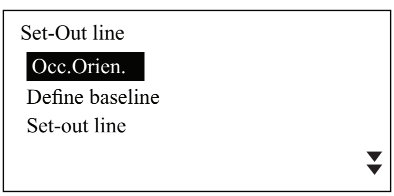
Setting-out line is used for setting out a required point at a designated distance from the baseline and for finding the distance from the baseline to a measured point.



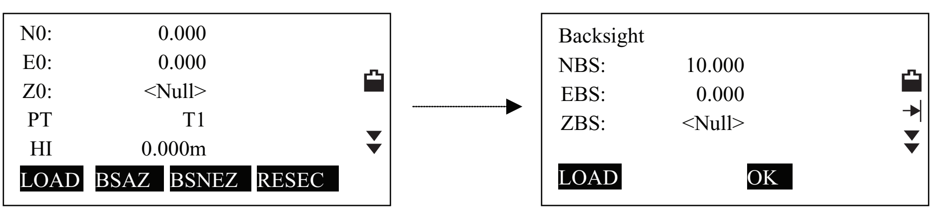
1. Press ‘MENU’ key, then select ‘S-O line’.



2. Select ‘Occ.Orien.’ to set the instrument station date



3. First, input coordinates of station date directly, or when you wish to read in the registered coordinate date, press ‘LOAD’ and select Station Number. Second, press ‘BS NEZ’ to calculate azimuth angle from backsight coordinates. Finally, press ‘OK’ to set the input values.



4. Set backsight and press 'YES'. If you would like to check out distance from instrument station to there, press 'MEAS' to measure the actual distance and check it out. If it is okay press 'OK'.

Backsight
Take BS
ZA 79°00'00"
HA-R 143°18'00"
Azmth 143°18'00"
REC MEAS NO YES

→

BS Hdist check
calc HD 10.000m
obs HD 0.500m
dHD -9.500m
REC HT OK

→

Set-Out line
Occ.Orien.
Define baseline
Set-out line

5. Select 'Define baseline' in Set-out line, and enter the first point data and press 'OK'. If you wish to read in the registered coordinate date, press 'LOAD' and select Station Number.

Define 1st PT
Np: 10.000
Ep: 10.000
Zp: <Null>
LOAD REC MEAS OK

6. Enter the second point data and press 'OK'. If you wish to read in the registered coordinate date, press 'LOAD' and select Station Number.

Press 'FUNC' key. 'OBS' is displayed. >>When not observing the first and second points, this is not necessary.

Press 'OBS' and Sight the first point, and then press 'MEAS'.

Define 2nd PT
Np: 10.000m
Ep: 15.000m
Zp: <Null>
P1
LOAD REC MEAS OK

→

Define 2nd PT
Np: 10.000m
Ep: 15.000m
Zp: <Null>
P2
OBS

→

Measure 1st PT
Np: 10.000m
Ep: 10.000m
Zp: <Null>
MEAS

7. Press 'YES' to use the measurement results of the first point.

Measure 1st PT
SD: 9.536m
ZA 74°00'15"
HA-R 14°10'05"
HR: 0.000m
NO YES

*You can input target height here.

*Press "NO" to observe the first point again.

8. Sight the second point and press 'MEAS'. Press 'YES' to use the measurement result.

Measure 2nd PT
Np: 10.000m
Ep: 15.000m
Zp: <Null>
MEAS

→

Measure 2nd PT
SD: 16.440m
ZA 85°03'03"
HA-R 37°44'33"
HR: 0.000m
NO YES

*You can input target height here.

*Press "NO" to observe the first point again.

9. The distance between the two measured points, the distance calculated from inputting the coordinates of two points and scale factors are displayed.

Azumth	90°00'00"
Hcalc	5.000m
Hmeas	14.971m
Scale X	1.000091
Scale Y	1.000091

Sy = 1 | Sy = Sx | OK

*[Sy=1]: Set scale factory to "1".

