Università degli studi di Catania

Doctoral Thesis

Atmospheric water vapour tomography for DInSAR application and effect of volcanic plume on the microwaves

Author:
Massimo Aranzulla

Tutor:
Prof.ssa J. Immè

Co-Tutor:
Dr. G. Puglisi

Geofisica e Fisica dell'Ambiente
Dipartimento di Fisica e Astronomia

December 2013

UNIVERSITA' DEGLI STUDI DI CATANIA

Abstract

Fisica
Dipartimento di Fisica e Astronomia

Doctor of Philosophy

Atmospheric water vapour tomography for DInSAR application and effect of volcanic plume on the microwaves

by Massimo Aranzulla

A particular synergy among GPS and SAR techniques, to improve the precision of the current ground deformation monitoring techniques, is investigated. The study of atmospheric anomalies in the GPS EM waves propagation is useful to extrapolate information about the wet refractivity field. Because of its height and the quite variable weather conditions, the estimation of Mount Etna atmospheric anomalies using GPS measurements have noticeable importance to calibrate the SAR interferograms and to establish the "effective" ground deformation of the volcanic edifice. In this study we presented a method to obtain a 3D electromagnetic waves velocity tomography, starting from the GPS output data analysis. Thanks to the agreement between the University of Catania and the INGV-OE, the GPS data used in this work come from "Etn@net" framework. The GPS processing has been carried out by using the GAMIT software, by adopting appropriate processing parameters. A new software was developed for deriving the tropospheric tomography from the GPS data. The code was validated by using synthetic tests which assume different structure of atmospheric anomalies and with random noise about twice severe than the typical errors of the GPS. The results of the tests proved that the tomography software is able to reconstruct the simulated anomalies faithfully. The code was applied to study the structure of the atmosphere in an actual case: the period of August 12, 2011 at 10.00 am. The results of the tomography indicate clearly important features of the refractivity field of the studied day. In conclusion, the synthetic tests and the application on actual data sets of the new software demonstrate that it is able to reveal the tropospheric anomalies and thus it is an useful tool to improve the results of the SAR interferometry.

An indirect outcome of the use of the GPS for the atmospheric sounding on an active volcanic area is that concerning the detection of volcanic products in the atmosphere. Due to the Mt. Etna persistent activity occurred during the last two years, the capability of GPS to detect the volcanic plume was investigated. The Etna volcano is particularly suited for an in-depth investigation into the aptitude of GPS observations to detect volcanic plumes, owing to both the high frequency of explosive episodes and also the well-developed GPS network. Two different approaches were tested, in order to examine the capability of the GPS network to detect volcanic plumes at Etna. The first approach is applied on the signal strength of the GPS L2 carrier phase data, the second approach, instead, is statistical, and analyzes the single difference post fit residual of elaboration signals to assert the hypothesis that the plume affects the GPS data. The proposed method has been tested for the September 4–5, 2007 activity of Mt. Etna. Results from nineteen GPS permanent stations show that during this explosive activity, the GPS residuals definitely include the contribution of the volcanic plume. In the future, data derived from the GPS stations located on Etna's flanks could be used to improve the alerting system of volcanic ash, already operating at the Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Etneo.

Contents

List of Figures

List of Tables

Abbreviations

DInSAR	Differential INnterferometric SAR
DOY	Day Of Year
DWS	Derivative Weight Sum
ECMWF	European Centre for Medium-Range Weather Forecasts
EM	Electro Magnetic
EO	Earth's Observations
Etn@net	Etna network
ESA	European Space Agency
EU	Europe Union
GLONASS	GLObal NAvigation Satellite System
GMF	Global Mapping Fuction
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPT	Global Pressure Temperature
IGS	International GPS Service
INGV-OE	Istituto Nazionale di Geofisica e Vulcanologia - Osservatorio Etneo
IRNSS	Indian Regional Navigational Satellite System
ITRF	International Terrestrial Reference Frame
KHIT	Knot HIT counter
LET	Local Earthquake Tomography
LOS	Line Of Sight
MEO	Medium Earth Orbit
NTP	Normal Temperature and Pressure
PCV	Phase Center Variation
RS	ReSolution

SA	Selective Availability
SF	Spread Function
SAR	Synthetic Aperture Radar
SLR	Side Looking Radar
SNR	Signal to Noise Ratio
VOXEL	VOlume piXEL

A mio padre, che ogni giorno

mi accompagna in questo mondo.

Introduction

The monitoring of the ground deformations is one of the geophysical methods used to study the volcanic and seismogenic areas. Before a volcanic eruption occurs, the stress of the Earth's crust caused by the magma uprising produces slow deformations, which are often used as precursor phenomena. The complexity of the seismic cycle prevents the use of ground deformations as precursors, but their monitoring is fundamental to study co-seismic displacements and post-seismic relaxation.

From over a century ground deformation monitoring is performed by using geodetic techniques, such as leveling, tilt and more recently GPS (Global Positioning System). In recent years also the Earth's Observations (EO) from satellites have gained an important role for the geophysical monitoring as well as for the protection of the environment, the prevention and recovery of disasters and the security homeland. Among the EO techniques, the differential interferometric analysis of Synthetic Aperture Radar (SAR) data is one of the most commonly used for ground deformation monitoring in combination with geodetic techniques, as the GPS.

In this thesis a particular synergy among GPS and SAR techn

investigation into the aptitude of GPS observations to detect volcanic plumes, owing to both the high frequency of explosive episodes and also the well-developed GPS network. Two different approaches were tested, in order to examine the capability of the GPS network to detect volcanic plumes at Etna. The first approach is applied on the signal strength (signal-to-noise ratio) of the GPS carrier phase data, the second approach, instead, is statistical, and analyze the single difference post fit residual of elaboration signals to assert the hypothesis that the plume affects the GPS data.

