LOST WITHOUT IT

How GPS is more than just navigation

Chris McAlister

University of Southern Queensland
Toowoomba

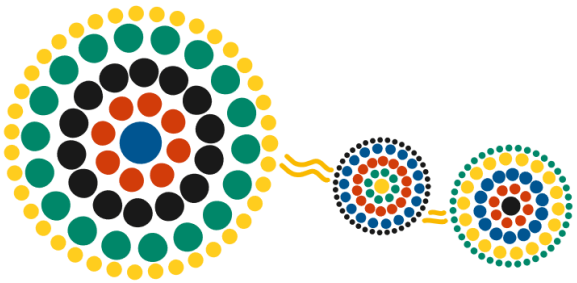
CONTENTS

ACKNOWLEDGEMENT OF FIRST PEOPLES

The University of Southern Queensland acknowledges the traditional custodians of the lands and waterways where the University is located. Further, we acknowledge the cultural diversity of Aboriginal and Torres Strait Islander peoples and pay respect to Elders past, present and future.

We celebrate the continuous living cultures of First Nations Australians and acknowledge the important contributions Aboriginal and Torres Strait Islander people have and continue to make in Australian society.

The University respects and acknowledges our Aboriginal and Torres Strait Islander students, staff, Elders and visitors who come from many nations.



ACCESSIBILITY INFORMATION

We believe that education should be available to everyone, which means supporting the creation of free, open and accessible educational resources. We are actively committed to increasing the accessibility and usability of the textbooks and resources we produce.

Accessibility features of the web version of this resource

The web version of this resource has been designed with accessibility in mind and incorporates the following features:

- It has been optimised for people who use screen-reader technology
    - all content can be navigated using a keyboard
    - links, headings and tables are formatted to work with screen readers
    - images have alt tags
- Information is not conveyed by colour alone.

Other file formats available

In addition to the web version, this book is available in a number of file formats, including PDF, EPUB (for eReaders), and various editable files. Look for the "Download this book" drop-down menu on the landing page to select the file type you want.

Third-Party Content

In some cases, our open texts include third-party content. In these cases, it may not be possible to ensure the accessibility of this content.

Accessibility Assessment

Below is a short accessibility assessment of key areas that have been assessed during the production process of this open text. The checklist has been drawn from the BCcampus Open Education Accessibility Toolkit. While a checklist such as this is just one part of a holistic approach to accessibility, it is one way to begin our work on embedding good accessibility practices in the books we support.

We hope that by being transparent on our current books, we can begin the process of making sure accessibility is top of mind for all authors, adopters, students and contributors of all kinds on all our open-text projects. As such, we welcome any feedback from students, instructors or others who encounter the book and identify an issue that needs resolving.

Accessibility Checklist

Category	Item	Status
Organising Content	Content is organised under headings and subheadings	Yes
Organising Content	Headings and subheadings are used sequentially (e.g. Heading 1, Heading 2, etc.)	Yes
Images	Images that convey information include Alternative Text (alt-text) descriptions of the image's content or function	Yes
Images	Graphs, charts, and maps also include contextual or supporting details in the text surrounding the image	Yes
Images	Images, diagrams, or charts do not rely only on colour to convey important information	Yes
Images	Purely decorative images contain empty alternative text descriptions. (Descriptive text is unnecessary if the image doesn't convey contextual content information)	Yes
Tables	Tables include column headers and row headers where appropriate	Yes
Tables	Tables include a title or caption	Yes
Tables	Tables do not have merged or split cells	Yes
Tables	Tables have adequate cell padding	Yes
Weblinks	The web link is meaningful in context, and does not use generic text such as "click here" or "read more"	Yes
Weblinks	External web links open in a new tab. Internal web links do not open in a new tab.	Yes
Weblinks	If a link will open or download a file (like a PDF or Excel file), a textual reference is included in the link information (e.g. '[PDF]')	Yes
Embedded Multimedia	A transcript has been made available for a multimedia resource that includes audio narration or instruction	No but videos have closed captions
Embedded Multimedia	Captions of all speech content and relevant non-speech content are included in the multimedia resource that includes audio synchronised with a video presentation	Yes
Embedded Multimedia	Audio descriptions of contextual visuals (graphs, charts, etc.) are included in the multimedia resource	No
Formulas	Formulas have been created using MathML	Have been created with LaText
Formulas	Formulas are images with alternative text descriptions if MathML is not an option	No
Font Size	Font size is 12 points or higher for body text	Yes
Font Size	Font size is 9 points for footnotes or endnotes	–
Font Size	Font size can be zoomed to 200%	Yes

Accessibility improvements

While we strive to ensure this resource is as accessible and usable as possible, we might not always get it right. We are always looking for ways to make our resources more accessible. If you have

problems accessing this resource, please contact the UniSQ Open Educational Practices team to let us know so we can fix the issue.

ACKNOWLEDGEMENTS

Although this book has my name on it, the foundations of it were built by many a great surveying lecturer before me, and it has been shaped by the ones I have the pleasure of working with still. To you all, thank you for answering the call of education and for everything you have done for me, each other, and our students.

As with any book, there's a cast of supporting actors who have made it possible. This is my attempt to remember them all. As you might know though I'm really bad with names, so if I forgot you, you probably need to send me more memes so I remember you for the next one!

First up, my never-ending thanks to the Bathurst crew – Bear, Swampy, Tony, Simon, Dicko, Krazy, Jonesy, Les, Keith, George, Doug, and still to this day one of the greatest geodesists I've known, Case. You guys shaped my future in a way I'll never be able to fully appreciate, and I hope I'm doing justice to all the amazing things you took the time to teach me.

To the people that showed me that being a lecturer didn't have to be boring – Craig, Bruce, Bill, Chris, Michael, Ed & Pat – if I have even half the impact on my students that you all had on me it'll have been worth it. Thank you for your energy and expertise. And for putting up with me a student. I'm really sorry!!!

To Matt – I'm not sure you'll ever understand how much of a Rockstar you are to me, but that you are as excited about showing me photos of your grandkids as you are about talking about GNSS shows just how much of a legend you truly are.

To my family – P, Fletch, Mum & Dad. I could never have imagined the chaos of our neurospicy life, but now I wouldn't have it any other way. There certainly wouldn't be a GPS textbook with a zombie on the cover in the world without your influence. You make me a better lecturer and a better person. Never stop being weird.

Darren. Words can't cover it. Insert all the things here.

The rest of my tribe who ensure things are never boring – Kristy, Catherine, Amber, Peta, Amy, Peter, Jem, Shayne, Hanka and my work family (you know who you are!) – you all rock. Thanks for putting up with me.

A big shout out to Emma and Nikki, and Samara for the amazing cover art – I still can't believe this is real! You're all incredible and I hope having a zombie textbook in the suite continues to make you laugh.

And finally, to my students – past and present. This is really all for you. Thank you for being committed, patient, inquisitive, collaborative, engaged and just plain funny. You're the reason I do this. Continue being awesome.

Cheers,

Chris

P.S. Gaz – I'm never getting your card working. Ever. Seriously. There is no escape.

ABOUT THE AUTHOR

When I was in my fourth year of university, I applied for a 12 month undergraduate position at what was then the Department of Lands, out in Bathurst NSW. That's the job that set me on the path of GPS, and ultimately is responsible for the fact that zombies, dinosaurs, Lego and lactose intolerant volcano gods are all a part of my work life, as well as in the life of hundreds of surveying students and graduates. Who all likely think I'm a touch mad.

Since Bathurst all those years ago, a series of events that have included, but are not limited to: surveying rural Queensland with a broken arm, monitoring dune erosion while 8 months pregnant, doing the survey from hell that was submitted as the first Qld ePlan, and avoiding being a cadastral surveyor; have led to the point where I've written a GPS textbook with a Zombie on the cover.

This probably all makes much more sense if you know that I'm a Professional Fellow in Surveying at UniSQ, which is a fancy way of saying I'm a lecturer that came from industry. UniSQ has been letting me loose on classes since 2018, which has resulted in me having the absolute privilege of delivering courses in all the things I love to people who seem to want to listen.

This book is born of my want to make learning fun, without compromising on the content. I am lucky to have inherited my approach to writing from my Dad, who had a never ending capacity to make even the most boring thing fun, and ensured I had a sense of humour that is most definitely reflected in this book. Add in a dash of the neurospicy variety, and you end up with something that resembles me: Chris, the slightly mad lecturer who has classes that are never ordinary, encourages memes as a legitimate response to assignment questions, and will never turn down an opportunity to have Lego in my class.

I hope you enjoy this book, and if you have any comments, suggestions or memes you want to throw my way, please get in touch.

About the Cover

Also, if you've made it this far, you really deserve an explanation about the Zombie cover. This book was originally a PDF for a course I teach called Introduction to GPS. For that course, I needed to come up with assignments that students who are all over the world (and sometimes even on ships or in Antarctica), with limited access to equipment fancier than their phone, could all do. And it had to be more interesting than measuring a handful of non descript marks in a park somewhere. Challenge accepted!

Around 2am one morning while I was writing the first version of what would become this book, the random idea of making the assignment about surviving a zombie apocalypse appeared. Students could use whatever app they liked to collect GPS points at locations that would provide them with water, food, security and exercise, and they could record short videos discussing what errors were influencing the GPS quality. They would also have to prepare a 'Survival Manual' – in case they got turned into zombies, those left behind would know how GPS worked. Once the students had completed the course, I'd send them a little paracord keyring to add to their zombie survival kit. It has become a bit infamous – I've heard students ask each other if they'd done the zombie course yet. The cover of this book is a testament to learning never being boring.

FOREWORD

Matt Higgins

The first edition of *Lost Without It*, this important text on satellite positioning, is being published in 2023, which also marks the 50th anniversary of the approval of the Global Positioning System (GPS) program by the U.S. Department of Defense. I have the great honour of serving on the U.S. Advisory Board for Space-Based Positioning, Navigation and Timing (PNT), with Dr. Bradford Parkinson, who was the Chief Architect for GPS. In the 50 years since Brad and his team designed and built that amazing system, GPS has become an indispensable part of modern life. So much so, that there are now 3 other Global Navigation Satellite Systems (GNSS), like GPS and 2 regional systems, along with various augmentation systems around the world.

While the impact of GNSS is truly pervasive, there are few professionals whose daily work has been influenced more than those in Surveying and the Spatial Sciences. GNSS has revolutionised both the efficiency and effectiveness of so many processes in the measurement, capture and management of spatial data.

I have known the author, Chris McAlister, for many years, having had the pleasure of working closely in Queensland Government with her partner Darren, another formidable member of the Surveying Profession. I was, therefore, very pleased to accept Chris' invitation to write this foreword for *Lost Without It*. Plus, I worked out years ago, that it is easier to say yes to Chris than summon the courage and energy to try to do the opposite.

Chris has drawn on her extensive experience in both the field and the lecture room to develop this excellent text. While GPS has now been around for 50 years, the field of Positioning, Navigation and Timing is hardly showing its age and is characterized by ongoing and extensive change. It is therefore great to see that this text has been published on a digital platform that enables ongoing refinement and improvement to adapt and reflect the myriad of changes that we are likely to continue to see in the next 50 years. The digital platform also allows the reader to easily access the many informative GNSS/PNT resources available on the Internet, now and into the future.

I hope you enjoy this wonderful new resource and soon come to the realisation, like the title says, that you would be *Lost Without It*.

Matt Higgins PSM
President IGNSS Association
Member of the U.S. Space-based Positioning, Navigation, and Timing Advisory Board
Honorary Member of the International Federation of Surveyors
Honorary Fellow of the Geospatial Council of Australia

INTRODUCTION

The overall objective of this book is for you to be able to explain **Global Navigation Satellite System (GNSS)** principles, uses and carry out GNSS activities. More specifically, you will be able to:

1. discuss the features and applications of GNSS and its importance in society today
2. define coordinates systems likely to be encountered by GNSS users and calculate and discuss GNSS coordinates
3. describe global satellite navigation systems, satellite orbital characteristics, and satellite signal structure
4. define the fundamental characteristics of GPS and other GNSSs and outline their development
5. discuss the principles of GNSS observations, make observations using a GNSS receiver, and calculate and analyse findings
6. explain GNSS observations techniques, and calculate and evaluate levels of accuracy associated with GNSS observations
7. demonstrate an understanding of error sources in GNSS observations, and explain the uses and critical factors of Differential GNSS techniques
8. identify and discuss project planning features when using GNSS, and discuss the key steps in planning a GNSS data collection project for asset mapping
9. explain GNSS data collection and processing procedures, including Differential GNSS, and evaluate collected and processed data
10. describe the use of GNSS for asset mapping, and other common uses.

PART I

FUNDAMENTALS

1.1 FUNDAMENTALS

Learning Objectives

After successfully completing this chapter you should be able to:

- Describe how GPS improved on earlier navigation systems
- Identify some general uses of GNSS
- Explain the main segments of GNSS
- Explain the hierarchy of satellite systems

In the beginning...

Throughout human history, we have sought simple ways of determining where we are, and where we're going on the Earth. Positioning, knowing where we are, and navigation, knowing where we're going and how to get there, have always been some of the most basic problems facing civilisations.

Developments in science and technology have provided us with a multitude of solutions in the last few centuries, although none have been as revolutionary and globally adopted as the United States' Global Positioning System (GPS). Utilising a combination of ground control and satellites to provide a location to a device capable of receiving a GPS signal, GPS set the stage for military and civilian use planet wide. No longer were we restricted by weather, daylight or location in being able to determine where we were!

The model used for GPS was quickly replicated by other countries and organisations, and now we have multiple systems providing us with positioning and navigation information. These are generically known as **Global Navigation Satellite System** or **GNSS** for short.

Understanding how these systems were developed, the fundamentals of how they work, the different applications and accuracies achievable using different methods of GNSS observation are important for any professional that provides positioning or navigation information as part of their work.

Almost everyone refers to all of the Global Navigation Satellite Systems collectively as "**GPS**". We even do it in the name of this subject!
This is what is known as a **Proprietary eponym** or **Generic trademark**
– when the name for a successful product becomes the name that a class of objects is known by.

Radio waves

A Scottish physicist named James Clerk Maxwell developed a unified theory of electromagnetism in the 1870s, and in it he predicted the existence of radio waves. The following decade, a German physicist called Heinrich Hertz proved that Maxwell was correct, and communication was changed forever.

Radio waves also became useful in measuring distance. The concept was to measure the time it took for a radio signal to travel from a transmitter to a receiver. This signal travel time was then multiplied by the speed of light (about 300 000 000 metres per second) to obtain a distance between the transmitter and receiver.

If you knew where the transmitter was (and when it sent the signal) then you could position yourself on a circle with radius equal to the measured distance and centred on the transmitter.

This method is called **trilateration** and is shown in **Figure 1.1(a).**

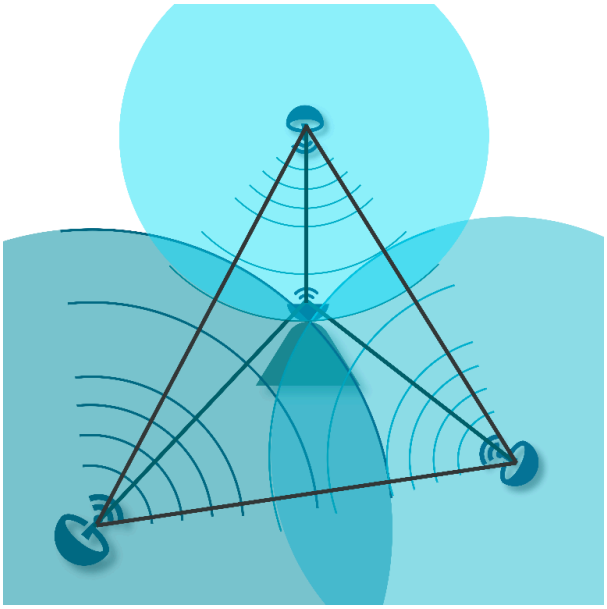


Figure 1.1(a): Trilateration using radio waves

LORAN

The LORAN (long range navigation) positioning system was developed and used during World War II, until the 1980s.

LORAN was based on triangulation from several radio transmitters on the coastline. Provided the user was within range of more than one transmitter, they could use the known position of the transmitters and the distance and angle between them to calculate position.

Unfortunately, this system was only available in some coastal areas and its accuracy varied, depending on electrical interference, atmospheric conditions, distance from transmitters and geographic variations. The LORAN system provided 2-dimensional (horizontal) positioning to an accuracy of about 300 metres.

Transit

Transit was a navigation and positioning system that was based on satellites. It was called the Transit System or 'Sat-Nav'.

The satellite-based positioning was centred around the Doppler shift of the satellite signal. The signal changed as the satellite rose in the sky and then set. The effect was much like a car horn changing frequency as it moves towards, and then away from, a listener. The Transit system used this Doppler Effect (or change in frequency) to calculate position.

It was possible for surveyors on land to get a 3-dimension position accurate to less than a metre by post-processing about 2–3 days of data.

GPS

GPS provides an answer to one of mankind's most troublesome questions: 'Where on earth am I?'

Finally, at the beginning of the 1970s, the satellite age began in earnest when the United States Department of Defence gained Congress approval to develop a reliable and accurate global navigation system.

To overcome some of the limitations of previous systems, radio transmitters were placed in

satellites that orbited the earth. The system was called the Global Positioning System (GPS), and it can provide 3-dimensional position fixes anywhere in the world. The system relies on using satellites as reference points.

Position accuracy can vary from tens of metres to a few millimetres depending on the measuring techniques, field procedures and quality of receivers used. We will discuss the different kinds of methods available to achieve different accuracies in the second part of this subject.

GPS not only provides an accurate and reliable position fix, it also provides very accurate time, and this has become critical to the banking sector.

GNSS

Once GPS became a globally useful technology, other countries and regions began investing in their own Global Satellite Navigation Systems. We will cover this in the next chapter.

1.2 WHAT IS A GNSS?

People have probably always wanted (and needed) a simple way of determining exactly where on the earth they are, and in which direction they are travelling. Positioning and navigation are such basic problems for civilisation that they have commanded a lot of effort over the centuries.

Even today the ability to know where you are and where you are heading is crucial to many activities. As we saw in the previous section, over the years many technologies have tried to provide us with this information, but only GNSS has changed navigation and positioning so dramatically.

GNSS signals are freely available to everyone and are available 24 hours per day, seven days per week, anywhere on earth and in all weather. All that is required is some form of technology with a GNSS chip, called a receiver.

GNSS receivers use GNSS satellites as space-based reference stations to calculate positions on earth. Because the satellites are in high orbits, they avoid most of the problems encountered by many of the earlier land-based navigation systems. And because GPS and several other subsequent systems were primarily designed for defence, significant effort has been made (and is continuing) to ensure they are reliable, robust and not easily susceptible to interference.

The elements of GNSS

Note: The three segments that make up a GNSS are derived from the segments names used in GPS, so there may be slight variations in the way things are named in different systems.

GNSS consists of three main segments:

1. GNSS are worldwide navigation systems based on a constellation of satellites orbiting the earth. This is called the **Space segment.**
2. The space segment works in conjunction with an Earth based network of ground stations, called the **Control segment.** It is also sometimes referred to as the **Ground segment.**
3. The last segment is called the **User segment** and is made up of anyone with a GNSS receiver.

The space segment

The space segment of a GNSS is made up of what is called a **constellation** of satellites. A constellation is a group of satellites that belong to a particular GNSS. They usually consist of a minimum of 24 operational satellites, often with spares in case one gets damaged or has operational issues.

Satellite is the shorthand name of the full **Satellite Vehicles**. This is often shortened to **SVs**.

Each satellite has several incredibly accurate atomic clocks, information about where it is meant to be at any given time, how 'healthy' it is, and information about where other satellites in its constellation are.

The information a satellite has about its own predicted orbit is referred to as an **ephemeris,** and it can best be thought of as a personal calendar.

The information that each satellite has about other satellites is called an **almanac**, and this is more like a shared or group calendar.

The GPS Navstar (Navigation Satellite Timing **a**nd Ranging) satellites have about a seven-year life span, so the constellation is constantly being replenished. These satellites are in orbits that are approximately 20 200 km above the earth and take about 12 hours to orbit the earth. From this we can see that the satellites travel at about 13 000 km/hr and are typically above the horizon for about 5 hours.

The GPS constellation is arranged into six orbital planes, each inclined at 55 degrees to the equator, and spaced evenly in longitude around the equator. This configuration is designed to ensure that users anywhere in the world have at least four satellites above the horizon at any time, day or night. See **Figure 1.2(a)** for an illustration of the arrangement.

The path that satellites take over the Earth is called their **ground track.**

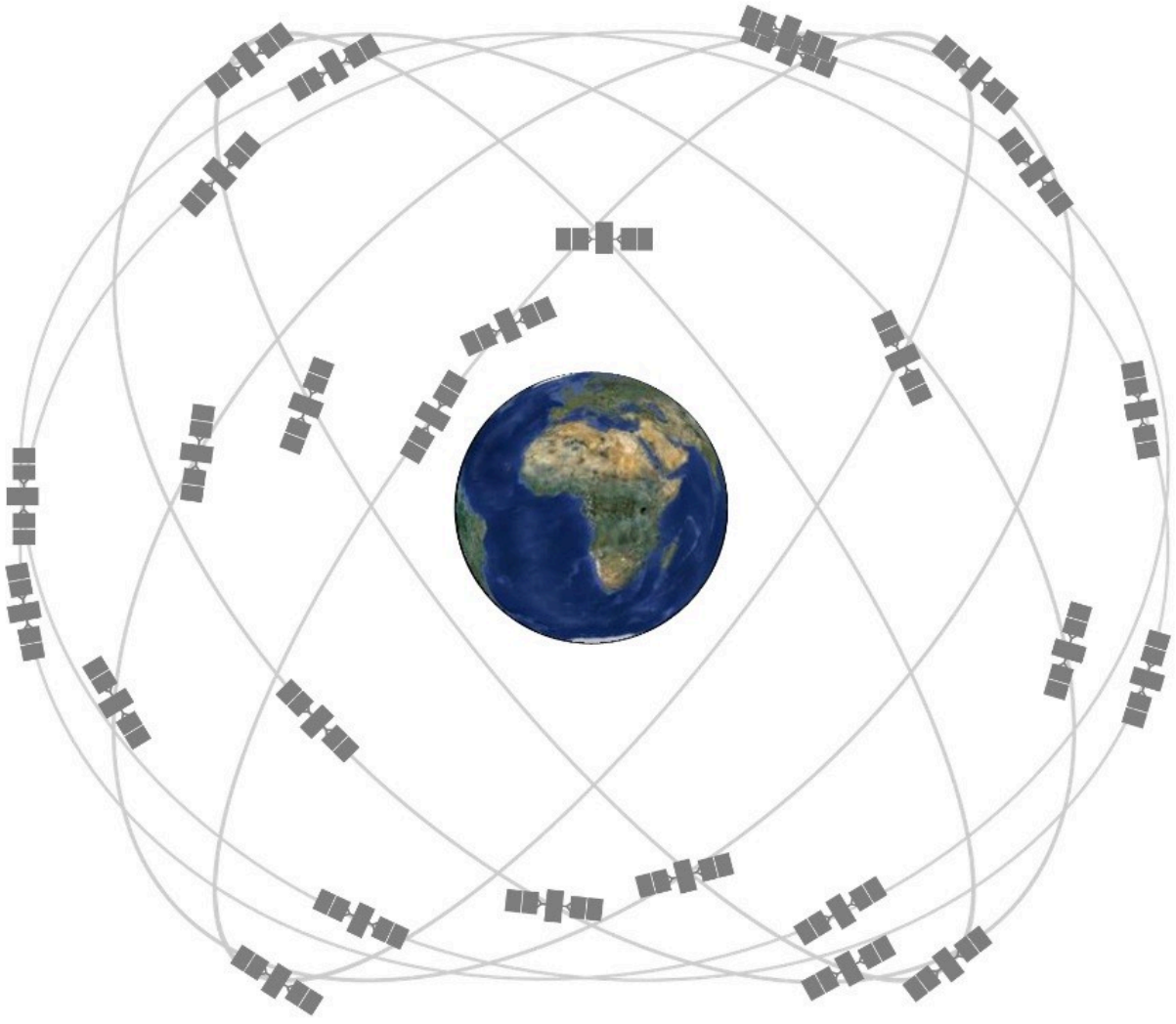


Figure 1.2(a): Image depicting the orbits of GPS satellites. Image in the Public Domain

The control segment

The **control segment** is the segment that maintains the GNSS and contains ground control stations that communicate with the satellites.

The GPS control segment consists of five tracking stations, the positions of which are accurately known. The control segment comprises a master control station and four other monitor stations.

The master GPS control station is located at Schriever Air Force Base in Colorado Springs, USA. **Figure 1.2(b)** shows the location of the other monitor stations, spaced fairly equally around the earth and close to the equator. Six additional National Geospatial Intelligence Agency monitory stations were incorporated in 2005.

At the master control station the information is processed to yield clock corrections and the ephemeris of predicted orbits of each satellite. This information is then uploaded to the appropriate satellites, which in turn broadcasts the data to user receivers as the message part of the signal.

This predicted ephemeris is referred to as the **broadcast ephemeris**. Another form of ephemeris is produced by post-processing the data from the tracking stations and this is referred to as the **precise ephemeris**. Apart from the GPS control segment, several organisations also track GPS satellites and produce ephemeris data. We will discuss ephemeris information in more detail in a later section.

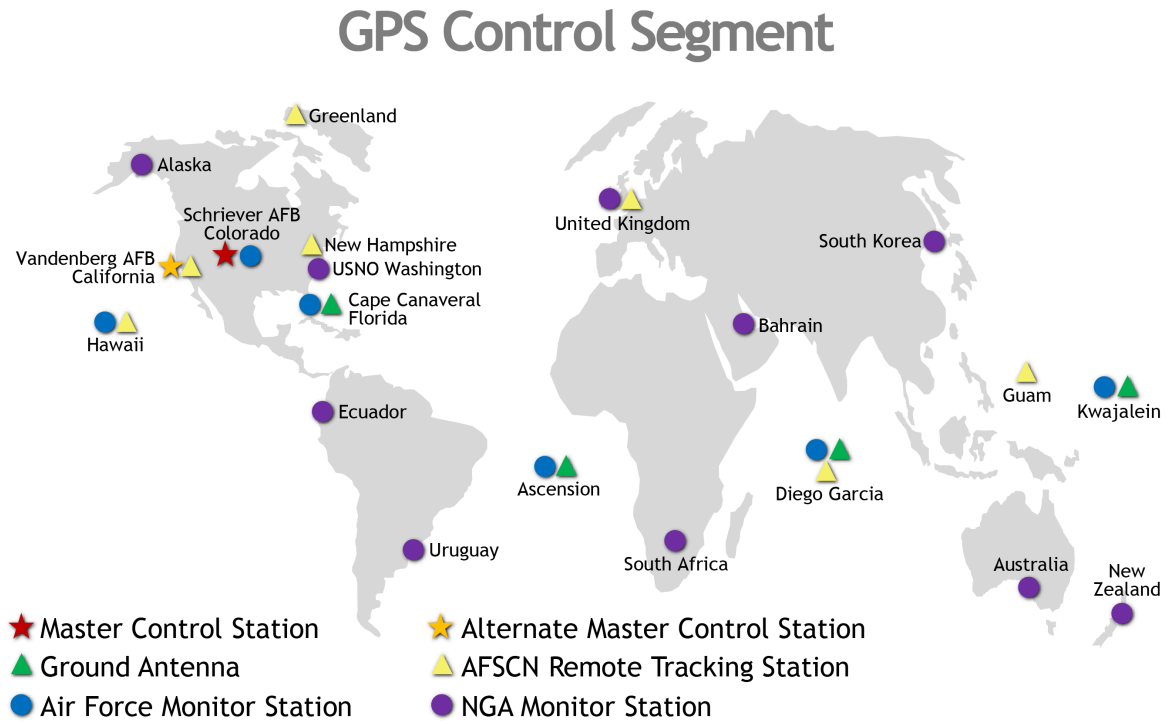


Figure 1.2(b): Location and types of GPS Control Segment. Image in the Public Domain

The user segment

The **user segment** comprises anyone who has a GNSS receiver!

This includes military and civilian users. Today GNSS chips are in every smartphone, and higher quality receivers are small enough and cheap enough to be carried by almost anyone.

The applications for GNSS are almost limitless and aren't just in navigation. Banks use the accurate atomic clocks to deal with transfers of money across accounts and stock markets in Nano seconds.

Positioning and location applications for GNSS include surveying, navigation, machine guidance and control, agriculture, mapping, emergency services, vehicle tracking and recreational uses.

1.3 GLOBAL SYSTEMS

Since GPS became available for civilian use, there have been any number of navigation systems that have leveraged the same technology and approaches. From full global systems, to local ground based systems, there are over 20 systems currently in operation.

Global Primary Systems or **GNSS** are those that are full satellite constellations; that have global coverage. **Regional systems** are systems with only a few satellites that are designed to cover a specific area.

Satellite Based Augmentation Systems (SBAS) use the positioning information from GNSS combined with corrections sent via communication satellites to provide a commercial service to users. **Global SBAS** have global coverage, while **Regional SBAS** are restricted to specific areas.

Ground Based Augmentation Systems (GBAS) work in a similar way to SBAS, however, rather than using the communications satellites for corrections, they utilise a ground based method of relay, such as radio waves or mobile phones.

These are best explained by the hierarchy of satellite systems, as shown in **Figure 1.3(a)**. We will discuss most of these systems in further detail throughout the book.

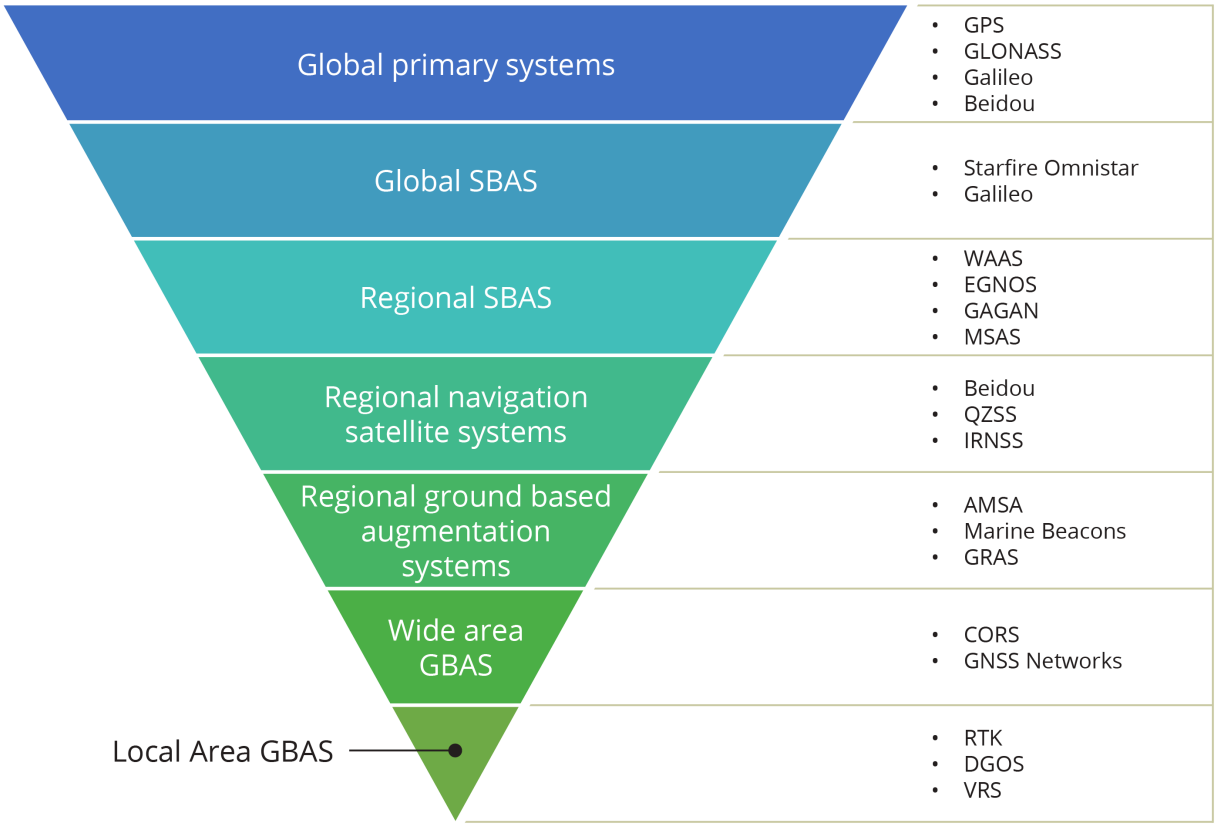


Figure 1.3(a): The Hierarchy of Satellite Systems

Global Primary Systems

There are currently four main global primary systems:

GPS

As previously discussed, GPS is run by the US Department of Defense, and was primarily developed for military purposes, however, since becoming available for civilian use has radically changed the navigation landscape.

GLONASS

The **GLO**bal **NA**vigation **S**atellite **S**ystem (GLONASS), deployed by the Russian Federation, is becoming an increasingly important system. The first GLONASS launch was in October 1982. Most satellites are launched three at a time on a Proton launch vehicle from Baikonur.

Overall, GLONASS is similar to GPS in terms of the satellite orbits and signal structure but there are some important differences. The GLONASS constellation uses three orbital planes rather than the six used by GPS.

While GPS satellites all broadcast the same radio frequency but different codes, GLONASS satellites all broadcast the same code but each satellite uses a slightly different frequency.

GLONASS also has two levels of service. Civil users have access to the Channel of Standard Accuracy which provides horizontal position accuracy of 60m (99.7% confidence) and vertical position accuracy of 75m (99.7 % confidence). Like the GPS, the GLONASS Channel of High Accuracy is only available to authorised users.

Galileo

Galileo is the European Union's GNSS. The European Union identified that they had several issues in relying on GPS or GLONASS for positioning systems, including:

- sovereignty and security of Europe
- present system not fully meeting civil users' performance requirements
- needing to ensure that users are not exposed to risks linked to a virtual monopoly
- capacity for European industry to compete on a fair basis in a fast growing market.

Identification of these issues led to the formal recommendation to the European Union to involve Europe in a new generation of Satellite Navigation Services, and Galileo was developed.

Galileo differed from the GPS and GLONASS systems in that it is a system under civil control.

The Galileo Constellation has three planes, circular orbits at a 56° inclination. The satellites orbit at a 23 616 km altitude.

Galileo also brought security to the EU in the wider sense, their technologically advanced and computer dependent society depends on GPS timing services for its services in energy, transport and financial transactions, and the cost of interruptions to the transport sector has been estimated at up to 300 million Euros per day.

Galileo Services
Galileo Services offer three levels of service:

Open Service

- provides position velocity and time services to mass market users
- similar to the future Standard Positioning System provided by modernised GPS
- accuracy better than 7 m worldwide and on an availability of 99%
- free of charge but, the quality of services is not guaranteed.

Commercial Service

- access commercially controlled
- provides improved service
    - accuracy
    - integrity
- ranging and timing service to knowledgeable professionals
- surveying
- meteorological forecasting
- time calibration.

Other Services

- The system is public regulated. There is no military involvement.
- Safety of life such as it is unencrypted; high integrity; confirmation of signal.
- The system provided near real-time; precise; and return link is feasible.

BeiDou

China has its own satellite navigation system that provides regional coverage of China and surrounds. This system is called BeiDou. It is named after the group of seven stars of the constellation Ursa Major, known in many cultures under different designations, in the UK 'The Plough', 'Big Dipper' in the US and 'Big Mother Bear' in Russia.

The first two BeiDou satellites were launched in 2000. Since 2001 China's army and others have thus had access to a domestic satellite positioning system.

The third BeiDou satellite was blasted into orbit two and a half years later in 2003. Nearly four years later in 2007, the fourth BeiDou SV was put into orbit and operates as spare. The fifth BeiDou, launched in 2007, was not positioned in an approximately geostationary orbit 35 800 km above earth's surface like the other four were. Instead, it is in an orbit of perigee 21 519 km and apogee 21 544 km.

The second generation of the BeiDou Satellite Navigation System (BDS), also known as COMPASS or BeiDou-2, will be a global satellite navigation system consisting of 35 satellites. The system was operational in China in December 2011, with 10 satellites in use.

BeiDou's ground segment includes the central control station and orbit tracking stations at Jamushi, Kashi and Zhanjiang.

Global and Regional SBAS

Satellite Based Augmentation Systems utilise the positioning signal from GNSS and communication satellites to send signal corrections to users. The structure of a typical system is outlined in **Figure 1.3(b).**

Similarly to the GNSS set-up, SBAS systems have independent communication satellites and ground facilities. Users generally pay a subscription fee to access the corrected signal, which gives them a more accurate position.

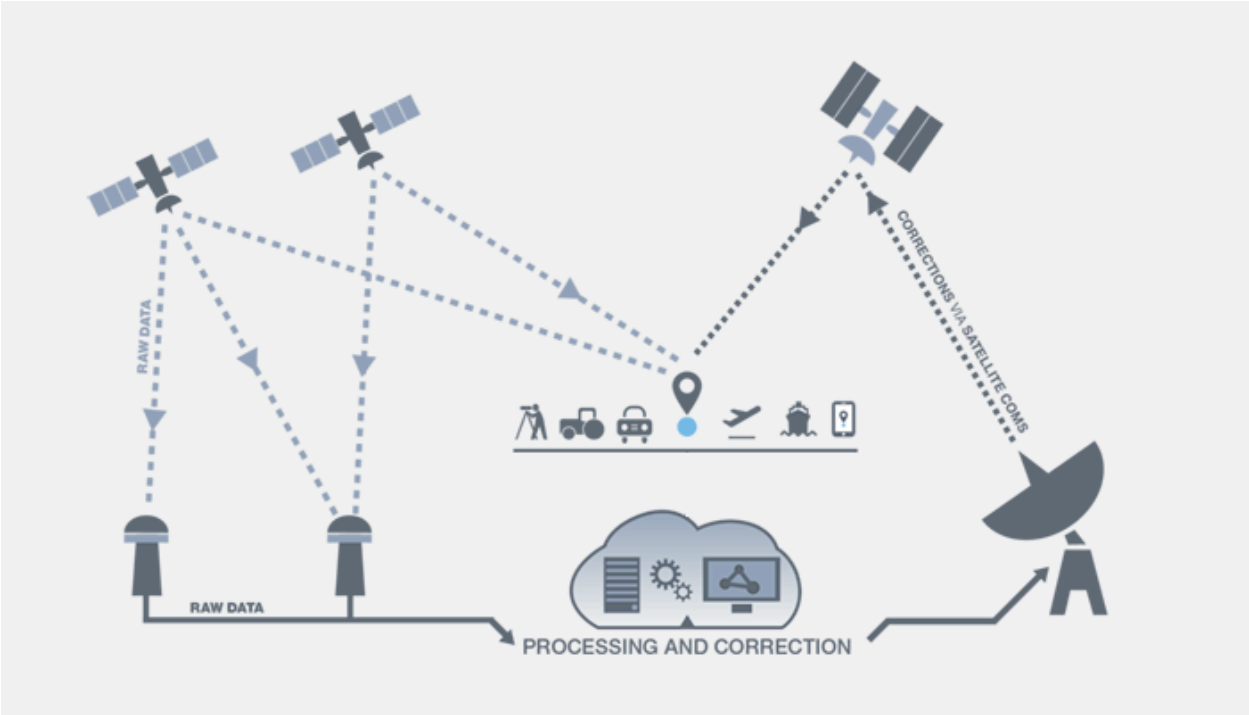


Figure 1.3(b): SBAS set-up. Source: Commonwealth of Australia (Geoscience Australia) used under CC-BY 2.0 Licence

Regional Satellite Systems

Several regions have launched their own satellites, that provide a 'mini' GNSS that focuses on their country or regional. The Japanese system QZSS is an example of a regional system.