*[1: **]: Change the grade display mode to "1:** = elevation: horizontal distance"

Grade % -2.669

1: ** | % | OK

10. Press 'YES' on the screen of step 9 to define the baseline.

Set-Out line
Point
Line

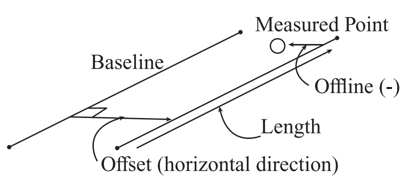
Move to setting-line measurement

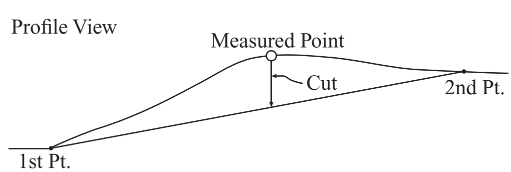
A) Set-out line point
B) Set-out line line

11. Select 'Line' in <Set-out line> and press 'ENT'.

Setting-out line measurement how far horizontally the measured point is from the baseline and how far vertically the measured point is from the connected line. The baseline can be offset in a horizontal direction if necessary.

- Before performing setting-out line, the baseline must be defined.





Set-Out line
Point
Line

12. Enter the offset value and press ‘ENT’.
    Sight the target and press ‘MEAS’ on the screen.
    The measurement result is displayed and press ‘STOP’ to stop the measurement.

```
Set-Out line

Offset   1.000 m

                              MEAS
```

→

```
Set-Out line
SD:              1.33m
ZA           85°03’03”
HA-R         37°44’33”
HR:             1.200m
                      NO  YES
```

Offset (+): On the right side of the baseline.

Offset (-): On the left side of the baseline.

13. Enter the target height, and then press ‘YES’ to use the measurement results.

```
Set-Out line
SD:              1.33m
ZA           85°03’03”
HA-R         37°44’33”
HR:             1.200m
                      NO  YES
```

*Press “NO” to observe the target again.

14. Display the difference between the measured point and the baseline.
    If you would like to record the result, press ‘REC’.

```
Set-Out line
Offline:       -0.004m
Cut:            0.006m
Length:        12.917m

REC             MEAS
```

*Offline: (+) The point is on the right side of the baseline.

(−) The point is on the left side of the baseline.

*Cut: The point is below the baseline.

*Fill: The point is above the baseline.

*Length: Distance along the baseline from the first point to the measured point.

15. Sight the next target and press ‘MEAS’ to continue the measurement.

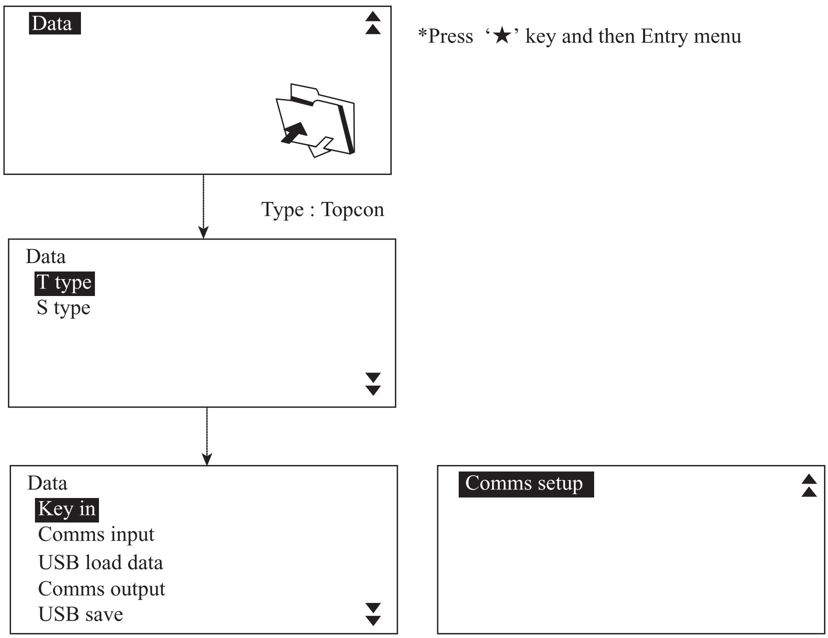
```
Set-Out line
Offline:       -0.004m
Cut:            0.006m
Length:        12.917m

REC             MEAS
```