Thesis outline

Chapter 1 provides an overview on the Global Positioning System involved in this research. Initially we explain the role of the three segments in the GPS system, with particular attention to the space segment for that concerning the detailed description of the GPS blocks and the story of launches. The next generation of GPS, the block III satellites, are examined for future perspectives involved by new frequencies. Brief descriptions of the GPS signal structure and the way in which it is generated are explained. Finally, an overview of the state of the art of the mathematical models concerning the analysis of codes and phases of GPS signals is reported, focusing on the main equations involved and on the errors limiting their accuracy.
Chapter 2 is entirely dedicated to the theory of atmospheric radio wave propagation. A brief review of the Earth atmosphere allows to introduce the electromagnetic wave propagation in the ionized atmosphere and in the neutral atmosphere. Since the interaction of the waves in the troposphere is a central topic of the thesis, is accurately examined. The comprehension of the delay occurring in the neutral atmosphere is a key topic because it is applied in the next chapters. Finally, some applications of space geodetic techniques for atmospheric studies are reported.
Chapter 3 introduces the reader to the experimental section by defining the laboratory (Mount Etna) and the available instruments (the GPS network Etn@net). The GAMIT processing strategy, the parameters and the output file are commented. The chapter concludes with a quality control check of the GPS station by inspecting the LC post-fit phase residuals.
Chapter 4 concerns one of the two main topics treated in this study: the detection of the volcanic plume by GPS signal. To estimate how much the plume affects the GPS data, two different approaches are examined. The first approach is applied on the signal strength of the GPS carrier phase data, also called signal-to-noise ratio. The second approach, statistical, is carried out by analyzing the single difference of the post-fit phase residuals of the GAMIT elaboration.
Chapter 5 deals with the second of the two main topics treated during the PhD: the

methods to obtain a tropospheric wet refractivity tomography.
The first part of chapter explains the methods to solve a finite linear system of equations to apply for tomographic studies. The purpose of this section is the knowledge of the analytical theory behind the tomography. The comprehension of the least square and minimum norm approaches led to apply a mixed approach called damped least square method.
Following we explain the methods to produce a tomography, starting by modifying software originally implemented for seismic tomography, SIMULPS12, to face with the atmospheric problem. To this aim we performed several simulations and tested different network configurations. Due to the modest results in the use of the modified version of SIMPLUS12, we decided to write new tropospheric tomography software.
The last part of the chapter contains the developed procedures and the synthetic tests carried out to assess the suitability of the new software to obtain an appropriate refractivity structure. Finally, a test case was carried out, by using actual GPS data relevant to a selected day used to perform a tomography, which revealed important features of the distribution of the refractivity during the studied day.

Chapter 1

The Global Positioning System

1.1 Overview of GPS

The Global Positioning System (GPS) is a space-based satellite navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an un-obstructed line of sight to four or more GPS satellites. The system provides critical capabilities to military, civil and commercial users around the world. It is maintained by the United States government and is freely accessible to anyone with a GPS receiver. The technology used in GPS was initially created for the United States Air Force, and was born out of the Air Force's need of an accurate guidance system. In 1960 a scientist, Ivan Getting, joint with Air Force and left Raytheon Company, and began developing the GPS. Together with a team of aerospace engineers and scientists, developed the three-dimension GPS concept. In 1957, at the height of the Cold War and the Space Race, the former U.S.S.R. launched Sputnik, the first ever Soviet satellite. While observing Sputnik, U.S. scientists have realized they could track the satellite's orbit by listening to the changes in its radio frequency using the Doppler effect. In 1960, using the knowledge gained from tracking Sputnik, the U.S Navy was able to launch TRANSIT, the first satellite navigational system, which was used to help guide the Navy's fleet of ballistic missile submarines. It used a constellation of five satellites. Since its release for commercial use in 1967, the TRANSIT positioning system found many applications in surveying and geodesy for the establishment of widely spaced network stations over large regions or even over the globe. TRANSIT satellite transmissions at 150 and 400 MHz are more susceptible to ionospheric delays and disturbances than the higher GPS frequencies. Satellites were visible for only about 20 min. A fix position could be obtained every 1-3 hours when a satellite comes into view. In 1967, the U.S. Navy developed the Timation technique, a system that "uses a highly

stable, synchronized clock in the satellite", a technology on which GPS is dependent; by 1974, the first atomic clocks were put into orbit. The GPS project was developed in 1973 to overcome the limitations of previous navigation systems, integrating ideas from several predecessors, including a number of classified engineering design studies from the 1960s. GPS was created and realized by the U.S. Department of Defense.

By 1978, the U.S. military had launched the first four GPS satellites. Until 1983, GPS technology was available for U.S. military use only, until a tragedy of such a magnitude occurred that the benefits of GPS technology could no longer in good conscience be restricted from civilian use. In 1983, Korean Air Lines Flight 007, a civilian Korean Air airplane carrying 269 passengers, was shot down after mistakenly entering Soviet airspace, killing all passengers on board. In response to this tragedy, President Ronald Reagan issued a directive that when GPS technology was operational, it would be free and available to the world. On Feb. 14, 1989, the first modern satellite was launched. By July 17, 1995, the GPS is complete.

This system consists of a constellation of 24 satellites, each orbiting the world once every 12 hours, from 20000 Km above the Earth. In 1996, President Bill Clinton made further advances in the use of GPS as a civilian tool when he issued a policy directive declaring that GPS was a dual-use system technology, meaning it was to be used for both peaceful and military aims.