QZSS

Quasi Zenith Satellite System (QZSS) has four satellites in different orbits, each at a 45° incline. Due to the high density of buildings in Japan, the satellites needed to orbit very high in the sky relative to the horizon. To allow for this, the satellites have a ground track of what is essentially a large figure of eight! A significant portion of this figure of eight tracks over Australia, and there are also three QZSS control stations in Australia.

GBAS

The main type of GBAS in use is also known as an Auxiliary system. They are designed to leverage the main GNSS or Regional Satellite System to provide improved local accuracy. They require a specialised receiver and are generally commercial services that require payment for access.

Table 1.3: GBAS Auxiliary Systems.

Country	Main System	Aux System	Acronym
USA	GPS	Wide Area Augmentation System Local Area Augmentation System	WAAS LAAS
Russia	GLONASS	Ground based Regional Augmentation System	GRAS
Europe	Galileo	European Geostationary Navigation Overlay Service	EGNOS
China	BEIDOU	Geosynchronous Earth Orbit Satellite System	GRIMS
Japan	QZSS	Multifunction Satellite Augmentation System	MSAS/MTSAS
India	GPS	GPS aided Geo augmented navigation Indian Regional Navigation Satellite System	GAGAN IRNSS

PART II

DATUMS AND COORDINATES

2.1 DATUMS AND COORDINATES

Learning Objectives

After successfully completing this chapter you should be able to:

- identify and describe the essential elements of ellipse geometry
- differentiate between the ellipsoid and geoid and identify when measurements relate to each of these
- define and compare common horizontal and vertical datums in use today
- outline the history of the development of these datums
- explain the fundamental characteristics of the Australian Height Datum (AHD)
- demonstrate an understanding of the construction and use of geoid models
- apply values from geoid models to attain orthometric heights from GNSS surveying observations.

In the beginning...

A guy looked down a well.

What came next

Eratosthenes was a Greek mathematician, geographer, poet, astronomer and librarian. He was the first known person to calculate the circumference of the Earth, in around 240BC. And he did it all without leaving his home in the city of Alexandria.

Eratosthenes knew that at noon on the summer solstice, in what is now known as Aswan in Egypt, if a person was standing over a well their shadow would block the reflection of the Sun on the water in the well. At the same time in Alexandria, he measured the shadow of a rod on the ground. Using the well-known distance between Aswan and Alexandria (which was surveyed every year) he combined this information with some simple geometry to determine the circumference of the Earth. He didn't get it exactly right, but given the technology of the day, he did a pretty good job!

Armed with this knowledge, Eratosthenes published a three-volume work called **Geographika,** and created the science we now know as Geography. Geographika contained the first model of the Earth that was covered with the grid we now recognise as latitude and longitude.

You can watch the NASA video explaining the history and applications of geodesy below.

Video 2.1: Looking down a well – a brief history of geodesy [2 mins, 25 secs]

Note: Closed captions are available by clicking the CC button in the clip below. This video is in the Public Domain.

Due to his enormous contributions to the field, Eratosthenes is often referred to as the **Father of Geography.**

All because some guy looked down a well.

The field of geodesy

Our understanding of the size and shape of the Earth has become known as **geodesy**, from the Greek word **ge** (earth) and **daiein** (divide).

The *Offshore Petroleum Amendment (Datum) Bill* defines geodesy as:

"**The branch of applied mathematics which determines the shape and area of large tracts of country, the exact position of geographical points, and the curvature, shape, and dimensions of the earth.**"

The surface of the Earth is very irregular, so when we measure things on the surface, we need to show them relative to some standard surface. A critical component of geodesy is developing these standard surfaces; mathematical models that represent the size and shape of the earth.

Ellipsoids

In the 17th century, Sir Isaac Newton reasoned that the surface of the Earth could best be represented by the surface of the oceans, but due to the rotation the Earth, the effect of centrifugal forces on the liquid oceans would make them bulge at the equator.

Measurements by geodesists proved this to be the case, showing that the Earth is about 43km wider at the equator than pole to pole, which is about 1/300 of the diameter.

This means that the Earth is not a perfect sphere, but an **oblate ellipsoid.**

The easiest way to visualise an oblate ellipsoid is to imagine what happens to a slightly flat soccer ball (or other ball that takes your fancy) that someone is sitting on. Ellipsoids can be described mathematically, as they have standard geometry.

Oblate means that a round shape has been flattened along the z axis. In the world of astronomy, this is the axis of rotation of the planet or other astronomical body.

Prolate means that the round shape has been flattened along the x & y axes, much like an AFL Sherrin football.

Ellipsoid geometry

An ellipsoid provides a simple mathematical approximation for the shape of the earth, either regionally or globally. The ellipsoid is a surface of revolution generated by rotating an ellipse about its minor (shortest) axis.

Figure 2.1(a) shows a typical ellipsoid and the geometric meaning of the parameters that define its size and shape

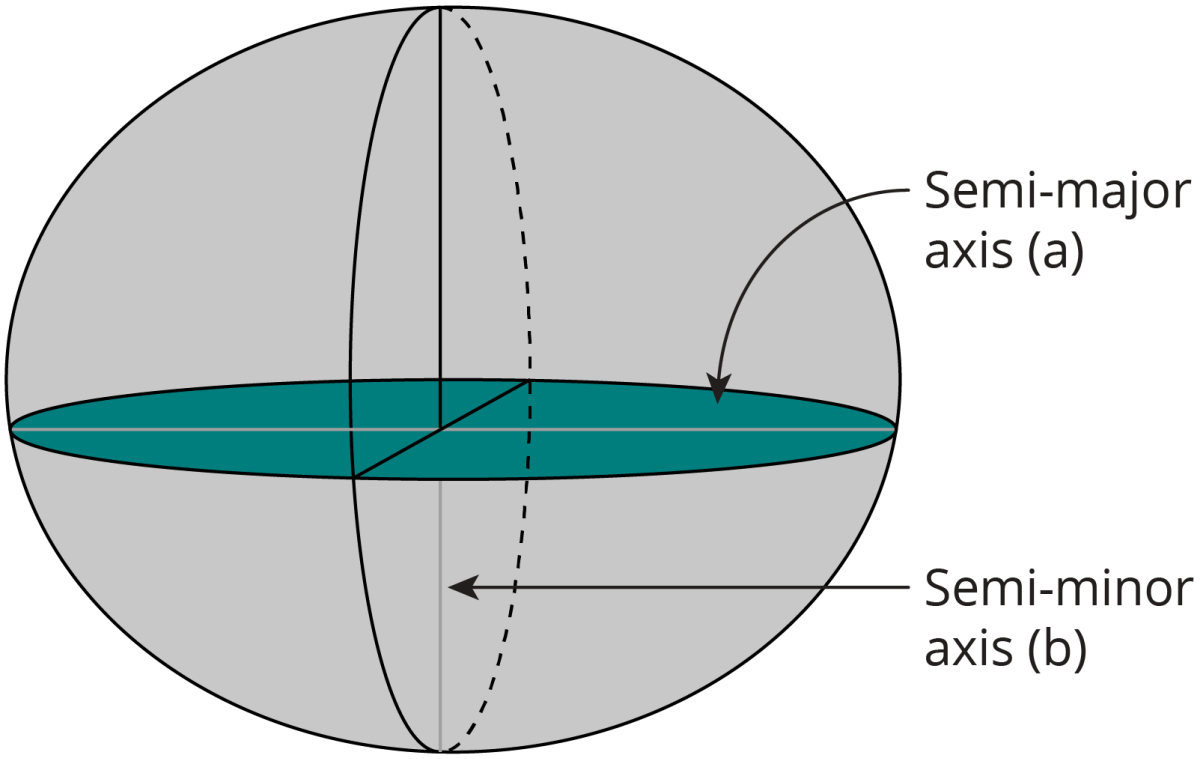


Figure 2.1(a): A Reference ellipsoid

There are three parameters used to describe an ellipsoid:

a = major semi axis of the ellipse
b = semi minor axis of the ellipse
f = flattening (approx $\frac{1}{300}$ for Earth).

The **flattening** determines the amount by which the ellipsoid is 'flattened' along the semi minor axis, which for the Earth is the axis of rotation. The flattening is defined by the equation:

$$f = \frac{(a-b)}{a}$$

The flattening is essentially a ratio of the two axes, so it has a value range of 0 to 1 where a value of zero would mean the two axes are equal (as in the case of a sphere).

Only two parameters are necessary to define the size and shape of a reference ellipsoid, and usually the parameters chosen are the length of the semi major axis and the inverse flattening (1/f).

The inverse of flattening is normally used when referring to ellipsoids to do with the Earth, as the flattening of the Earth is quite a small number: 0.003353.

As an ellipsoid is a mathematical model, it can never be directly measured, however, we are able to measure things relative to it.

The geoid

When we measure points on the irregular surface of the Earth, we need to show them relative to a reference surface, which as we have just discussed, can be mathematically defined by an ellipsoid.

However, this requires us to '**reduce**' them from the location we took the measurement, to the reference surface, which can be quite time consuming, particularly without the benefit of modern

day calculators and computing power. An obvious choice of a common surface that could be used everywhere was the sea level surface.

The level of the oceans at any given point on the earth is determined by the force of gravity that keeps the oceans from flying off into outer space. Gravity is a function of mass, and since the material of which the earth is made is not of uniform density, the sea level deviates from a regular ellipsoid shape from place to place.

The inconsistent density of the Earth is due to land masses, deep sea trenches and other materials. This means the exact shape of the earth, as indicated by the level of the oceans, can only be described as a 'blob'. It does not conform to any regular mathematical formula or regular geometric shape. Since this terminology is not very scientific or aesthetically pleasing, geodesists invented a new shape and they gave this new shape the name **geoid**.

Because sea level is affected by gravity, the geoid is considered an **equipotential surface,** meaning it is perpendicular to gravity everywhere, as shown in **Figure 2.1(b)** The geoid extends worldwide, across the continental land masses.

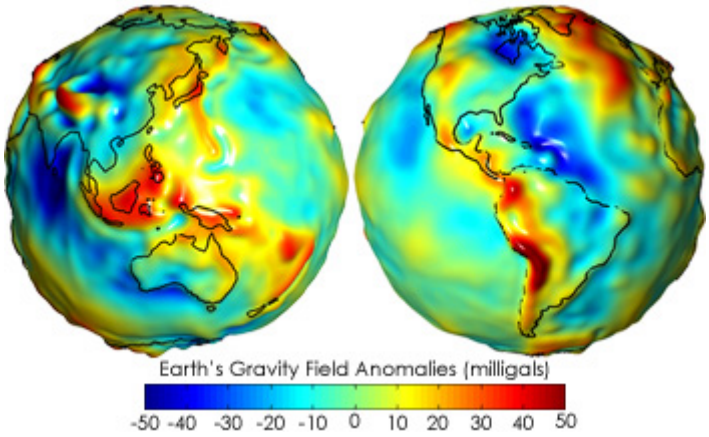


Figure 2.1(b): The geoid undulations of Earth, using units of gravity in colour. Image in the Public Domain

It is important to note the geoid only considers the influences of the Earth's own rotation and gravity on sea level, ignoring all other influences, like the impact the moon and winds have in creating tides.

The geoid and heights

As the geoid is an equipotential surface, this means that it is defined by gravity. The benefit of gravity and the geoid is in measuring heights. We can use a surveying technique called levelling to measure the differences in heights between places, as shown in **Figure 2.1(c).**

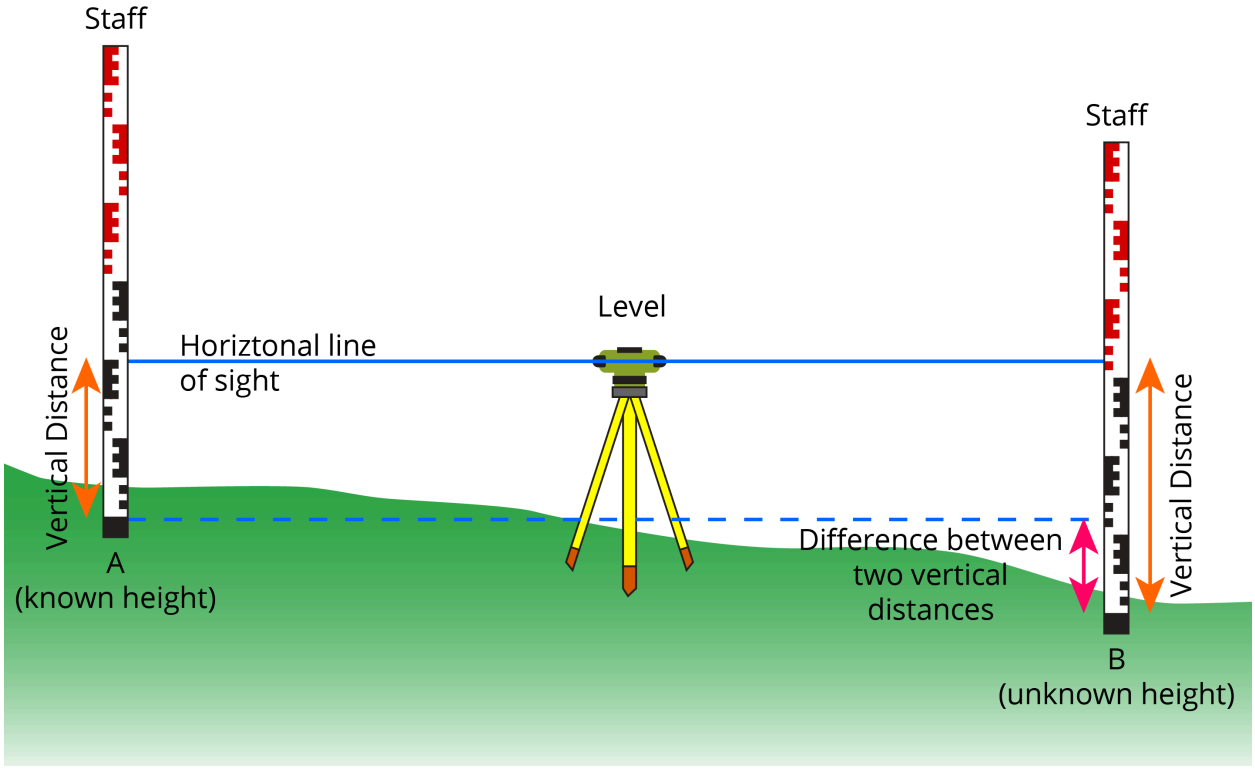


Figure 2.1(c): The principle of levelling

While the variation in the Earth's surface, called the **topography**, ranges from -11,034m in the Marianas Trench to +8,848m at Mt Everest, the deviation of the geoid from a global ellipsoid only ranges from approximately +85m in Iceland to -106m in India. So, while the geoid is an undulating surface due to changes in gravity, it is considerably more regular than the Earth's physical surface.

The deviation of the geoid from the ellipsoid will change at different locations, and is called the **geoid-ellipsoid separation**, and is represented by **n**. It is sometimes referred to as the geoid height.

But alas, measuring the topography of the Earth is what we need to do, and the geoid provides a reference surface that we can measure heights relative to. The distance from the geoid to the topography of the Earth is known as the **orthometric height**, and is represented by the symbol **H**. The distance from the ellipsoid to the topography is known as the **ellipsoid height** and is represented by the symbol ***h***. The relationship between the three heights can be shown by the equation:

H = *h* -N

Where H = orthometric height

h = ellipsoid height

N= geoid – ellipsoid separation or geoid height

This relationship is shown in **Figure 2.1(d)**.

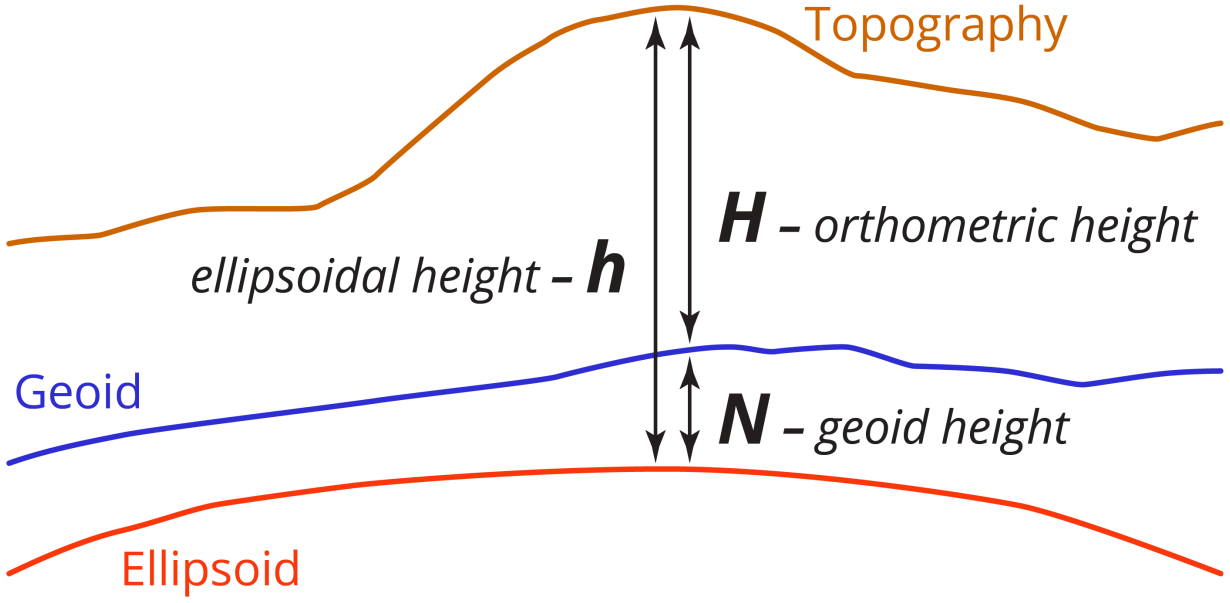


Figure 2.1(d): The relationships between different surfaces. Source: Department of Natural Resources, Mines & Energy used under a CC BY 4.0 licence

When calculating the different heights, it is critical that the user understands which surface they are measuring to, and particularly which ellipsoid they are using.

A warning note: If we use ellipsoid heights in some situations (like engineering design) instead of geoid heights, we can end up in a situation where we are trying to get water or fluids to flow uphill!

Geoid models

As gravity is different across the globe, it is common for countries and regions to produce their own **geoid model,** where various scientific methods are used to approximate the geoid, most often by observing gravity using gravimeters, or by observing **mean sea level** (**MSL**). The later technique is discussed further in the section of this module on height datums. The critical component of a geoid model is the value of the difference between the geoid and the ellipsoid.

The geoid and horizontal positions

However convenient the geoid might initially appear to be as a reference surface, particularly for its very close approximation to sea level, it is unsuitable as a reference surface for horizontal positions.

The use of an ellipsoid that most nearly matches the topographic surface of the earth (or the geoid), to which calculations can be related is required. Such models of the surface of the earth are called **geodetic datums** and can be used in navigation, surveying and mapping.

Geodetic datums

A datum is also called a **reference surface** or **reference system**. It provides a model that lets people measure locations on the Earth's surface, relative to a particular ellipsoid that best fits the country, region or continent where the measurements are taking place, at a particular point in time.

There are three main parts to a geodetic datum as shown in **Figure 2.1(e)**; the ellipsoid, epoch and reference frame (which is also called a realisation).

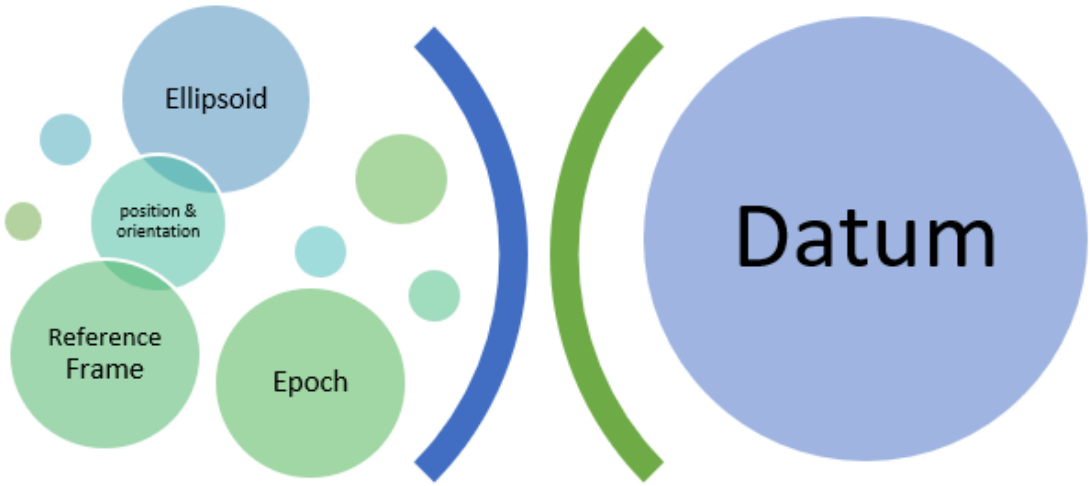


Figure 2.1(e): The key components of a datum

The ellipsoid has the usual ellipsoid geometry properties of semi minor and semi major axes, as well as flattening, that define its **shape** but also needs to have a **position** and **orientation,** which are determined by the **reference frame.**

The position of the ellipsoid refers to where its centre (or origin) is relative to the Earth, and the orientation refers to the axis of rotation. The axis of rotation is determined by the location of the north and south poles and the equator, which subsequently determines how the lines of latitude and longitude are orientated.

A **reference frame**, or **realisation** is where physical measurements at known points on the Earth are taken to determine the position and orientation of the ellipsoid. The realisation commonly use known points that are known in surveying as p**ermanent marks**, which are physical survey marks that are maintained by various State and Territory governments in Australia.

Datums are also referenced to a point in time, called an **epoch**. This essentially means that the parameters for that datum were defined at this date, and any measurements made on that datum are relative to the positions at that point in time. The epoch is the third of the three components that define a datum.

There are many different datums, and they are all uniquely named. An example is the global datum developed for GPS, called the World Geodetic System of 1984, shortened to **WGS84.** The two parts that make up the name are explained in **Figure 2.1(f)**.

Before technology like GNSS, local and regional models were very common, as there was no real need to relate data on a global scale. These datums are called **local datums**, and have huge variations in their ellipsoid geometry, position and orientation.

Once GPS arrived, a **global datum** was needed – a datum that was the best fit for the geoid in a global sense. **Figure 2.1(g)** shows the concepts of local and global datums.

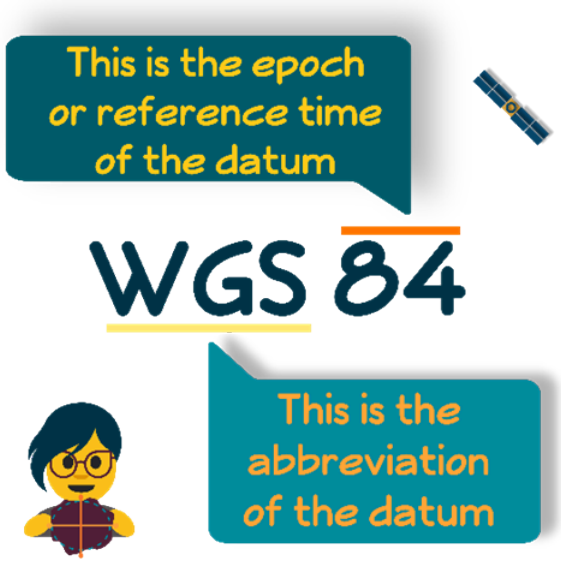


Figure 2.1(f): The naming convention for datums

Figure 2.1(g): Local and global datums

GNSS and Geocentric datums

Because satellites are affected by Earth's gravity field, their orbits are related to the centre of mass of the earth (we will cover orbits in Chapter 3 when we discuss Kepler's laws of planetary motion).

This means that calculations of the positions of satellites need to be referred to a datum based on the earth's centre of mass. Datums that are centred on the centre of mass of the earth are known as **geocentric datums**.

A geocentric datum provides direct compatibility with GNSS measurements and mapping that is based on a geocentric datum

Ideally we would like a single ellipsoid, or one worldwide standard, that best represents the entire globe, so for world-wide applications and to achieve compatibility with GPS, it was decided to define a geocentric datum. This ellipsoid would provide a best fit of the earth on a global scale. Today, the GPS datum is WGS84 (World Geodetic System of 1984) and the GRS80 (Geodetic Reference System of 1980) ellipsoids are the ones preferred by surveyors and cartographers worldwide and represent the best fit of the entire geoid in a global sense.

Since any ellipsoid is only a best fit to the geoid: just how good is this best fit? WGS84 is approximately 60 metres above and 100 metres below the geoid in Australia. Since the radius of the earth is about six million metres, the maximum deviation of the ellipsoid and the geoid is about one part in 100 000.

Note that since the WGS84 ellipsoid represents a best fit for the entire earth, a local datum would provide deviations significantly smaller than these maximums.

Height datums

Generally, an approximation of **Mean Sea Level** (**MSL**) is used to make height datums, as MSL can be considered an approximation of the geoid. MSL is observed through a network of tide gauges – currently 19 years is understood as the appropriate amount of time for data to be collected to determine MSL.

Permanent survey marks, known as **benchmarks** are installed near the tide gauges, and the height (relative to MSL) is transferred from the gauge to the marks by surveying levelling techniques. The '**level**' or height above MSL is assigned to the mark, and this is used to determine the level of other marks across land.

2.2 AUSTRALIAN DATUMS

The current Australian datum is a geocentric datum called the **Geocentric Datum of Australia 2020**, or **GDA2020** for short. Prior to GDA2020, we had GDA94 and before that **AGD66** and **AGD84**.

Australian Geodetic Datum

Australia's first official national geodetic datum was called the **Australian Geodetic Datum** (AGD). AGD was initially introduced in 1966, and was based on the local ellipsoid known as **Australian National Spheroid** or **ANS**. AGD had two versions: AGD66 and AGD84.

The fundamental characteristics of the AGD datum were:

- **ellipsoid** – the ellipsoid used was the Australian National Spheroid
- **position** – the position of this ellipsoid was determined to provide the best fit to the Australian regional geoid
- **orientation** – the National Mapping Council decided that the ANS should be parallel with the earth's mean axis of rotation at the start of 1962, and that the meridian plane of zero longitude should be parallel to the zero meridian plane near Greenwich.

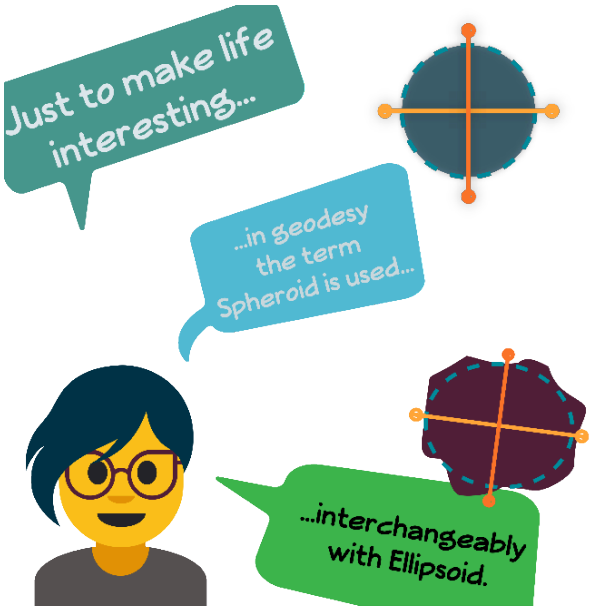


Figure 2.2(a): A note on terms

Note: the concept of longitude will be fully explained in the chapter on coordinate systems.

The passing of time revealed inaccuracies in the AGD66 coordinates, and this led to a new adjustment of the geodetic network in 1982. The basic definition of the datum was not changed for this purpose – the same ellipsoid was used (ANS) and the origin was held fixed. However, the addition of more observations and the use of improved error modelling and adjustment procedures gave rise to noticeably different coordinates. The resultant coordinate set was termed the AGD84. The comparison of the two datums is shown in **Table 2.2(a).**

Table 2.2(a): AGD66 compared to AGD84

Designation	**Australian Geodetic Datum (AGD)**	
Datum	AGD66	AGD84
Geographic coordinate set	AGD66	AGD84
Grid coordinates	AMG66	AMG84
Reference Frame	N/A	
Epoch	1966.0	1984.0
Reference Ellipsoid	Australian National Spheroid	
Semi major axis (a) value	6378160.0	
Inverse Flattening	298.25	

GDA94

A new and improved datum was adopted by Australia on 1st January 2000. This was known as the **Geocentric Datum of Australia 1994** (**GDA94**).

GDA is a geodetic geocentric datum, based on a model of the whole earth, which has its centre coincident with the earth's centre of mass. As the words imply, 'geocentric datum' is a datum that has its origin at the centre of mass of the earth.

The key advantage of the GDA over Australia's earlier datum (AGD) is that GDA is totally compatible with satellite-based navigation systems such as GPS and with major international geographic systems.

GDA is the datum used in Australia, while GDA94 was the geodetic coordinate set (latitudes and longitudes) computed in terms of the datum at 1 January 1994. The reference frame for GDA94 was the **International Terrestrial Reference Frame** (**ITRF1992**), at epoch 1994.0.

The fundamental characteristics of the GDA datum, adopted by a meeting of the Intergovernmental Committee on Surveying and Mapping (ICSM) held in Canberra on 28–29 November 1994, are shown in **Table 2.2(b)**.

The principal benefit of Australia adopting a geocentric datum was that it allowed people to use GNSS measurements, while also providing a single standard for collecting, storing and using geographic data. This ensures compatibility across various geographic systems at the local, regional, national and global level. For this reason the GDA forms the basis of the Australian Spatial Data Infrastructure (ASDI) – the infrastructure to manage Australia's key spatial data sets.

Table 2.2(b): GDA94 parameters

Designation	The Geocentric Datum of Australia (GDA)
Datum	GDA94
Geographic coordinate set	GDA94
Grid coordinates	MGA94
Reference Frame	ITRF1992 (International Terrestrial Reference Frame 1992)
Epoch	1994.0
Reference Ellipsoid	GRS80
Semi major axis (a) value	6378137m
Inverse Flattening	298.257222101
Position	GRS80 is geocentric (centred on the earth's centre of mass)
Orientation	The rotation axis is aligned with the direction of the Conventional International Origin (CIO) for the Polar motion and the zero meridian is as defined by the International Earth Rotation Service (IERS).

The issues with GDA94 have arisen over time:

- Due to tectonic plate movement, Australia has been moving approximately 7cm per year, meaning the GDA94 positions have continued to move away from the ITRF92 positions. By 2020 the difference will be 1.8m.
- Improvements in global reference frames (ITRF2014 was created) have given a better definition of the shape of the Earth, and this has meant ellipsoidal heights have changed approximately 9cm.
- Parts of the Australian crust have changed.

To address these issues, GDA2020 was developed.

GDA2020

GDA2020 is the latest Australian Datum, and at the time of this module being written, is still being implemented in some Australian states and territories.

The main difference between GDA94 and GDA2020 is that GDA2020 has a future epoch – 2020.0. This means that Australia is moving towards the positions in GDA2020, rather than away from them like in GDA94.

The shift from GDA94 to GDA2020 is approximately 1.8 m and is show in **Figure 2.2(b)**.

The summary of the GDA2020 parameters are given in **Table 2.2(c)**.

Figure 2.2(b): Approximate shift from GDA94 to GDA2020 locations across Australia. Source: Commonwealth of Australia used under a CC BY 2.0 Licence

Table 2.2(c): GDA2020 parameters

Designation	The Geocentric Datum of Australia (GDA)
Datum	GDA2020
Geographic coordinate set	GDA2020
Grid coordinates	MGA2020
Reference Frame	ITRF2014
Epoch	2020.0
Reference Ellipsoid	GRS80
Semi major axis (a) value	6378137m
Inverse Flattening	298.257222101
Position	GRS80 is geocentric
Orientation	The rotation (minor) axis is aligned with the direction of the Conventional International Origin (CIO) for the Polar motion and the zero meridian is as defined by the International Earth Rotation Service (IERS).

Australian Height Datum

The geoid is a real surface, however, it is not a physical surface that we can directly observe. The level surfaces of the earth's gravity field can be sensed, for example, by fluid in a vial (as in a spirit bubble).

On the other hand, surveyors have preferred a vertical datum that not only has physical reality,

but is also universally recognised as a 'sensible' one, being prominent and obvious to non-surveyors. This surface is the mean level of the ocean.

In 1971 Geoscience Australia, the national agency responsible for datums (among other things), developed the **Australian Height Datum 1971** (**AHD71**) for mainland Australia, and in 1979 developed AHD79 for Tasmania. AHD71 and AHD79 are both generally referred to as just **AHD**.

This was the first attempt to define a continent-wide height datum. Prior to this, Australia had many different levelling datums, all of them regional, and adopted by some government department or organisation to suit its own purpose. These local height datums were usually based on a local tide gauge as a zero level point.

AHD was tied to an estimate of **Mean Sea Level** (**MSL**), which was meant to approximate the geoid. The MSL estimate was calculated by having 30 tide gauges, distributed fairly uniformly around the Australian mainland coast, collect two years worth of data from 1966-1968. The 30 MSL estimates acted as the **constraint** for the calculations of heights, meaning they were given a set or **fixed value** of 0.0m. In surveying we generally abbreviate fixed value to just **fixed** – "This mark is fixed at 0.0m AHD".

In Tasmania, the AHD is connected to MSL at two tide gauges, Burnie and Hobart. The AHD (Tasmania) is independent of AHD (mainland), with no connection ever being made between them. The tide gauges are shown in **Figure 2.2(c).**

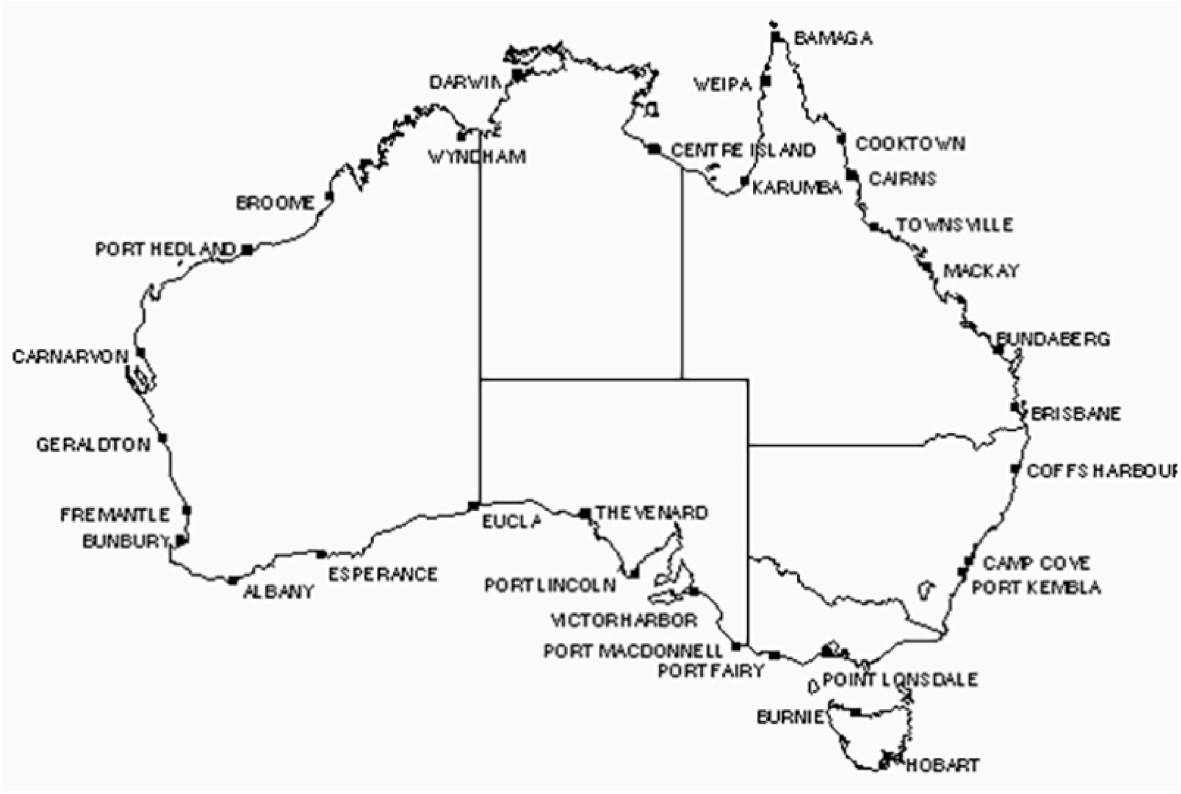


Figure 2.2(c): The tide gauge network used to create AHD in 1971. Source: Commonwealth of Australia used under a CC BY 4.0 licence

Once MSL was determined, a number of benchmarks were installed near the tide gauges and the height transferred from the gauge to the marks by surveying levelling techniques.

From 1945-1971, surveyors had been completing a massive campaign of levelling all across Australia, called the **Australian National Levelling Network**. They completed 97 230 km of two way levelling. Two way meant that the surveyors would complete a loop, where they finished measuring on the same mark they started on. Each loop would '**run**' through permanent benchmarks, hence lengths of levelling are often referred to as a **run**.

The data collected by the levelling runs were combined with the MSL data on the benchmarks,

and it was all put through a statistical process that surveyors regularly use called **least squares adjustment.** The adjustment used the differences in heights determined by the levelling, and the known heights at the benchmarks, to assign an AHD height to all the marks in the network.

The location of the AHD tide gauges, and the levelling runs is shown in **Figure 2.2(d)**. Also shown in this figure are the National Tidal Centre (formerly known as the National Tidal Facility) Australian Baseline Sea Level Monitoring Array (ABSLMA) locations. These locations are the more modern stations used to monitor MSL.

There have been a number of issues found with the initial AHD determination, including:

- The two year period for determining MSL is now considered to be inadequate.
- The temperature of the water in northern Australia compared to southern Australia was not considered, resulting in an almost 1m north-south tilt.
- Errors were found in the levelling runs.

Attempts to correct for these issues have, in par,t been made by a geoid model correction product called **AUSGeoid**.

AUSGeoid

While not a datum in its own right, the geoid is a critical part of height datums, and is most often approximated by MSL for this purpose.

The geoid model currently used in Australia is called **AUSGeoid** and is essentially a grid of geoid ellipsoid separations, (values) in terms of the GRS80 ellipsoid and AHD, as shown in **Figure 2.2(e)**. Newer AUSGeoid models also have considerations for known issues with AHD.

- AUSGeoid98 model provides ellipsoidal to gravimetric geoid values, but can have offsets with respect to AHD of up to 0.5m.
- AUSGeoid09 model provides ellipsoidal to AHD values.
- AUSGeoid2020 provides ellipsoidal to AHD values with uncertainty.

Geoscience Australia provides online tools to covert between ellipsoidal and AHD heights, based on a GDA94 or GDA2020 position.