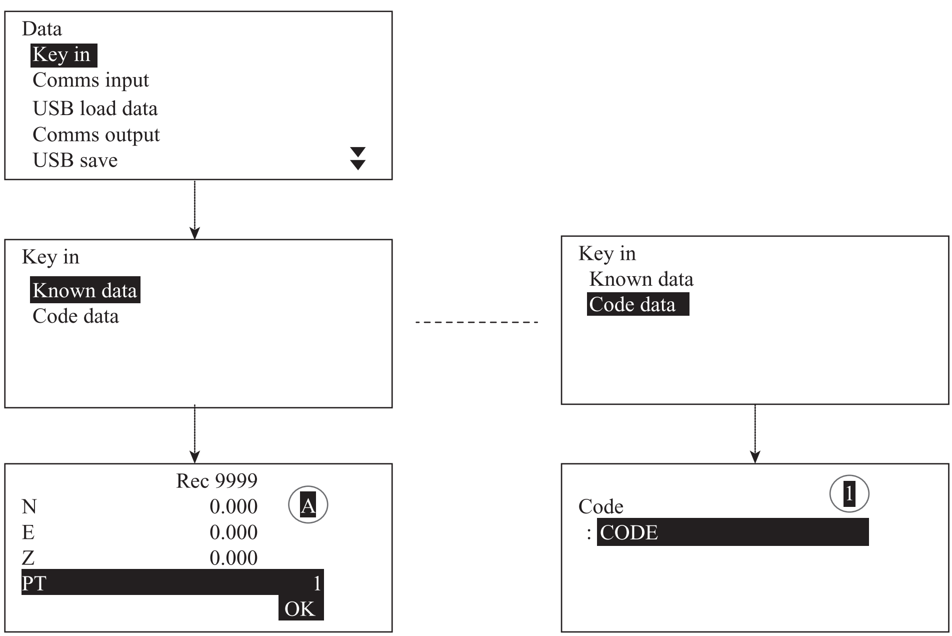
END

C.2. STAR KEY MOD

C.2.1 Data



Key in



Input can be modified by pressing "SHIFT". *Note: A, captial letter; a, small letter; 1, numerals

C.2.2. USB Save

Data
T type
S type

Press ‘USB’ on the status screen

Type: Topcon

USB
Save data
Load known PT
Save code
Load code
File status

Quick format

JOB1	12
JOB3	22
JOB2	32
JOB4	42
JOB5	52
	OK

In the list of JOBs, select the JOB to be recorded and press ‘ENT’.

JOB1	OUT
JOB2	22
JOB3	32
JOB4	42
JOB5	52
	OK

‘OUT’ is displayed to be the right of the selected JOB.

Save date
GTS (Obs)
GTS (Coord)
SSS (Obs)
SSS (Coord)

Save date
Obs date
Reduced date

Select output format.

JOB01. raw
Date : Jan/01/2012
Time : 08:00
Format : GTS (Obs)
3.7GB / 3.7GB OK

Enter the file name. Press ‘ENT’ to set the date.

Press ‘OK’ to save the JOB to the external memory media.

After saving a JOB, the screen returns to the JOB list.

*To cancel or change the JOB file in the middle of workflow, return to status screen.

C.2.3. USB Load Data

Data
Key in
Comms input
USB load data
Comms output
USB save

Select 'USB load data' and press 'ENT'.

Type: Topcon

USB load data
known data
Code data

Select 'known date' and press 'ENT'.

Load known pt.

Job. JOB1

OK

Press 'OK'.

If you would like to change the JOB which date is loaded, return 'JOB select' in the Entry menu.

Load known pt.
GTS(Coord)
SSS(Coord)

Select the format and press 'ENT'.

TEST

Select the actual file name and press 'ENT'.

TEST
272byte
Jan/01/2011/00:41
Format: GTS (xxx)
Confirm?
NO OK

Confirm the content and press 'YES.'