In 2000, the U.S. military stopped its practice of intentionally blurring the signals for security purposes, and as a result, the civilian demand for GPS technology exploded. As clear, the US Government had a monopoly of knowledge on the GPS. Due to the public desire for these products and technologies, GPS devices immediately became 10-times more accurate, and over time have become cheaper and cheaper to buy and produce. Announcements from Vice President Al Gore and the White House in 1998 initiated these changes. Advances in technology and new demands on the existing system have now led to efforts to modernize the GPS system and implement the next generation of GPS III satellites and Next Generation Operational Control System. In 2000, the U.S. Congress authorized the modernization effort, GPS III. Finally, in 2004, President Bush further ensured the availability and accessibility of personal GPS units when he issued a policy that would ensure that civilian GPS would be free of direct user fees. Soon after, in 2005, the first modernized GPS satellite was launched. This modern satellite began transmitting a second civil signal that "provided civil users with an open access signal on a different frequency." In addition to GPS, other systems are in use or under development. The Russian Global Navigation Satellite System (GLONASS) was developed contemporaneously with GPS, but suffered from incomplete coverage of the globe until the mid-2000s. There are also the planned European Union Galileo positioning system, Chinese Compass navigation system, and Indian Regional Navigational Satellite System (IRNSS).

1.2 GPS segments

The major components of GPS are: the space segment that consists of the GPS satellites, the control segment that consists of a Master Control Station and the user segment. The space segment transmits signals on two phase modulated frequencies. These transmissions are carefully controlled by highly stable atomic clocks inside the satellites. The satellites also transmit a navigation message that contains orbital data for computing the positions of all satellites. The control segment consists of a Master Control Station located near Colorado Springs and several monitoring stations located around the world. The purpose of the control segment is to monitor continuously the satellite transmissions, to predict the satellite ephemeris, to calibrate the satellite clocks, and to update the navigation message periodically. The user segment simply stands for the total user community. The user will typically observe and record the transmissions of several satellites and will apply solution algorithms to obtain position, velocity, and time (fig. 1.1).

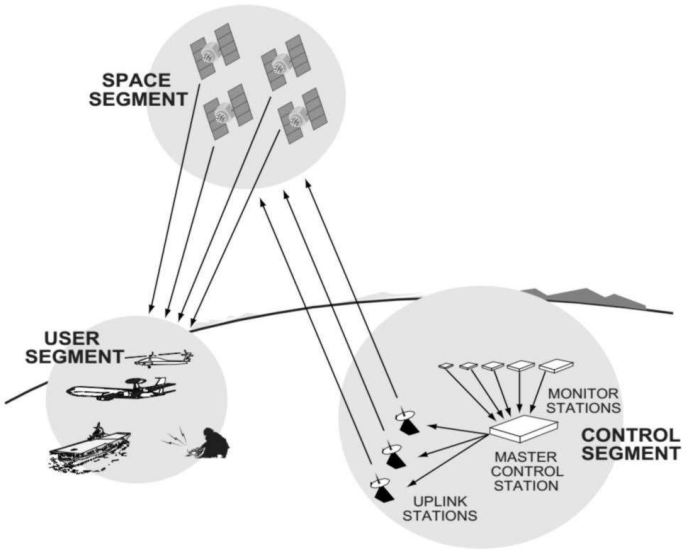


FIGURE 1.1: NAVSTAR GPS Segments

1.2.1 The space segment

In general the satellites orbit the Earth with a speed of 3.9 km per second and have a circulation time of 12 h sidereal time, corresponding to 11 h 58 min Earth time. This means that the same satellite reaches a certain position about 4 minutes earlier each day. The mean distance from the middle of the Earth is 26560 km (fig. 1.2). With a mean Earth radius of 6360 km, the height of the orbits is then about 20200 km. Orbits in this height are referred to as medium Earth orbit (MEO). In comparison, geostationary

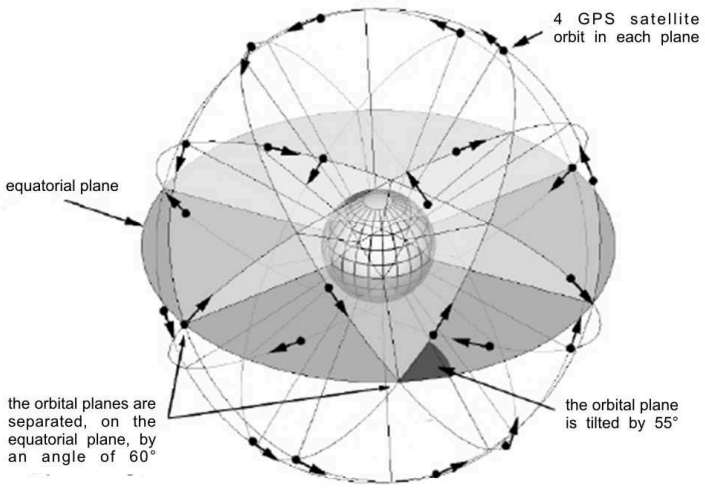


FIGURE 1.2: Orbit plans of the GPS-Satellites (distances are to scale)

satellites like ASTRA or Meteosat satellites orbit the Earth at 42300 km, which is about twice the distance of GPS satellites. The satellites are arranged on 6 planes, each of them containing at least 4 slots where satellites can be arranged, however, to optimize global satellite visibility, the satellites are not evenly spaced within the orbital plane. Today, typically more than 24 satellites orbit the Earth, improving the availability of the system. The inclination angle of the planes towards the equator is 55°, the planes are rotated in the equatorial plane by 60° each other. This means that the orbits range from 55° north to 55° south (fig. 1.2). Block I satellites had an inclination of 63° over the equator. An example of the orbital constellation is shown in Figure 1.4. The