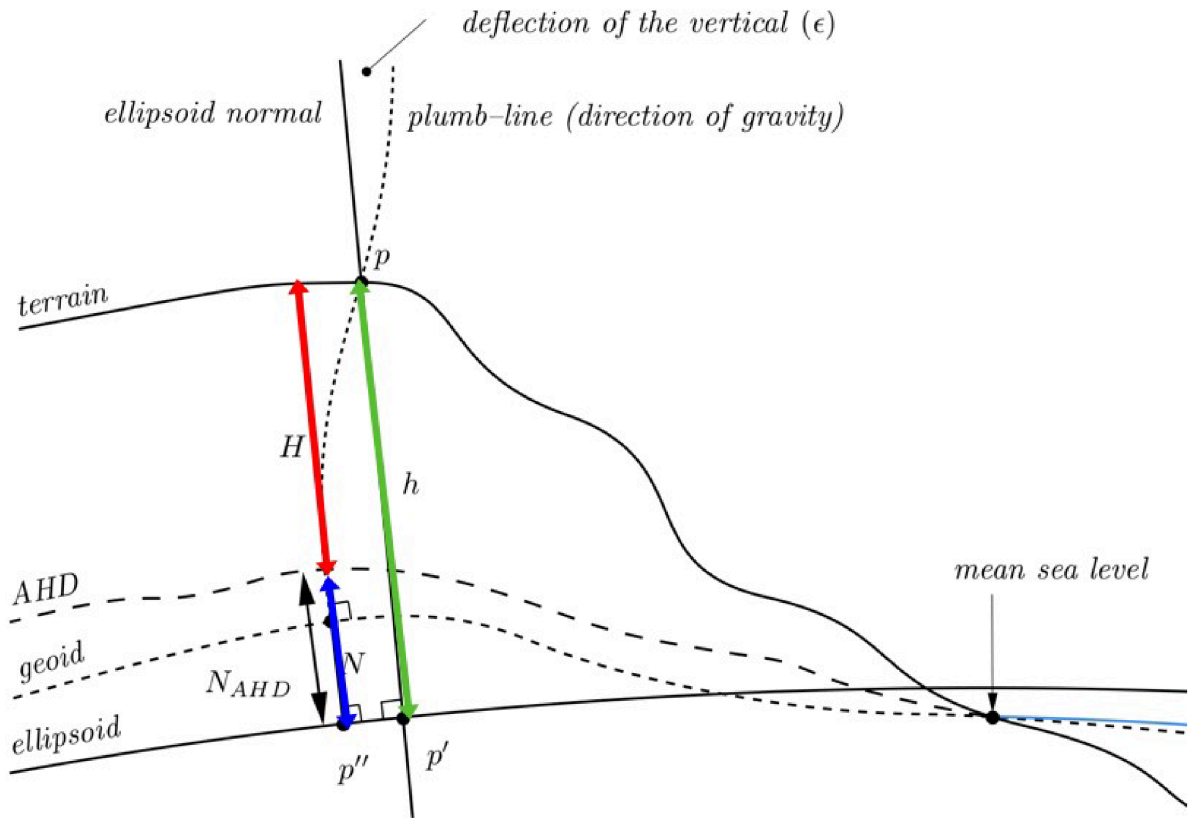


Figure 2.2(d): The AUSGeoid model (blue) enables users to convert ellipsoidal heights (green) to derived AHD heights (red). Source: Commonwealth of Australia used under a CC BY 4.0 licence

New Zealand Datums

New Zealand uses the New Zealand Geodetic Datum (NZGD), NZGD2000. The parameters are shown in **Table 2.2(d)**. The NZ datum is different from WGS84 and other ITRF datums, which are Earth centred and fixed, as it is a **plate fixed** datum, moving with the land mass as it moves and changes.

A plate fixed datum is used where a land mass is being subject to regular changes due to tectonic plate movement. The datum moves with the land mass, changing as the land does. This change is managed through a time dependent **deformation model** which models the deformation of the NZ land mass.

Being plate fixed means that the datum provides constant and unchanging positions, even though the land itself is moving. If you were to use positions in WGS84 for NZ, they would move as shown in **Figure 2.2(e)**.

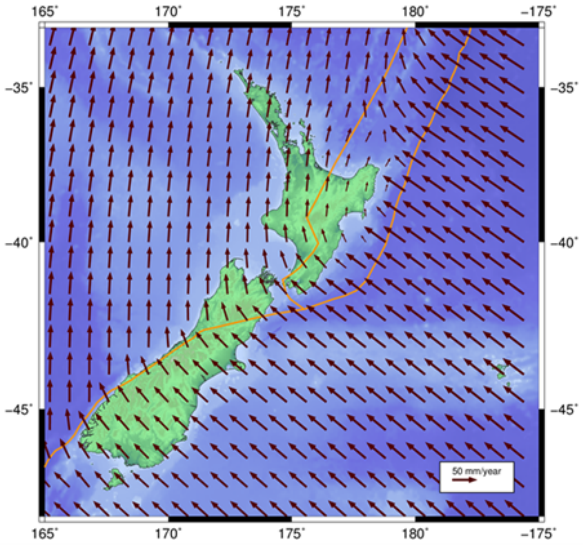


Figure 2.2(e): Land Information New Zealand (LINZ). Source: LINZ used under a CC BY 4.0 licence

Table 2.2(d): NZGD2000 parameters.

Designation	New Zealand Geodetic Datum 2000
Datum	NZGD2000
Reference Frame	ITRF96
Epoch	2000.0
Reference Ellipsoid	GRS80
Semi major axis (a) value	6378137m
Inverse Flattening	298.257222101
Position	GRS80 is geocentric
Deformation model	LINZ deformation model

2.3 COORDINATES

Learning Objectives

After successfully completing this chapter you should be able to:

- identify and describe the different coordinate system models
- define and compare common coordinate systems in use today
- explain how coordinate systems link to ellipsoid models and datums
- demonstrate an understanding of the basic concepts of projections and how they apply to datums
- outline the projections in use in Australia.

Coordinate system

We learnt in the that if we want to measure a position with GNSS, we need to have an understanding of the datum that we're using – this part of the module covers how we describe that position using a **coordinate system.**

Most people are familiar with coordinates from a topographic map, or even from their GPS navigation system in their car or smartphone; they are essentially a numerical way of describing the position of something relative to a datum or other reference surface.

However, there are some differences between coordinates on a round shape like the Earth and a flat map!

The relationship between a datum and the positions on it can be represented by three different types of coordinate systems relevant to geodesy;

- geographic coordinate systems: Latitude, longitude and ellipsoidal heights
- earth centred Earth fixed (ECEF) Cartesian coordinates
- projection coordinates.

Geographic coordinate systems

Lines of latitude and longitude are a familiar sight to anyone who's ever looked at a globe of the Earth, and they're a concept that has been around for a long time.

Our ability to use them effectively is less recent though. While latitude is relatively easy to understand and use for navigation due to the equator and shadows (remember the well?!), longitude is a more tricky concept, particularly when navigating across vast oceans. An incredible amount of very smart people tried to solve the problem of longitude as far back at the 1400s. Eventually it was an English clockmaker named John Harrison who invented a device called the marine chronometer that proved to be the key piece in solving the problem. But back to coordinate systems…

A **geographic coordinate system** is used to define locations on the earth's surface in three dimensions. The coordinates comprise angular measures (latitudes and longitudes) and heights referenced to a defined ellipsoidal datum surface.

Latitude, Longitude and ellipsoidal height are all dependent on the size, shape and position of the reference ellipsoid being used. This means that a position of a point described by a latitude, longitude and ellipsoidal height in one datum, say WGS84, will have a different latitude, longitude and ellipsoidal height on a different datum, like GDA2020. This is important to remember when measuring anything using GNSS!

Latitude and longitude are known as **curvilinear coordinates,** meaning they are expressed in angular units such as **degrees, minutes and seconds of arc**.

Latitude

A position's **latitude** is measured relative to the equator – points in the southern hemisphere of the Earth will have a negative latitude value, while points in the northern hemisphere have a positive latitude value. At the equator latitude has a value of 0°, and at the poles is has a value of ±90°. Latitude is described by the lowercase Greek letter phi .

Technically, latitude is defined as the angle between the major axis of the ellipsoid and the normal to the tangent plane at that point, measured at the point of intersection of the normal with the equatorial plane. In more simple terms, latitude is essentially measuring the angle from the equator to the position of the point you are trying to measure. **Figure 2.3(a)** shows the semi major axis a, semi minor axis b, and the angle of latitude. Note the perpendicular to the ellipsoid line (also known as the normal) doesn't intersect at the origin due to the ellipsoidal shape.

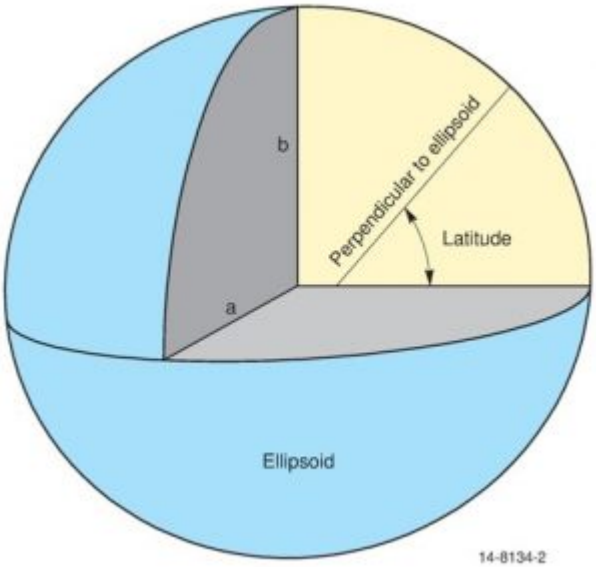


Figure 2.3(a): Latitude in a Geographic coordinate system. Source: ICSM, GDA94 Technical Manual v2.3 used under a CC BY 4.0 licence

Lines of constant latitude are called **parallels of latitude**, as they are parallel to each other. These parallels are circles of different radius, where all the points on them are at the same angle from the equator.

Latitudes in Australia range from about -10° or 10° South (i.e. south of the equator) to about 45° South; and in mainland New Zealand from about 36° South to about 47° South.

Longitude

Longitude is the method we use to measure positions on a spheroid or ellipsoid in what is the East and West direction on the Earth. It is described using the lowercase Greek letter Lambda – Λ

The lines of longitude are also called **meridians of longitude**, meaning that they all run from pole to pole – they're the ones that look like segments of an orange. These meridians are all the same size and shape, and are also known as **great ellipses** if they have the same centre as the ellipsoid used in their datum.

Longitude is described relative to the Royal Observatory in Greenwich England (pronounced Gren-itch) which is the meridian of 0°, making it the **Prime Meridian**. The other meridians are measured relative to the Prime Meridian along the equator. To the west of Greenwich the meridians are considered negative, through to -180°, while to the East they're considered positive through to +180°. **Figure 2.3(b)** shows the angle of longitude relative to the Greenwich Meridian of 0°.

Longitudes in Australia range from about 110° East to about 155° East, and in New Zealand from about 167° East to 179° East, relative to the Greenwich Prime Meridian.

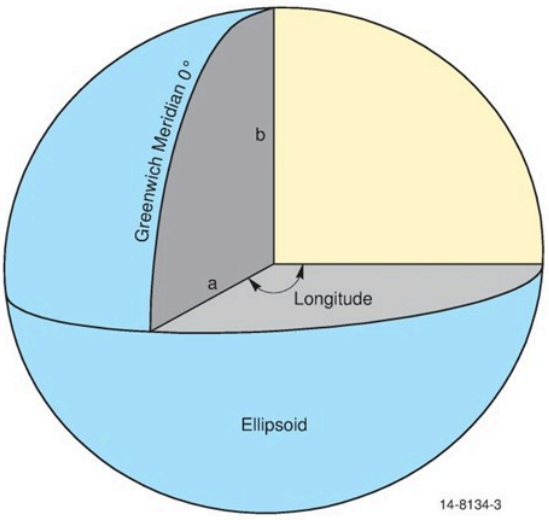


Figure 2.3(b): Longitude in a geographic coordinate system. Source: ICSM, GDA94 Technical Manual v2.3. used under a CC BY 4.0 licence

Distances from latitude and longitude

The distances subtended by one unit of measure (say one minute of arc) in latitude and longitude are not equal. For example, this means that the distance on the ellipsoid between latitudes of 10°13' and 10°14' (one minute of arc) won't be equal to the distance on the ellipsoid between the longitudes of 10°13' and 10°14'. At the equator the distance represented by one minute of latitude and one minute of longitude are approximately equal. This is because the equator is the only parallel of latitude whose radius is approximately equal to that of the meridians.

As the longitude meridians get closer together towards the poles (called converging) the distance between the meridians decreases. This is known as the **meridian convergence**, and is represented by the lowercase Greek letters delta and alpha – $\Delta\alpha$. Meridian convergence is shown in **Figure 2.3(c).**

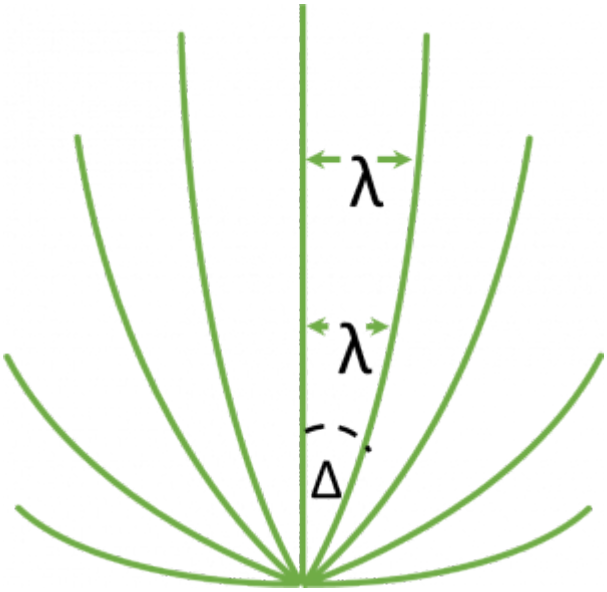


Figure 2.3(c): Meridian convergence in the southern hemisphere

Ellipsoidal heights

The third component of a geographic coordinate system is the ellipsoidal height, represented by lowercase (h). As covered in the section on ellipsoids and datums, ellipsoidal height is the difference between a point on the topography and the ellipsoid, which can be measured using GNSS, or calculated if you know the orthometric height (H) and the N value for the geoid-ellipsoidal separation.

Ellipsoidal height is measured along the ellipsoidal normal, from the surface of the ellipsoid to the point on the surface. The point on the ellipsoid is generally represented by ρ'while the point on the surface is represented by ρ . Refer to the diagrams in the **Datum chapter**.

2.4 EARTH CENTRED EARTH FIXED CARTESIAN COORDINATES

The system of Cartesian coordinates was developed by a French mathematician named René Descartes in the 17th century. A Cartesian coordinate system has three axes – **X, Y & Z**, which are in pairs perpendicular to each other, to represent the three dimensions.

A 3D Cartesian coordinate system must:

- share a common point, called the **origin**
- have an orientation for each axis
- use the same unit of length, for example, metres or kilometres.

The requirements for a Cartesian coordinate system are not that different to a datum, so using it as a coordinate system for the Earth becomes reasonably simple.

An **Earth centred Earth fixed (ECEF)** system has its centre at the centre of mass of the Earth, and the orientation of the axes are aligned with the Earth's natural axis system. An example of a point, shown by P, and it's location in an ECEF Cartesian coordinate system is shown in **Figure 2.4(a)**.

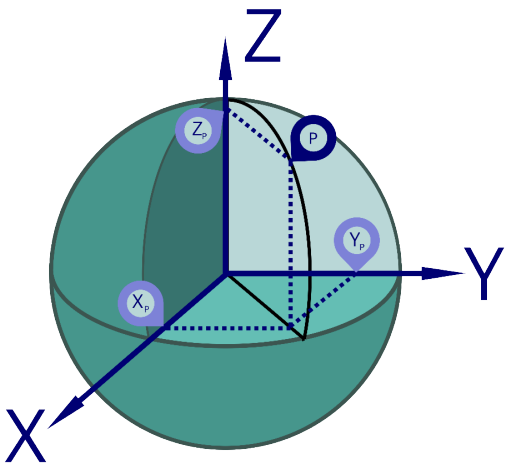


Figure 2.4(a): An ECEF Cartesian coordinate system showing the (X, Y, Z) coordinate locations

Generally, the following is true for an ECEF Cartesian coordinate system that relates to a geocentric global datum:

- The origin is at the centre of mass of the Earth.
- The X axis and Y axis are perpendicular to each other, and form a plane that is the same as the plane formed by the equator, (called the **equatorial plane**).
- The X axis is at the prime meridian of Greenwich at 0° longitude.
- The Z axis is perpendicular to the X & Y axes, and aligns with the N-S pole, or axis of rotation, of the Earth.
- The axes are fixed to the Earth's motion, so the coordinates of points are not constantly changing, apart for tectonic plate movement. This is why it is called "Earth fixed".

ECEF Cartesian coordinates are generally shown as a set like (X, Y, Z), where they are all in the same units.

Importantly, an ECEF Cartesian coordinate system does not need a defined datum to express the

coordinates of a point, however, you do need to know the origin and orientation of the axes of the coordinate system itself.

ECEF Cartesian coordinate systems and GNSS surveying

In more traditional surveying approaches, the horizontal position of something was most often dealt with separately to its vertical position or height. When GNSS came along, which is inherently a three dimensional system, a broader view of how to survey things had to be taken.

We will discuss the different GNSS techniques in Modules 4-6, however, it's worth pointing out that ECEF Cartesian coordinates are utilised in a GNSS survey technique where we have two or more GNSS receivers collecting data at once. We are able to compare the positions of the GNSS receivers to calculate the three dimensional differences between them, which is known as a **baseline.** A baseline is usually represented by the Greek letter delta (which is often used to show that a value is the difference between two things), followed by the axis letter, for example, $(\Delta X, \Delta Y, \Delta Z)$.

Coordinate transformations

Often data is collected on one reference surface, and for one of many reasons, may need to be changed to another reference surface, or even from one type of coordinate system to another. The process that moves the coordinates of points from one surface or system to another is called a **transformation**.

A transformation can involve a number of factors, including an origin change, rotation and scale to name a few, and these factors are known as **transformation parameters,** describing how the transformation needs to occur.

A type of transformation where we 'slide' all the points a consistent distance and direction is called a **translation**, which is often referred to in surveying as a **block shift**.

Moving between GDA94 and GDA2020 required a coordinate transformation. The transformation parameters for new datums to old datums are usually published at the same time as the definition of the new datum.

2.5 PROJECTION COORDINATES

Although computing power has made it much easier to model 3D shapes like the Earth, this hasn't always been the case. For almost as long as humanoid species have existed, we have captured location information in the only way available to us – a two dimensional map format.

Cartography, the practice of making maps, has long needed methods to represent the 3D size and shape of the Earth on a 2D surface like paper, and it achieves this using models called **map projections**.

Map projections use the idea that a 2D surface can be manipulated to wrap around, and intersect or touch a sphere or ellipsoid, and then points can be 'projected' outwards from the sphere or ellipsoid onto the 2D surface.

This is easier to visualise with an example…

Imagine if you could put a torch inside a model of the Earth that shone in 360°, and then made pinpricks around the coastlines of all the countries. You then wrapped a piece of paper around the equator of the model like a cylinder, and drew a dot everywhere a dot of light landed on the paper. When you unwrapped the paper and lay it flat, you would have a representation, or map, of all of the coastlines on the Earth. Congratulations! You have used a **cylindrical projection** to transform three dimensional coordinates (positions on a globe) to a two dimensional system (a paper map).

Representing a 3D surface on a 2D surface isn't without its problems though, as you move further away from where the 2D surface touches or intersects with the 3D surface, the distortion of distance, shape and area becomes larger and larger.

In our example above, the pinpricks that are closest to the equator are going to going to appear on the paper relatively close to where they are on the globe, and their distance, shape and area will be pretty close to reality. However, when we move closer to the poles, the distance from the pinprick to the paper is much larger, so the distances, shape and area of the coastlines are all going to be grossly exaggerated. This is exactly what happens on most world maps that use a cylindrical map projection – countries near the pole look much larger than they are in reality.

Types of projections

There are three main types of projections; cylindrical, conical and plane. Other types exist, but these are the most common for mapping purposes.

Cylindrical projections

Cylindrical projections are like the one in the example above, however, sometimes the cylinder passes through the ellipsoid or sphere, as shown in **Figure 2.5(a).** This is probably the most common type of projection used, since being made famous by the **Mercator projection**.

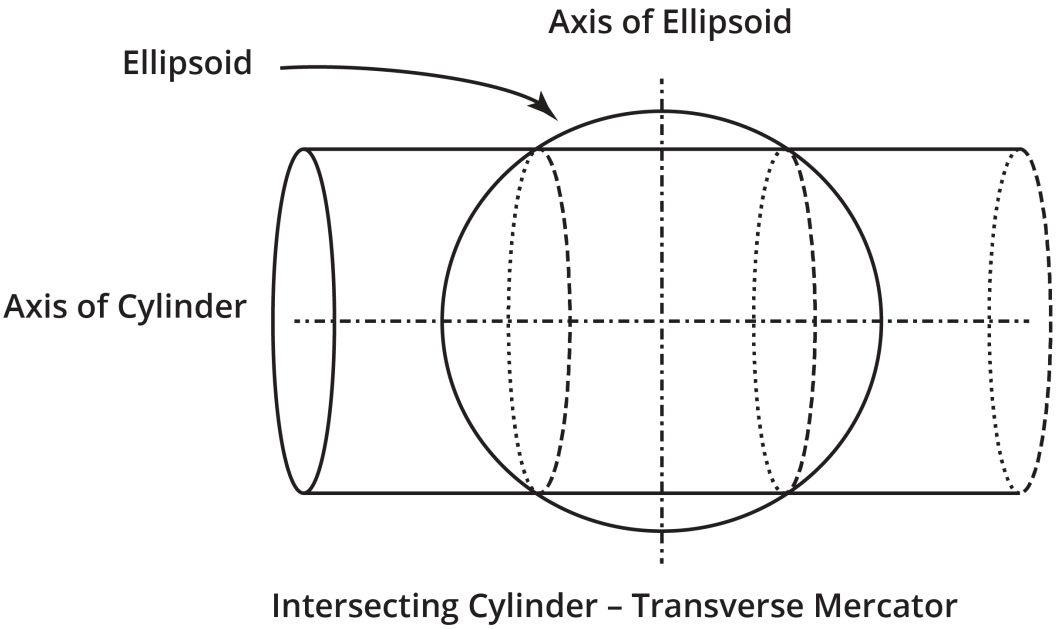


Figure 2.5(a): Intersecting Cylinder – Transverse Mercator map projection

The **Universal Transverse Mercator (UTM)** projection, shown in **Figure 2.5(b)**, a more modern version of the Mercator projection, is possibly the most common in use today. UTM works by dividing the world into a series of 60 consecutively numbered zones that are each 6 degrees of longitude wide, and the zones extend from 80 degrees North to 80 degrees South.

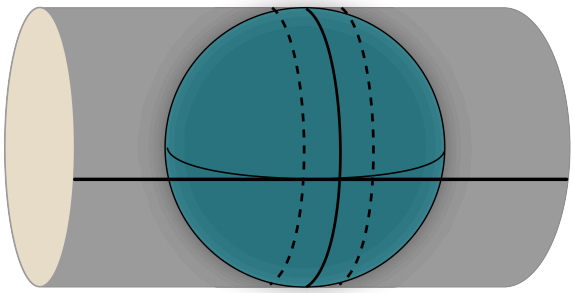


Figure 2.5(b): Universal Transverse Mercator projection

Conical projections

Conical projections are generally only used when you need to represent a region of one hemisphere of the Earth; they are not well suited for mapping large areas. Conical projections maintain the accuracy of an area but distort the shape of countries. The most common conical map projections are the Lambert or Albers projections, shown in **Figure 2.5(c)**.

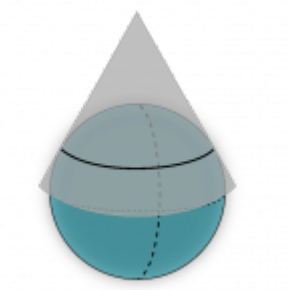


Figure 2.5(c): A conical projection, showing the intersection of the projection and the ellipsoid

Plane projections

Plane projections remove the curving of the 2D surface that happens in the cylindrical and conical map projections, and simply keep the projection plane flat. See **Figure 2.5(d).**

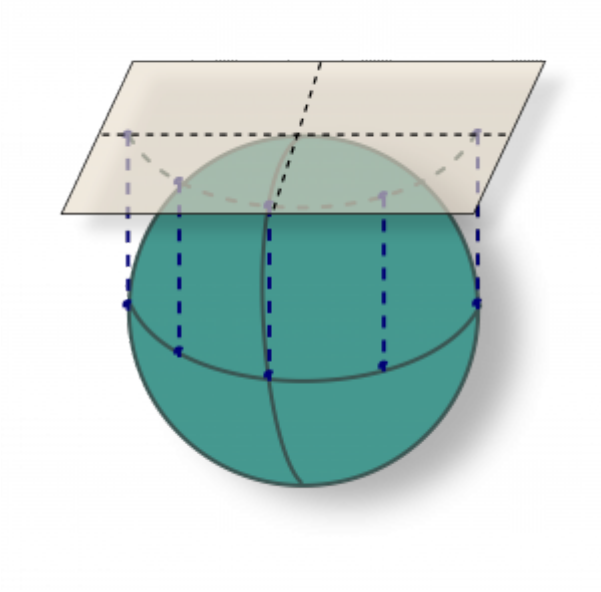


Figure 2.5(d): A Plane map projection, showing where points on the equator project onto the 2D plane

Imagine if we chopped the model of the Earth from earlier in half along the equator, and then placed it equator side down onto a glass table top that was lit from underneath. We then place a piece of paper on top of the North Pole, so that the paper is parallel to the table top. The pinpricks now have light coming through them that is perpendicular to the table top and also to the piece of paper – this is how a plane projection works.

In some sense the UTM is a combination of cylindrical and plane projections – the 60 zones mentioned earlier are essentially planes that have a central meridian of longitude. The central meridian is in the middle of the zone, which is 6 degrees of longitude wide, meaning the central meridian has 3 degrees of longitude either side of it. The central meridian is used as the **true origin** of the zone, but to avoid negative numbers, a false origin is used.

Map Grid of Australia

The map projection used in Australia is called the **Map Grid of Australia (MGA)** and is based on the cylindrical UTM projection. The units of MGA are in metres, making them easier to use in a practical sense.

Coordinates in MGA are described as **eastings** and **northings**, indicating the distance in metres from the origin in a positive eastwards direction (easting) and northward direction (northing).

The UTM zones that apply to Australia (west to east) are zone 46, at the Cocos (Keeling) Islands, through to zone 59 at Norfolk Island. Zones 50-56 cover the main continent of Australia, with zone 50 being the west coast and 56 the east coast of the mainland.

MGA coordinates are calculated relative to a GDA datum, so **MGA2020** is the projection used for GDA2020, and **MGA94** is used with GDA94. The origin of MGA2020 is a false origin, and has a value of (+500,000m, +10,000,000m).

Geoscience Australia has tools that incorporate mathematical formula, called **Redfearn's Formula**, to determine grid coordinates (easting and northing) from geographic coordinates (latitude and longitude) and vice versa, for Australia.

New Zealand Transverse Mercator

The **New Zealand Transverse Mercator 2000 (NZTM2000)** map projection is based on the NZGD2000 datum, however, is only used for the main New Zealand island group – separate projections are defined for the offshore islands and the continental shelf.

PART III

GNSS BASIC PRINCIPLES

3.1 GNSS BASIC PRINCIPLES

Learning Objectives

After successfully completing this chapter you should be able to:

- explain the basic principles of satellites' motion using Kepler's laws of planetary motion
- describe satellites' position in space using ECEF Cartesian coordinate system
- explain the different satellite orbits parameters for different GNSS
- define and explain the basic components of a GNSS signal structure
- explain the fundamental concepts and importance of GNSS time systems
- explain the function and structure of navigation data messages
- differentiate between the almanac and an ephemeris, and explain how and why they are used
- list error sources in GNSS positioning
- list accuracy levels in GNSS.

In the beginning...

In 1942 the German rocket (technically it was a guided long range ballistic missile) V-2 was the first man-made object to reach space, reaching around 80 km above MSL. In 1957, the Union of Soviet Socialist Republics launched the **Sputnik 1** ("Satellite 1" in English) successfully, and it orbited the Earth at around 250 km. Sputnik 1 had two radio transmitters on board, and it emitted beeps that could be heard on radios across the planet. The radio signals that were transmitted were used to gather information about the Earth's atmosphere, as well as understanding temperature and pressure in space.

The idea that we could use radio signals from satellites for positioning wasn't that big of a leap from the use of radio waves on the ground for measurement. And once Sputnik 1 carried transmitters successfully, it was really only a matter of time before a military started experimenting with the idea. As with most technological advances, the military involvement in GPS meant a scale of investment that would have been almost impossible in the civilian sector.

It turns out the key ingredients for GPS were Scottish & German physicists, a German mathematician, a World War, a Cold War, some astrophysicists and electrical engineers, a space race and a huge amount of money!

3.2 THE MOTION OF SATELLITES

Humans have been observing celestial bodies and their motion since ancient times, with some of the earliest records dating back to around 1200 BC.

The term s**atellite** is used in astronomy to describe an astronomical body (like a moon) that orbits a planet. They are also known as **natural satellites** since the invention of **artificial satellites** like those used in GNSS. Regardless of whether they're natural or artificial, all satellite orbits are governed by a set of scientific laws, most commonly known as **Kepler's laws of planetary motion**.

Johannes Kepler was a German mathematician and astronomer in the 17th century, and he published a series of works that outlined how the Earth and other planets orbited the Sun. His work built on the theories of Nicolaus Copernicus, who had suggested that planets orbited the Sun in circular orbits. Kepler realised that the orbits were actually elliptical and proposed three laws that explained how these orbits worked.

In GNSS, Kepler's laws of planetary motion are used to predict the position of satellites, which is a critical component of positioning, and is included in the signals transmitted by GNSS satellites. This information is in the ephemeris, which we will discuss further in the signals section of this module.

We already know that ellipses are important to GNSS, and how to define them, the fact that satellites have an elliptical orbit makes it pretty easy to understand the maths in Kepler's laws.

However, before we discuss Kepler's laws, it's helpful to get our heads around some basic astronomical terms that relate to elliptical orbits. Because we're going to be talking about satellites orbiting the Earth, we'll look at the basics from this perspective.

The basics of elliptical orbits

Kepler proved that the orbit, or path that a satellite took around a body, was an ellipse. As we already know, an ellipse is defined by the semi major (a) and semi minor (b) axes, as shown in **Figure 3.2(a).** The centre of the ellipse is represented by *C*.

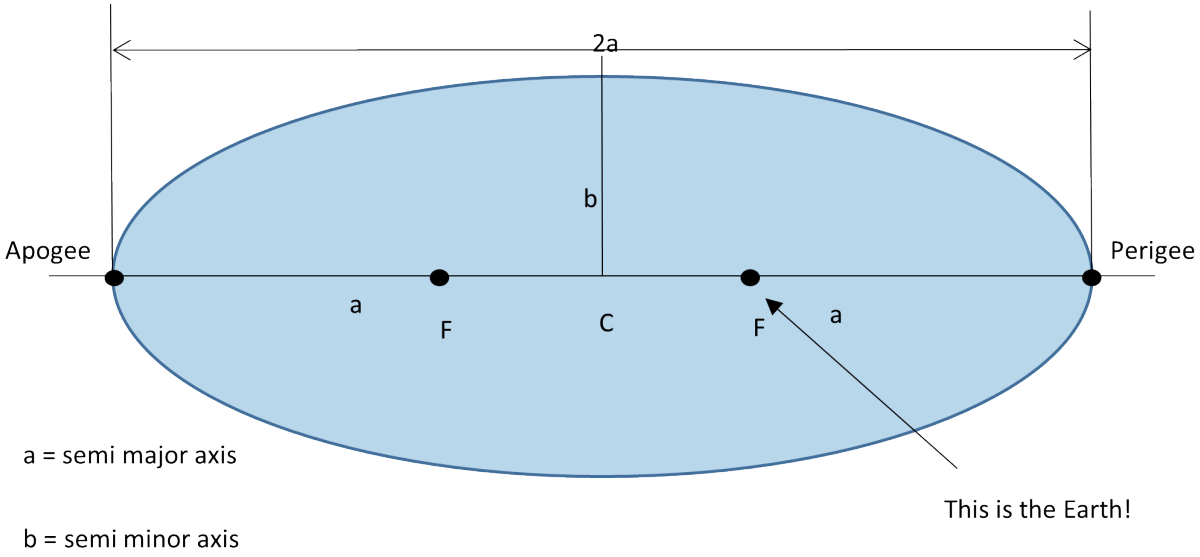


Figure 3.2(a): The fundamentals of elliptical orbits.

Perigee and apogee

The **perigee** and the **apogee** are points on the ends of the major axis, as shown in **Figure 3.2(a)**. The prefix **Peri** means "near", so the perigee is the closer of the two points to the Earth, while **Ap** means "away from", so the apogee is the furthest point on the major axis from the Earth.

Depending on what the body being orbited is, depends on what these concepts are called. When orbiting the Sun we use the term "–hellion", giving us aphelion and perihelion, while stars use "-astron" – apastron and periastron. The generic version is "-apsis", giving apapsis and periapsis.

Foci

An ellipse also has two **foci** points (**focus** is the singular), represented by , as shown in **Figure 3.2(b)**. The positions of the foci are such that at any point on the outside of the ellipse, the sum of the distances from that point to each of the foci will always be the same.

$$d_1 + d_2 = 2a$$

Where d_1 = distance from one foci to the satellite
d_2 = distance from the other foci to the satellite
$2a$ = major axis

An easy way to visualise this is if you put two nails in a piece of wood a distance apart, and join them with a piece of string about two and a half times longer than the distance between them. If you pulled the string tight with a pencil and then moved it around 360 degrees, keeping the string tight, it would draw an ellipse. The total length of the string didn't change, but the distance from each focus did. This is shown in **Figure 3.2(b).**

The distance from the centre of the ellipse to either of the foci is represented by

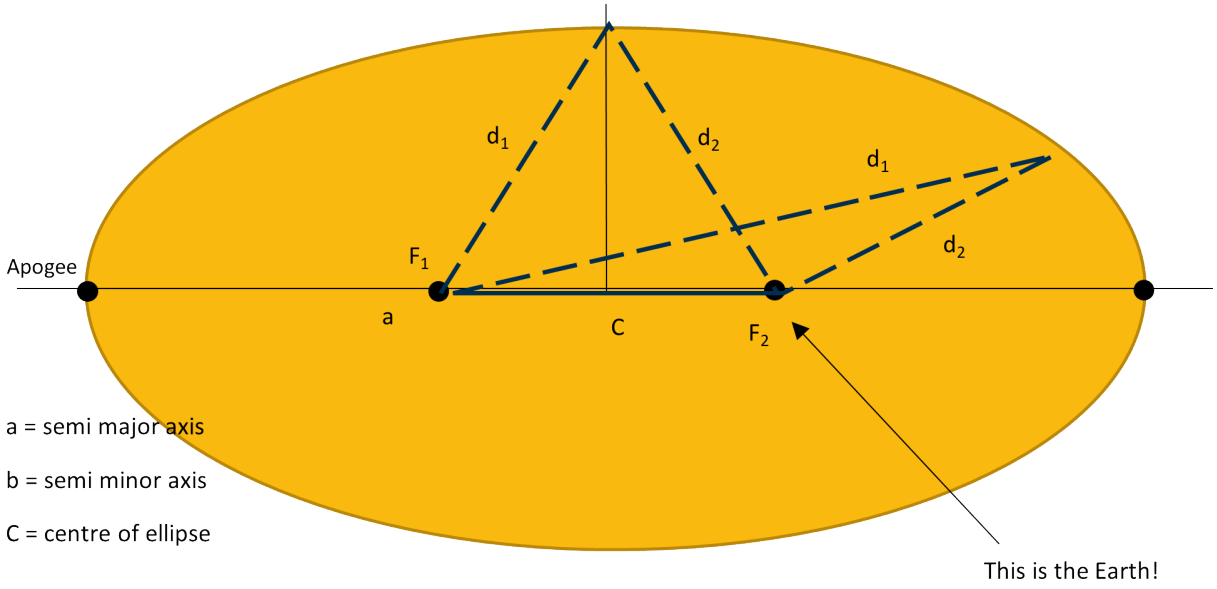


Figure 3.2(b): The sum of the distances from the foci to a point on the outside of the ellipse are equal.

Eccentricity

How much an ellipse is away from being a circle is measured by its **eccentricity.** The distance between the two foci is defined in terms of eccentricity of the ellipse, which is represented by . Eccentricity is mathematical concept that has to do with conical sections, and for an ellipse, eccentricity will have a value greater than zero but less than one.

If you know where the centre of the ellipse is, the eccentricity is simply the distance from the centre to either of the foci, divided by the length of the semi major axis, given by the equation:

$$e = \frac{c}{a}$$

where e = eccentricity
c= distance from the centre to either foci
a = semi major axis

If you don't know the centre, then it is defined as the ratio of the distance between the two foci to the length of the major axis, given by the formula:

$$e = \frac{2c}{2a}$$

where e = eccentricity
2c= distance between the foci
2a = length of the major axis

Eccentricity is important because it helps us calculate the position of satellites in their orbit relative to the Earth.

Kepler's first law – the Law of Orbits

The Kepler's first law, also known as the **Law of Orbits** states that the orbit of every planet will be an ellipse, with the centre of mass of the Sun being at one of the foci of the ellipse, as shown in **Figure 3.2(a)**. Remember Kepler was talking in terms of our solar system, so he was describing how planets

were orbiting the Sun. For GNSS, the equivalent is that satellites will have an elliptical orbit where the Earth is at one of the foci of the ellipse.

The equation for Kepler's first law is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^6} = 1$$

where x = distance from the sun in the major axis
a= semi major axis
y= distance from the sun in the minor axis
b = semi minor axis

When dealing with the orbit of a GNSS satellite around the Earth:

- x = distance from the earth in the major axis
- y = distance from the earth in the minor axis

To calculate the x and y values in Kepler's equation, we need to do some calculations to determine the position of a satellite in its orbit. The best way to understand this is through an example.

Let's assume we have a satellite 'S' that is in an elliptical orbit around the Earth, as shown in **Figure 3.2(c)**. We need to determine where the satellite is relative to the Earth, but as well as relative to its orbit. We can determine the values from Kepler's equation using our knowledge of plane coordinates, as shown in **Figure 3.(c).**

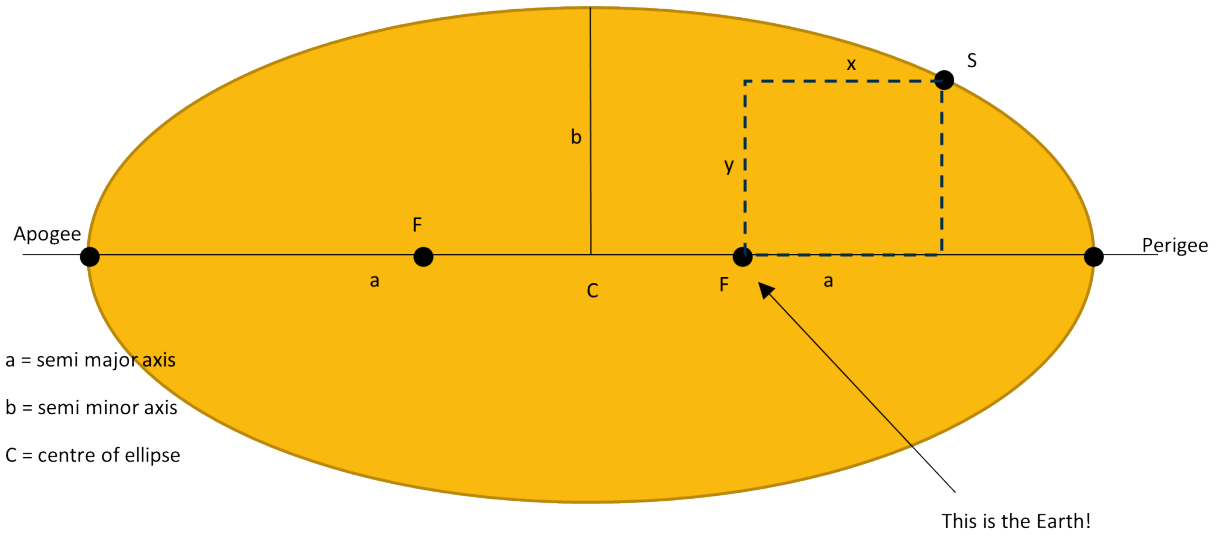


Figure 3.2(c): The position of a satellite in orbit.