TOPCON TOTAL STATION GTS/100N, CYGNUS, SET0N SERIES

Appendix D

Field procedure for executing topographic survey, stakeout side shot and data comm

1. TOPO
2. Stake Out
3. Side Shot
4. Data Communication

1. Topo

1.1 Press the Menu key	[MENU]	MENU 1/3 F1:DATA COLLECT F2:LAYOUT F3:MEMORY MGR. P↓
1.2 Press the F1	[F1]	
1.3 Press Input to make a new file or press List to use the memoried file		SELECT A FILE FN:__________ INPUT LIST --- ENTER
1.4 Press OCC.PT F1 from the data collect menu	[F1]	DATA COLLECT 1/2 F1:OCC.PT# INPUT F2:BACKSIGHT F3:FS/SS P↓
1.5 Press F4 key	[F4]	PT# →PT-01 ID : INS.HT: 0.000 m INPUT SRCH REC OCNEZ

1.6 Decide the OCC point from the list or input the new point and enter the NEZ coordinates

```
OCC.PT
 PT#:PT-01

INPUT LIST NEZ ENTER
```

1.7 Press enter

```
OCC.PT
 PT#=PT-01

[ALP][SPC][CLR][ENT]
```

1.8 Enter ID, INS.HT and press F3 to REC then F3 to Yes [F3]

```
PT#    :PT-11
ID     :
INS.HT→    1.335 m
INPUT SRCH REC OCNEZ
>REC ?       [YES][NO]
```

1.9 Press F2 to get the Back sight [F2]

```
DATA COLLECT    1/2
 F1:OCC.PT# INPUT
 F2:BACKSIGHT
 F3:FS/SS         P↓
```

1.10 Press F1 to input the point, PCODE and R.HT [F1]

```
BACKSIGHT
 PT#:

INPUT LIST NE/AZ ENT
```

1.11 Decide the Back sight point [F1]
Collimate Back sight point [F2]
Press F2 0SET
Then press F3 MEAS

```
BS#    →PT-22
PCODE  :
R.HT   :    0.000 m
INPUT 0SET MEAS BS
```

1.12 Press F2 SD [F2]

```
BS#    →PT-22
PCODE  :
R.HT   :    0.000 m
*VH    SD   NEZ  ---
```

```
 V :    90°00'00"
 HR:     0°00'00"
 SD*[n]      <<< m
> Measuring...
```

1.13 Press F3 to TOPO [F3]

```
DATA COLLECT      1/2
 F1:OCC.PT# INPUT
 F2:BACKSIGHT
 F3:FS/SS            P↓
```

1.14 Decide the Point Name, PCODE and R.HT Then press F3 MEAS	[PT] [PCODE] [R.HT] [F3]	PT# →PT-01 PCODE :SOKKIA R.HT : 1.200 m INPUT SRCH MEAS ALL VH *SD NEZ OFSET
1.15 Collimate the target point Press the Type of measuring VH, SD, NEZ (Example: SD) Star mark next to SD means the status of currently using	[Collimate] [F2]	PT# →PT-01 PCODE :SOKKIA R.HT : 1.200 m INPUT SRCH MEAS ALL VH *SD NEZ OFSET
1.16 Measuring status		V : 90°10'20" HR: 120°30'40" SD*[n] < m > Measuring... < complete >
1.17 The measuring data is memorized and the display changes to the next point. Point is automatically incremented.		PT# →PT-02 PCODE :SOKKIA R.HT : 1.200 m INPUT SRCH MEAS ALL
1.18 Continue measuring the same way. (Press F3 MEAS and select the type of measuring VH, SD, NEZ) To finish the mode, press ESC key	[ESC]	PT# →PT-03 PCODE :SOKKIA R.HT : 1.200 m INPUT SRCH MEAS ALL

2. Stake Out

2.1 Press Menu Press F2	[MENU] [F2]	MENU 1/3 F1:DATA COLLECT F2:LAYOUT F3:MEMORY MGR. P↓
2.2 Select a file Same steps 1.4–1.12		SELECT A FILE FN:________ INPUT LIST SKP ENTER
2.3 After the procedure of getting the backsight press F3 to start Stake out	[F3]	LAYOUT 1/2 F1:OCC.PT INPUT F2:BACKSIGHT F3:LAYOUT P↓