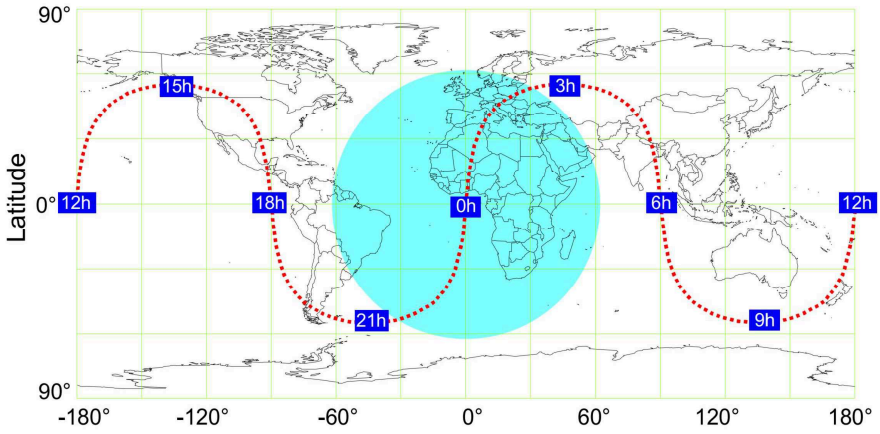


FIGURE 1.3: 24 hour tracking of a GPS satellite with its effective range

six orbital planes, labeled A—F, are evenly spaced in right ascension and are inclined by 55° with respect to the equator. Because of the flattening of the Earth the nodal regression is about −0.04187° per day; an annual orbital adjustment carried out by the control center keeps the orbit in place. The orbital path is close to circular, with a

semi-major axis of about 26000 km. The satellites will complete two orbital revolutions while the Earth rotates 360° (one sidereal day, fig. 1.4). This means the satellites will rise about 4 min earlier each day. Because the orbital period is an exact multiple of the

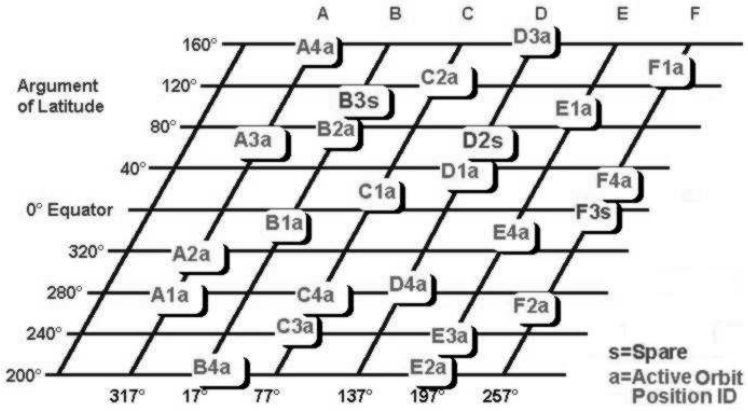


FIGURE 1.4: Approximate GPS satellite constellation slots

period of the Earth's rotation, the satellite trajectory on the Earth (i.e., the trace of the geocentric satellite vector on the Earth's surface) repeats itself daily. Because of their high altitude, the GPS satellites can be viewed simultaneously from a large portion of the Earth. Usually the signals of the satellites are used only once the satellite are above a certain vertical angle, referred to as the mask angle, usually $10 - 15°$. The reason for using a mask angle is that tropospheric effects on the signal propagation are especially unpredictable for altitudes within the mask region.
The additional advances in technology and new demands on the existing system led to the effort to modernize the GPS system. A brief description follows of the GPS Blocks and the story of launch for each satellite that occurred during the last thirty years.

The buildup of the satellite constellation began with a series of space vehicles called the **Block I** satellites. These satellites were launched into 63 declined orbital planes and were positioned within the planes such that optimal observing geometry was available over certain military proving grounds in the continental United States. Dual solar arrays supplied over 400 watts of power, charging NiCd batteries for operations in Earth's shadow. S-band communications were used for control and telemetry, while a UHF channel provided cross-links between spacecraft. The payload included two L-band navigation signals at 1575.42 MHz (L1) and 1227.60 MHz (L2). The final Block I launch was conducted on 9 October 1985, but the last Block I satellite wasn't taken out of service until 18 November 1995, well past its 5 year design life. The Block I series consisted of the concept validation satellites and reflected various stages of system development. Lessons learned from the 11 satellites in the series were incorporated into the fully operational Block II series.

The **Block II** satellites were the first full scale operational GPS satellites, designed to provide 14 days of operation without any contact from the control segment. Block II spacecraft were 3-axis stabilized, with ground pointing. Two solar arrays supplied 710 watts of power, the same S-band and UHF bands were used for control and cross-links between spacecraft. Also the same two L-band navigation signals were used. Each spacecraft carried 2 rubidium and 2 cesium clocks. The first of the nine satellites in the initial Block II series was launched February 14, 1989; the last was launched October 1, 1990. Today's there are no operating Block II satellites.

The **Block IIA** satellites were slightly improved versions of the Block II series, designed to provide 180 days of operation without contact from the control segment. 19 satellites in the Block IIA series were launched, the first on November 26, 1990 and the last on November 6, 1997. Designed to last 7.5 years, on May 2013, eight satellites of this series are still in service. Two of the satellites in this series are equipped with laser retro-reflectors, allowing them to be tracked independently of their radio signals, providing unambiguous separation of clock and ephemeris errors.