Next, we need to determine the angle between the satellite and the Earth, represented by the lowercase Greek letter phi, ϕ, and we'll call the distance from the satellite to the Earth R (keeping in mind its value will change depending on where the satellite is in its orbit), as shown in **Figure 3.2(d).**

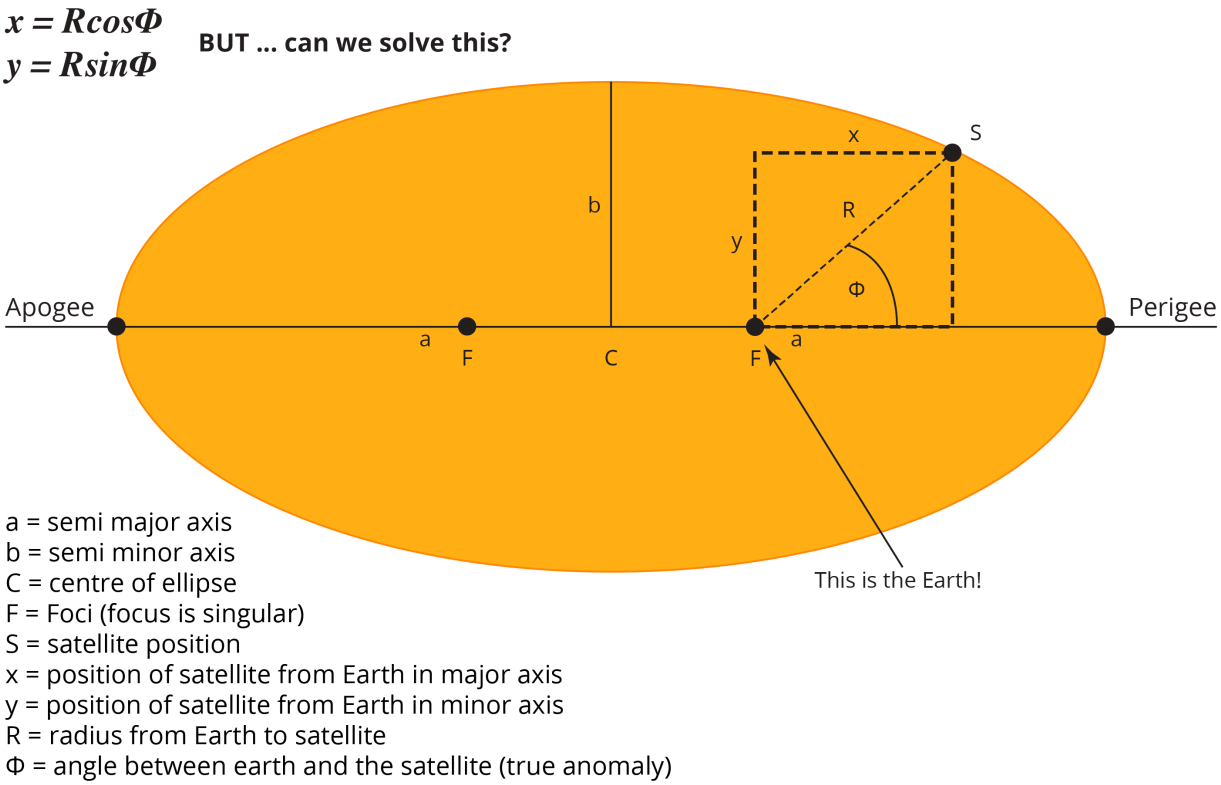


Figure 3.2(d): Describing the position of a satellite in orbit using the true anomaly and R.

We can then use trigonometry to determine x the y values, however, doing the trigonometry on a shape that has a constant radius is significantly easier, so we project the position of the satellite onto a circle with a radius equal to the semi major axis, as shown in **Figure 3.2(e)**. In the calculations on the circle, our x and y values become X and Y as shown.

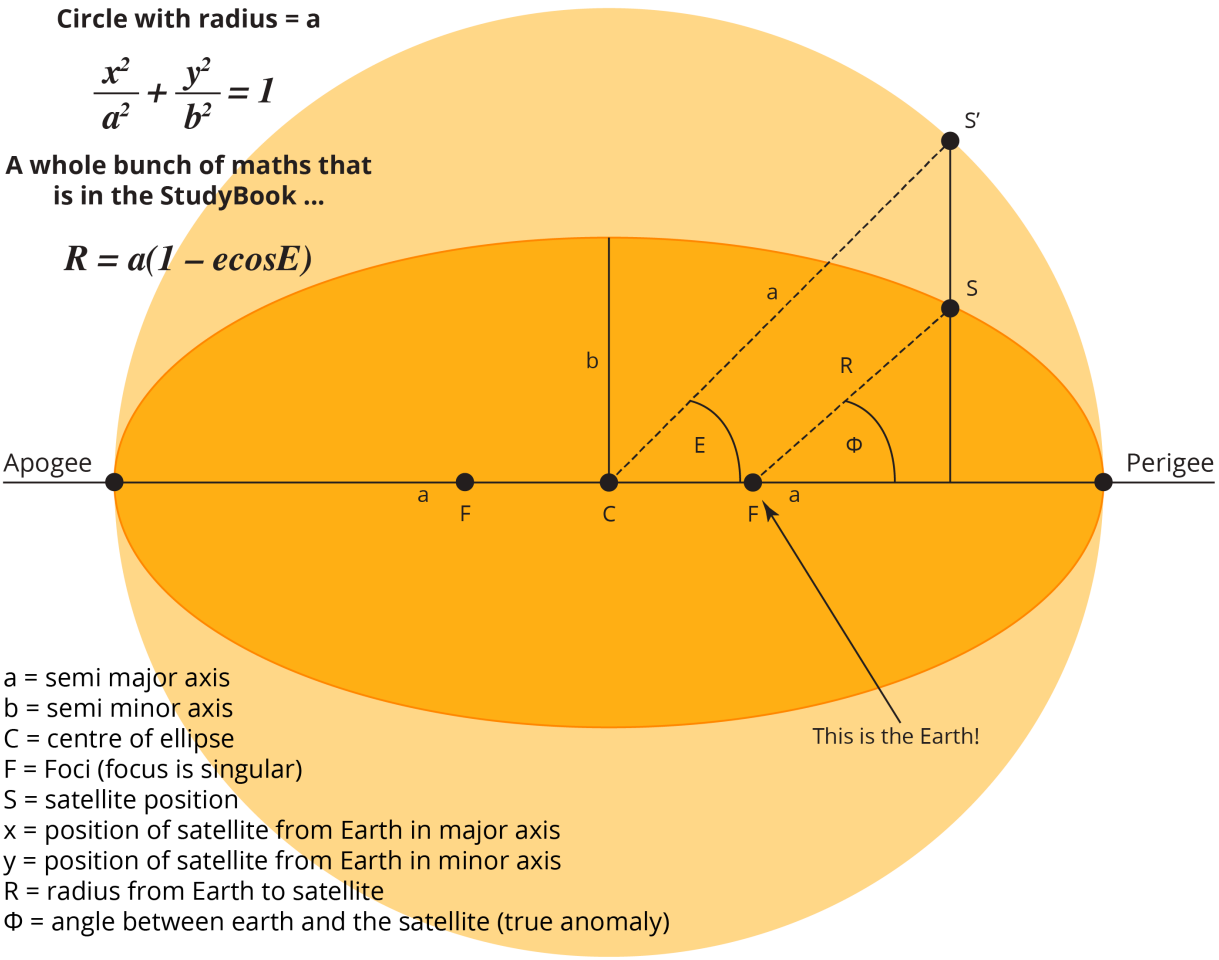


Figure 3.2(e): Determining the motion of a satellite.

In astronomy, the angle between the satellite (or planet) given by ϕ, is called the **true anomaly**, while the angle between the centre of the ellipse and the satellite given by E is called the **eccentric anomaly**. These can be averaged to generate a **mean anomaly**, represented by M.

Considering the eccentricity of the ellipse and the mean anomaly, the version of Kepler's first law that applies for GNSS satellites is:

$$M = E - e\sin E$$

Where M = Mean anomaly
e = eccentricity of the ellipse
E= angle between the centre and S'

You can watch a YouTube clip on how the Law of Orbits works below.

Video 3.2: Kepler's First Law of Motion – Elliptical Orbits [3 mins, 18 secs]

Note: Closed captions are available by clicking on the CC button in the video below.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://usq.pressbooks.pub/gpsandgnss/?p=172#oembed-1

It is worth noting that at the point where the satellite is at the **perigee**, there is more atmospheric friction compared with when the satellite is at the **apogee**. The orbit will tend to become more circular and closer to the earth until eventually the atmospheric friction becomes too large and burns up the satellite. This is important to know when we discuss the satellite orbit information, the ephemeris, later in this chapter.

Kepler's second law – the Law of Areas

The second law is called the **Law of Areas** and showed that for a given time period a satellite will sweep out an equal area, regardless of where it is in its orbit. This in turn proved that the speed of a satellite changes at different points in its orbit, travelling faster when it is near the perigee, and slower near the apogee.

For satellites orbiting the Earth, this law is given by the equation:

$$\frac{1}{2}RV\,sin\delta = \frac{1}{2}\sqrt{GM.a(1-e^2)}$$

Where R = distance from the Earth to satellite
V = speed of the satellite
δ = tangent angle at S.
G = Universal constant
M = mass of the Earth plus the satellite
a = semi – major axis
e = eccentricity

In **Figure 3.2(f)** the time between S_1 and S_2 is the same for both sections of the orbit, and the area scribed out are also equal, despite the distance travelled by the satellite being different.

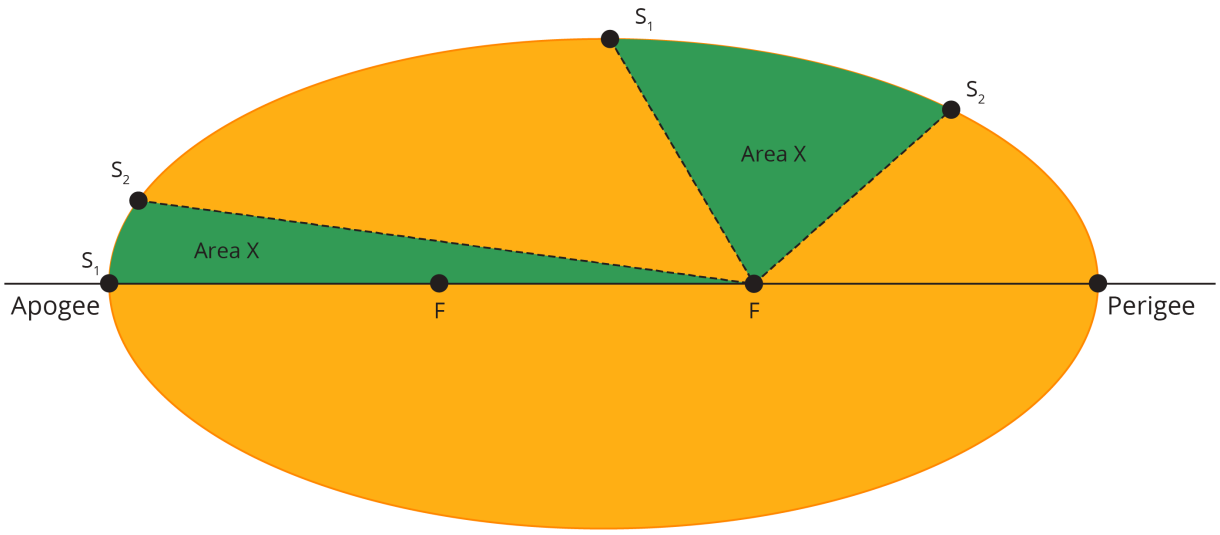


Figure 3.2(f): Kepler's Law of Areas.

You can watch this YouTube clip on how the Law of Orbits works below.

Video 3.3: Kepler's Second Law of Motion – Equal Area in Equal Time [3 mins, 36 secs]

Note: Closed captions are available by clicking on the CC button in the video below.

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://usq.pressbooks.pub/gpsandgnss/?p=172#oembed-2

Kepler's third law – the Law of Periods

The third law of Kepler is also known as the **Law of Periods** and shows the size and shape of the elliptical orbit, the time taken for a satellite to complete an orbit and the mass of the Earth and satellite are all related. The equation states that the square of the orbital period is proportional to the cube of the semi major axis of the orbit. The equation for this law for satellites is:

$$T^2 = \frac{4\pi^2 a^3}{GM}$$

Where T = time to complete a full orbit
a = semi-major axis
G= Universal constant
M= Mass of the Earth and Satellite

You can watch a YouTube clip on how the Law of Periods works below.

Video 3.4: Kepler's Third Law of Motion – Law of Periods [5 mins, 54 secs]

Note: Closed captions are available by clicking on the CC button in the video below.

One or more interactive elements has been excluded from this version of the text. You can view them online here:
https://usq.pressbooks.pub/gpsandgnss/?p=172#oembed-3

3.3 THE POSITION OF SATELLITES

Now we understand how to determine the motion or orbit of a satellite, the next step is to understand how we describe this relative to the Earth. As we learnt in **Chapter 2**, the best way to do this is using a coordinate system. To describe the position of satellites, we use an Earth centred, Earth fixed Cartesian coordinate system.

The system is defined by:

- centre of the Earth is fixed at (0, 0, 0)
- X axis aligns with the zero meridian, which is the first point of Aries, represented by *y*. This is the point of the Sun at the vernal equinox in the northern hemisphere, around 20th – 22nd March.
- Y axis is perpendicular to the X axis in the equatorial plane, also called the plane of the celestial equator
- Z axis is the mean axis of rotation for the Earth.

However, this isn't enough information to fully describe the position of a satellite, as we need to understand how the orbit of a satellite is positioned relative to the Earth as well. We do this by using the six Keplerian elements that define an orbit.

Orbital elements

From Kepler's laws, there are six elements or parameters that are needed to describe the position of a satellite, three of them we already know:

- Eccentricity of the ellipse
- The semi major axis
- The true anomaly

The remaining three are to do with how the satellite orbit is positioned relative to the Earth, and to do this we measure the three different angles between the orbital plane and the Earth's equatorial plane. These are shown in **Figure 3.3(a)** In this figure:

- The **line of nodes** is the line where the two planes intersect. The edges of the orbital plane on this line are points called **nodes**; the **ascending node** and the **descending node**
- The **first point of Aries** is represented by the lowercase Greek letter gamma, . This is the point of the Sun at the vernal equinox in the Northern Hemisphere, around 20th – 22nd March

The three remaining elements are:

- The **inclination** of the orbital plane is the angle between the equatorial plane and the orbital plane, and is represented by i

- The **longitude of the ascending node** is the angle from the vernal equinox to the ascending node in the equatorial plane. It is represented by the uppercase Greek letter omega, Ω
- The **argument of the perigee** is the angle from the line of nodes to the perigee in the orbital plane. It is represented by the lowercase Greek letter omega ω

These are shown in **Figure 3.3(a).**

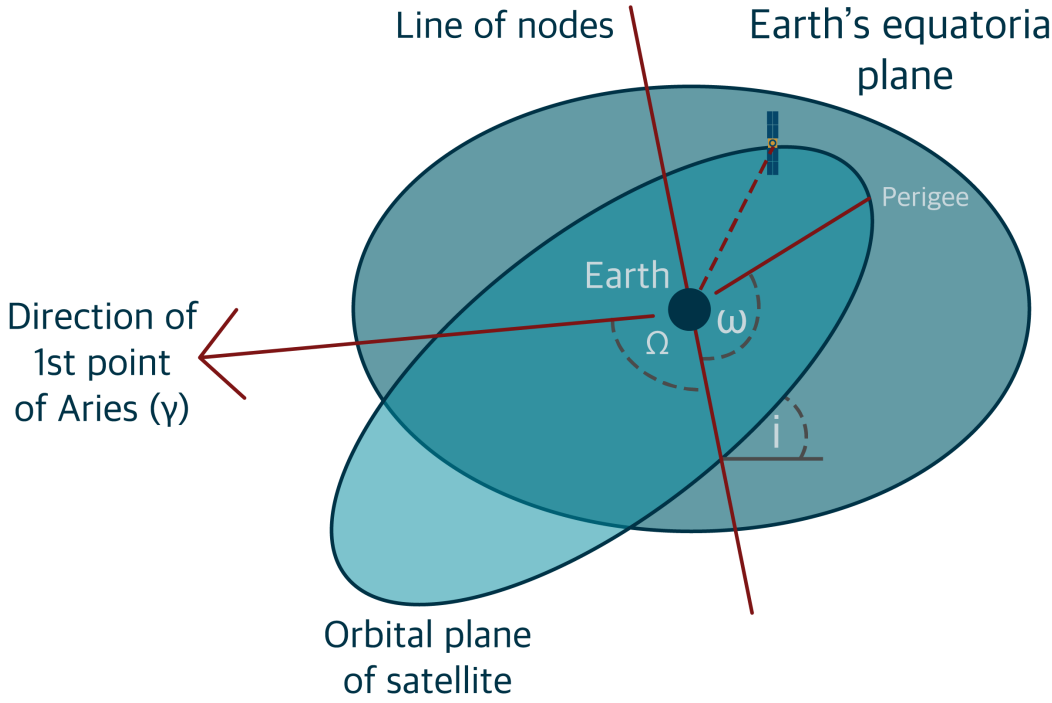


Figure 3.3(a): The Keplerian elements used to describe orbits.

Combining these three parameters with the ECEF Cartesian coordinate system, we are able to describe the position of a satellite around the Earth. **BUT….**

The Earth is an ellipse that isn't homogenous – gravity, air resistance, radiation pressure from the Sun all impact the orbit of satellites. This means the Kepler elements are constantly changing, so our predictions of satellites' orbits have to be constantly updated. This is where the ephemeris comes into play, but more on that later.

Satellite constellations

Now we have a good understanding of how the motion and position of satellites are measured, we can review the three main Global Systems of GNSS and their specific satellite configurations.

These are all correct at the time of writing, however, the number of functional satellites will vary due to decommissioning of older satellites, and the launch of new ones.

GPS

The GPS satellite constellation consists of a minimum of 24 operational satellites arranged in six orbital planes with four satellites in each plane.

Although the nominal GPS operational constellation consists of 24 satellites that orbit the earth in about 12 hours, there are often more than 24 operational satellites as new ones are launched to replace older satellites.

The orbit altitude is approximately 20,000km, and is such that the satellites repeat almost the same ground track and configuration over any point in just short of every 24 hours (actually about four minutes earlier each day).

There are six orbital planes (with nominally four SVs in each), equally spaced (60 degrees apart), and inclined at 55 degrees with respect to the equatorial plane. This constellation often provides the user with between five and eight GPS SVs visible from any point on the earth.

Galileo

The Galileo constellation has 22 operational satellites, with another two used for testing and two that are decommissioned. They are arranged in three orbital planes, with an inclination of 56°. The planes are designated as A, B and C, and each have eight 'slots' for satellites, all of which are currently filled.

Galileo satellites are at an altitude of 23 616km.

GLONASS

The GLONASS constellation has developed in two main phases;

- Initial development through to full constellation deployment happened during the 1970s and 1980s, with a decline in operations in the 1990s after the collapse of the Soviet Union. The satellites launched during this phase are all decommissioned.
- Since the late 1990s and early 2000s, led by the Russian Space Agency. The current constellation has 24 satellites, with an additional satellite under testing and several more commissioned.

The inclination of the GLONASS satellite orbital plane is 64.8° and there are three orbital planes. Satellites orbit at 19 100 km.

3.4 GNSS SIGNALS

GNSS satellites transmit information by radio waves, and there are three main components to the radio wave signals they transmit; the carrier wave, ranging code and navigation data.

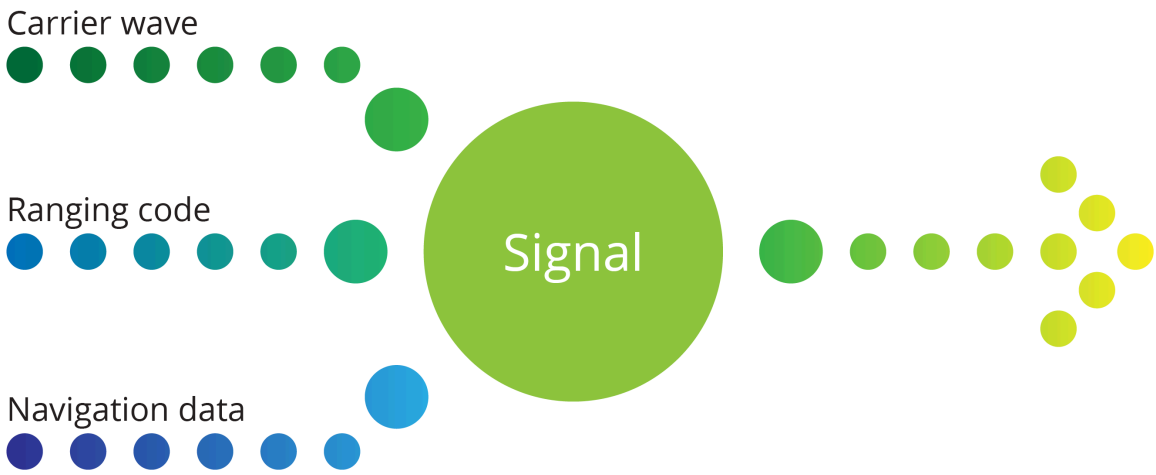


Figure 3.4(a): The three main components of a GNSS signal.

The basics

While this chapter will not go into great depth on how radio waves function, enough of the basics are included for you to understand how GNSS signals are created and what information they contain.

General signal structure

Radio waves are three dimensional, however, they are generally depicted in two dimensions. Radio waves are described by five main physical quantities, as outlined in **Table 3.4(a).**

Table 3.4(a): Physical quantities used to describe radio waves

Quantity	Symbol	Dimension
Circular frequency	f	Cycles per second
Phase	ϕ	cycle
Wavelength	λ	Metres per second
Period	P	Seconds
Speed of light	c	Metres per second

Carrier waves and modulation

A **carrier wave** is simply a wave of a constant frequency, like a sine wave. Carrier waves are kind of like a blank piece of paper, on their own they're not particularly interesting. But like drawing something on a piece of paper, adding an **input signal** to a carrier wave makes it a lot more interesting. The process of adding an input signal to a carrier wave is called **modulation**, and in most GNSS, we modulate a carrier wave with a **code**. This is called Code-division multiple access (**CDMA**). GLONASS is the exception to this – it uses a process called FDMA, however, its newer satellites include signals that use the CDMA technique.

Ranging codes

Ranging codes are binary, meaning they are a string of zeros and ones that represent different messages depending on the sequence of numbers and method of encoding that is used. An example of a ranging code is shown in **Figure 3.4(b).**

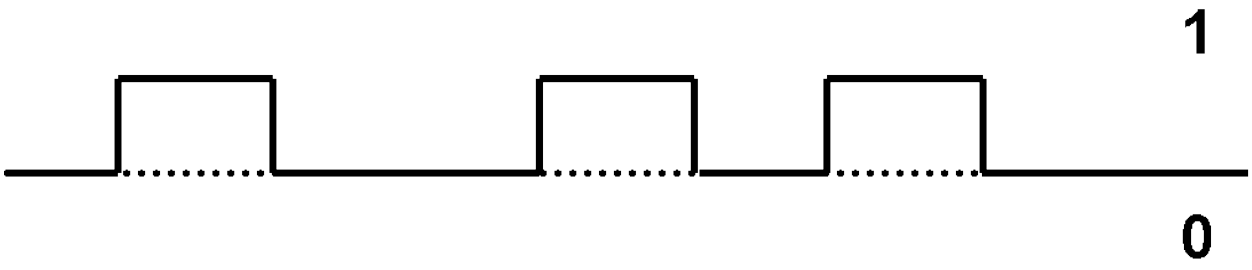


Figure 3.4(b): An example of a ranging code section.

Ranging codes in GNSS are also known as **pseudo-random noise (PRN) ranging codes**, even though they're not random, they appear to lack any definite pattern. PRN ranging codes repeat after a period of time.

Fundamental frequencies

Each GNSS satellite has a number of atomic clocks on board, and these have a **fundamental frequency**. The fundamental frequency is used to generate carrier waves, the PRN ranging code and the navigation message.

Navigation messages

The navigation message is a critical component in any GNSS signal and is unique to each satellite in a constellation. A navigation message contains information about:

- satellite health
- GPS week number
- the atomic clock data and corrections
- orbital data parameters (this is the ephemeris data)
- ionospheric data
- almanac data

Algorithms are inbuilt into most GNSS receivers to use the information in the navigation message to improve the accuracy of their position. The navigation message is maintained and managed by the control segment, who are constantly checking the satellite parameters of operation against the antenna, signal, time clocks, atmospherics, communications and power level information data held by the master station.

Different GNSS measurement processing techniques require the information in the navigation message to deal with different errors – this will be discussed in more depth in Chapters, 4, 5 and 6 when we discuss different observation techniques.

Almanac

As briefly covered in **Chapter 1**, the **almanac** is like a group calendar for all the satellites in a constellation. It contains the approximate orbital parameters of each satellite, and all satellites have the almanac in their navigation message. It is useful for the rapid acquisition of satellites by GNSS receivers.

Ephemeris

The **broadcast ephemeris** is like a personal calendar for each satellite – it provides the orbital parameters of individual satellites. Each satellite has its own ephemeris, which is broadcast in the navigation message, however, each satellite only broadcasts its own ephemeris. The ephemeris contains information including:

- age of ephemeris data
- satellite PRN number
- satellite health status
- current GPS week
- reference epoch
- five of the orbital parameters:
    - semi major axis
    - eccentricity
    - argument of perigee
    - longitude of ascending node at weekly epoch
- the Mean anomaly at reference epoch
- clock phase bias
- clock frequency bias

The orbits of satellites in a GNSS are constantly monitored by the control segment, and once a satellite has finished its orbit, the ephemeris data and other tracking information can be used to produce a **final** or **precise ephemeris.** These **ephemerides** (plural of ephemeris) are produced by organisations like the **International GNSS Service**, and can be used to correct GNSS data to gain very accurate positioning results, and are usually available between 1-2 weeks after the observations are taken.

GNSS signal recipes

So how do all of these things add up to be a GNSS signal?

Each of the global systems have their own set of signals, but essentially they work in the same way. GPS and Galileo, (and some of the newer GLONASS signals) have a fundamental frequency, which generates a carrier wave of a specified frequency, and then a PRN ranging code and a navigation message are modulated on to the carrier wave to create a **modulated carrier wave.** When a change in the code value happens, the carrier wave phase is shifted 180°, as shown in **Figure 3.4(c).**

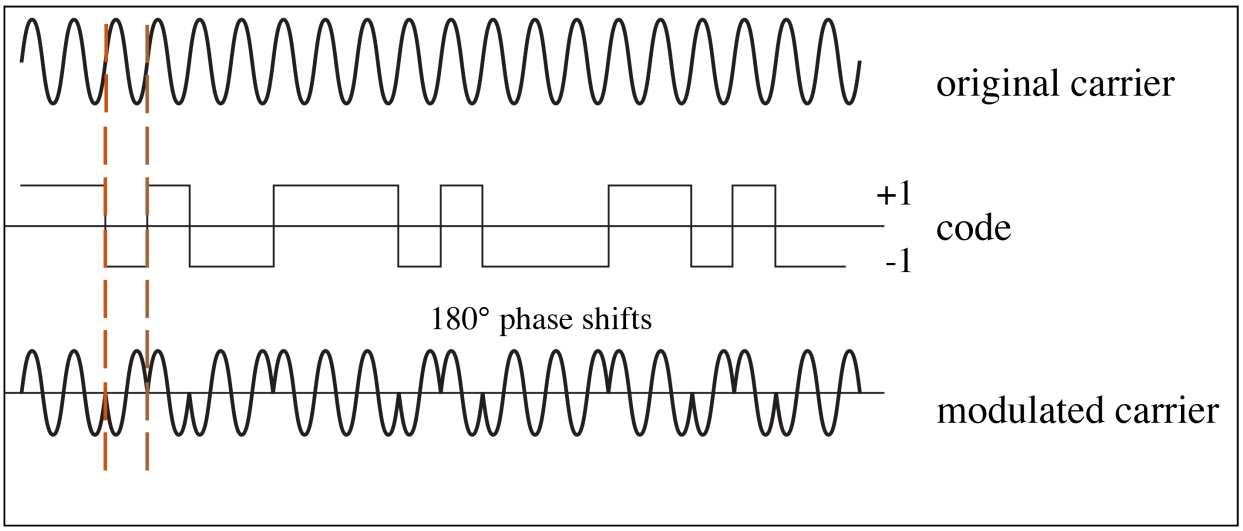


Figure 3.4(c): Carrier wave modulation with the code. Source: Reproduced with permission from Princeton University.

For GPS and Galileo, the fundamental frequency is 10.23MHz, and this is used to generate several different signals. Each satellite has its own unique PRN number, and this is how GNSS receivers are able to determine that data being received is coming from a specific satellite.

GPS signals

GPS has three main signals, **L1, L2** and **L5.**

L1

L1 is the signal used by almost every GNSS receiver chip in existence. It is the GPS signal that contains the PRN ranging code called the **coarse acquisition code**, known as **C/A code**, which is accessible by civilian GNSS receivers. L1 is one of the two original signals.

L1 contains a military code called the **precise code**, known as the **P(Y) code**, which is only accessibly by the US military**.** It is actually the P code encrypted with the W code to make the Y code, however, it's most commonly known as the P(Y) code.

L1 uses the **Civilian navigation data message**, known as **CNAVDATA**, which is a code that is transmitted at 50 **bits per second (BPS)**.

L1C is a newer version of the L1 signal, but it broadcast on the same frequency as L1. It contains the C/A code and the P(Y) code, and 3 new components:

- L1C code – the updated version of the C/A code
- M code – Military code, the new military code
- CNAV-2 – the new navigation data message format for GPS, transmitted at 100BPS

L2

L2 was one of the original signals along with L1, however, it was largely a military signal to begin with, only having the P(Y) code and the NAVDATA message (the precursor to CNAVDATA). People quickly figured out that by using information in the L2 NAVDATA and components of the P(Y) code that using L2 with L1 allowed significantly more accurate positions to be determine. It is on this basis that the majority of high-level accuracy GNSS surveying is able to be undertaken.

L2C is an updated version of L2, and like L1 & L1C, is broadcast on the same frequency as L2. It now contains a civilian code, L2C (made up of CM and CL codes), as well as the M, P(Y) and CNAVDATA codes.

L5

L5 is what is referred to as a **safety of life** signal, as it is broadcast in the section of the radio spectrum that is used by aviation safety services, however, it is available for all civilian users. It contains the I5, Q5 and the CNAVDATA codes.

The process of generating the 5 GPS signals from the fundamental frequency is shown in **Figure 3.4(d)**, showing the frequency multipliers and dividers as well as the factors. A summary table is also provided in **Table 3.4(b).**

Table 3.4(b): Summary of GPS signals

Signal	**L1**	**L1C***	**L2**	**L2C***	**L5**
Fundamental Frequency factor	X154	X154	X120	X120	X115
Frequency (MHzs)	1575.42	1575.42	1227.60	1227.60	1176.45
Wavelength	19cm	19cm	24cm	24cm	25cm
Code Info	L1	L1C*	L2	L2C*	L5
Civilian	C/A Code	C/A Code L1C		CM Code CL Code (L2C Code)	I5 Code Q5 Code
Military	P(Y) Code	M Code	P(Y) Code	M Code	
Navigation Message	CNAVDATA	CNAV-2	CNAVDATA	CNAVDATA	CNAVDATA

*L1C and L2C broadcast on the same frequencies as L1 and L2 respectively

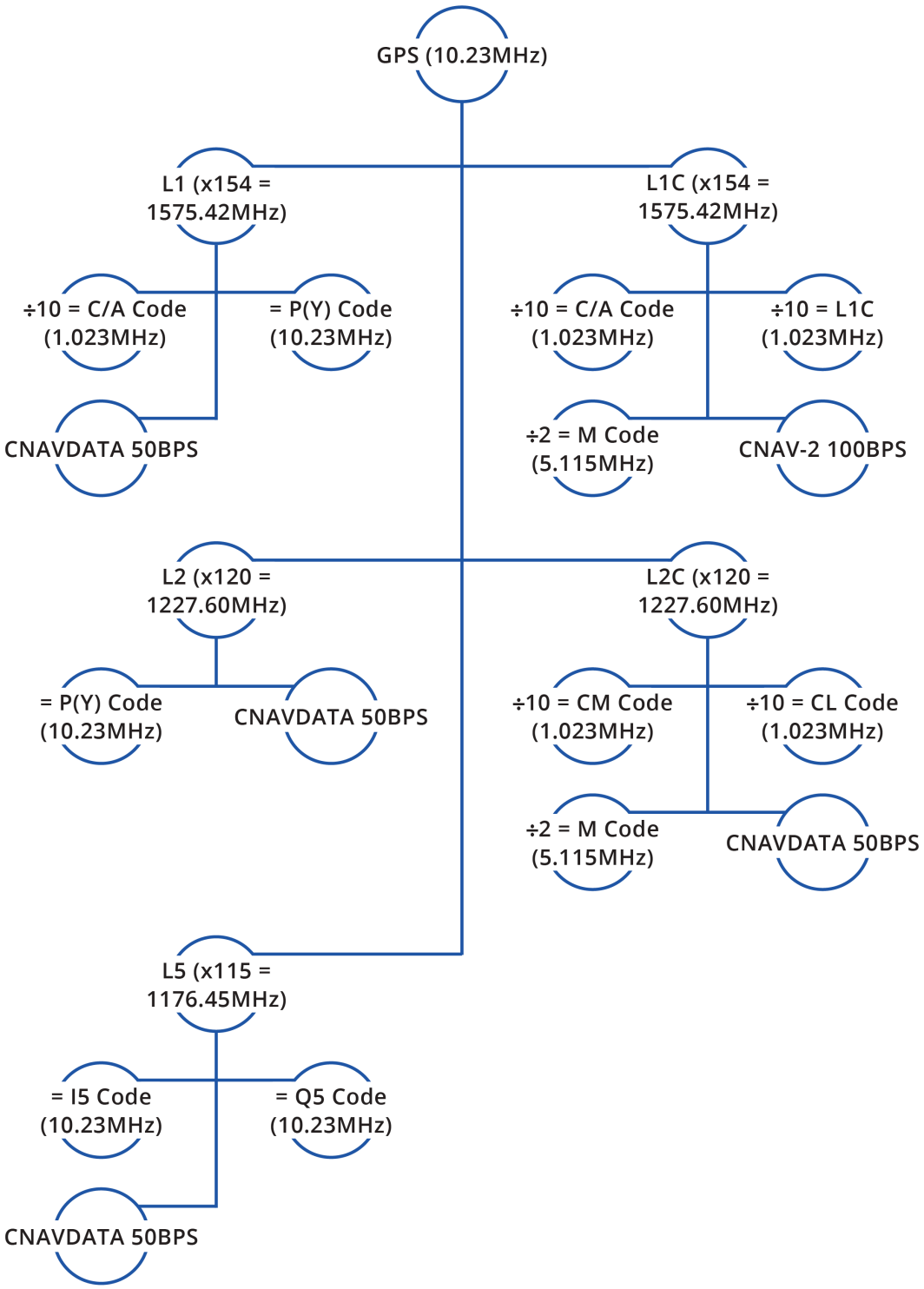


Figure 3.4(d): GPS signal combinations structure.

Galileo signals

Galileo has four signals E1, E6, E5a and E5b. Because it is a civilian operated system, has no military codes, however, it has a Public Regulated Service (PRS) code that is encrypted for governmental authorised users and sensitive applications.

The E1 and E5a signals overlap with the GPS L1 and L5 signals.

The E1 signal has two codes, the E1 Open Service code, and the PRS code.

The E6 signal has the Commercial Service code that is only accessible through paid services.

The E5a signal is similar to the L5 signal – it is primarily considered a safety of life signal.

The process of generating the four Galileo signals from the fundamental frequency for is shown in **Figure 3.4(e)**, showing the frequency multipliers and dividers as well as the factors.

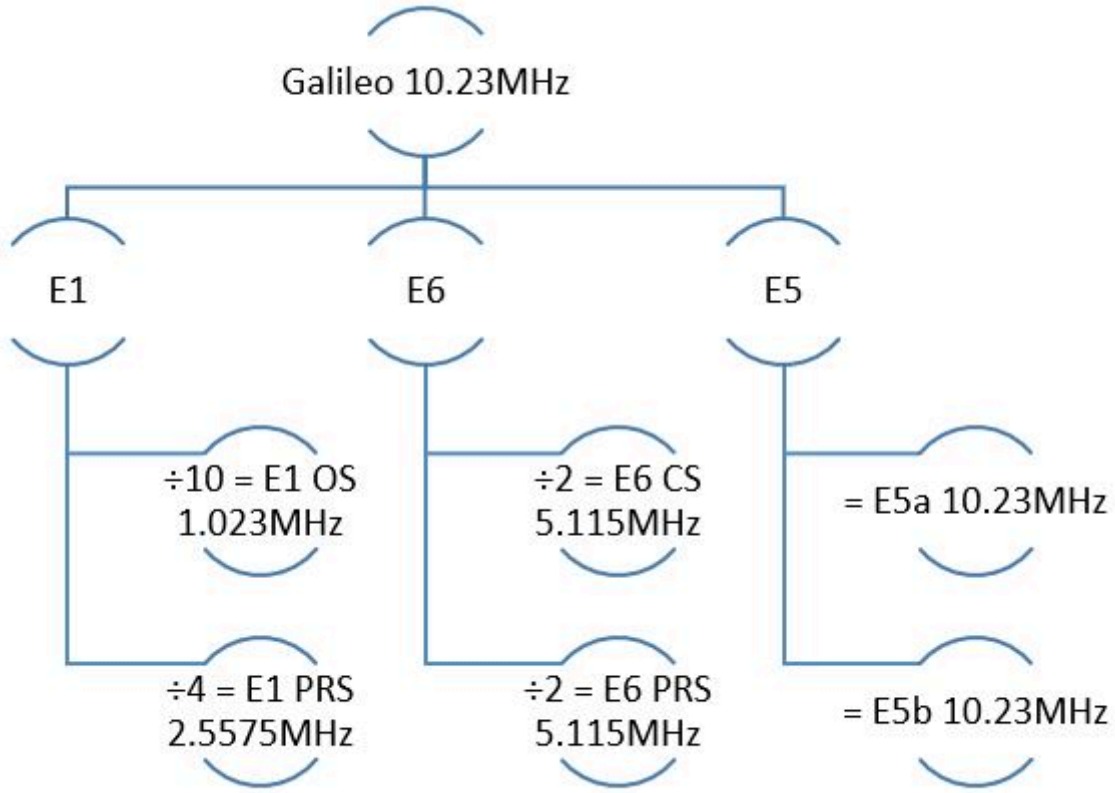


Figure 3.4(e): Galileo signal combinations structure.