2.4 Decide the place to stake out Choose the point from the list or input the new point	[F1] or [F2]	LAYOUT PT#:________ INPUT LIST NEZ ENTER
2.5 Enter the R.HT Press F4	[F4]	REFLECTOR HEIGHT INPUT R.HT : 1.500 m --- --- [CLR] [ENT]
2.6 The calculation of the point Press F1 to set the angle		CALCULATED HR= 90°10'20" HD= 123.456 m ANGLE DIST --- ---
2.7 Collimate the prism and press [F1] or [F3]	[F1] or [F3]	PT#: LP-100 HR: 6°20'40" dHR: 23°40'20" DIST --- NEZ ---
2.8 HD: Actual horizontal distance dHD: Horizontal distance to be turned to the layout point		HD* 143.84 m dHD: -13.34 m dZ: -0.05 m MODE ANGLE NEZ NEXT
It means Actual horizontal distance minus Calculated horizontal distance. dZ : Vertical distance to be turned to the layout point It means Actual vertical distance minus Calculated vertical distance		

3. Side Shot

(Set up the instrument at a known point, and measure the coordinate of the new po

3.1 Side-shot method Press MENU and press F2 LAYOUT Press F4 to get to Page 2/2	MENU] [F2] [F4]	LAYOUT 2/2 F1:SELECT A FILE F2:NEW POINT F3:GRID FACTOR P↓
3.2 Press F2	[F2]	LAYOUT 2/2 F1:SELECT A FILE F2:NEW POINT F3:GRID FACTOR P↓

3.3 Press F1 Side shot

```
NEW POINT
 F1:SIDE SHOT
 F2:RESECTION
```

3.4 Select a file
then Enter

```
SELECT A FILE
 FN:__________

INPUT LIST ----ENTER
```

3.5 Enter the new point name

```
SIDE SHOT
 PT#:________

INPUT SRCH --- ENTER
```

3.6 Enter R.HT
then press Enter

```
REFLECTOR HEIGHT
INPUT
R.HT  =   0.000 m
 ---  ---  [CLR][ENT]
```

3.7 Collimate the new point and prism
Distance measuring starts

```
 HR:    123°40'20"
 HD*[n]        < m
 VD:             m
> Measuring...
    < complete >
```

3.8 Press F3 to store into COORD DATA. [F3]

```
 N :  1234.567 m
 E :   123.456 m
 Z :     1.234 m
>REC ?      [YES][NO]
```

3.9 The input menu for next new point is displayed.
Point name is automatically incremented.

```
SIDE SHOT
 PT#:NP-101

INPUT SRCH --- ENTER
```

4. Data Communication

4.1 Sending data

4.1 Press MENU key [F3]
then press F3 [F3]

```
MENU                1/3
 F1:DATA COLLECT
 F2:LAYOUT
 F3:MEMORY MGR.     P↓
```

4.2 Press F4 [F4]

```
MEMORY MGR.         1/3
 F1:FILE STATUS
 F2:SEARCH
 F3:FILE MAINTAN    P↓
```

4.3 Press F1

[F1]

```
MEMORY MGR.        3/3
 F1:DATA TRANSFER
 F2:INITIALIZE
                    P↓
```

4.4 Select data format
GTS format: the conventional data
SSS format: Including PCODE data
This case we select F1

[F1]

```
DATA TRANSFER
 F1:GTS FORMAT
 F2:SSS FORMAT
```

4.5 Select the type of data
This case we select F1

[F1]

```
DATA TRANSFER
 F1:SEND DATA
 F2:LOAD DATA
 F3:COMM. PARAMETERS
```

4.6 Enter file name you want to send

```
SELECT A FILE
 FN:_________

INPUT LIST --- ENTER
```

4.7 Press F3 and send the data

[F3]

```
SEND MEAS. DATA

 >OK ?
 ---  ---  [YES][NO]
```

4.8 See the data that is sent

```
SEND MEAS. DATA

 < Sending Data!>
               STOP
```

4.2 Loading Data

4.2.1 Same steps 4.1–4.4
then press F2 to load data

[F2]

```
DATA TRANSFER
 F1:SEND DATA
 F2:LOAD DATA
 F3:COMM. PARAMETERS
```

4.2.2 Select the type of data this case we select F1	[F1]	LOAD DATA F1:COORD. DATA F2:PCODE DATA
4.2.3 Press the F1 and enter new file name you want to receive	[F1]	COORD. FILE NAME FN:__________ INPUT --- --- ENTER
4.2.4 Press F3 and Loading starts	[F3]	LOAD COORD. DATA >OK ? --- --- [YES][NO]
4.2.5 See the data that is received		LOAD COORD. DATA < Loading Data!> STOP

INDEX

Q

R

This page is intentionally left blank