The **Block IIR** satellites weight 2,030 kg at launch (1,080 kg once on orbit). The first attempted launch of a Block IIR satellite failed on January 17, 1997 when the Delta II rocket exploded 12 seconds into flight. The first successful launch was on July 23, 1997. Twelve satellites in the series were successfully launched and all operating today's.

The **Block IIR-M** satellites include a new military signal and a more robust civil signal, known as L2C. There are eight satellites in the Block IIR-M series, the first Block IIR-M satellite was launched on September 26, 2005. The final launch of a IIR-M was on August 17, 2009.

State/Block	Launch period	Launched	Operational	Retired	Launch Failures
Block I	1978–1985	11	0	10	1
Block II	1989–1990	9	0	9	0
Block IIA	1990–1997	19	8	11	0
Block IIR	1997–2004	13	12	0	1
Block IIRM	2005–2009	8	8	0	0
Block IIF	From 2010	4	4	0	0
Block IIIA	From 2014				
Block IIIB	Unknown				
Block IIIC	Unknown				
Total		64	32	30	2

TABLE 1.1: Summary of the GPS satellites launch scheduled.

The **Block IIF** is an interim class of GPS satellites, which will be used to keep the Navstar GPS operational until the GPS Block IIIA satellites become operational. They

will be the final component of the Block II GPS constellation to be launched. The spacecraft have a mass of 1,630 kilograms and a design life of 12 years. Like earlier GPS satellites, Block IIF spacecraft operate in semi-synchronous medium Earth orbits, with an altitude of approximately 20,460 kilometres, and an orbital period of twelve hours. The satellites will replace the GPS Block IIA satellites.

Block/ SatID	I	II	IIA	IIR	IIR-M	IIF
01						•
02				•		
03			•			
04			•			
05					•	
06			•			
07					•	
08			•			
09			•			
10			•			
11				•		
12					•	
13				•		
14				•		
15					•	
16				•		
17					•	
18				•		
19				•		
20				•		
21				•		
22				•		
23				•		
24						•
25						•
26			•			
27						•
28				•		
29					•	
30					•	
31					•	
32			•			

TABLE 1.2: Cross reference of Satellite ID to blocks

The **Block III** is the next generation of GPS satellites which will be used to keep the new Navstar GPS operational. The GPS modernization program is an effort to extend the great success achieved in the past three decades. The GPS modernization program

started with the cancellation of Selective Availability (SA) in 2000. It will be followed by the addition of a new military signal (M-code) and a second civil code on L2, then a third civil frequency L5. Further modernization consists of the assessment and design of a new generation of satellites to meet military and civil requirements through 2030. Table 1.1 includes a summary of the launch scheduled of the GPS satellites. Table 1.2 shows the correspondence between the satellite identification number and the belonging block. The purpose of the GPS III program is to deliver major improvements in accuracy, assured service, integrity, and flexibility for civil and military users. The United States Air Force plans to purchase up to 32 GPS III satellites. GPS IIIA-1, the first satellite in the series, is projected to launch in 2014. The first and second GPS III satellites are on schedule for launch in 2014 and 2015.

Block/Sat ID	I	II	IIA	IIR	IIR-M	IIF	IIIA
L1-C/A	•	•	•	•	•	•	•
L1-C							•
L1-P	•	•	•	•	•	•	•
L1-M					•	•	•
L2-C					•	•	•
L2-P	•	•	•	•	•	•	•
L2-M					•	•	•
L5						•	•

TABLE 1.3: summary of the launch schedule of the modernized GPS satellites

Future Block III variants are planned to incorporate additional capabilities. One of the first announcements was the addition of a new civilian-use signal to be transmitted on a frequency other than the L1 frequency used for the existing GPS Coarse Acquisition (C/A) signal. Ultimately, this became known as the L1/2C signal because it is broadcast both on the L1 and L2 frequency. It is in principle transmitted by all block IIR-M and later design satellites.

The L1/2C signal is tasked with providing improved accuracy of navigation, providing an easy-to-track signal, and acting as a redundant signal in case of localized interference. The immediate effect of having two civilian frequencies being transmitted from one satellite is the ability to directly measure, and therefore remove, the ionospheric delay error for that satellite. A major component of the modernization process, a new military signal called M-code was designed to further improve the anti-jamming and secure access of the military GPS signals (tab.1.3). The M-code is transmitted in the same L1 and L2 frequencies already in use by the previous military code, the P(Y) code. The new signal is shaped to place most of its energy at the edges (away from the existing P(Y) and C/A carriers). Specifications of the modulated signals are listed in Table 1.3.

1.2.2 The control segment

The control segment, also referred to as ground segment or operational control system, is responsible for the proper operation of the GPS system. The main tasks performed by the control segment are the following:

- Monitoring and control of satellite orbital parameters.
- Monitoring health and status of the satellite subsystems (solar arrays, battery power and level of propellant used for maneuvers).
- Activation of spare satellites.
- Update of parameters in the navigation message (ephemeris, almanac and clock corrections).
- Resolving satellite anomalies.
- Controlling Anti-Spoofing (A/S).
- Passive tracking of the satellites.