GLONASS signals

As mentioned previously, GLONASS signals are a combination of CDMA and FDMA signals. An outline of the main signals in each type is provided in **Figure 3.4(f)**.

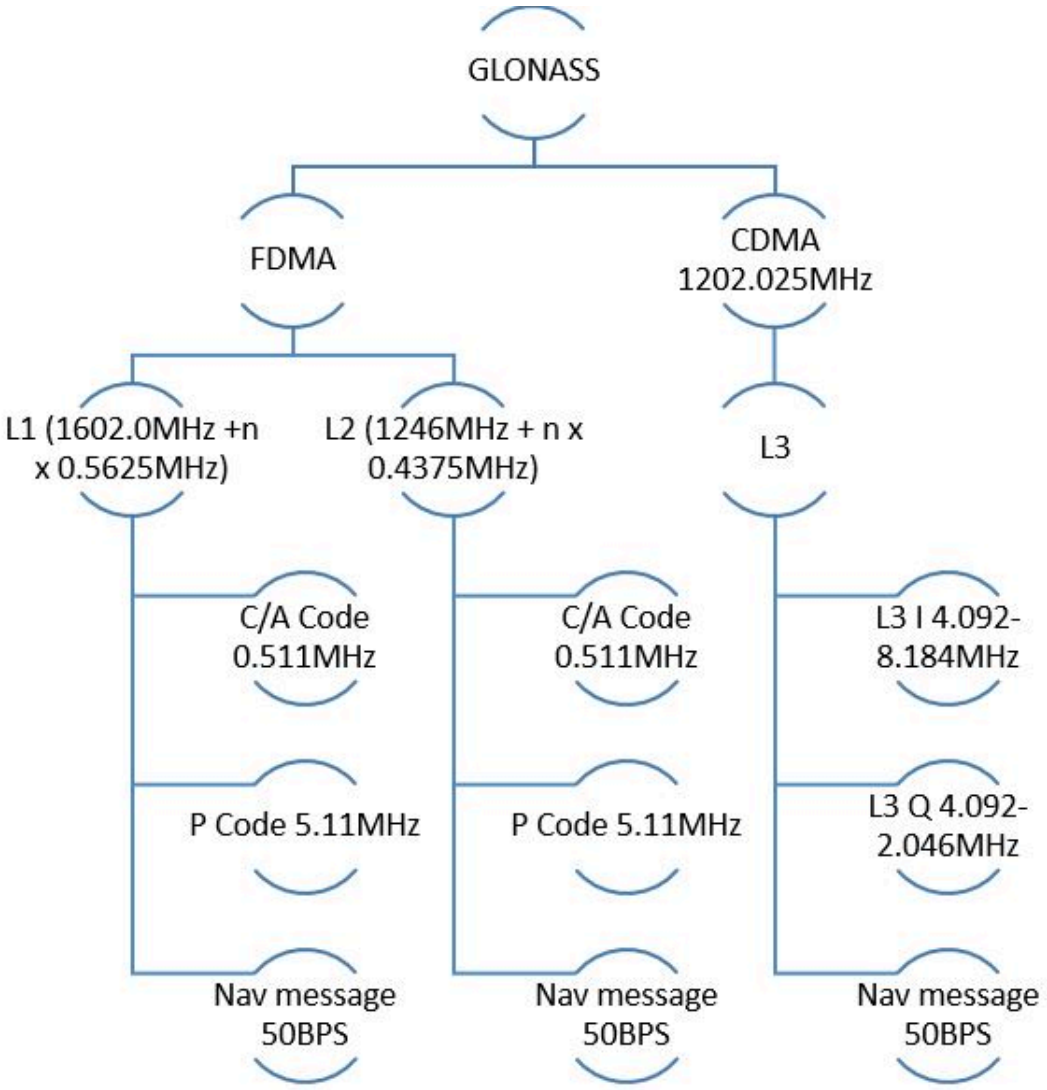


Figure 3.4(f): GLONASS signal structure.

GNSS Time

All GNSS relies on accurate measurement of the distance between a GNSS receiver and satellites, which requires incredibly accurate measurements of the time it takes a signal to travel from a satellite to a receiver – remember that GNSS satellites are orbiting at approximately 20 000 km, and radio waves travel at the speed of light, which is 299 792 458 metres per second! This means GNSS signals take between 60 and 90 milliseconds to reach Earth. For any GNSS to work properly, all the satellites must have their atomic clocks synchronised.

Coordinated Universal Time (UTC) is the standard the world uses to regulate time, and GNSS link their timing systems to UTC at different epochs.

GNSS Time is generally expressed in seconds, for example GPS time is expressed as a week number and then the number of seconds into the week it is.

The different global systems use different types of atomic clocks, and most satellites will have multiple types of clocks on board. Rubidium and Caesium are standard on GPS SVs, Galileo has Hydrogen maser and Rubidium.

3.5 ERRORS IN GNSS

The types of errors in GNSS will be discussed in more detail in Modules 4-6 when discussing specific observation techniques, however, it is important to have an overview of the different errors that do happen in GNSS with respect to the fundamentals learnt in Modules 1, 2 and 3. An understanding of these errors is also important in understanding the different kinds of GNSS measurement techniques available, as each attempts to make corrections for different types of errors, leading to different levels of accuracy.

Errors in GNSS, as shown in **Figure 3.5(a)**, can be grouped into three main categories: satellite, signal and receiver errors. Some errors are combinations of more than one area, however, they have been listed in the one that is the primary influence. Human error is listed with receivers, however, the capacity for us to cause errors in GNSS data extends far beyond using a receiver incorrectly or poorly!

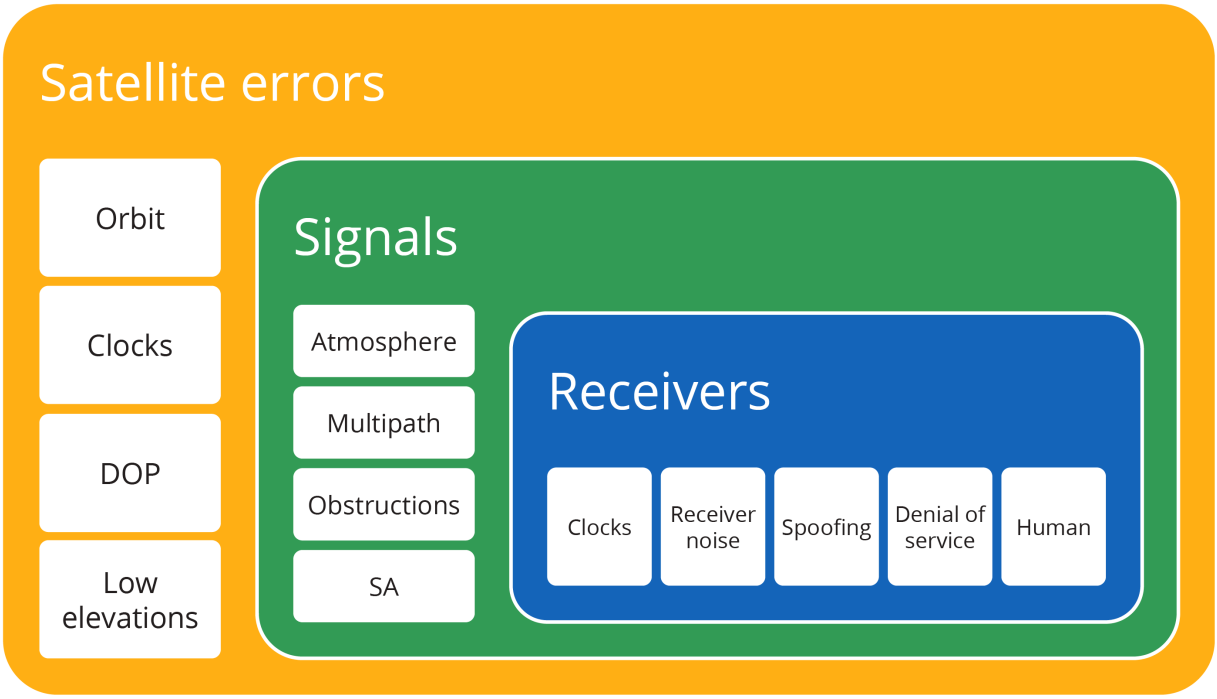


Figure 3.5(a): Errors in GNSS positioning

Some errors are specific/unique to a particular site or to a particular receiver, and these errors are classified as location and receiver errors respectively. The other kinds of errors are common to all GNSS receivers in the same area, and are referred to as system wide errors. **Figure 3.5(b)** outlines how the errors above fall into each category.

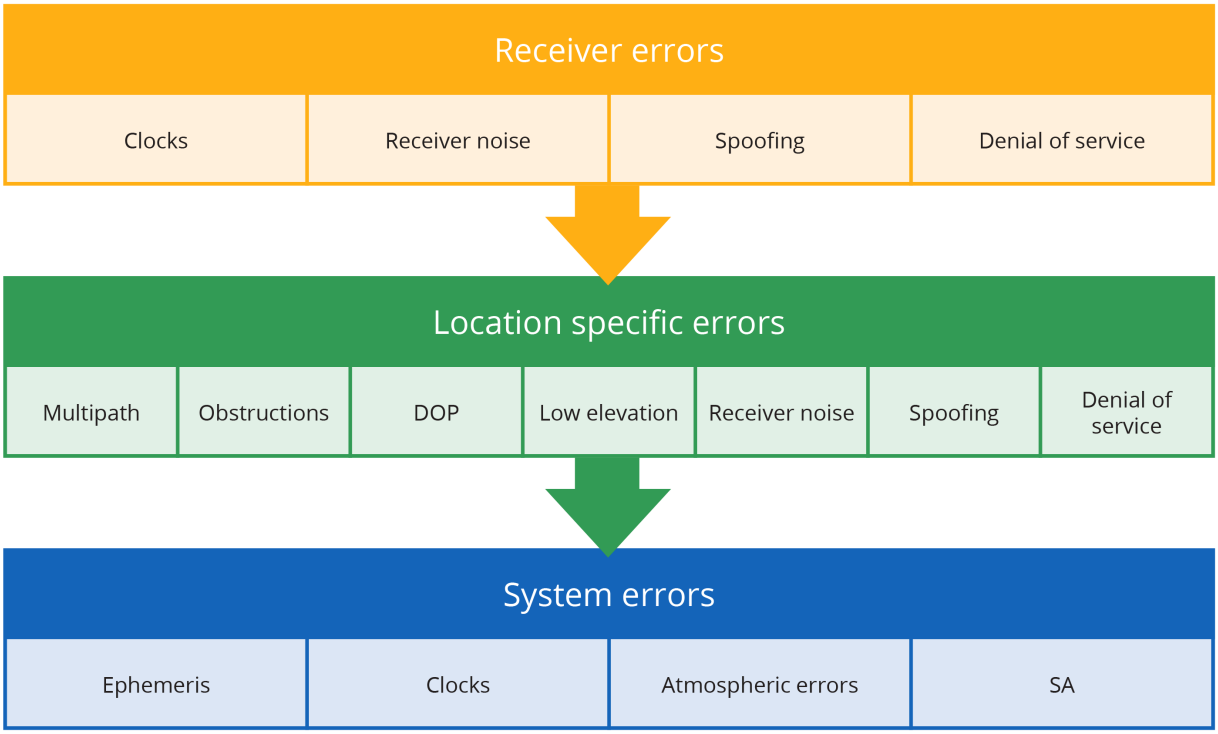


Figure 3.5(b): Error categories for GNSS positioning

Satellite errors

Orbit errors

As discussed in previous sections of this chapter, the orbital parameters of a satellite need to be constantly monitored and updated by the control segment due to orbit error, as shown in **Figure 3.5(c)**. This means that positions calculated using broadcast ephemerides will be less accurate than those calculated using final ephemerides.

Some GNSS surveying techniques require the final ephemerides to be used to guarantee the best solution for a position. Usually these surveys have observations of longer than an hour on multiple permanent marks.

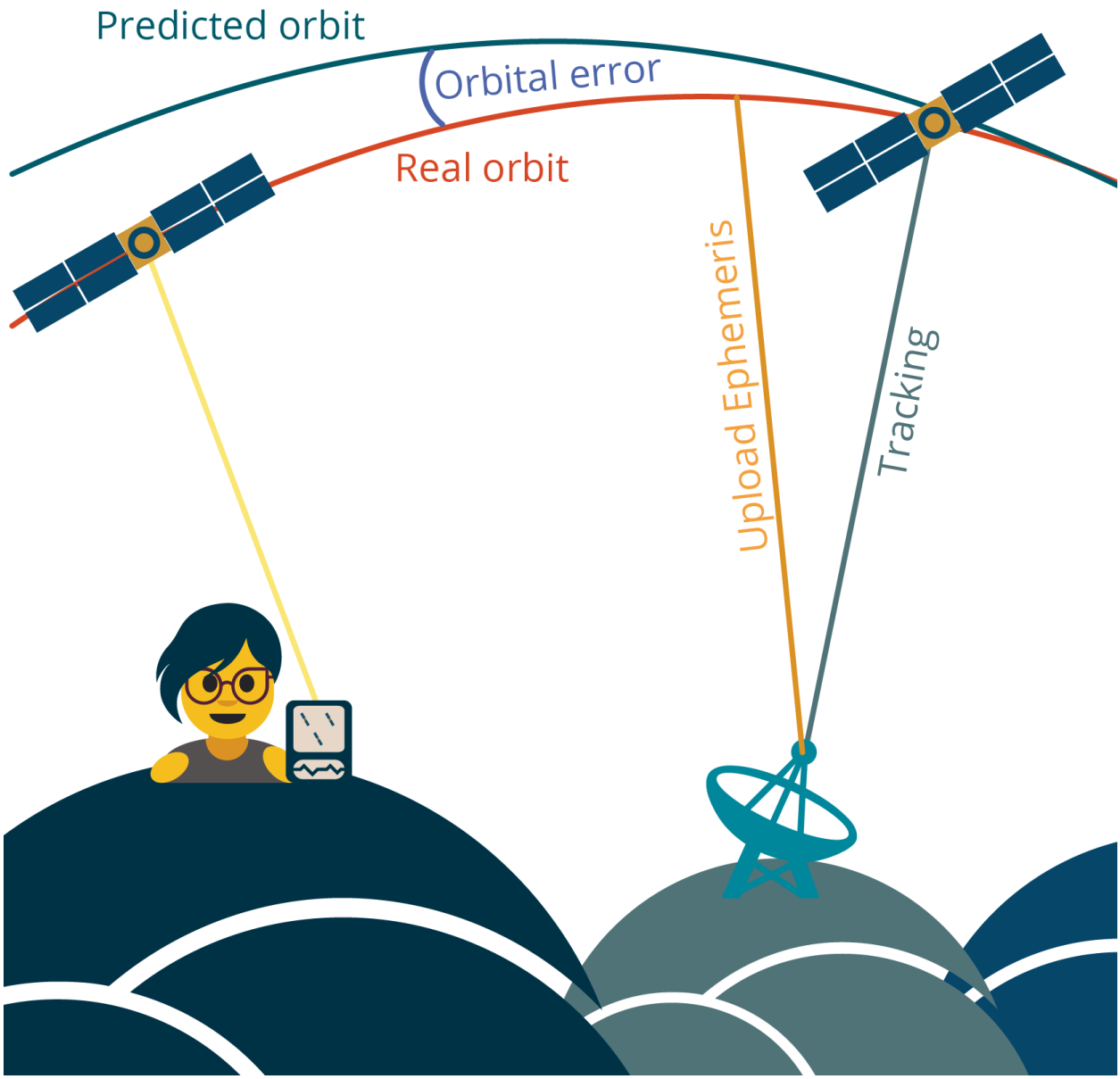


Figure 3.5(c): Orbit error

Satellite clock errors

As previously mentioned, the different global systems use different types of atomic clocks, and most satellites will have multiple types of clocks on board. Rubidium and Caesium are standard on GPS SVs, Galileo has Hydrogen maser and Rubidium.

These clocks are incredibly accurate, but they aren't perfect, and experience 'drift'. This means they lose or gain one nanosecond for every three hours of time.

The drift of atomic clocks is well documented and understood in GNSS, so adjustments are regularly made by the control segment, and these corrections are included in the navigation data message broadcast by each satellite.

Dilution of precision

Note: While DOP is not technically solely a satellite error (it is more correctly an error relating to satellite geometry and the measurements taken to satellites), it is included in this category for completeness.

In the early day of GPS, there were limited numbers of satellites, so the minimum standard for an ideal satellite configuration was determined – four satellites spaced evenly at 90 degree intervals around the horizon, at 45 degree elevation. Any variation from this configuration can be calculated, and is known as **dilution of precision**, more commonly referred to as **DOP**.

DOP can vary in value from 1 (good) to infinity (very bad), where anything below four is generally consider sufficient for collecting GNSS positions. There are several variations of DOP:

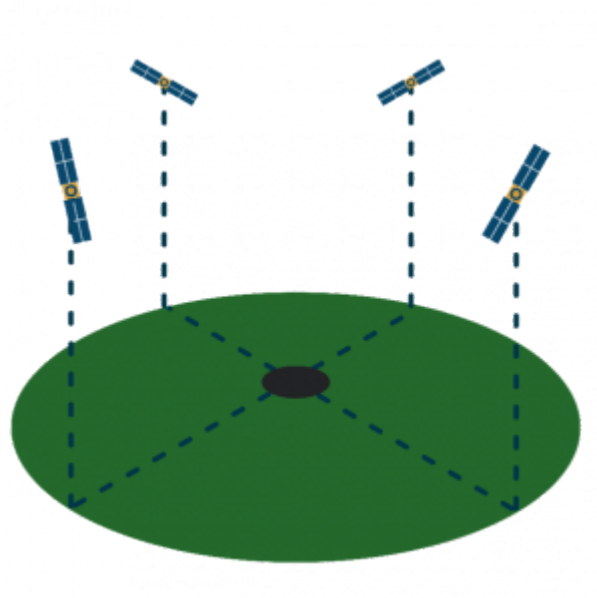


Figure 3.5(d): Ideal satellite configuration

- PDOP – Position dilution of precision. This is still the most commonly used DOP value
- HDOP – Horizontal dilution of precision
- VDOP – Vertical dilution of precision

The use of DOP values is largely redundant since the addition of satellite constellations beyond GPS, but many receivers that only track GPS SVs will continue to calculate and record PDOP values.

Low elevations

Once satellites become low on the horizon, it is more likely that their signals will suffer from atmospheric errors and multipath (see the next section for these errors), and thus it is better if these satellites are not included in our measurements.

To prevent the signals from these satellites being included in measurements, GNSS receivers have a function called an **elevation mask** which prevents the low elevation satellites from being included once they get below a certain angle to the horizon. The elevation mask angle can be adjusted depending on project requirements.

Signal errors

Atmospheric errors

Radio waves behave in similar ways to light waves, and when they pass through the different layers of the atmosphere they are bent, or refracted, from their original path. This causes delays in the GNSS signal reaching the receiver, and thus a positioning error, as shown in **Figure 3.5(e)**. Atmospheric delays are the largest during the heat of the day, and during sunspot activity or solar flares.

The troposphere is approximately 50km 'thick' and is the layer of the atmosphere that contains water vapour, and is where our local weather is generated. Generally, each receiver will have a troposphere model that is able to correct for this error.

The ionosphere sits above the troposphere, and is about 200km 'thick'. It contains charged particles, called **ions**, and the level of activity of the ions impacts GNSS signals.

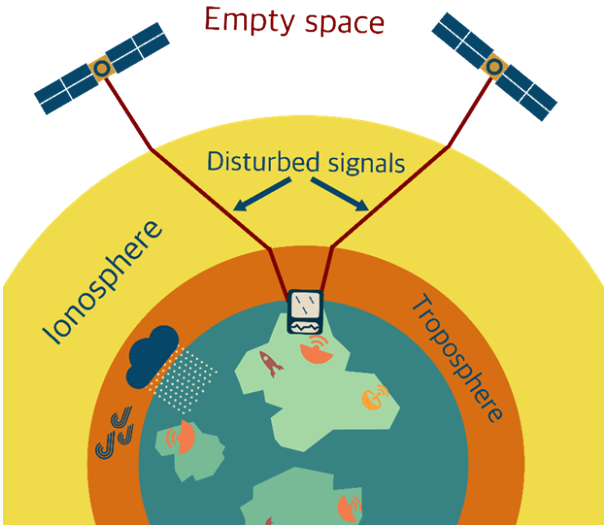


Figure 3.5(e): Atmospheric interference of signals

Over shorter distances and areas, atmospheric delays can generally be corrected, however, when using particular techniques of GNSS data collection that use multiple GNSS receivers over longer distances, the atmospheric errors may need to be modelled at more sophisticated levels. Modelling the troposphere is usually reasonably simple, however, the ionosphere is highly variable and more difficult to model. Using dual frequency GNSS receivers; ones that can process two GNSS signals at a time, is beneficial in reducing atmospheric errors.

Multipath

Multipath occurs when the signal doesn't make it directly to the GNSS receiver, but is reflected off other surfaces first, as shown in **Figure 3.5(f)**. This could be buildings, trees, vehicles or something else nearby the receiver. Similar to the issues of the signal being influenced by the atmosphere, the issue with multipath is the delay in the signal reaching the receiver, and the positioning error this causes due to the time delay.

Multipath is increasingly being dealt with by the signals processing technology in the GNSS receiver's antenna, however, situations where multipath could be generated should be avoided where possible.

Figure 3.5(f): Multipath error

Obstructions

Radio waves are blocked by physical obstructions such as mountains, buildings or even vegetation. This prevents the GNSS signals making it to the receiver, or if they do, it is likely they have experienced some level of multipath.

Selective Availability

Selective availability (SA) was an intentional degradation of the GPS signals by the US military to prevent hostile forces using GPS for positioning. With SA turned on, the horizontal accuracy of measurements was about 100m, while vertical was +/- 200m. Once it was turned off in early 2000, horizontal and vertical accuracy was approximately +/- 20m.

The US has committed to not including SA capacity on its future generations of satellites, however, the possibility of it happening again can't be discounted as GPS is a military system.

Receiver errors

Receiver clock errors

While satellites have expensive and very accurate atomic clocks, it would be impractical (and too expensive!) to equip each GNSS receiver with an atomic clock. GNSS receivers have an inexpensive, good quality quartz clock that has a drift of about 1 000 nanoseconds every second.

While using quartz clocks in receivers could seem to be an issue, receivers are able to adjust their time using the GPS system time when they are receiving data from satellites.

Receiver noise

All GNSS receivers are subject to other radio waves that occur naturally or through other processes. These other radio waves, along with other types of electrical interference are what is known as **background noise,** and impact the ability of the receiver to 'hear' the GNSS signals.

This is best understood through an example.

Imagine if you're having a conversation with a friend, and then someone starts blasting music out of a speaker at twice the volume that you're speaking. You can no longer hear your friend talking (the signal), because the strength of the noise coming out of the speaker (the background noise) is now greater than your friend's voice.

This imbalance of the signal to the background noise is called the **signal to noise ratio (SNR)** and for GNSS is the strength of the GNSS signal to the background noise. If the SNR is large (generally greater than 10), the signal is strong compared to the background noise. If the SNR for a GNSS receiver drops below 6, this would usually generate an error and stop the receiver collecting that signal.

Spoofing

GNSS signal **spoofing** is when a false GNSS signal is sent to a GNSS receiver to make it think it is somewhere different to where it is. While spoofing is illegal in most countries, it is still prevalent in some parts of the world. GPS L1 is the signal most likely to be spoofed, and spoofed data can usually be detected at close inspection – often the almanac and other typical GNSS signal characteristics are missing. Most new GNSS signals are being designed to prevent spoofing attacks.

Denial of service

Most commonly called **jamming**, denial of service is when a GNSS receiver is prevented from receiving GNSS signals through intended interference.

Human error

The greatest **potential** source of error in any GNSS observations are humans. We might measure the wrong point, use the wrong datum, write coordinates down incorrectly, observe for too short of a time or some other error in our process. To combat this we use multiple systems of checks in our processes, as well as building in measurement redundancy to our work.

3.6 GNSS ACCURACY

The two components of GNSS signals allow us to generate positioning information in different ways, as shown in **Figure 3.6(a).**

The code section of the signal, referred to as the **code observable**, gives us the first two levels of accuracy – point positioning and DGPS:

- **Point positioning –** this is normally done with a device that has a low cost single frequency GPS chip. It can achieve an accuracy of ±10m relatively quickly, and uses **code observations**, which are discussed in **Chapter 4.**
- Differential range positioning, known more commonly as **Differential GPS**, or **DGPS.** This technique also uses low cost, single frequency receivers. One receiver is set up on a known point – this is called the **base station**, while the other receiver is used to collect positions of points. The information collected at both sites is used to determine corrections for the GNSS errors that can be applied to the collected data in **real time**, or afterwards, called **post processing**. DGPS is generally accurate to 0.5-5m relative to the known point. This method also uses code observations, and is discussed in **Chapter 5**.

The carrier section, referred to as the **phase observable,** gives us the last level:

- High precision GNSS surveying – This technique requires more sophisticated multiple frequency receivers, collecting phase data from the carrier wave part of the signal from satellites. This data is used to determine the three dimensional positions of multiple points relative to each other, called **baselines**. These baselines can have precisions of better than five parts per million, giving us millimetre to centimetre accuracies. This is discussed in **Chapter 6**.

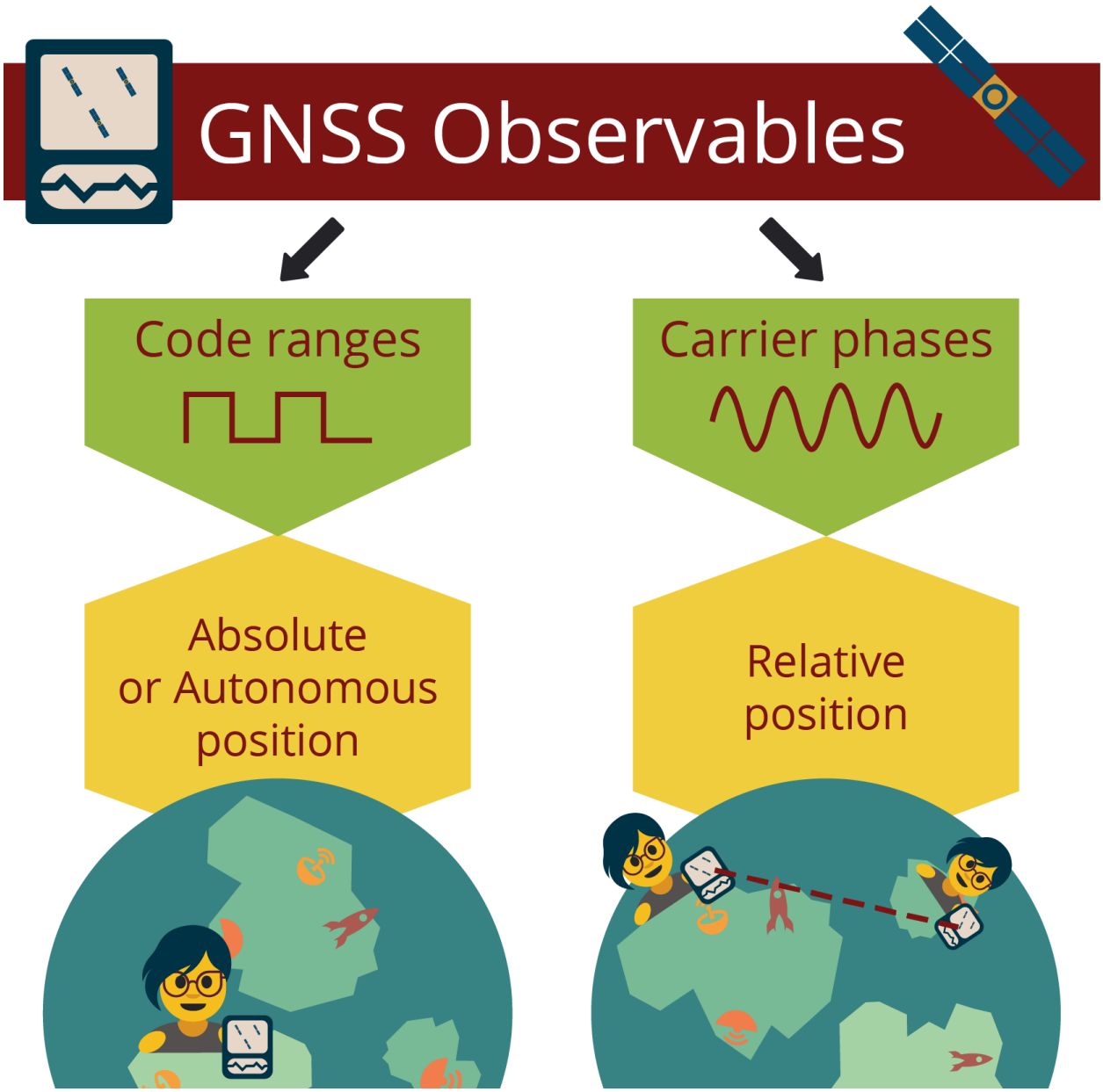


Figure 3.6(a): GNSS observables

PART IV

CODE OBSERVABLE

4.1 CODE OBSERVABLE

Learning Objectives

On the successful completion of this chapter you should be able to:

- explain what point positioning is
- explain the general theory behind code pseudo-ranges
- list and explain each of the 5 steps for code pseudo-range/point positioning
- explain the errors associated with point positioning and how to minimise them.

In the beginning...

01010111 01101000 01100101 01101110 00100000 01000111 01010000 01010011 00100000
01100110 01101001 01110010 01110011 01110100 00100000 01100010 01100101 01100011
01100001 01101101 01100101 00100000 01100001 01110110 01100001 01101001 01101100
01100001 01100010 01101100 01100101 00100000 01100110 01101111 01110010 00100000
01100011 01101001 01110110 01101001 01101100 01101001 01100001 01101110 00100000
01110101 01110011 01100101 00101100 00100000 01110100 01101000 01100101 00100000
01101101 01100001 01101001 01101110 00100000 01101101 01100101 01110100 01101000
01101111 01100100 00100000 01101111 01100110 00100000 01110000 01101111 01110011
01101001 01110100 01101001 01101111 01101110 01101001 01101110 01100111 00100000
01110101 01110011 01100101 01100100 00100000 01110100 01101000 01100101 00100000
01100011 01101111 01100100 01100101 00100000 01101111 01100010 01110011 01100101
01110010 01110110 01100001 01100010 01101100 01100101 00100000 01100011 01101111
01101101 01110000 01101111 01101110 01100101 01101110 01110100 00100000 01101111
01100110 00100000 01110100 01101000 01100101 00100000 01110011 01101001 01100111
01101110 01100001 01101100 00101100 00100000 01110100 01101000 01100101 00100000
01100010 01101001 01101110 01100001 01110010 01111001 00100000 01000011 00101111
01000001 00100000 01100011 01101111 01100100 01100101 00101100 00100000 01110100
01101111 00100000 01100111 01100101 01110100 00100000 01100001 01101110 00100000
01100001 01110101 01110100 01101111 01101110 01101111 01101101 01101111 01110101
01110011 00100000 00101000 01101111 01110010 00100000 01100001 01100010 01110011
01101111 01101100 01110101 01110100 01100101 00101001 00100000 01110000 01101111
01110011 01101001 01110100 01101001 01101111 01101110 00100000 01110111 01101001
01110100 01101000 00100000 01100001 00100000 01000111 01010000 01010011 00100000
01110010 01100101 01100011 01100101 01101001 01110110 01100101 01110010 00101110
00100000 01001001 01110100 00100000 01110111 01100001 01110011 01101110
10000000011001 01110100 00100000 01110000 01100001 01110010 01110100 01101001
01100011 01110101 01101100 01100001 01110010 01101100 01111001 00100000 01100001
01100011 01100011 01110101 01110010 01100001 01110100 01100101 00100000 00101000
01100010 01100101 01100011 01100001 01110101 01110011 01100101 00100000 01101111

01100110 00100000 01110011 01100101 01101100 01100101 01100011 01110100 01101001
01110110 01100101 00100000 01100001 01110110 01100001 01101001 01101100 01100001
01100010 01101001 01101100 01101001 01110100 01111001 00101001 00100000 01100010
01110101 01110100 00100000 01101001 01110100 00100000 01110111 01100001 01110011
00100000 01110011 01110100 01101001 01101100 01101100 00100000 01110100 01101000
01100101 00100000 01100110 01100001 01110011 01110100 01100101 01110011 01110100
00100000 01110111 01100001 01111001 00100000 01100001 01101110 01111001 01101111
01101110 01100101 00100000 01100011 01101111 01110101 01101100 01100100 00100000
01100111 01100101 01110100 00100000 01100001 00100000 00110011 01000100 00100000
01110000 01101111 01110011 01101001 01110100 01101001 01101111 01101110 00100000
01110101 01110011 01101001 01101110 01100111 00100000 01110010 01100001 01100100
01101001 01101111 00100000 01110111 01100001 01110110 01100101 01110011 00101110
0001010 01010100 01101000 01101001 01110011 00100000 01100110 01100001 01110011
01110100 00100000 00110011 01000100 00100000 01110000 01101111 01110011 01101001
01110100 01101001 01101111 01101110 01101001 01101110 01100111 00100000 01110111
01100001 01110011 00100000 01110000 01101111 01110011 01110011 01101001 01100010
01101100 01100101 00100000 01100010 01100101 01100011 01100001 01110101 01110011
01100101 00100000 01110100 01101000 01100101 00100000 01000111 01010000 01010011
00100000 01110010 01100101 01100011 01100101 01101001 01110110 01100101 01110010
01110011 00100000 01100001 01101100 01110011 01101111 00100000 01101000 01100001
01100100 00100000 01100001 00100000 01100011 01101111 01110000 01111001 00100000
01101111 01100110 00100000 01100001 01101100 01101100 00100000 01110100 01101000
01100101 00100000 01100100 01101001 01100110 01100110 01100101 01110010 01100101
01101110 01110100 00100000 01110011 01100001 01110100 01100101 01101100 01101100
01101001 01110100 01100101 00100000 01100011 01101111 01100100 01100101 01110011
00101100 00100000 01100001 01101110 01100100 00100000 01110100 01101000 01100101
01111001 00100000 01100011 01101111 01110101 01101100 01100100 00100000 01100110
01101001 01100111 01110101 01110010 01100101 00100000 01101111 01110101 01110100
00100000 01101000 01101111 01110111 00100000 01101101 01110101 01100011 01101000
00100000 01101111 01100110 00100000 01100001 00100000 01100100 01100101 01101100
01100001 01111001 00100000 01110100 01101000 01100101 01110010 01100101 00100000
01110111 01100001 01110011 00100000 01100010 01100101 01110100 01110111 01100101
01100101 01101110 00100000 01110111 01101000 01100101 01101110 00100000 01110100
01101000 01100101 01111001 00100000 01110010 01100101 01100011 01100101 01101001
01110110 01100101 01100100 00100000 01110100 01101000 01100101 00100000 01100011
01101111 01100100 01100101 00100000 01100001 01101110 01100100 00100000 01110111
01101000 01100101 01101110 00100000 01110100 01101000 01100101 01111001 00100000
01110100 01101000 01101111 01110101 01100111 01101000 01110100 00100000 01110100
01101000 01100101 01111001 00100000 01110011 01101000 01101111 01110101 01101100
01100100 00100000 01101000 01100001 01110110 01100101 00100000 01110010 01100101
01100011 01100101 01101001 01110110 01100101 01100100 00100000 01101001 01110100
00101100 00100000 01100010 01100001 01110011 01100101 01100100 00100000 01101111
01101110 00100000 01110100 01101000 01100101 00100000 01100011 01101111 01110000
01111001 00100000 01101111 01100110 00100000 01110100 01101000 01100101 00100000
01100011 01101111 01100100 01100101 00100000 01110100 01101000 01100101 01111001
00100000 01101000 01100001 01100100 00100000 01101001 01101110 01110011 01110100
01100001 01101100 01101100 01100101 01100100 00101110 0001010 01000010 01110101
01110100 00100000 00101000 01110100 01101000 01100101 01110010 01100101
10000000011001 01110011 00100000 01100001 01101100 01110111 01100001 01111001

01110011 00100000 01100001 00100000 01100010 01110101 01110100 00101001 00101100
00100000 01101101 01100001 01110100 01110100 01100101 01110010 01110011 00100000
01100111 01100101 01110100 00100000 01100011 01101111 01101101 01110000 01101100
01101001 01100011 01100001 01110100 01100101 01100100 00100000 01100010 01100101
01100011 01100001 01110101 01110011 01100101 00100000 01101111 01100110 00100000
01100001 00100000 01110111 01101000 01101111 01101100 01100101 00100000 01100010
01110101 01101110 01100011 01101000 00100000 01101111 01100110 00100000 01100100
01101001 01100110 01100110 01100101 01110010 01100101 01101110 01110100 00100000
01100101 01110010 01110010 01101111 01110010 01110011 00101110 0001010

4.2 POINT POSITIONING

Most GNSS chips available in devices like smartphones, tablets and in car systems, only use the L1 GPS signal for positioning. These types of receivers are called **single frequency receivers.** They use the **code observable** – the PRN codes, C/A and P(Y), on L1 to obtain a position of a point quickly and in three dimensions, using just one GNSS receiver.

This is where we get the name **point positioning** from; it is the fastest, cheapest, easiest, but least accurate method of GNSS positioning.

The basics of point positioning

By now, we have covered the basics of how GNSS works, so you should be aware that the **time** GNSS signals (radio waves) take to get from a satellite to a GNSS receiver are a critical component of measuring the distance from a satellite to a GNSS receiver. This distance is referred to as a **range**, and taking these measurements is referred to as **ranging**.

We've also briefly covered the fact you need a minimum of four satellites to get a position using GNSS, but why is this?

Well, it's essentially trilateration over some big distances, plus correcting for clock errors using the satellite ranges. Let's look at this in more detail.

We can measure a satellite range by knowing the amount of time the GNSS signal takes to get to our receiver, but if we can only determine one satellite range, all this really means that we're somewhere on a sphere with a radius of approximately 20,000km, as shown in **Figure 4.2(a)**. This isn't particularly useful on its own!

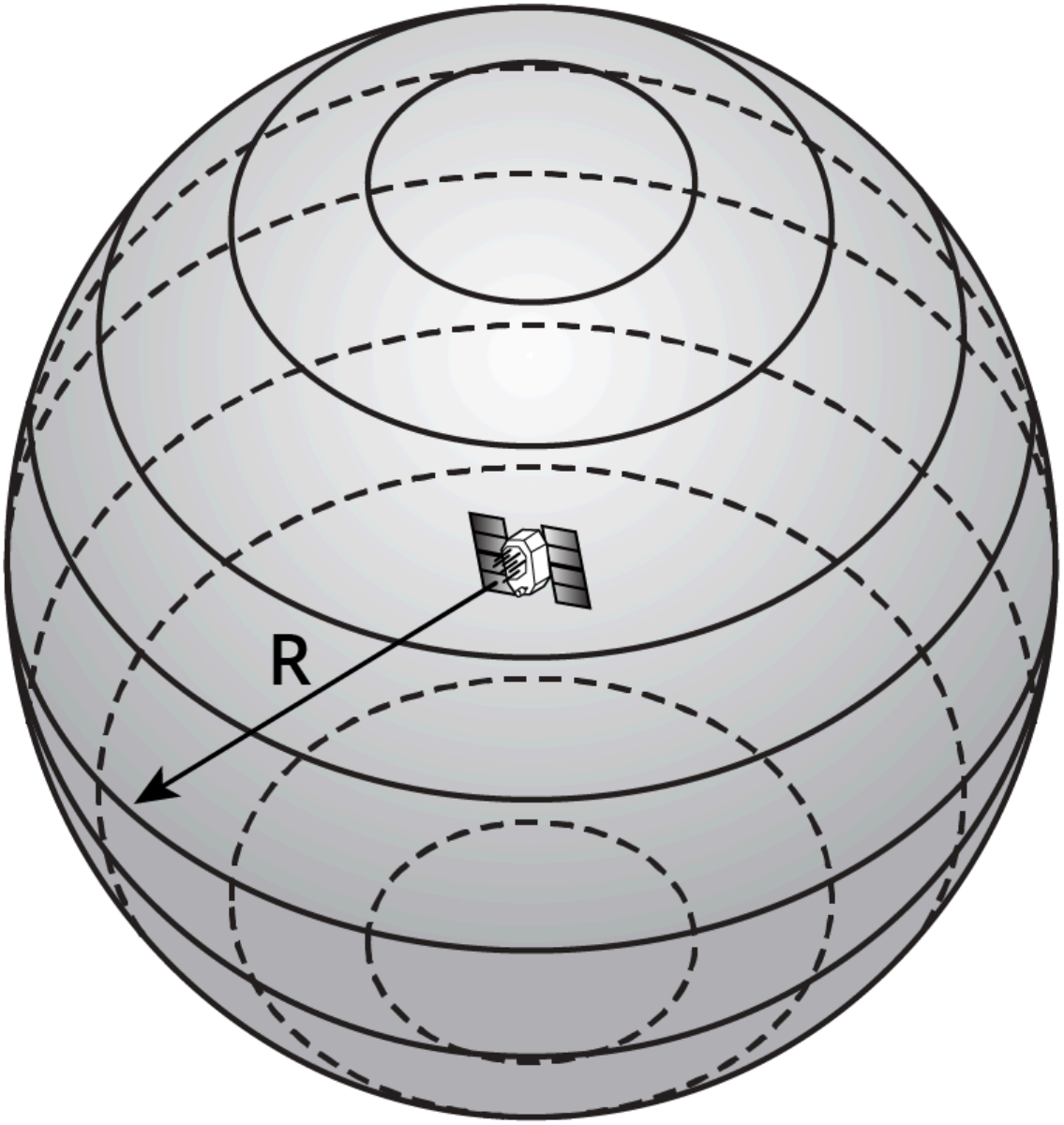


Figure 4.2(a): Determining position – One satellite radius

By adding in a second satellite range, we now have two spheres of similar radius that have an intersecting plane. We know we're somewhere on the circle where the two spheres intersect, as shown in **Figure 4.2(b)**.

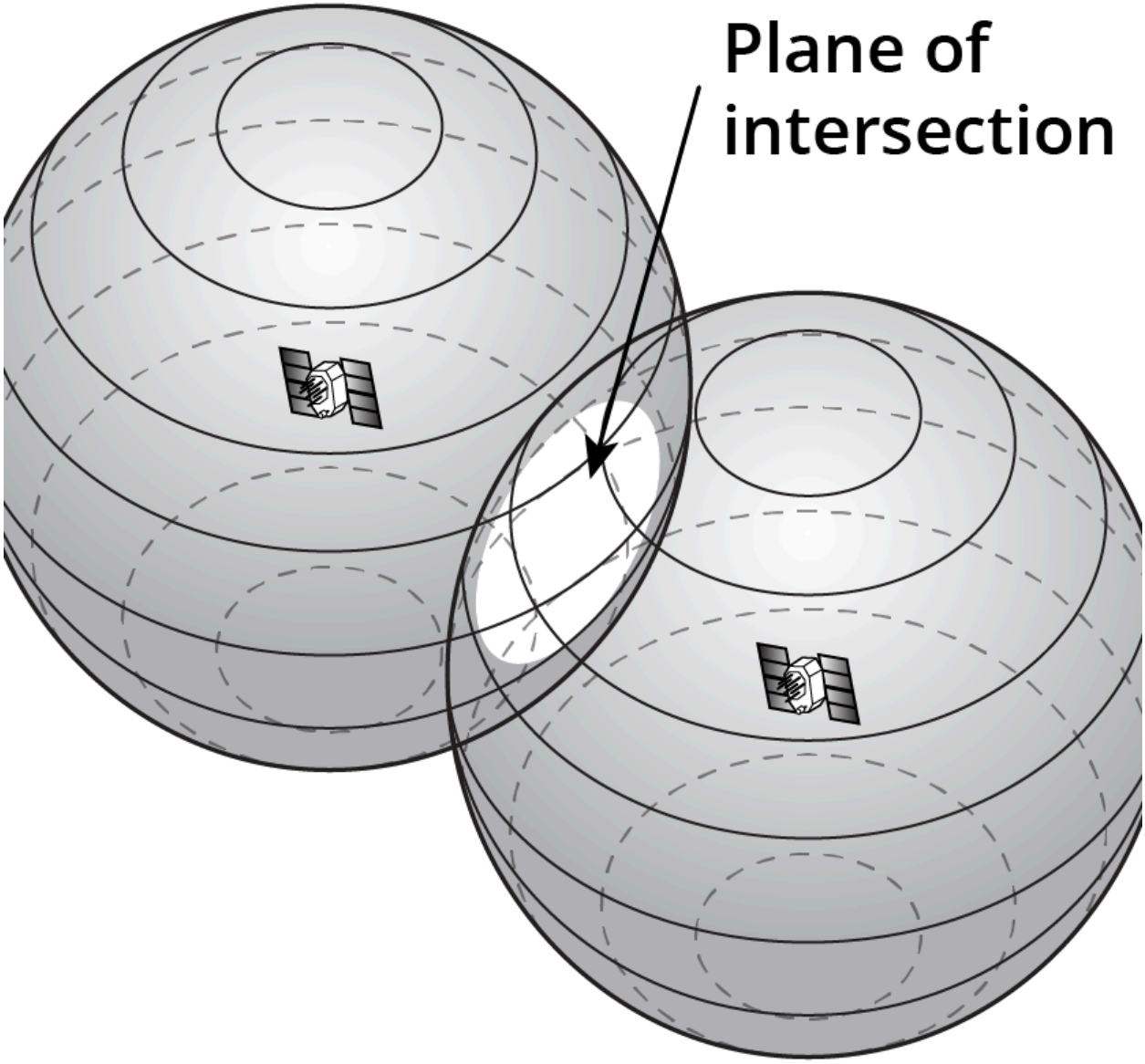


Figure 4.2(b): Determining position – two satellites

The third satellite range gives us a third sphere, and now we're down to two possible spots on the circle where the first two spheres intersect, as shown in **Figure 4.2(c)**.

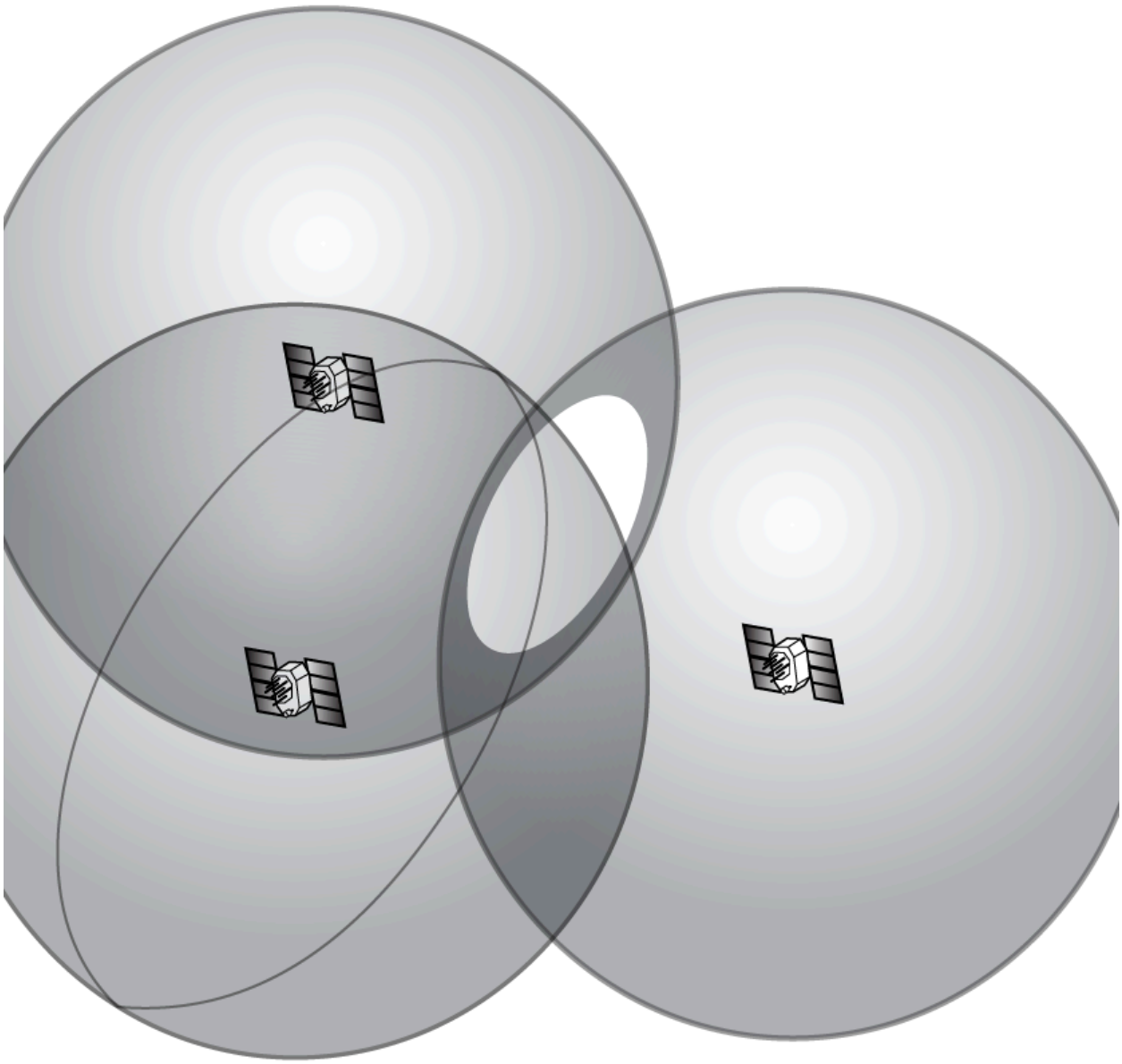


Figure 4.2(c): Determining position – three satellites

Adding in satellite range number four means we can eliminate one of the two points, as shown in **Figure 4.2(d)** and now we have a point position in an ECEF coordinate system.

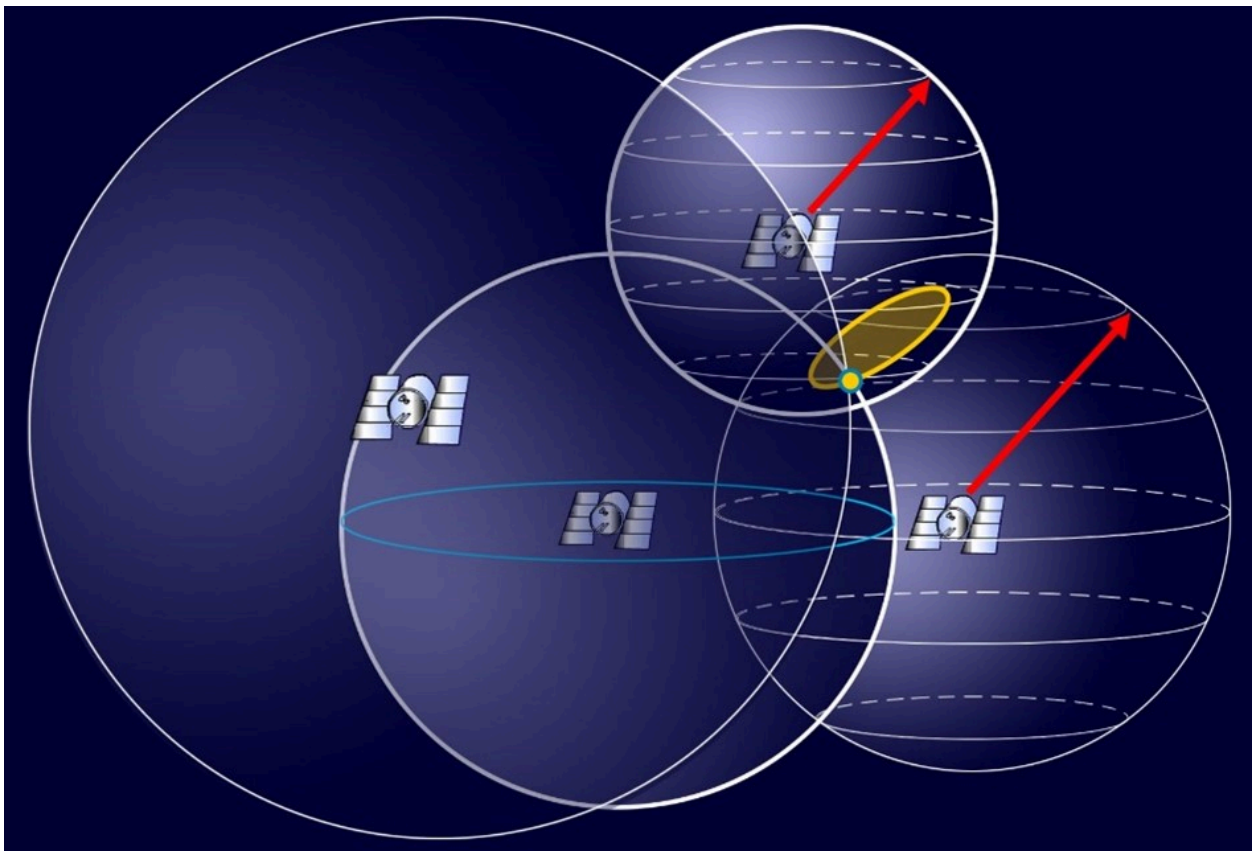


Figure 4.2(d): Determining position – four satellites

In reality though, our receiver can figure out our position with only three satellite ranges, as one of these points is going to be nowhere near the Earth. The GNSS receiver has various techniques it uses to allow it to figure out the correct point from the incorrect one. This is essentially how point positioning works, as shown in **Figure 4.2(e)**.

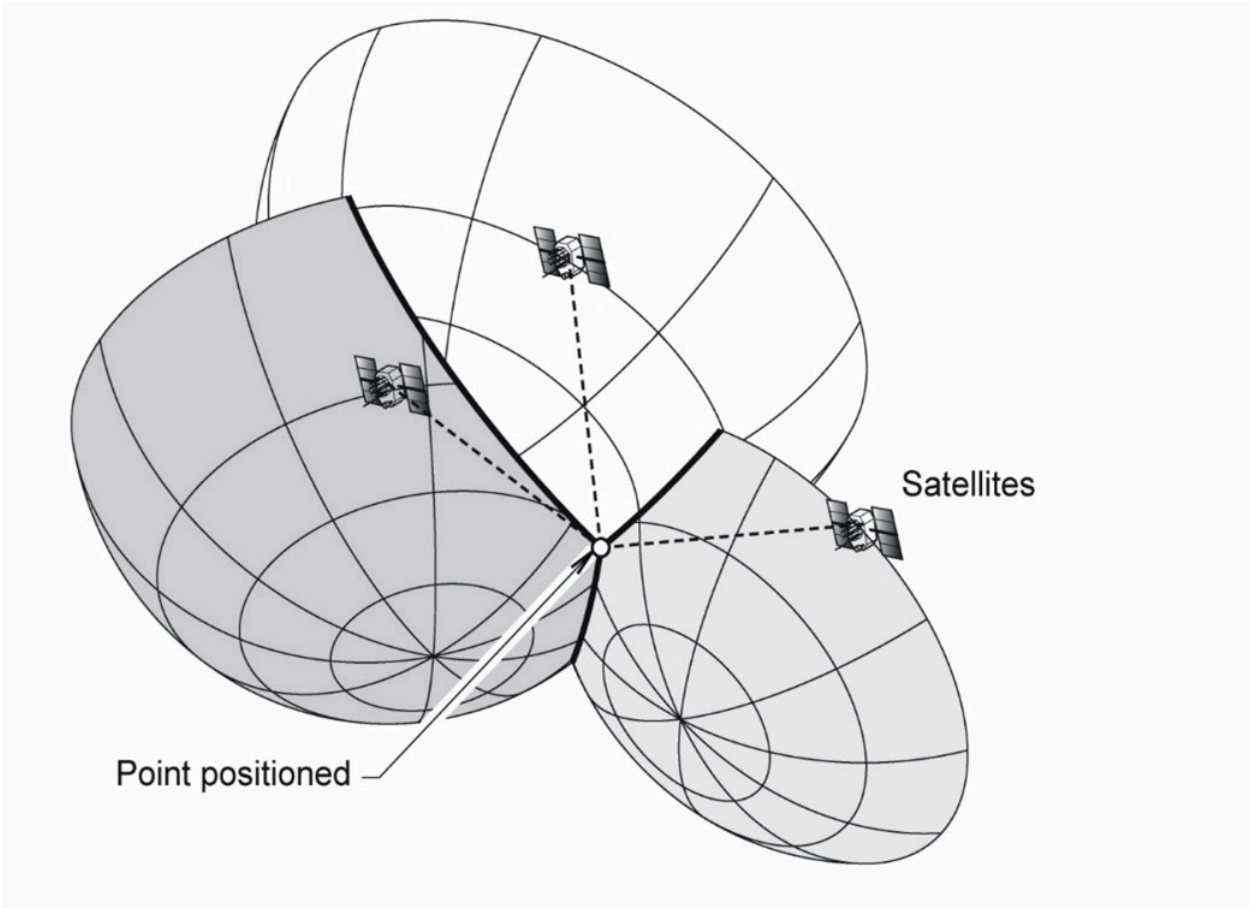


Figure 4.2(e): Point positioning

If you know your elevation, you could also eliminate one of the satellite ranges and replace it with a sphere that has a radius of your distance from the centre of the Earth, but thankfully we no longer have to do this. So, what's with needing the fourth satellite then?

We know there are errors in GNSS measurements due to receiver, location and system errors. The fourth satellite essentially acts as a check on the other three measurements, and helps us deal with clock errors. In point positioning we do this through a process called **code pseudo range positioning**.

4.3 CODE PSEUDO RANGE POSITIONING

In point positioning, we use the code part of the signal to determine the satellite range, however, for a series of reasons we'll discuss soon, this range isn't actually the correct range, so we refer to it as a **code pseudo range.**

Code pseudo range positioning works on the simple idea that each GNSS receiver has a copy of the codes each satellite transmits (remember that each satellite, apart from GLONASS, have unique codes) and that the receiver can compare the code it receives with the code it has on board to determine the difference between them. This is shown in **Figure 4.3(a).**

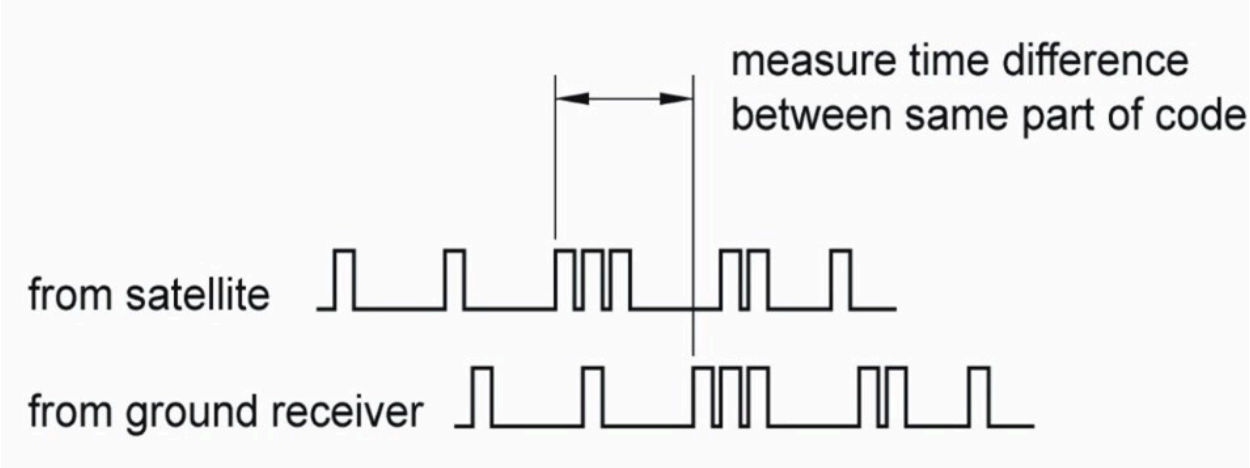


Figure 4.3(a): The difference in code travel time from satellite to receiver

The receiver identifies the code it should be hearing from a particular SV at a particular time, and compares it to what it has received. The difference between the two codes is the **time delay**, which can then be used to calculate the distance to the SV.

This assumes that the signal generated by the GNSS receiver is produced at exactly the same time as the satellite generates and sends the signal, however, GNSS receivers only have quartz clocks, which are nowhere near as accurate as the atomic clocks that SVs have.

Now we have two problems to solve, the signal delay and clock differences. BUT using the fourth satellite can help us resolve this! BUT…satellite orbits and atmospheric errors also impact our signal!

This is why code pseudo ranging has five fundamental components.

Step 1: Trilateration

The fundamental concept that allows us to generate a position in point positioning is the technique of trilateration, as discussed in section 4.2 – we can determine our position using three satellites ranges.

Step 2: Satellite code pseudo ranging

The satellite and receiver should be generating the same code at the same time, and by understanding the delay in the satellite code in getting to the receiver, we can determine the distance/range to the satellite. This can be shown by the equation:

$$R = c \times \Delta t$$

Where R = range distance
c= speed of light (299792458ms^{-1})
Δt = time delay

However, as we've already recognised, there can be issues accurately determining ranges due to clock errors.

Step 3: Four satellites for X, Y, Z & time

The issues with accuracy and time of the clocks in SVs and GNSS receivers is referred to as **clock error** or **clock bias**, and both satellites and receivers are subject to these errors. Clock errors impact the ability to accurately measure the range, as the time a signal takes to reach a receiver cannot be accurately determined.

Before GNSS chips started appearing in smartphones and devices connected through telecommunications networks, the quartz clocks in GNSS receivers had no central system to align themselves to. This meant they needed an external source to confirm their time, much like when a watch has a new battery put in it – it can tell time, but until it is set relative to a time zone, it isn't particularly useful. This meant that GNSS receivers relied on receiving satellite signals to adjust their clocks. Now that most devices capable of point positioning are connected to telecommunications networks, their capacity to be aligned to time systems is much more sophisticated, which results in a lower level of clock error, but does not remove it completely.

The way we are able to resolve the clock errors in point positioning is by using the fourth satellite measurement we discussed earlier.

Modelling clock errors

Note: The equations included in this section are not required to be memorised, but are included to assist in explaining the process of removing clock errors in code pseudo ranging.

For this next section, let's assume we're using GPS time.

The first clock error is the difference between the satellite clock reading and GPS time at the moment of signal emission, and the second is the difference between the receiver clock reading and GPS time at the moment of signal reception.

We can determine the first by knowing the time the satellite clock reads when the signal is sent, the delay of the clock relative to GPS time, to give us the true time that a signal is sent. In equation format, this looks like:

$$t_s(\text{true}) = t_s\,(\text{GNSS}) - \delta_s$$

Where t_s(true) = true time of signal emission
t_s (GNSS) = satellite clock reading at time of signal emission
δ_s = clock delay (offset) at signal emission with respect to GPS time

We can now determine the second clock error by knowing the reading of the receiver clock when it receives the signal, the delay of the clock relative to GPS time, to give us the true time that a signal is received. In equation format, this looks like:

$$t_u(\text{true}) = t_u(\text{GNSS}) - \delta_u$$

Where $t_u(\text{true})$ = true time of signal reception
$t_u(\text{GNSS})$ = receiver clock reading at time of signal reception
δ_u = clock delay (offset) at signal reception with respect to GPS time

From our previous time delay equation, we already know that the time delay is equal to the difference between the time the satellite emits the signal to when it is received, so we can substitute in our clock error equations:

$$\Delta t(\text{true}) = t_u(\text{true}) - t_s(\text{true})$$
$$= (tu(\text{GPS}) - \delta u) - (ts(\text{GPS}) - \delta s)$$
$$= \Delta t(\text{GPS}) + \Delta\delta$$

To determine range, we now have the equation:

$$R = c \times \Delta t(\text{true})$$
$$= c(\Delta t(\text{GPS}) + \Delta\delta)$$

The atomic clocks on SVs are monitored by the control segment, and are kept aligned to a specific time, so the corrections for the SV clock offsets can be assumed to be compensated for (there is some residual offset, but for the purposes of this course we can assume it is zero).

This now means that when a receiver is first turned on, the time difference between the receiver generated code and the received code from the satellite will have two components, the measure signal travel time $\Delta t(\text{GPS})$) and the receiver clock error (δ_u). This now means our previous range equation can be once again reorganised: $R = c\Delta t(\text{GPS}) - \delta_u$).

This range equation can now be used to determine the position of our receiver, and solve the clock bias.

Determining position using pseudo ranges

As covered in previous modules, the position of satellites is described in an ECEF Cartesian coordinate system, as shown in **Figure 4.3(b).**

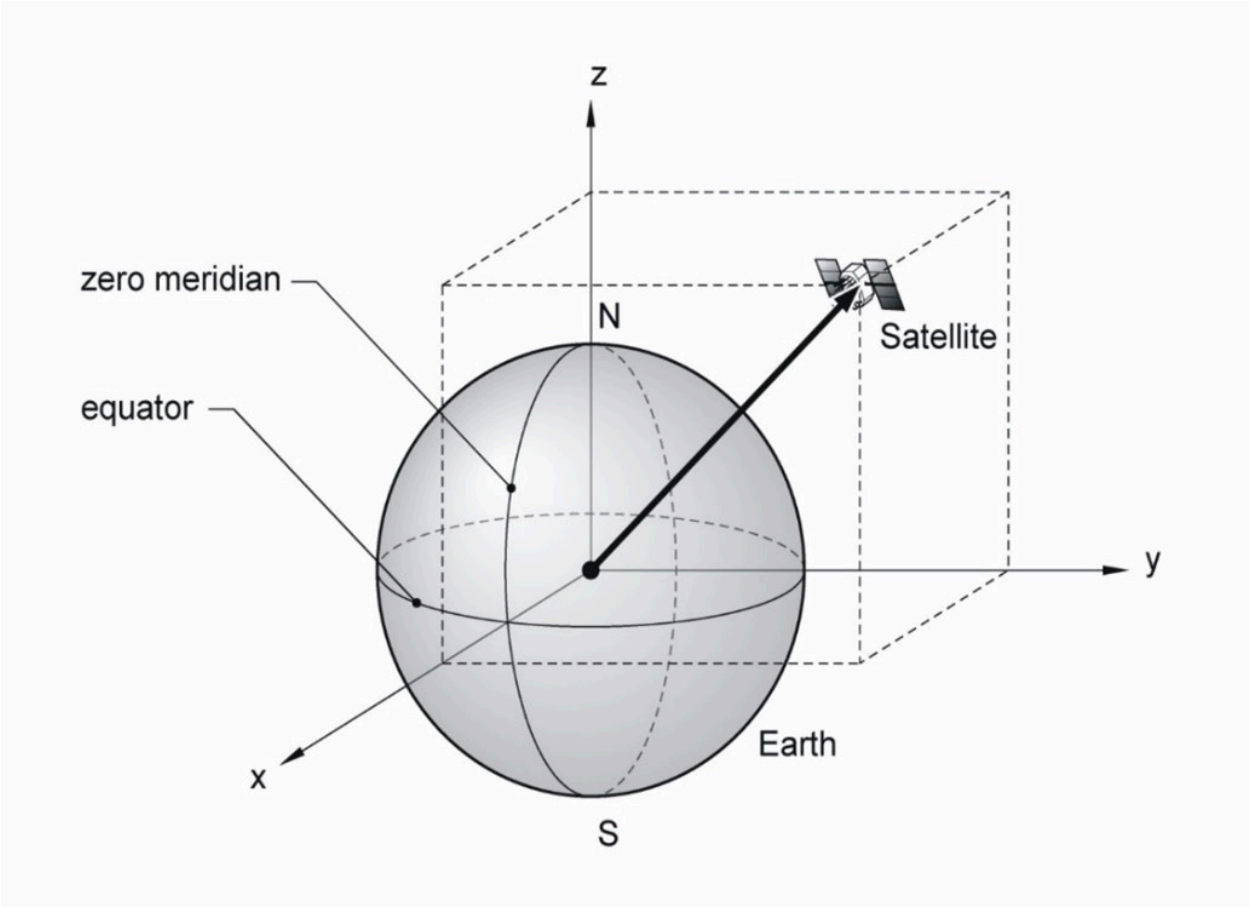


Figure 4.3(b): Satellite position in an ECEF Cartesian coordinate system

The range from a GNSS receiver to a satellite, as shown in **Figure 4.3(c)**, can be given by the equation:

$$R = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2}$$

Where R = Range from receiver to satellite

ΔX = difference in the X axis of satellite and receiver

ΔY = difference in the Y axis of satellite and receiver

ΔZ = difference in the Z axis of satellite and receiver

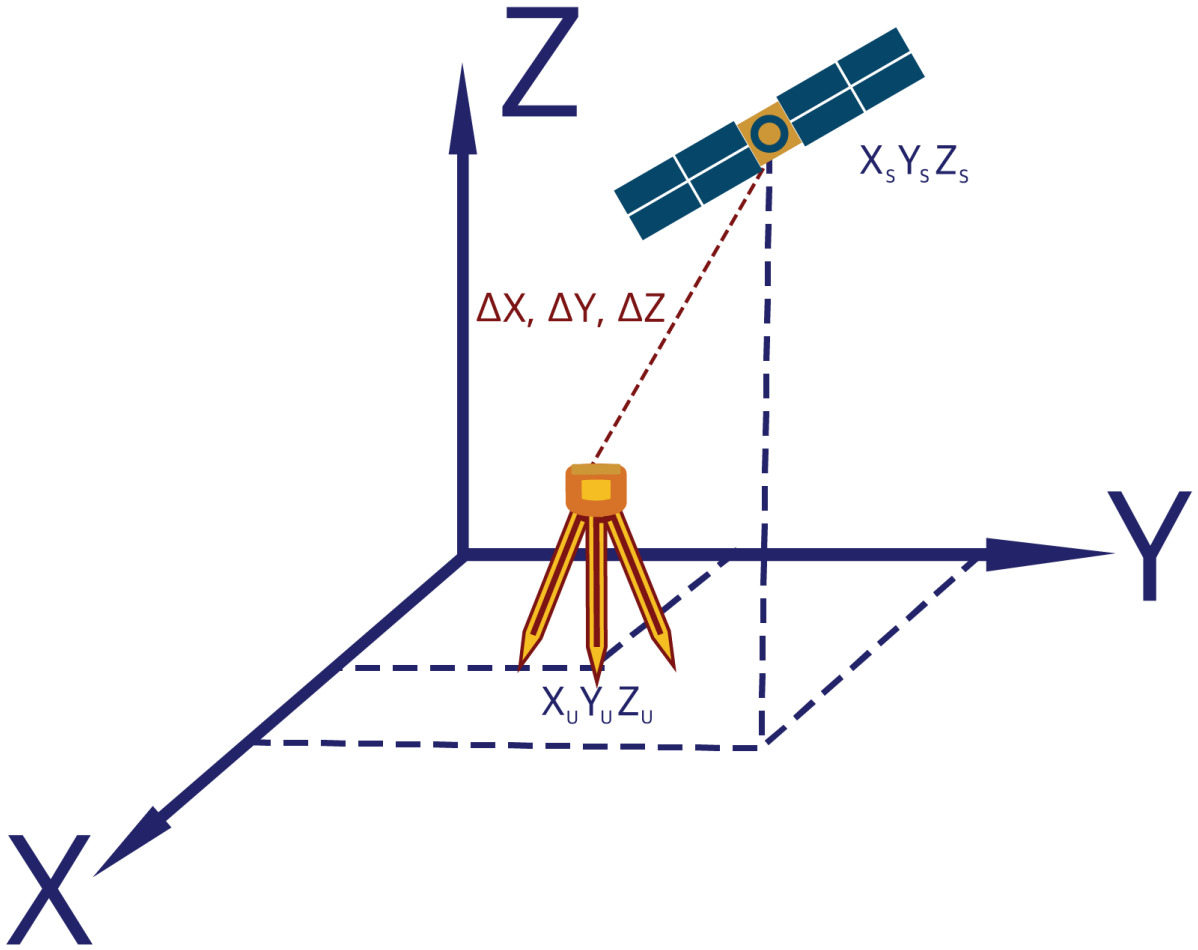


Figure 4.3(c): Range from receiver to satellite

Taking the clock error from the previous section into account, the equation can be given as:

$$R = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} - c\delta_u$$

We now have four unknowns in our equation; the position of the receiver in three axes, and the clock error.

Each satellite range will have its own equation, and we know we are able to solve the receiver position with three satellites (three range equations with three variables), so with four satellites we are able to solve for the clock bias as well.

Step 4: Orbit errors

But the satellites are moving! And so are we!

In simplistic terms, we account for the fact satellites and the receiver are moving by:

- assuming the position of the satellite from the code pseudo range values and the ephemeris data
- use the time delay to rotate the Earth the amount it would have moved in that time, and recalculate the position of the receiver
- range is then recalculated between the new receiver position and the estimated satellite position, to give a new estimate of time delay.

The process continues until the range or the time stop changing (this is called 'converging' in a maths context), giving a new range value, and thus a more accurate satellite position.

Step 5: Atmospheric corrections

As discussed in the errors section of **Chapter 3**, the ionosphere and troposphere delay GNSS signals by causing them to refract due to charged particles and water vapour respectively.

In code pseudo range positioning, the GNSS receiver will make estimated corrections for these delays, and apply them when calculating position.

Code pseudo range positioning summary

As you have seen throughout the discussion of code pseudo ranging, the range measured is adjusted and corrected for a variety of factors, but can still not be considered the true range from satellite to receiver, hence the reason it is called **pseudo range.**

Point positioning using the technique of code pseudo ranging utilises the code observable from four satellites to determine a 3D position by applying a series of mathematical processes:

1. trilateration
2. satellite code pseudo ranges
3. utilising four satellites to solve four unknowns
4. orbit error correction
5. atmospheric error correction

4.4 POINT POSITIONING ERRORS

Point positioning is impacted by the errors discussed in Module 3. The main error that point positioning resolves is the receiver clock error.

The errors that impact point positioning are:

- Satellite clock errors – these are primarily dealt with by the control segment through the application of corrections.
- Satellite orbit area – while point positioning can resolve a level of orbit errors, it relies on the accuracy of the ephemeris, which is managed by the control segment.
- DOP – GNSS receivers tolerance levels can be set to only accept data when DOP values are below certain levels.
- Low elevations – similarly to DOP, GNSS receivers may have an elevation mask setting that can be adjusted depending on local obstructions.
- Atmospheric delays – point positioning attempts to resolve some atmospheric delays through receiver models, however, single frequency observations like point positioning make resolving these errors difficult. Dual or multiple frequency methods of observation are significantly better at resolving atmospheric errors.
- Obstructions and multipath – these are site dependent and impact point positioning. Users should consider choosing a different location if possible.
- Receiver noise – tolerance levels for SNR can be set in most receivers. Avoidance of electrical influence, such as power lines or similar can assist in reducing receiver noise issues.
- Spoofing – point positioning is highly susceptible to spoofing as the majority of attacks are in the L1 band.
- Human error – our capacity to introduce error into point positioning is infinite, however, it can be mitigated by understanding of how GNSS operates and through development of quality systems to manage observations of positions.

Minimising errors in point positioning

While reducing the errors in point positioning is virtually impossible through observation manipulation, there are observation techniques that allow errors to be minimised or averaged out. These techniques also assist in addressing human error.

Point averaging

GNSS errors can be significantly reduced by averaging individual point positions over time. The amount of time a point should be observed is dependent on the accuracy looking to be achieved, however, in point positioning 10-30 seconds is usual sufficient for 10m accuracy in open areas.

Redundant observations

Revisiting a point multiple times generates **redundant observations**. These are most easily thought of as checks on the observations taken to assure the user of the data that it is of the appropriate accuracy. Redundant observations should be done at the end of each observation session, before and after rest periods on projects (including overnight) or even periodically throughout a project. Observations should be separated by an appropriate amount of time to ensure systematic errors are reduced.

Measuring known marks

Periodically or throughout a project, a GNSS receiver being used for point positioning should be checked against a mark of known coordinates. This allows the user to compare the receiver coordinates to the known coordinates, checking for any ongoing bias or issues with positioning.

PART V

DGPS

5.1 DGPS

Learning Objectives

On the successful completion of this chapter you should be able to:

- explain what a differential technique is
- explain what the different methods of DGPS are
- explain the difference between real time and post processed corrections
- explain the difference between DGPS and point positioning
- explain the errors associated with DGPS and how to minimise them
- debate the need for, and explain what constitutes redundant or check observations with DGPS.

In the beginning...

Not satisfied with one GPS receiver, some surveyor somewhere figured out that using two receivers could help cancel out some of the errors that impact point positioning.

Enter differential GNSS, which has stubbornly hung onto the name **differential GPS** or **DGPS** in the majority of cases.

DGPS uses two GNSS receivers using code observable to reduce the errors that impact point positioning, and can be used in real time, or the correction process can be saved for later when back in the office.

If one GNSS receiver is good, two **MUST** be better right?!

Note: There are other differential techniques, like RTK, that rely on the carrier phase component of the GNSS signal. These are not covered in this module, they are covered in **Chapter 6.**

DGPS

Point positioning provides a quick and cheap method for getting autonomous positions, however, the errors that limit its accuracy, limit its applications outside of navigation.

Unless you have more than one GNSS receiver.

Differential GPS, most commonly known as **DGPS**, is a technique that uses code observable to make point positioning more accurate, by using two GNSS receivers in different locations. DGPS has a variety of applications in asset management, basic data collection and improved navigation.

DGPS utilises two GNSS receivers in different locations, but same area, to model errors and apply corrections to observations. One receiver is set up over a mark (usually a permanent survey mark) that has a known coordinate – this receiver is called the **base station**, as it stays in place for the entire time observations are being undertaken by the second receiver. The second receiver is called the **rover**, as it is moving around taking the GNSS observations.

By observing a point of known coordinates with a base station, we are able to determine the difference between the known coordinates and the collected coordinates. Essentially a base station tells us the difference between where a point position says it is, and where it knows it is. This

difference is then used as a basic model of the errors for that area, and this model can be applied as **corrections** to the data collected by other GNSS receivers in the same area at the same time.