The control segment is comprised of four major subsystems:the master control station, a backup master control station, a network of four ground antennas and a network of globally-distributed monitor stations. The **master control station** is located at Colorado Springs and represents the central control node for the GPS satellite constellation. The master control station is responsible for all aspects of constellation command and control, to include: routine satellite bus and payload status monitoring, satellite maintenance and anomaly resolution, monitoring and management of precise positioning service signal in space performance in support of all performance standards, navigation message data upload operations as required to sustain performance in accordance with accuracy and integrity performance standards. The **distributed monitor stations** are located around the world (red points in figure 1.5) and equipped with standard atomic clocks and GPS receivers to continuously collect GPS data for all the satellites in view from their locations. The collected data are sent to the Master Control Station where they are processed to estimate satellite orbits (ephemeris) and clock errors, among other parameters, and to generate the Navigation Message. Prior to the modernization program, the monitor stations network comprised five sites, the blue points in figure 1.5. In order to increase performance and accuracy, new stations were incorporated into the control segment providing greater visibility of the constellation, the red points in figure 1.5. With this configuration, each satellite is seen from at least three monitor stations, which allows computing more precise orbits and ephemeris data, therefore improving

FIGURE 1.5: Map of distributed monitor stations (from http://www.kowoma.de/).

system accuracy. The **ground antennas** up-link data to the satellites via S-band radio signals. These data include ephemeris and clock correction information transmitted within the Navigation Message, as well as command telemetry from the master control station. This information can be uploaded to each satellite three times per day, i.e., every 8 hours; nevertheless, it is usually updated just once a day. The ground antennas are co-located in four of the monitor stations.

1.2.3 The user segment

The user segment is composed of hundreds of thousands of U.S. and allied military users of the secure GPS Precise Positioning Service, and tens of millions of civil, commercial and scientific users of the Standard Positioning Service. In general, GPS receivers are composed of an antenna, tuned to the frequencies transmitted by the satellites, receiver-processors, and a highly stable clock (often a crystal oscillator). They may also include a display for providing location and speed information to the user. A receiver is often described by the number of channels: how many satellites it can monitor simultaneously. Originally limited to four or five, this has progressively increased over the years so that, today's, receivers typically have between 12 and 20 channels. Many GPS receivers can relay position data to a PC or other device using the NMEA protocol. Although this protocol is officially defined by the National Marine Electronics Association (NMEA), references to this protocol have been compiled from public records, allowing open source tools like "gpsd" to read the protocol without violating intellectual property laws. Other proprietary protocols exist as well, such as the SiRF and MTK protocols.

1.3 GPS signal structure

The radio signals intercepted by the antenna of a GPS receiver are amazingly complex, but this complexity gives GPS its versatility. Once fully operational, the NAVSTAR GPS will be available continuously, 24 hours a day, anywhere on or near the Earth, providing high accuracy positions and velocities in real time to both military and civilian users. To be able to provide such service, the GPS satellites must transmit signals that contain a number of different components. The features that the GPS signal is called upon to provide are summarized in the following:

- Work with one-way measurements(receive only).
- Serve tens of thousands to millions of both military and civilian users.
- Provide accurate, unambiguous, real-time range measurements.
- Provide accurate Doppler-shift measurements.
- Provide accurate carrier-phase measurements.
- Provide a broadcast message.
- Provide ionospheric delay correction.
- Allow simultaneous measurements from many satellites.
- Have interference protection.
- Have multipath tolerance.

GPS is a one-way ranging system: signals are only transmitted by the satellites. This system characteristic is dictated primarily by the fact that GPS is, first and foremost, a military system. Military security requires that users not announce their position by sending signals to the satellites. However, the requirement that GPS be a passive system means that both the GPS receiver and the satellites must contain their own clocks that must be very closely synchronized. This synchronization is achieved using the signals themselves and is the reason why signals from four, rather than three, satellites are normally required to determine a three-dimensional position. Another reason why GPS is a passive system is the need to serve a large number of users. The military requirement alone involves tens of thousands of users. Some projections place the eventual number of civilian users in the billions. For GPS satellites to process two-way communications from each user, the system would have to be much more complicated, which would likely reduce its versatility.

1.3.1 The carriers

The GPS satellite signal belonging to blocks I/II/IIA/IIR/IIR-M, consists of two components, *L*ink 1 or L_1, at a center frequency of 1575,42 [MHz],*L*ink 2 or L_2, at a center frequency of 1227,6 [MHz]. In addition, satellites of blocks IIF/III add a new component, called *L*ink 5 or L_5, at a center frequency of 1176,45 [MHz]. Each of the center frequencies is a coherently selected multiple of 10,23 [MHz] master clock, generated by the atomic clock on board of the satellite. In particular the link frequencies are the following:

- $L_1 = 1575,42[MHz] = 154 \times 10,23[MHz]$, referred to as Link 1 $\Rightarrow \lambda_{L_1} \approx 19[cm]$
- $L_2 = 1227,60[MHz] = 120 \times 10,23[MHz]$, referred to as Link 2 $\Rightarrow \lambda_{L_2} \approx 24[cm]$
- $L_5 = 1176,45[MHz] = 115 \times 10,23[MHz]$, referred to as Link 5 $\Rightarrow \lambda_{L_5} \approx 26[cm]$