The errors that impact an area aren't random, and can generally be modelled as relatively smooth movements over time. Any time a GNSS receiver is collecting positioning data, it is creating a position and a time record for a point, and by matching the time of a correction to a rover measurement, we can apply the right correction to the point collected by the rover.

There are two methods of applying DGPS corrections, the first is by using coordinates, and the second (and more accurate) method is by using pseudo range corrections.

Coordinate corrections

In **Figure 5.1(a)**, we have a DGPS set up, where a base station is set up on a mark with known coordinates, and a rover is placed on a point. At time , the base station is logging a position away from the known coordinates, and it continues to log a position every second after the initial position (etc.). At the same time, the rover receiver is logging positions every second as well.

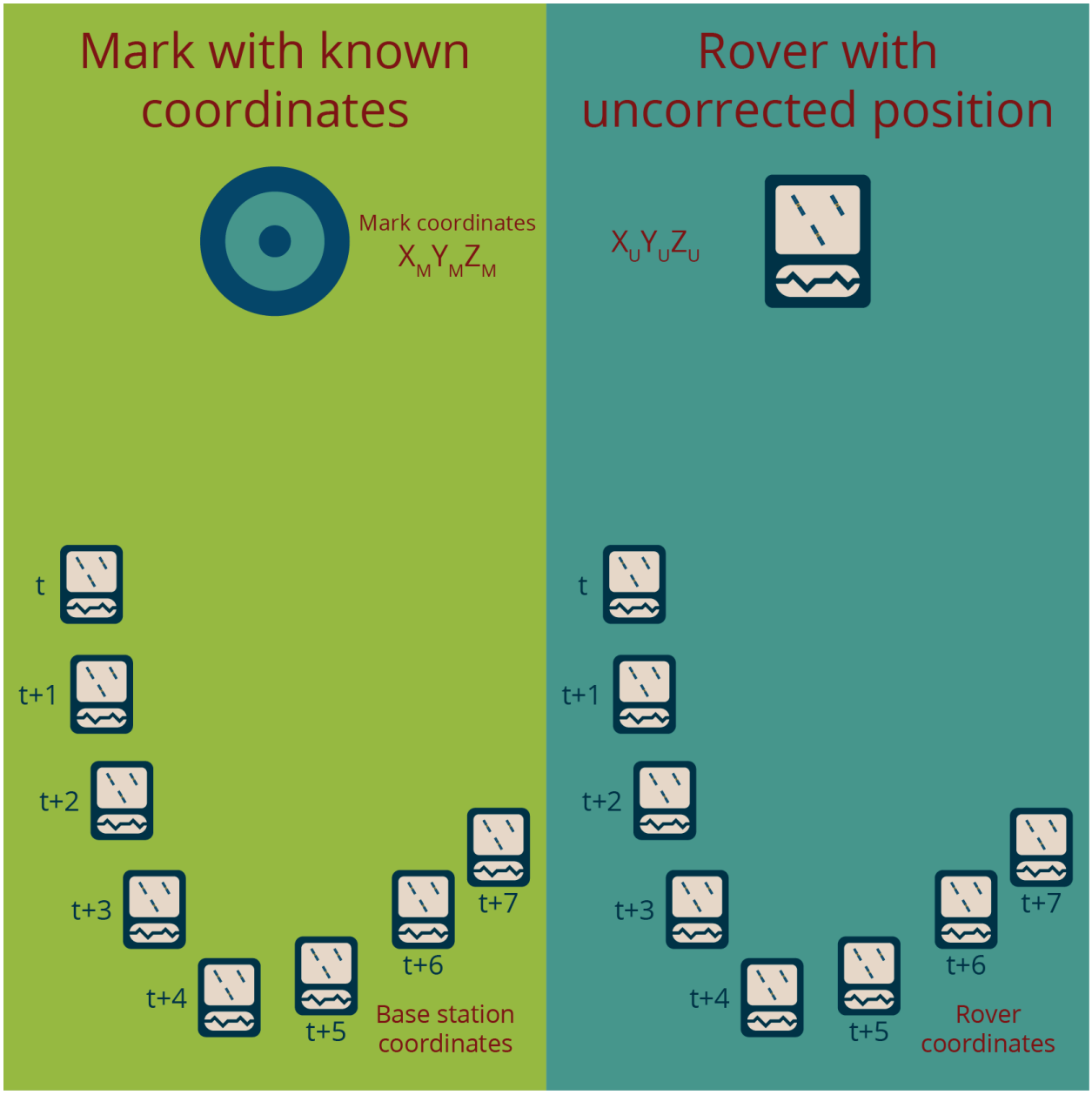


Figure 5.1(a): DGPS base station and rover collecting positions

In **Figure 5.1(b)**, the differences between the known coordinates of the mark and the logged position at the various points in time are represented by the baselines . These base station baselines are the **corrections** that are then applied to the rover data to calculate its **corrected position.** This situation assumes that the rover is experiencing the same errors as the base station.

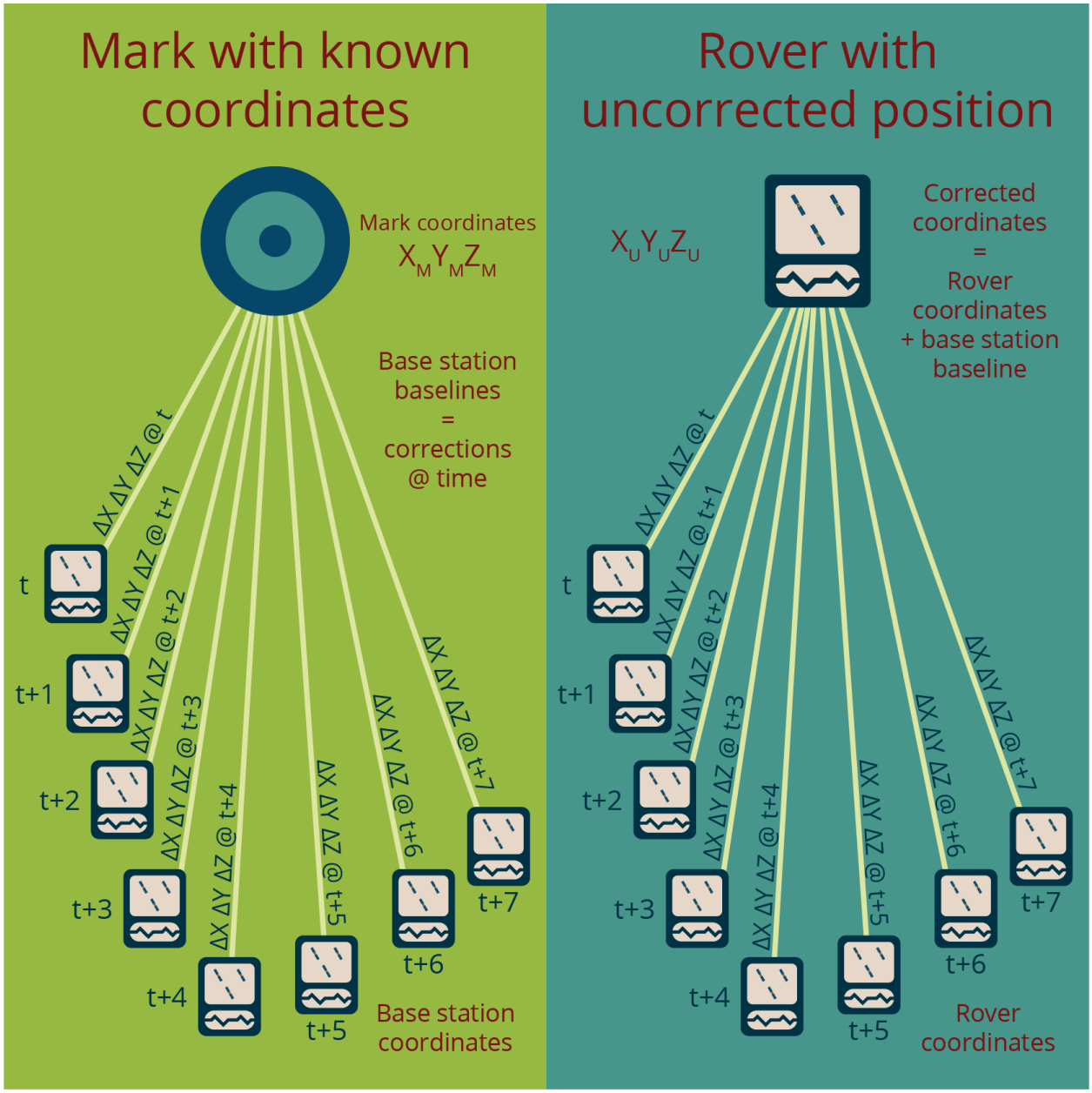


Figure 5.1(b): DGPS rover corrected positions from base station baselines

The simplest method of applying corrections to coordinates in DGPS is by using a translation or **block shift** method, where all coordinates are moved in the same direction, the same amount. This is shown in **Figure 5.1(c).**

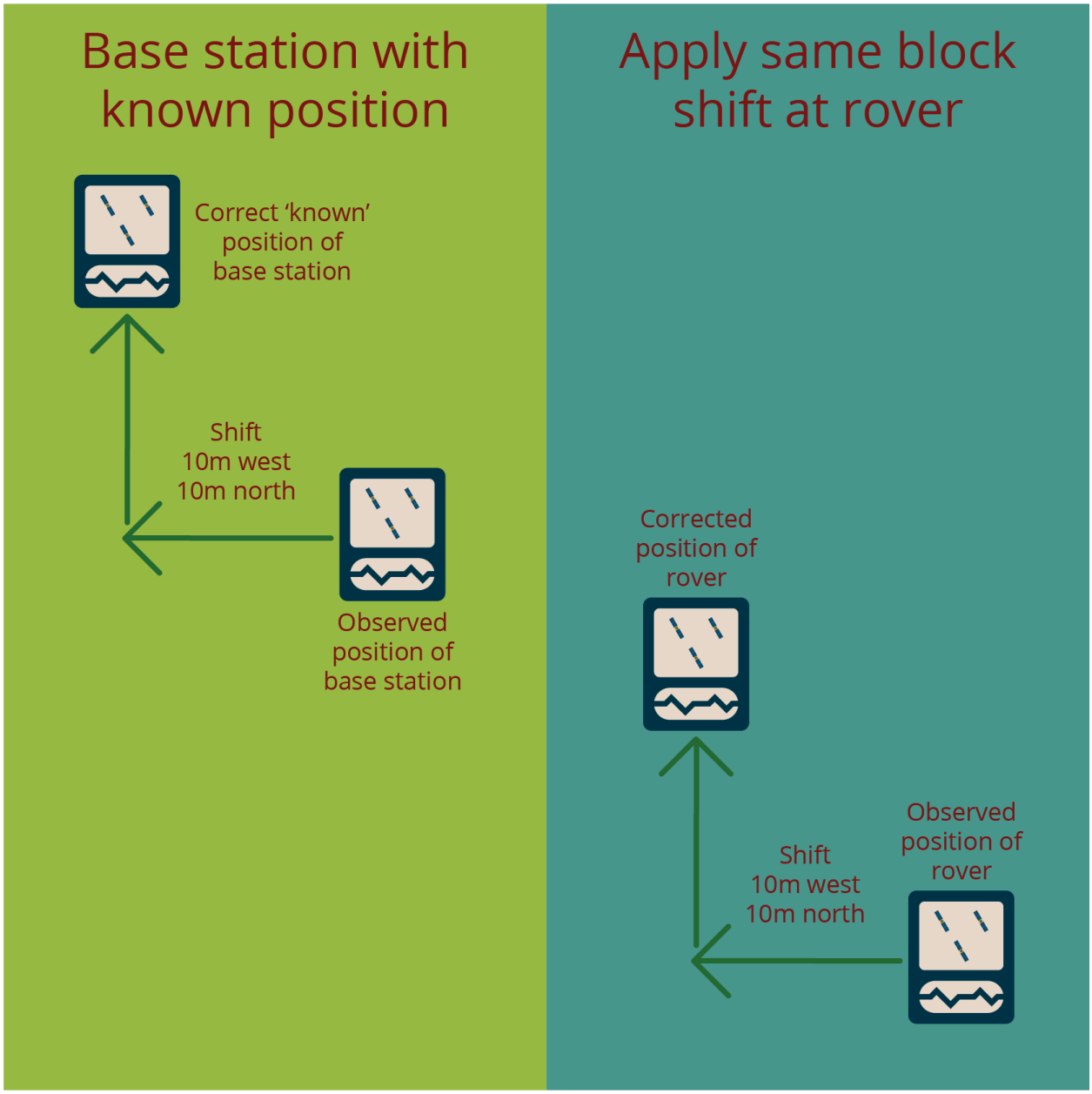


Figure 5.1(c): DGPS block shift

For corrections from a base to be applicable to another GNSS receiver's data, each receiver must be tracking a common set of satellites, meaning they need to be in relatively close proximity to each other, with sites that allow for good satellite visibility by the GNSS receivers.

Pseudo range corrections

In the **pseudo range correction** method of DGPS, the base station doesn't use the normal point positioning technique of using time to determine range. The base station already knowns exactly where it is, and where the satellites are meant to be, so it knows the ranges already, and uses this to determine the times. It compares this calculated time to the time the signals actually took, minus any error or delay in the satellite's signal.

This error correction is then transmitted to other receivers in the area, and they can use the error correction message to correct their position pseudo ranges and positioning solutions, as shown in **Figure 5.1(d)**. Because the distance between the receivers is tiny compared to the distances to the

satellites, errors measured by the base station will be almost exactly the same for the other receivers in the area.

The pseudo range correction method is the more common method of DGPS.

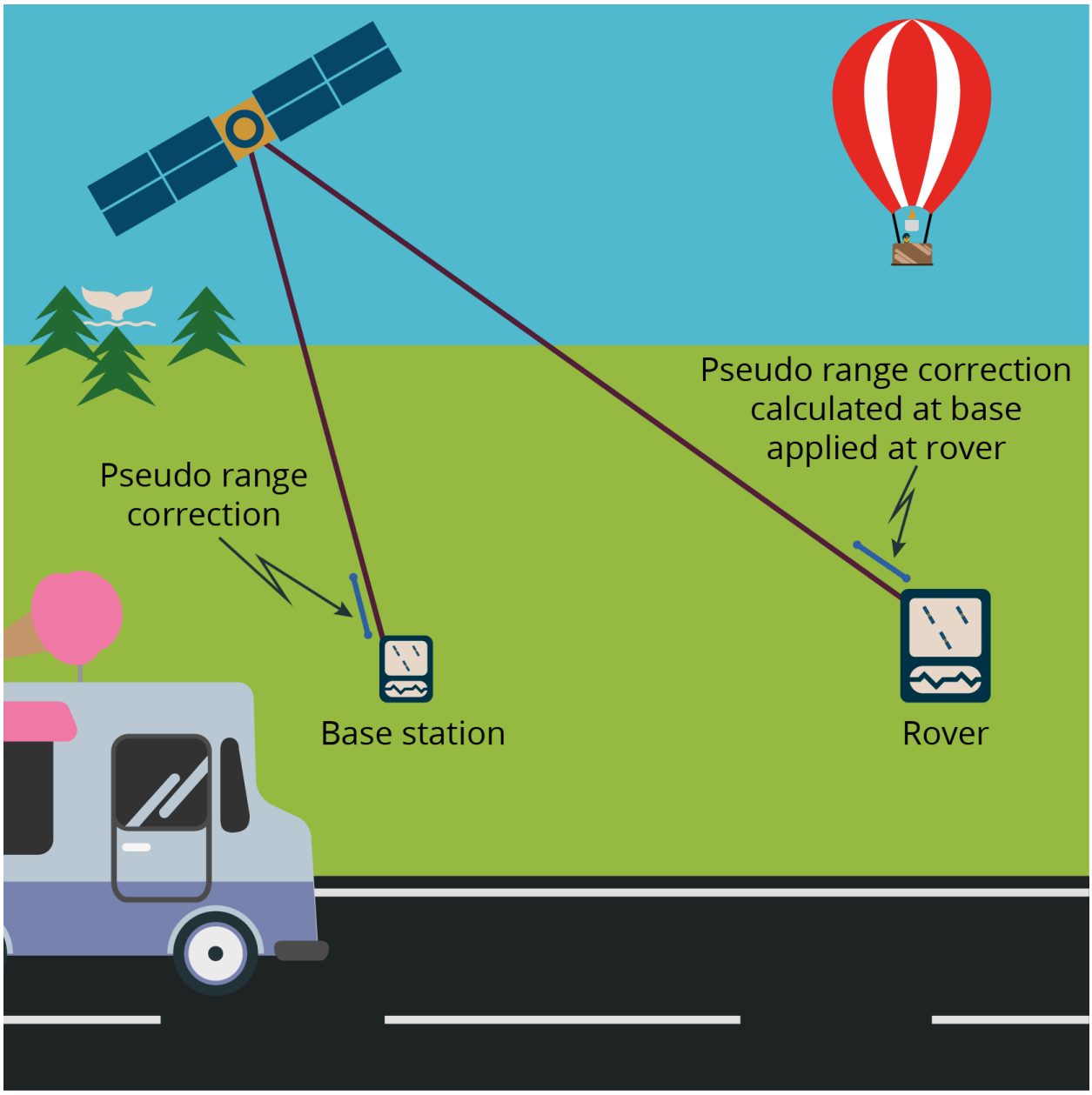


Figure 5.1(d): Pseudo range correction

Base station requirements

The following are requirements for the base station for the DGPS technique to be utilised:

- The base station must record data during the same time period as the rover is recording data. It should be set up to start recording data prior to the rover commencing measurements, with enough time for point positioning point averaging to resolve any localised errors.
- The logging interval (epoch) of the base station must match the rover epoch.
- The known coordinates of the base station must be accurate and reliable.
- The base station must be carefully centred over the ground mark and the height of the

antenna must be accurately measured and recorded.
- The base station must have a clear, unobstructed view of the sky as it must track the same satellites that the rover tracks.
- The distance between the base station and the rover should not be excessive.
- Any elevation mask set at the base station must be less (lower) than the elevation mask of the rover.

5.2 DGPS CORRECTION METHODS

There are two different ways the corrections in DGPS can be applied.

If they are applied at the time they are created, these are called **real time corrections**, while corrections applied later, usually in the office, are called **post processed corrections**.

There are advantages and disadvantages to both approaches.

Real time corrections

Real time corrections can be applied by communicating the correction message through a radio link or a telecommunications connection, such as a mobile phone, that is configured to communicate with the GNSS receiver.

These corrections may come from your own base station, or they may be broadcast from a commercial provider, where users pay a subscription fee to access the correction service. Usually commercial services require you to use a specialised receiver that is programmed to process the correction service signals.

These DGPS commercial systems sit at the bottom of the GNSS hierarchy; they are generally ground based augmentation systems (GBAS), where the corrections are broadcast from a ground system such as radio or a telecommunication network. Occasionally, they will be satellite based augmentation systems (SBAS), where the correction is broadcast via a communications satellite.

The obvious advantage of real time corrections is getting the corrections immediately. Real time corrections also allow the rover to be used for navigation activities, such as finding a particular mark or location of an object, and also marking out coordinates of something (at the appropriate accuracy).

The disadvantages of real time corrections are a generally higher expense – the cost of the radio or communications link, and potentially the commercial service cost, along with any specialised receiver costs. Real time corrections can also mask system errors that happen over longer timeframes.

Post processed corrections

Post processed corrections are applied at some point after the collection of positions is completed, and requires specialised software to determine the corrections at the base and then apply them to the rover dataset. This processing of applying the corrections is referred to as **post processing**.

Both the base station and the rover must log its data to memory if users are planning on post processing their data.

Post processing has the benefit of allowing a model of the entire time frame that the GNSS receivers were logging data to be created, which will often remove more errors than real time corrections. The disadvantage of post processed corrections is the lack of corrected data available in real time, so activities like navigating to specific points or using coordinates to mark out something are not able to be undertaken.

5.3 DGPS ERRORS

As with point positioning, DGPS is subject to the errors discussed in Module 3. However, the system errors that impact point positioning, including satellite clock errors, orbit errors and atmospheric errors can be minimised or eliminated using DGPS.

Satellite clock errors

Satellite clock errors are resolved in a similar fashion to the technique in point positioning. The additional receiver observing the same satellites provides additional redundancy, and the pseudo range correction method allows for the calculation of time rather than range, which also minimises the clock errors.

Orbit errors

Using the pseudo range technique for DGPS reduces the impact of the orbit errors. Because the distance between the receivers is tiny compared to the distances to the satellites, errors measured by the base station will be almost exactly the same for the other receivers in the area.

Atmospheric errors

GNSS signals are delayed as they pass through the atmosphere due to the ionosphere and troposphere. While the satellite ephemeris and receiver have some basic atmospheric models to adjust for this delay, one of the most effective way to adjust observations for this type of error is through differential techniques (observing multiple frequencies is another, but we'll get to that in **Chapter 6**).

Differential techniques assume that the atmospheric interference will be similar across an area, so the corrections determined at the base station can be applied to the rover. This means atmospheric errors are essentially removed from the positions at the rover.

Errors not resolved in DGPS

The errors that are not resolved or removed by DGPS are:

- DOP – GNSS receivers tolerance levels can be set to only accept data when DOP values are below certain levels
- Low elevations – similarly to DOP, GNSS receivers may have an elevation mask setting that can be adjusted depending on local obstructions.
- Obstructions and multipath – these are site dependent and impact DGPS observations. Users should consider choosing a different location if possible.
- Receiver noise – tolerance levels for SNR can be set in most receivers. Avoidance of electrical influence, such as power lines or similar can assist in reducing receiver noise issues.

- Spoofing – DGPS is still susceptible to spoofing as the majority of attacks are in the L1 band, however, the use of multiple receivers does reduce the risk marginally.

Human error – as with point positioning, our capacity to introduce error is infinite, however, can be mitigated by understanding of how GNSS operates, and also through development of quality systems to manage observations of positions.

Minimising errors in DGPS

The same techniques of error minimisation that apply to point positioning generally apply to DGPS, however, there are some variations.

Point averaging

GNSS errors can be significantly reduced by averaging individual point positions over time. The amount of time a point should be observed is dependent on the accuracy looking to be achieved, however, in DGPS 10-30 seconds is usual sufficient for 0.5m accuracy in open areas.

Redundant observations

Revisiting a point multiple times in point positioning generates redundant observations, however, in DGPS measuring a point multiple times cannot be considered as creating truly redundant observations, as the same baseline correction has been applied each time.

If there was an initial error in the baseline calculation, this error would carry through all measurements that used that baseline correction, so additional measurements on different days, from a different base station at a different time (to allow a different satellite configuration) or by different techniques are required to generate truly redundant measurements.

Measuring known points

While the base station is by definition on a known point, it can be useful to use the rover to observe known points periodically throughout an observation session as an additional check on the position information or real time corrections.

PART VI

PHASE OBSERVABLE

6.1 PHASE OBSERVABLE

Learning Objectives

On the successful completion of this chapter you should be able to:

- explain and describe baselines
- explain the general theory behind phase observable, including measurements and differencing
- explain the errors associated with GNSS surveying and how to minimise them
- explain the common methods of GNSS surveying that use phase observable
- discuss the ways that accuracy is determined in GNSS surveying
- explain the difference between point positioning, DGPS and GNSS surveying.

In the beginning...

Understanding the extents of land, and the location of the objects on them is a critical requirement for any country, as the security of land plays a significant role in how economies can develop and thrive.

This is why most explorers were also considered surveyors, and why Mt Everest is named after a surveyor (despite the fact he didn't want it named after him because he didn't want people to mispronounce his name. But they ignored him, and his nightmare came true!). Surveyors have been measuring the size, shape and location of land for what seems like forever – well before that guy that looked down a well.

At the national or state level, surveyors would have to survey huge areas that would take decades to cover, setting up trigonometric stations (known by surveyors as **trig stations**) on top of the highest land around so they could take meticulous measurements over and over again. In the beginning they used theodolites for triangulation, as they didn't have a quick, reliable method of measuring distance. With the invention of **Electronic Distance Measurement**, trilateration became achievable as distances could be measured in tiny timeframes compared to previously. However the need for trig stations continued as a visual line of sight was still needed for observations.

But then GPS became available to the public.

While autonomous positioning through GPS was a revolution for navigation and lower level accuracy activities, the need for surveyors to achieve high levels of accuracy remained. So some smart electrical engineers teamed up with some smart surveyors to develop the GNSS surveying technique we now know as **phase observable.**

In the early days surveyors had to get up at all sorts of crazy hours to make sure they could get four satellites, and the planning of the locations that had the best chance of getting accurate positions was meticulously and thorough.

Thankfully, a lot's changed since those days, including the development of some even smarter phase observable differential techniques that make life loads easier for surveyors.

Phase observable

If you asked a group of people what the biggest benefit of GNSS has been, it would be a fair bet that being able to know your position really quickly would make the list. But as you've learnt in the last two modules, people are constantly figuring out ways to get more accurate positions from GNSS measurements. Phase observable techniques are no different from the code observable techniques in this regard – our pursuit to remove (or at least minimise!) errors to squeeze the last bits of accuracy into our position is relentless.

Surveyors are no exception; when you make a living off knowing how to measure things incredibly accurately, it's pretty much your job to find better ways to do things – to remove as much of those pesky errors as possible!

So while the majority of the population is incredibly happy with being able to get an autonomous positon accurate to around five metres in a matter of second using the code observable, surveyors' need for millimetre accuracy means that they are willing to put out huge numbers of GNSS receivers over permanent survey marks for hours, days or even years at a time. Then some of them will spend more hours, days, years and even entire careers, pouring over the **relative positions** of these marks, at which point we still might not have a final position of the marks! Seems insane right?!

Well, in some ways it is, but in other ways phase observable is actually an incredibly smart and accurate way to measure huge distances and small movements in equal measure, simply by having a whole bunch of permanent survey marks.

In phase observable we are able to measure the differences in position of two or more of these survey marks, relative to each other, which you may recall are call **baselines.** If one or more of these marks has an accurate position, we can then use the baselines to 'transfer' those coordinates to other marks. This is done using essentially the same differential technique that was discussed in **Chapter 5**, but with receivers and antennae that can also read the carrier wave parts of GNSS signals, to within a couple of millimetres.

6.2 BASELINES

Baselines

We've discussed baselines previously – they are essentially a vector that represents the difference in position of two points.

Imagine you have two GNSS receivers that are capable of reading the phase observable, one is on mark **1** and one is one mark **2.** You could get an autonomous position on them using code pseudo ranging or DGPS, but we know these are subject to significant errors. Getting a more accurate position means we need to use the phase observable to create a baseline between the two marks. But for now, let's assume that we have collected the positions of the two marks.

The two receivers are shown in **Figure 6.2(a)**, in a Cartesian coordinate system, and they each have a known position. To calculate the baseline, it's a simple case of calculating the hypotenuse of two right angled triangles.

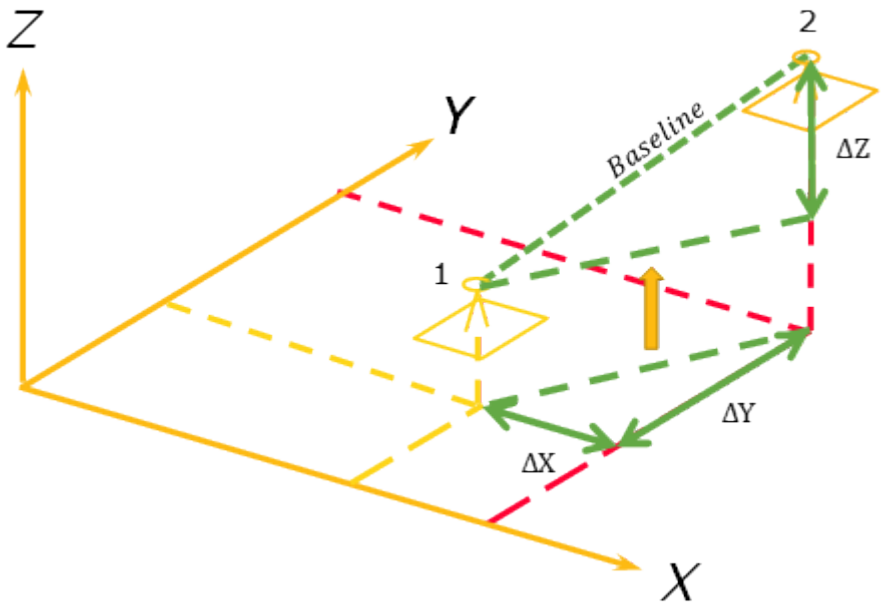


Figure 6.2(a): Calculating baselines using Pythagoras' Theorem

The first hypotenuse (or distance) we need to calculate is the horizontal distance between marks 1 and 2, which is simple given we can quickly determine our ΔX and ΔY values by subtracting the position of mark 1 from the position of mark 2. We can then use Pythagoras' Theorem to solve for the hypotenuse:

$$c^2 = a^2 + b^2$$

Where c = Hypotenuse

a = ΔX

b = ΔY

Substituting into Pythagoras' Theorem this gives:

$$Hypothenuse^2 = \Delta X^2 + \Delta Y^2$$

$$Hypotenuse = \sqrt{\Delta X^2 + \Delta Y^2}$$

We can now project or push this hypotenuse up to the level of mark 1, and by using the value – the difference between the heights of the marks in this case, we can make another right angle triangle. This time we have:

$$Hypotenuse^2 = \left(\sqrt{\Delta X^2 + \Delta Y^2}\right)^2 + \Delta Z^2$$

$$Hypotenuse = \left(\sqrt{\Delta X^2 + \Delta Y^2}\right)^2 + \Delta Z^2$$

Thus, our baseline between mark 1 and mark 2 can be described by the equation:

$$Baseline = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2}$$

But because we are attempting to measure the positions of the marks more accurately than the code observable can provide, we need another way to determine the baseline without starting positions of the marks.

Imagine if we knew a way to determine the distance between two objects using the actual GNSS signals? Oh wait….

The phase observable

Note: If you need a refresher on signals, head back to **Chapter 3** as we won't be discussing all the basics again here.

The **phase observable** section of GNSS signals is the carrier wave – the blank wave that is modulated with the binary code to become the **modulated carrier wave.** The code is unique to each satellite in most systems (remember that GLONASS is the exception).

The carrier wave is a 3D wave, and is a right hand polarised wave. This means if we were able to look down the centre of the wave in the direction it was travelling, it would appear to rotate in a clockwise direction, as shown in **Figure 6.2(b).**

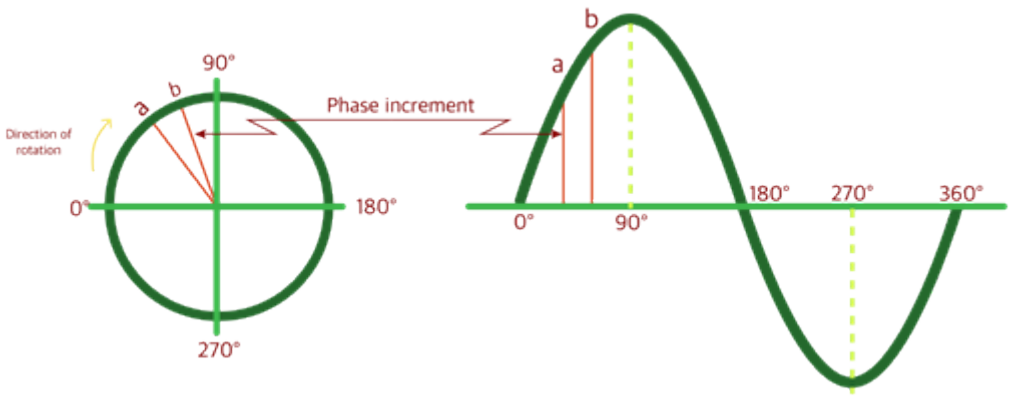


Figure 6.2(b): Phase increment, side and front view

When the carrier wave gets to the GNSS receiver, the first measurement that it observes is a **partial wave**, which is also referred to as a **partial phase**, and is represented by or by upper case Greek letter delta (used to indicate change or difference) and lower case Greek letter Lambda (wavelength) $\Delta\lambda$.

The GNSS receiver is in theory able to measure this partial phase quite accurately – to around of the wavelength. This equates to around 2mm for GPS L1 or L2 signals.

The remainder of the signal is a number of full wavelengths, but unlike code, it's pretty difficult to know exactly how many of the full wavelengths there actually are as they all look the same, and there's no distinct start or end. This unknown number of full wavelengths has to be a whole number (as the GNSS receiver already read the partial phase) and is known as the **ambiguity**, represented by , as shown in **Figure 6.2(c).**

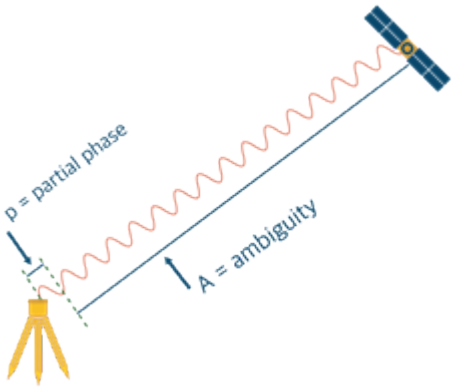


Figure 6.2(c): Ambiguity and partial phase

Combining the ambiguity and partial phase, we now have a range equation for phase observable:

$$R = (p + A)\lambda$$

Where R = range
p = partial phase
A = ambiguity
λ = wavelength

Solving for A is the critical component in GNSS surveying.

So how do we go about this?

In theory the carrier wave wavelength for each signal is a known value, and travels at the speed of light, so the range between the satellite and the receiver could be calculated easily, however, we know that there are a number of errors that impact this calculation.

Imagine if we knew a way to determine the relative distance between two objects another way? Oh wait…

Now we have an equation for range that uses the baseline approach, as well as the signal information:

$$\sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} = (p + A)\lambda$$

Differencing

As you've no doubt started to realise, most of GNSS positioning involves having more than one of something – satellites, receivers, epochs – to remove errors and solve unknowns from a single thing – satellite, receiver, epoch – and phase observable techniques are no different in solving for ambiguity. With the added bonus of removing a whole lot of errors.

In phase observable we refer to this technique as **differencing** – the process of using the difference between two things to create baselines. We're not going to cover the maths behind this in this subject, however, it is important to recognise that each type of differencing has its own equation to describe the baseline (remember back to the range equations in DGPS for a rough indication of how these look). The combinations of these equations help us solve the final baseline equation by getting rid of different variables.

There are three main types of differencing.

Single differencing

Single differencing can happen between two receivers and a satellite, or two satellites and a receiver. Essentially, it's two observations that are assumed to be happening at the same epoch. Let's look at the two receiver scenario first.

As shown in **Figure 6.2(d)**, a common satellite is broadcasting a signal that is being received by two receivers. If we assume we know the partial phase at each receiver, and assume the range from the satellite to the receiver on the left is fixed, we can remove this range from the longer range of the receiver on the right, leaving us with a **phase difference**. As you should remember from the DGPS module, this combination of satellite and receivers means we can remove the satellite clock error from the range as well

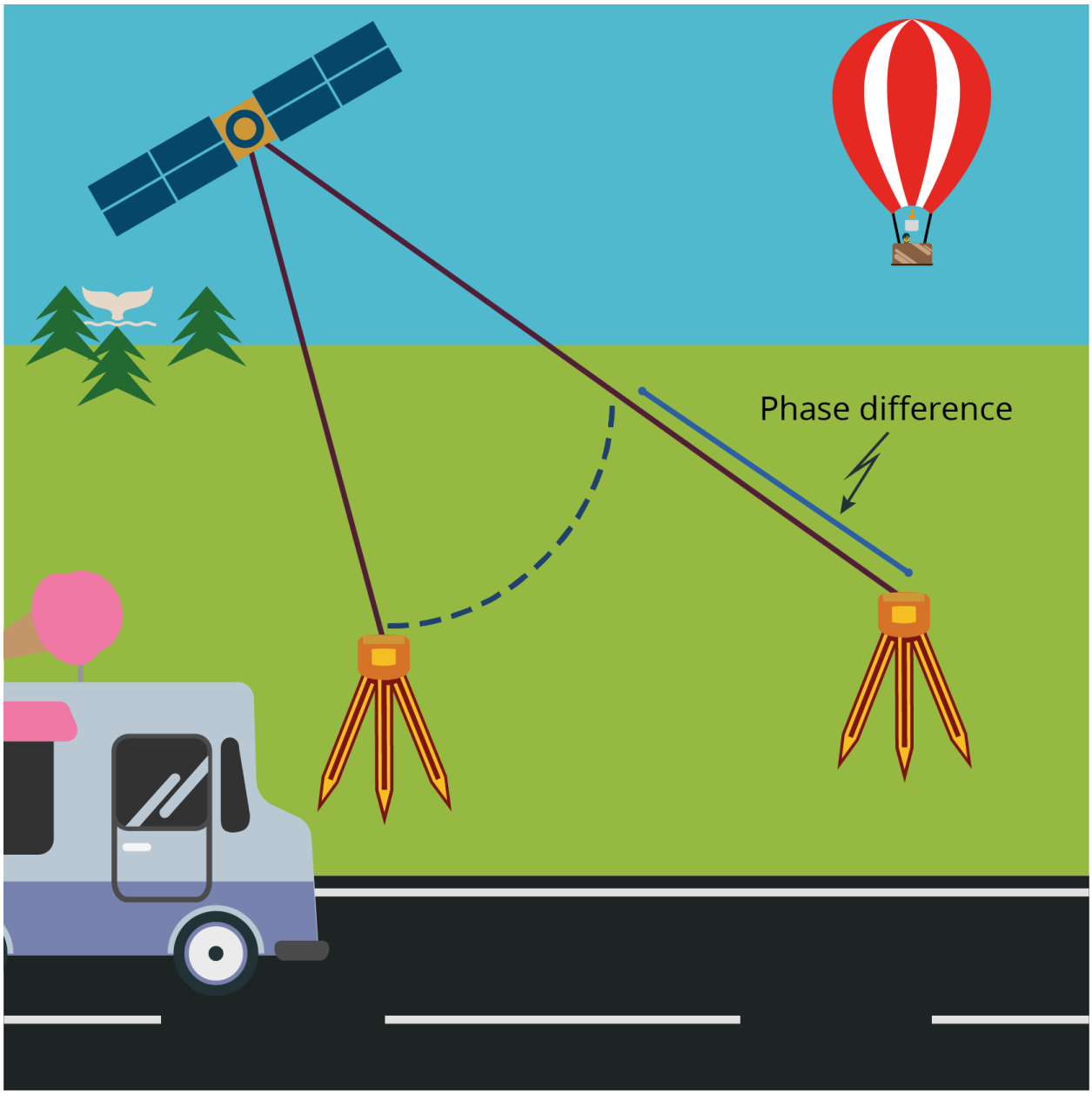


Figure 6.2(d): Single differencing between receivers

Depending on the locations of the receivers, this method of single differencing may also have a significant impact in reducing the atmospheric errors if the receivers are relatively close. The signal is passing through similar atmospheric conditions so the error is considered to be similar for both measurements, and thus can be considered resolved or cancelled out.

Orbit errors are also reduced in this method of single differencing, for the same reasons as the atmospheric errors.

The next type of single differencing is when we have two satellites and one receiver, as shown in **Figure 6.2(e)**. As with the one satellite, two receiver single differencing, we can determine the phase difference, but this version removes the receiver clock error.