These channels lie in a band of frequencies known as the L-band, which starts just above the frequencies used by cellular telephones. The International Telecommunications Union, the radio regulation arm of the United Nations, has set aside special sub-bands within the L-band for satellite-based positioning systems. The L_1, L_2 and L_5 frequencies lie within these bands. GPS signals must provide a means for determining not only high-accuracy positions in real time, but also velocities. Velocities are determined by measuring the slight shift in the frequency of the received signals due to the Doppler effect. In order to measure velocities with centimeter-per-second accuracies, centimeter wavelength (microwave) signals are required. Another reason for requiring such high frequencies is to reduce the effect of the ionosphere. The range between a satellite and a receiver derived by measuring travel times of the signal will therefore contain errors. The size of errors gets smaller as higher frequencies are used. But at the L1 frequency the error can still amount to 30 meters for a signal arriving from directly overhead. For some GPS applications, an error of this size is tolerable. However, other applications require much higher accuracies. This is why GPS satellites transmit on two frequencies. If measurements made simultaneously on two well-spaced frequencies are combined, almost all of the ionosphere's effect can be removed. Although high frequencies are desirable for the reasons just given, they should not be too high. For a given transmitter power, a received satellite signal becomes weaker as the frequency used becomes higher. The L-band frequencies used by GPS are therefore a good compromise between this so-called space loss and the perturbing effect of the ionosphere (ionospheric delays are enormous for frequency ranges below 100 MHz and above 10 GHz). GPS signals, like most radio signals, start out in the satellites as pure sinusoidal waves or carriers. But pure sinusoids cannot be readily used to determine positions in real time. Although the

phase of a particular cycle of a carrier wave can be measured very accurately, each cycle in the wave looks like the next one, so it is difficult to know exactly how many cycles lie between the satellite and the receiver.

1.3.2 The codes

For a user to obtain positions independently in real time, the signals must be modulated; that is, the pure sinusoid must be altered in such a way that time-delay measurements can be made. This is achieved by modulating the carriers with pseudo-random noise (PRN) codes. These PRN codes consist of sequences of binary values (0 and 1) that appear to have been randomly chosen. However, using a mathematical algorithm or special hardware devices, we can generate sequences that do not repeat until after some chosen interval of time. Such sequences are termed pseudo-random. The apparent randomness of these sequences makes them indistinguishable from certain kinds of noise. Although noise in a communication device is generally unwanted, in this case the noise is very beneficial. Exactly the same code sequences are independently replicated in a GPS receiver. By aligning the 'replicated sequence with the received one and knowing the instant of time the signal was transmitted by the satellite, the travel time, and hence the range, can be estimated. Each satellite generates its own unique code, so a GPS receiver can easily identify which signal is coming from which satellite, even when signals from several satellites arrive at its antenna simultaneously. Two different PRN codes are transmitted by each satellite: the C/A, or coarse/acquisition, code, and the P, or precision, code. The C/A code is a sequence of 1023 binary digits, or chips, which is repeated every millisecond. This means that the chips are generated at a rate of 1023 million per second and that one chip has a duration of about one microsecond. Each chip, riding on the carrier wave, travels through space at the speed of light. We can therefore obtain a unit of distance by multiplying the time interval by this speed. One microsecond translates to approximately 300 meters (293 [m]). This is the wavelength of the C/A-code. The precision of a range measurement is determined in part by the wavelength of the chips in the PRN code. Higher precisions can be obtained with shorter wavelengths. To get higher precisions than those afforded by the C/A-code, GPS satellites also transmit tire P-code. The wavelength of the P-code chips is only 30 meters, one-tenth the wavelength of the C/A-code chips; the rate at which the chips are generated is correspondingly 10 times as fast: 10.23 million per second. The P-code is an extremely long sequence. The pattern of chips does not repeat until after 266 days, or about $2.35 \cdot 10^14$ chips! Each satellite is assigned a unique one-week segment of this code, which is initialized at Saturday/Sunday midnight each week. The GPS PRN codes have additional useful properties. When a receiver is processing the signals from one

satellite, it is important that the signals received simultaneously from other satellites do not interfere. The GPS-PRN codes have been specially chosen to be resistant to such interference. Also, the use of PRN codes results in a signal that is essentially impervious to unintentional or deliberate jamming from other radio signals, a possibility that the U.S. Department of Defense, the owner of the system, has to worry about.

1.3.3 The broadcast message

To convert the measured ranges between the receiver and the satellites to a position, the receiver must know where the satellites are. To do this in real time requires that the satellites broadcast this information. Accordingly, there is a message superimposed on both the L1 and L2 carriers along with the PRN codes. Each satellite broadcasts its own message which consists of orbital information (the ephemeris) to be used in the position computation, the offset of its clock from GPS system time, and information on the health of the satellite and the expected accuracy of the range measurements. The message also contains almanac data for the other satellites in the GPS constellation, as well as their health status and other information. The almanac data are used by the receiver to determine the position of each satellite. The receiver uses this information to quickly acquire the signals from satellites that are above the horizon but are not yet being tracked. So, once one satellite is tracked and its message decoded, acquisition of the signals from other satellites is quite rapid. The broadcast message also contains another very important piece of information for receivers that track the P-code. As mentioned earlier, the P-code segment assigned to each satellite is seven days long. A GPS receiver with an initially unsynchronized clock has to search through its generated P-code sequence to try to match the incoming signal. It would take many hours to search through just one second of the code, so the receiver needs some help. It gets this help from a special word in the message called the *h*and-over word (HOW), which tells the receiver where in the P-code to start searching. The GPS broadcast message is sent at a relatively slow rate of 50 bits per second, taking 12.5 minutes for all the information to be transmitted. To minimize the time it takes for a receiver to obtain an initial position, the ephemeris and satellite clock offset data are repeated every 30 seconds. The C/A-code and P-code chip streams are separately combined with the message bits using modulo 2 addition (fig. 1.8). This is just the binary addition that computers and digital electronics do so well. If the code chip and the message bit have the same value (both 0 or both 1), the result is 0. If the chip and bit values are different, the result is 1. The carriers are then modulated by the code and message composite signal. This process is readily accomplished for the L2 channel because it carries only the P-code. But the L1 channel has to carry both the P-code and the C/A-code. This

result is achieved by a clever technique known as *p*hase quadrature. The P-code signal is superimposed on the L1 carrier in the same way as on the L2 carrier. To get the C/A-code signal on the L1 carrier, the un-modulated carrier is tapped off and shifted in phase by 90 degrees. This quadrature carrier component is mixed with the C/A-code signal and then combined with the P-code modulated in-phase component before being transmitted by the spacecraft antenna.