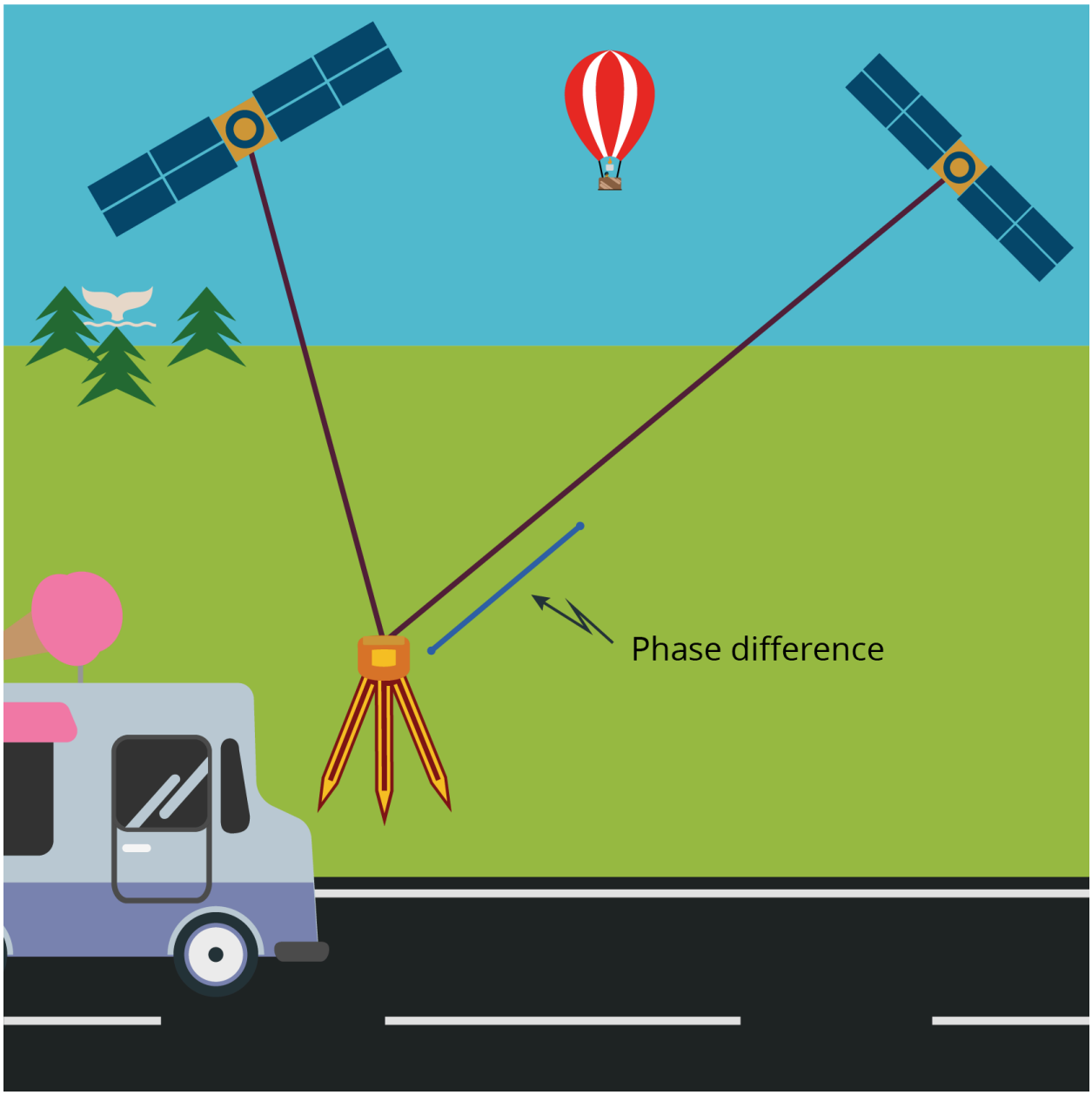


Figure 6.2(e): Single differencing between satellites

Double differencing

When we combine these two versions of single differencing, we get **double differencing**, as shown in **Figure 6.2(f)**. Double differencing is four observations, which are again assumed to be happening at the same epoch. We now have sufficient measurements that the satellite and clock errors are dealt with, and the phase differences can be used to determine a baseline between the two receivers, but only if we can solve the ambiguity.

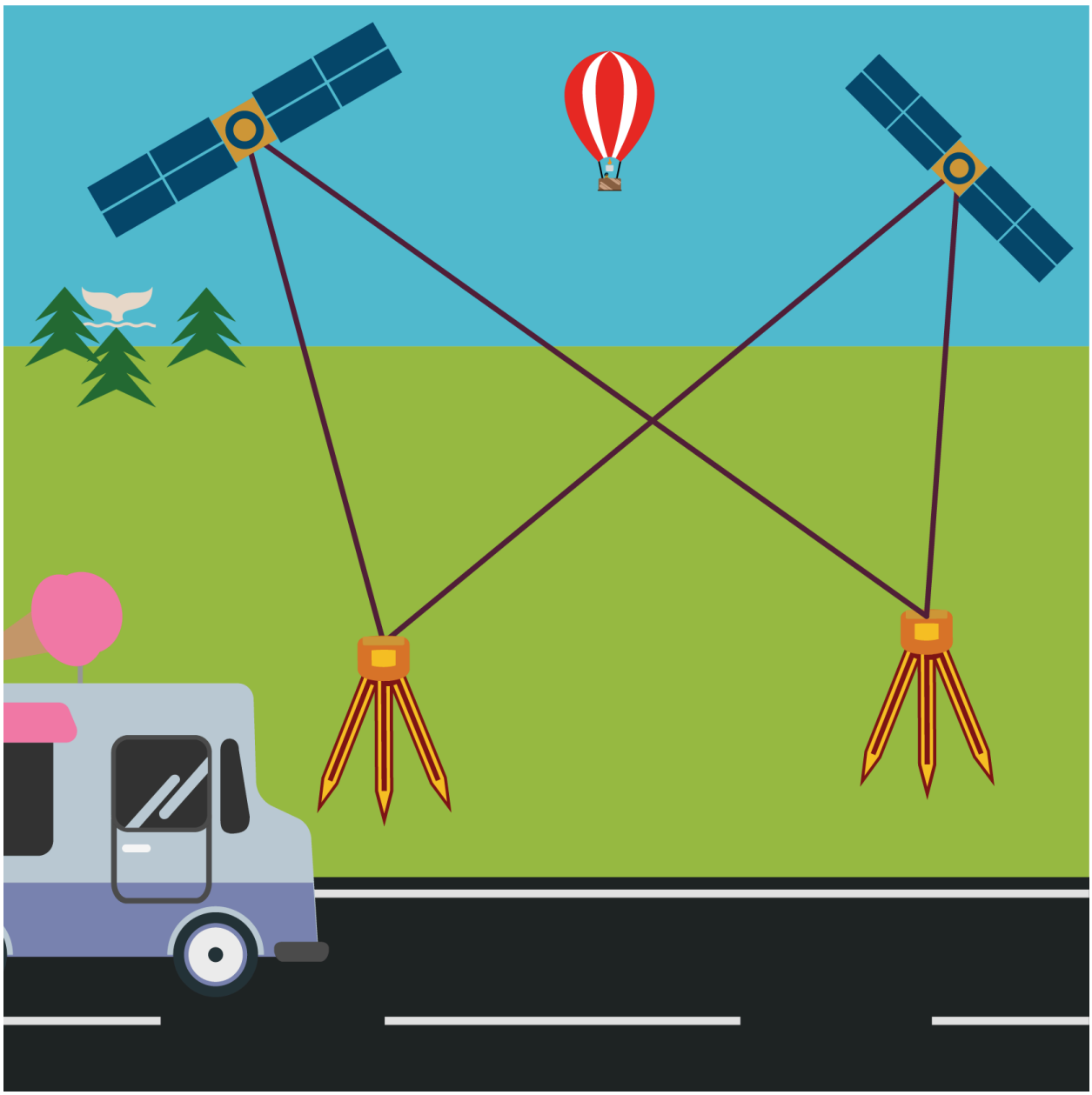


Figure 6.2(f): Double differencing

Triple differencing

All that we really have left to help solve for ambiguity is epochs, so if we took two lots of double differencing, at two different epochs, you guessed it, **triple differencing**! This is shown in **Figure 6.2(g)**.

While we're not going to get into the maths of it in this subject, the triple differencing maths is quite beautiful, as all of the ambiguities of the different signals cancel each other out. This leaves us with an equation for the baseline between the receivers that can be used to provide an estimated value of the baseline.

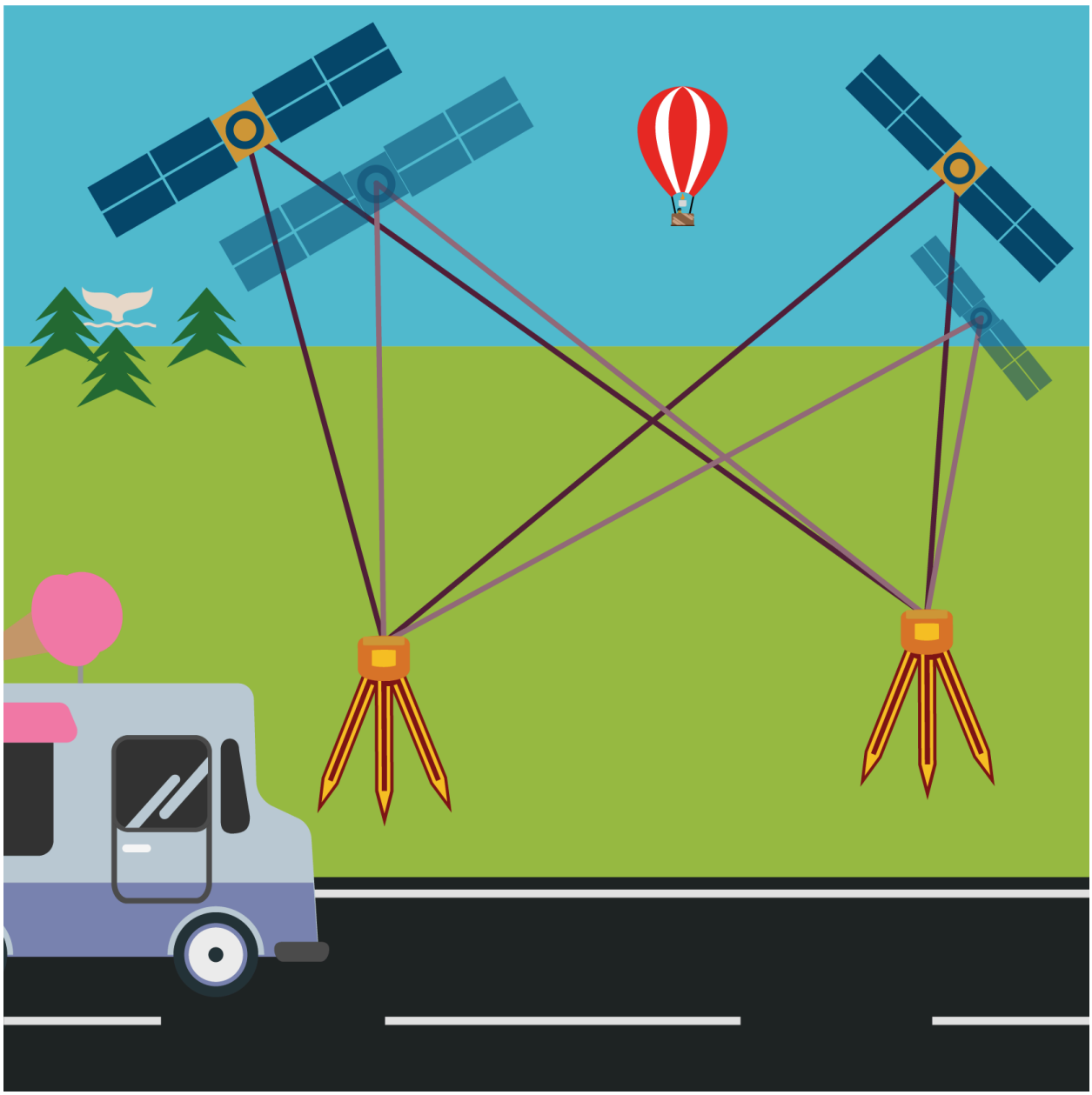


Figure 6.2(g): Triple differencing

Solving ambiguity

The baseline equation in double differencing isn't solved by triple differencing, but the estimated baseline value can be put into the double differencing baseline equation. This allows the receiver to re-estimate the ambiguities as it now has a rough baseline. This refined ambiguity value then can be used to improve the baseline estimate, and so on. This is the mathematical process of **iteration.**

While these iterations are occurring, the receiver is tracking more satellites to be able to solve the double differencing baseline equation. This equation has four unknowns; the three position unknowns – $(\Delta X, \Delta Y, \Delta Z)$ as well as the ambiguity value A. To solve for four unknowns the receiver needs four independent observations – four sets of double differencing. This means we need five satellites to solve the baseline.

This process of resolving the ambiguity using double differencing happens in two stages:

1. An initial estimate of the ambiguity is made using the triple differencing baseline estimate. This estimate is a number with a decimal place, which in computer science or programming language is called a **float value** or **float data type**. Using this estimate, a

baseline solution can be estimated, but because it's not very accurate, it is referred to as a **float solution**.

2. Once the float solution has been determined, and the receiver and software has measured enough independent double differenced solutions, it turns the float value solution of into an **integer value**, which is a whole number. At this point the ambiguities are considered to be resolved, and the receiver can provided a **fixed solution** as it has solved the unknowns.

A fixed solution is about 10 times better than a float solution in terms of accuracy. Given the number of satellites in orbit from different GNSS, a fixed solution should be the preferred option when undertaking phase observable techniques, particularly for those techniques that utilise a GNSS rover for undertaking measurements.

In summary

- **Single differencing** is two observations at the same epoch,
    - two receivers and a satellite eliminates the satellite clock errors, and significantly reduces the atmospheric and satellite orbit errors
    - two satellites and a receiver eliminates the receiver clock errors.
- **Double differencing** is a combination of the two single differencing methods, and removes satellite and receiver clock errors, as well as reducing the orbit and atmospheric errors as mentioned in single differencing.
- **Triple differencing** is two sets of double differencing with the same receivers and satellites, but at different epochs. It eliminates the ambiguity term and allows for the baseline equation to be solved once 4 independent double differenced solutions are observed.
- **A float solution** is when the ambiguity value is estimated as a float value i.e. with a decimal place.
- **A fixed solution** is when the ambiguity is resolved as an integer value – a whole number.

Carrier phase errors

While it outside the scope of this course, it is important to note that an error know as **cycle slip** can occur in phase observable techniques. This is when the signal is lost temporarily and the counting of the number of full cycles needs to commence again. This causes a jump in the data, as the integer changes. This is covered in the Geodetic Surveying A and B courses in more detail.

Baselines + Coordinates

The phase observable measurement technique in GNSS has allowed surveyors to measure to levels of accuracy that were previously unachievable without decades of measurement.

By applying coordinates to relatively accurate networks of GNSS baselines, the state and national **survey control networks** – those networks of permanent survey marks that have accurate coordinates on them, have significantly improved in a relatively short timeframe.

The combination of this level of accuracy, and the significant distances that baselines can be generated over has meant surveyors have an additional suite of tools and techniques at their disposal to complete their work. Not only has GNSS surveying made certain activities faster, but also significantly safer in some application areas, as drones have in recent years for engineering and mining. The next section will discuss how accuracy is measured in GNSS surveying, followed by a summary of the different techniques used in GNSS surveying.

6.3 GNSS ACCURACY

There are different levels of accuracy of GNSS measurements, and these are directly related to the technique and equipment used, as well as the length of the observation times.

SP1

In Australia, the way we describe and determine accuracy is outlined in the **Intergovernmental Committee for Surveying and Mapping (ICSM) Special Publication 1** document, which is referred to as **SP1**. ICSM is a group made up of surveyors and other spatial science professionals from each of the State and Territory Governments, along with representatives from Geoscience Australia.

SP1 is the document that sets out the standards for survey control (permanent survey marks) in Australia, and has six guidelines that cover the different methods of surveying that can be used to generate coordinates for permanent survey marks. These guidelines provide information on the methods and techniques needed to achieve different levels of accuracy, which is defined by the principle of **uncertainty.**

Uncertainty

Uncertainty is quite a simple idea – it is the measure of how wrong a value (in the case of GNSS, the coordinates) may be. It is expressed in the same units as the values, so in the case of GNSS it is usually expressed in metres.

For example, if the plane projection coordinates of a permanent mark were 1000m E and 50,000m N, with an uncertainty value of 0.05m in the horizontal that would mean the coordinates of the mark could be wrong by up to 0.05m.

Uncertainty is expressed as a standard deviation, but one that has been expanded to the 95% confidence interval. Remember from Module 0 the area under the curve of the normal distribution is about 68% at one standard deviation from the mean.

What **95% confidence interval** actually means is that we are 95% confident that a sample of data will contain the true mean of the population of the data, even though we don't know what the whole population looks like.

We achieve the expanded confidence interval by multiplying one standard deviation by a **coverage factor**, which is basically a scaling factor to take one standard deviation (which is about 68%) up to 95%. The coverage factor is represented by k.

The formula for uncertainty is:

$$95\%\ CI = 1\sigma \times k$$

Where 95% CI = 95% Confidence Interval
σ = standard deviation of sample
1.996 = coverage factor for one dimension

Uncertainty in two dimensions

For horizontal coordinates, we have two dimensions, so we can describe uncertainty as an **ellipse**, where one axis is representing the uncertainty of one horizontal direction (such as Easting), and the other ellipse axis represents the other horizontal direction (such as Northing).

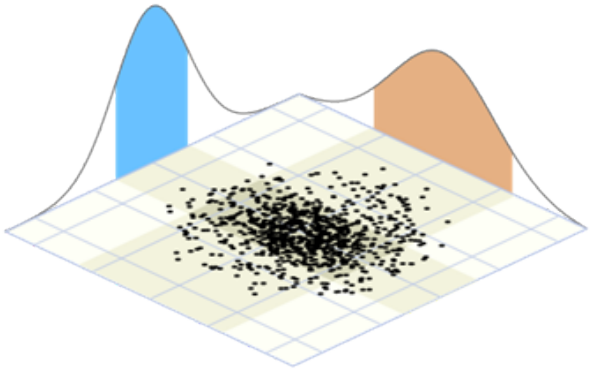


Figure 6.3(a): Diagram showing the interaction of two dependent Gaussian variables. Illustration for Gaussian correlation inequality. Derivative work of the illustration by Lucy Reading-Ikkanda in Quanta Magazine. Used under a CC BY-SA 4.0 licence.

Because we are dealing with data in two dimensions, and each has its own standard deviations, we have two normal distribution graphs that combine to mean the area under them is three dimensional. The easiest way to visualise this is to imagine two normal distribution curves are placed at a right angle to each other, their x axes are now the axes of the ellipse, like in figure 6.3(a). This is known as a **Gaussian bivariate normal distribution.** When there are more than two dependent variables, it is known as a **multivariate normal distribution.**

At one standard deviation, the area under these two normal curves is about 39%, so a larger coverage factor is needed. Thus, the equation for uncertainty for two dimensions is:

95% CI = 1σ x 2.448.
Where 95% CI = 95% Confidence Interval
σ = standard deviation of sample
2.448 = coverage factor for two dimensions

Uncertainty for three dimensions

To describe the uncertainty of a combine horizontal and vertical coordinate, we need a three-dimensional shape. Enter our old friend the **ellipsoid**!

To deal with the third dimension, the third normal distribution can be added to the two dimensional normal distributions, much like the axes for a Cartesian coordinate system. The area under the three curves at one standard deviation is about 20%, so the coverage factor is 2.796 to expand to the 95% confidence interval. From above, this is known as a **multivariate normal distribution.**

The equation for uncertainty in three dimensions is:

95% CI = 1σ x 2.796.
Where 95% CI = 95% Confidence Interval
σ = standard deviation of sample
2.796 = coverage factor for three dimensions

Types of uncertainty

Regarding survey data (including GNSS data), there are three main types of uncertainty. Survey, relative and positional. Each is used to describe a particular type of uncertainty that is useful in surveying, and all are expressed at the 95% confidence interval.

Survey uncertainty is the uncertainty of the coordinates (vertical and/or horizontal) of a permanent survey mark within a survey, without considering any uncertainty of the coordinates of the datum realisation. This means that we are looking at the uncertainty of the survey independent of the datum.

Relative uncertainty is the uncertainty of the coordinates (horizontal and/or vertical) between any two permanent survey marks. They would usual be from different surveys.

Relative uncertainty is useful in understanding the accuracy of different types of surveys, potentially using different techniques, by comparing their uncertainty values.

Sometimes relative uncertainty may be expressed as a ratio.

Positional uncertainty is the uncertainty of the coordinates (horizontal and/or vertical) of a permanent survey mark when its coordinates are considered with respect to the datum. This is basically the final level of uncertainty that a permanent mark can get, it indicates it has been included in some kind of large survey network adjustment. It is the uncertainty value that is recorded any of the government survey control databases.

Survey uncertainty for GNSS

Now that we have a framework to discuss accuracy of GNSS data, we can examine the survey uncertainty of each type of GNSS observation techniques we have learnt about in Chapters 4 to 6.

Figure 6.3(b) is taken from the *Guideline for Control Surveys by GNSS – SP1 Version 2.1*, and it outlines the survey uncertainty of all the code and phase observable techniques we have discussed.

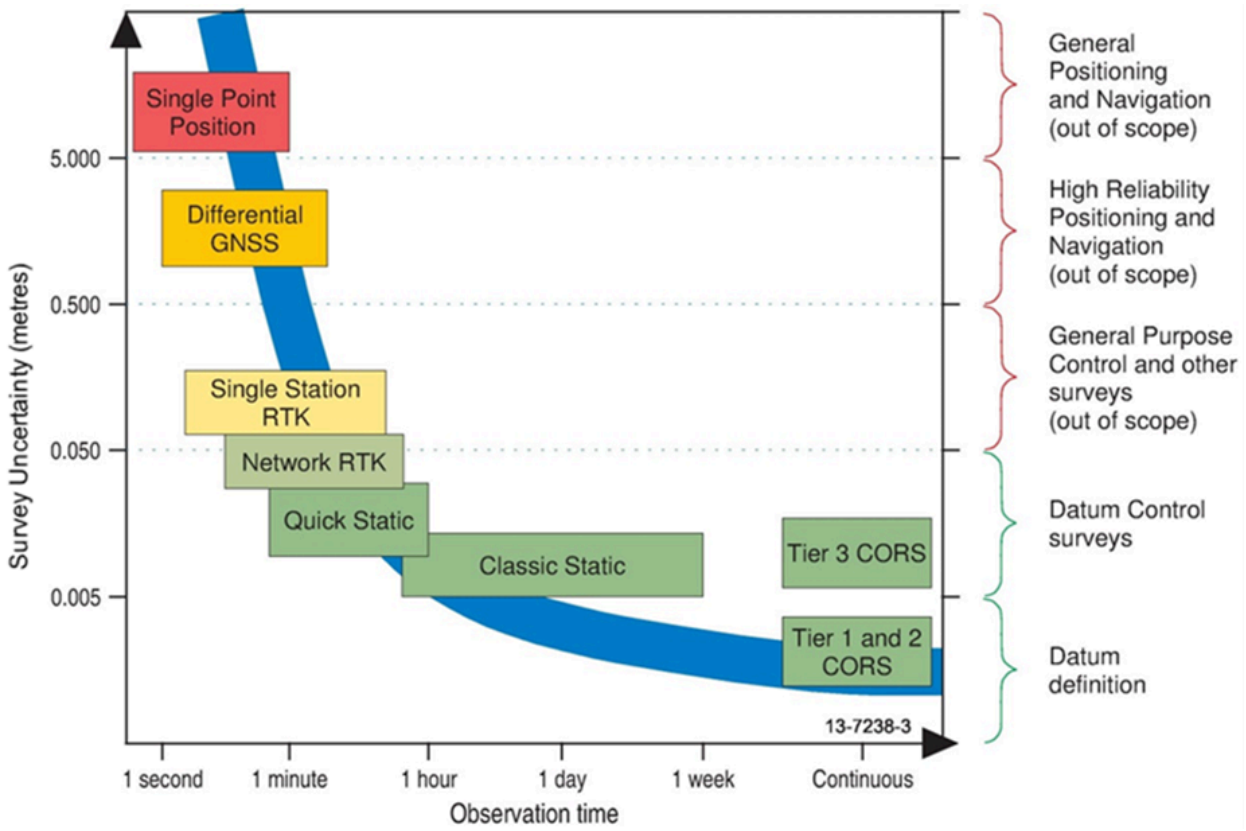


Figure 6.3(b): GNSS accuracy. Commonwealth of Australia – on behalf of the Intergovernmental Committee on Surveying & Mapping. Used under a CC BY 4.0 licence.

As SP1 is dealing with survey control, the code observable techniques are noted as out of scope. This means they are not discussed in SP1 as they are unable to achieve sufficient accuracies to contribute to assigning coordinates to permanent survey marks to be included in control surveys.

6.4 STATIC SURVEYING

Wanted

For the collection of GNSS data

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=422#h5p-3

Note: Image from Pixabay used under CC0 license.

6.5 FAST STATIC SURVEYING

Wanted

For the collection of GNSS data

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=426#h5p-4

Note: Image by Aavaaz home builder used under CC BY-SA 4.0 licence.

6.6 CORS SURVEYING

Wanted

For the Collection of GNSS Data

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=429#h5p-5

Note: Image by IXTA9839 used under CC BY-SA 4.0 licence.

6.7 REAL TIME KINEMATIC SURVEYING

Wanted

For the Collection of GNSS Data

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=432#h5p-6

Note: Image by North Surveying used under a CC BY-SA 4.0 Licence.

6.8 POST PROCESSED KINEMATIC

Wanted

For the Collection of GNSS Data

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=435#h5p-7

Note: Image by Paulbr75 from Pixabay usde under CC0 1.0 Licence.

PART VII

GNSS PROJECTS

7.1 GNSS PROJECTS

Learning Objectives

On the successful completion of this chapter you should be able to:

- explain what kinds of metadata should be collected for GNSS projects
- discuss what features, attributes and attributes values are, and how these are used in GNSS projects
- explain and describe basic data types
- outline and explain the considerations for preparing for GNSS project field work, including selection and management of equipment
- outline and explain the considerations in undertaking GNSS project field work, including advanced data collection techniques and finishing field work.

In the beginning...

Maps have been annotated with information about a location since the first ones were produced. Linking the location of something to the information about it meant that people could understand what was at a location, how to get there and why they might go there. This information is generically referred to as **location or spatial intelligence**.

Aboriginal and Torres Strait Islanders maintained this information about the land through stories, and connected these stories to the land through paintings, effectively creating the first location intelligence databases.

Before computers, this location intelligence was captured in field books, diaries and other records of explorers and surveyors, and linked to maps and plans through filling systems and library records. Anyone wanting to access it would have to physically visit a library or records store and pour over pages and pages of information.

Combining GNSS for accurate positioning with field computing power to collect location intelligence has streamed lined the data collection process significantly, and the inclusion of photos and even 360 degree scans has made the capacity for data collection virtually limitless.

This integrated capacity requires significant planning and preparation for field work – where teams of 20 or more people used to accompany surveyors in undertaking project field work, now days it's more likely to be one or two people.

7.2 DATA

There are three main types of data that we need to collect in any GNSS survey:

- Metadata – Data about our survey
- Position – GNSS Positional information about points or locations
- Attributes – location intelligence

Metadata

Our understanding where data is coming from, why it has been created, who collected it, what errors might be impacting it and how accurate it is are all critical in ensuring it is used appropriately. The creation of all this information about data is an important part of any GNSS survey. This "data about data" is referred to as **metadata**, and should be recorded for every GNSS survey.

Metadata allows any person to look at any dataset and understand how it was created, by whom and a whole bunch of other information. There are a number of international standards about what metadata needs to be created for spatial datasets.

In GNSS surveying, standards like SP1 and a concept called **legal traceability** provide a combined framework for what metadata might need to be collected for a GNSS survey. Legal traceability is the concept that there is a way to link a physical measurement to a standard.

More generally, metadata can most easily be understood as being the WHERE, WHEN, WHY, HOW, WHO and WHAT of a dataset:

- Where was the data collected?
    - the physical location
    - which datum or projection the data was collected in
- When was the data collected?
- Why was the data collected?
    - Was it for a high accuracy survey, or just to locate the footpaths in a park?
    - The WHY will give an indication of the accuracy and quality of a dataset
- How was the data collected?
    - What kind of technique was used?
    - What kind of equipment was used?
    - What were the serial numbers, brand and model of the equipment?
    - Was redundancy of measurements considered, and how was it undertaken?
- Who collected the data?
    - What were the names of the people involved in the collection?
    - What were their qualifications or experience in undertaking this kind of data collection?
    - Who did they work for?
    - Who was the client?
    - Who owns the data?
- What was the data?
    - What format was it collected in?

- What file types were used and what translations or transformation between file types occurred?
- Are there supporting files, like scans of field notes or photos?
- How large are the files?
- Where is the data stored?

An example of basic metadata that could be collected for a GNSS project is shown in **Table 7.1(a)**.

Table 7.1(a): Survey metadata

Survey site:	USQ Springfield	**Name of surveyor/s:**	Jane Smith
Date:	01/05/2054	**Project name:**	SVY1110 Tut 1
GNSS antenna type:	SVMax	**GNSS antenna serial number**	123456
GNSS Receiver type:	SVRX	**GNSS receiver serial number**	789456123
Datum:	GDA2054	**Weather conditions:**	Fine, 30, Slight wind
PM used for datum:	PM123456	**Raw data file name:**	20540501_1110.ab

Position information

As previously discussed, the most important component of positional information is understanding the reference frame in which it is collected, and the accuracy of the data created. Both of these components are influenced by decisions made prior to any observations even being collected, and should be recorded as part of the metadata.

Showing positions to an appropriate accuracy is incredibly important, and the number of decimal places used, or the rounding used is usually the easiest way to communicate this information.

Attributes

The information we collect about what is at a position is often as important as the position itself. This location intelligence may be stored alongside the position information in a Geographic Information System or even a financial asset register – the uses for this data is incredibly varied. It can combine with other spatial datasets for the purposes of research, planning and analysis – from where schools need to go in new developments, to where the closest restaurant is to your location.

The way we describe the components that make up location intelligence is as **features** and **attributes**. In a computer science world, these could also be referred to as objects or classes, but depends on the programming language being used.

Features

A **feature** is an object that has both a position and a type that is either a **point**, a **line** or a **polygon**, and also has location intelligence. Generally, features are also used to group particular kinds of objects together, such as trees, building types or road widths.

A point feature has a single location.

A line feature is a series of point features joined together.

A polygon feature is a series of line features joined to create a closed shape or area.

The decision on which type of feature should be used to represent an object is largely reliant on how the data will be used or displayed, and to what accuracy the data needs to be collected at.

For example, if you wanted to represent all the trees in a park, you might collect them as a point if

they were spread out, or if there was a dense grouping of them you might collect them as a polygon. Equally a park might be shown as a polygon if you were zoomed in closely on a map, but might be shown as a point if you were zoomed out to a city level.

Attributes and attribute values

Attributes are the information about a feature, and a feature can have multiple attributes attached to it. They are most easily considered the questions we want to ask about a feature.

For example, if we were undertaking a survey of a park, and we wanted to know whether it had a carpark, and if it does, how many carparks it had, and how many were for disabled access, the attributes could be:

1. Carpark
2. Number of carparks
3. Number of disabled carparks

The answers to these questions are called the **attribute values**, the individual object's information.

In our example of the carpark above, the attribute values could be:

1. Yes or True
2. 12
3. 2

Attributes are sometimes referred to as **aspatial data –** while they don't have a position in their own right, they are linked to a position, which is considered **spatial data**.

It is important to note that features might be entirely horizontal – they might not have a height attached to them.

Data formats

The way that data is collected is directly related to the way it is stored, and in most systems this is done electronically in some form of database. The allowable format of the data in a database is described by a **data format**.

There are any number of database systems, and each uses specific programming languages to read and write information. The language used also governs the data types used, however, some of the more common ones are covered in **Table 7.1(b)**.

Table 7.1(b): Common data formats

Data Type	Description	Example
Byte	Data	1
Short	Data	65
Int or Integer	Whole number, ranging from -2 billion to +2 billion (short scale)	1,123,465,798
Long	Whole number from -9 trillion to +9 trillion (short scale)	1,123,456,789,123,456,789
Char or Character	A single Unicode character (alphabet)	A
Float or Floating Point	Number with decimal places (the number of places can be specified)	1.1234567
Double	A number with up to 16 decimal places	1.1234567891234567
Boolean	True of False. Sometimes True is represented by 1 and False is represented by 0	TRUE
String	Multiple characters	This is a string
Class	A group of objects defined by a class	Tree
Array	A single object that contains multiple values of the same type. Also referred to as a list.	Gum Oak

The type of data format that we give to an attribute is called an **attribute type**. Attribute types allow us to generate forms that can be used to allow any user with the form to collect data in a predetermined way. This then allows us to ensure the data collected is in an appropriate format to allow for database storage and analysis.

Data priorities

While attribute types are important, deciding whether we require a particular attribute to be collected or not is also important. This decision determines whether the attribute is **mandatory** or **optional**, and is sometimes referred to as the priority of the data**.**

An example of all of the concepts of feature, attribute, attribute name, attribute type, attribute value and data priority are shown in **Table 7.1(c)**.

Feature Name		Attribute			Data
Name	Type	Name	Type	Value	priority
Tree	Point	Species	Menu	Gum Oak	Mandatory
		Height	Float		Mandatory
		Diameter	Float		Optional
		Comments	String		Optional

Table 7.1(c): Example of data structure for attribute collection

7.3 PREPARATION FOR GNSS PROJECT FIELD WORK

Note

Before commencing any kind of field work using GNSS (or any other kind of equipment) it is necessary to ensure that you abide by any Workplace Health and Safety requirements of your organisation.

In preparing equipment for GNSS projects, it is important that users are able to assemble, configure and operate any equipment they will be using appropriately and correctly.

Choosing a GNSS unit

The accuracy that is required for the project will generally be the most critical factor in selecting which GNSS technique you will use, and thus which equipment might be used. The list of equipment available for the various techniques is far too extensive to list here, however, a simple internet search will yield many results.

It should be noted that a number of **survey grade** (a GNSS unit capable of being used in GNSS surveying) receivers are integrated antenna and receiver units, however, geodetic GNSS units will usually consist of a separate antenna and receiver.

In selecting a GNSS unit that is capable of achieving the required accuracy, users should also give consideration to:

- **Durability** – if it is likely to be left outside for long durations, it will need to be able to withstand extreme temperatures and a variety of weather. Users in in areas of high humidity may also need to consider GNSS units that are waterproof to avoid damage to the electronics.
- **Security** – if the GNSS unit is likely to be in place for longer durations unattended, it is more likely to be stolen in certain locations. Considerations for site security should also include disruption by livestock – temporary electric fences might be needed.
- **Weight** – depending on the capacity of the user to operate or transport the GNSS unit, a more suitable unit might need to be selected, or additional people may be required to assist.

Batteries

Different GNSS units will have different batteries, and those batteries will have different discharge lengths (i.e. how long it takes a battery to run out). Users must ensure that they have sufficient batteries to last the entire duration of the field session. In some cases an external battery may need

to be connected to a GNSS receiver to ensure it can observe a length of time uninterrupted. In the case of CORS, additional batteries will be needed on site to ensure that if there are any mains or solar power interruptions, the battery will need to be able to provide sufficient power to avoid any short-term outages.

Batteries also reduce in their charge over time, so a battery that lasted 6 hours when purchased, may only last for 4 hours after one or two years. This will largely depend on the type of battery and its ability to have this 'memory' cleared.

Data storage

The size of the storage required in a GNSS unit will depend on a number of factors:

- whether a unit is using code or phase observable
- the rate (epoch) at which data is being stored
- whether a unit is also collecting attribute data, and what attribute types are being stored
- whether other data, such as photos or scans are being collected.

Equipment settings

Most GNSS units will have default settings that represent some kind of standard configuration, however, often you will want to use a different setting better suited to the location you're visiting, or because of the accuracy you are looking to achieve.

It is advisable to check these settings before leaving home or the office, as adjusting settings on the fly is asking for mistakes to be made if there is any confusion about setting meanings.

The main settings you should consider in any GNSS unit are discussed below.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=398#h5p-1

Field reconnaissance

While the focus of this subject is GNSS, consideration of the other components of GNSS project field work are necessary to ensure that you gain a holistic understanding of field work.

Some field work might be undertaken in urban areas that are only 10 minutes' walk or drive from your workplace, however, some others may require days of driving to reach. In each situation your organisation should have a comprehensive suite of resources that will ensure you are able to undertake the required work in a safe and efficient manner regardless of the location.

Field reconnaissance, often referred to as **rece** (pronounced wrecky) by surveyors, is when you visit the site you will be surveying to understand and number of factors that will include (but not be limited to): the job site, points of access, topography, vehicle access, obstructions, concerns, location of permanent survey marks and accommodation options.

The planning that can be undertaken as a result of field reconnaissance can result in time efficiencies in the field that far outweigh the cost of sending people to undertake the reconnaissance.

Use of online tools such as Google Earth and Maps have provided the opportunity for field reconnaissance to be done virtually, however, the benefits of undertaking field reconnaissance in person cannot be understated as they provide up to date information and conditions.

7.4 UNDERTAKING GNSS PROJECT FIELD WORK

Note

Before commencing any kind of field work using GNSS (or any other kind of equipment) it is necessary to ensure that you abide by any Workplace Health and Safety requirements of you organisation.

So, you've determined what accuracy you need, meaning you have your GNSS technique and equipment all selected, you've done your field reconnaissance and now you're all kitted up and ready to go in the field – congratulations!

Reaching this part of any project can feel like a huge amount of work in its own right, and careful planning and consideration in the planning stage will usually pay off in terms of ensuring the field work component of data collection goes as smoothly as possible.

Field forms

In discussing metadata, position data and location intelligence, we discussed the idea of capturing various pieces of information about our GNSS survey to ensure that we can explain our work. We most often capture this information in a **field form** or **booking sheet**, which will provide a number of predefined fields that are required to be completed.

In GNSS surveying, static techniques still tend to use paper field forms to collect this information, as the amount of information and checks required to be undertaken are extensive. RTK will usually have a digital version of this process that is captured on the field controller.

Data capture techniques

While we might plan to use GNSS to collect information for our project, our best intentions are sometimes not enough to make GNSS work in certain places – like under trees!

To combat this, there are a number of surveying techniques we can use to still let us collect information about objects that can't be collected directly by GNSS. We refer to these as **alternative** or **advanced data capture techniques.**

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/gpsandgnss/?p=404#h5p-2

Finishing GNSS project field work

After capturing data in the field, the job is only half done!

The process of downloading data, post processing (if needed), scanning of paper documents and

storing of files can be lengthy and involved. Organisations should have process to manage this part of any project, and should have consideration for:

- downloading processes
- file naming conventions and file formats
- storage of files in particular storage structures
- scanning resolutions of documents
- resolution of photos and scanning data to be stored
- deleting data from GNSS units
- provision of data to other organisations.

Checklists will assist these processes greatly, and will often be included as part of well-designed field forms.

Data download

To avoid any issues of finding out that no data has been collected or saved once you're back in the office or home, you should ALWAYS check that there is the required data in the GNSS unit. If it is missing for some reason, you may be able to collect it again while you're at the site.

It is also advisable that data be downloaded from the GNSS unit and a copy stored on an external device for redundancy. Accidentally running over your GNSS before you've downloaded it, but after you've finished collecting for the day is never going to be a fun experience!