1.3.4 The binary biphase modulation

Carrier waves can be modulated in a number of ways. The amplitude of the carrier, the frequency, or the phase can be varied (fig. 1.6). Phase modulation is the approach

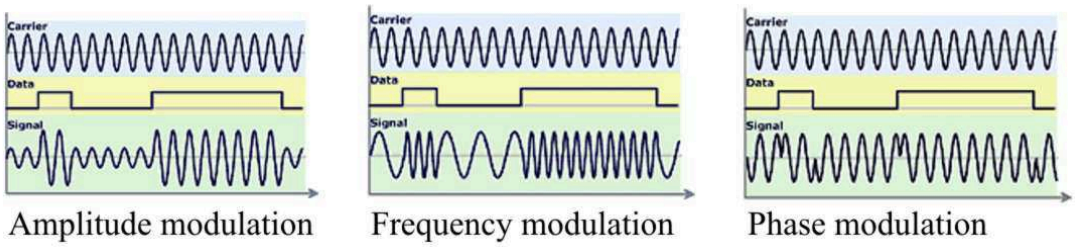


FIGURE 1.6: Amplitude, frequency and phase modulation techniques

used for the GPS signals. Because the PRN codes and the message are binary streams, there must be two states of the phase modulation. These two states are the normal state, representing a binary 0, and the mirror-image state, representing a binary 1 (fig. 1.7). The normal state leaves the carrier unchanged. The mirror-image state results in the unmodulated carrier being multiplied by -1. Therefore, a code transition either from 0 to 1 (normal to mirror image) or from 1 to 0 (mirror image to normal) involves a phase reversal or a phase shift of 180 degrees. This technique is known as "binary bi phase modulation". The composite GPS signal in all different components is illustrated in Figure 1.7, 1.8, 1.9. Forgetting for a moment that GPS is a ranging system, we could consider the satellites to be simply broadcasting a message in an encoded form. The bits of the message have been camouflaged by the PRN code chips. The effect of this camouflaging is to increase the bandwidth of the signal. Instead of occupying only a fraction of one kilohertz, the signal has been spread out over 20 megahertz. Inside a GPS receiver, the code-matching operation "de-spreads" the signal, allowing the message to be recovered. Clearly, this can only be done if the receiver knows the correct codes. The de-spreading operation conversely spreads out any interfering signal, considerably reducing its effect. Common in military circles for ensuring security and combating interference, this technique is known as direct-sequence spread-spectrum communication. Spread-spectrum signals have the additional property of limiting the interference from signals

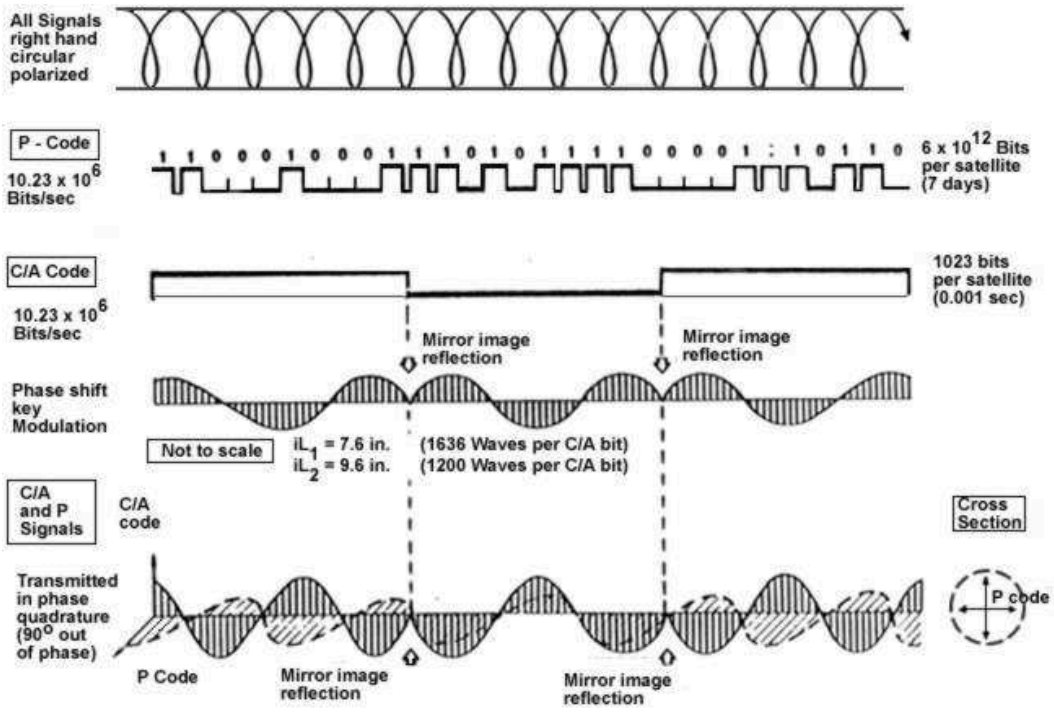


FIGURE 1.7: Structure of various types of GPS signals [Kulshrestha, 1997]

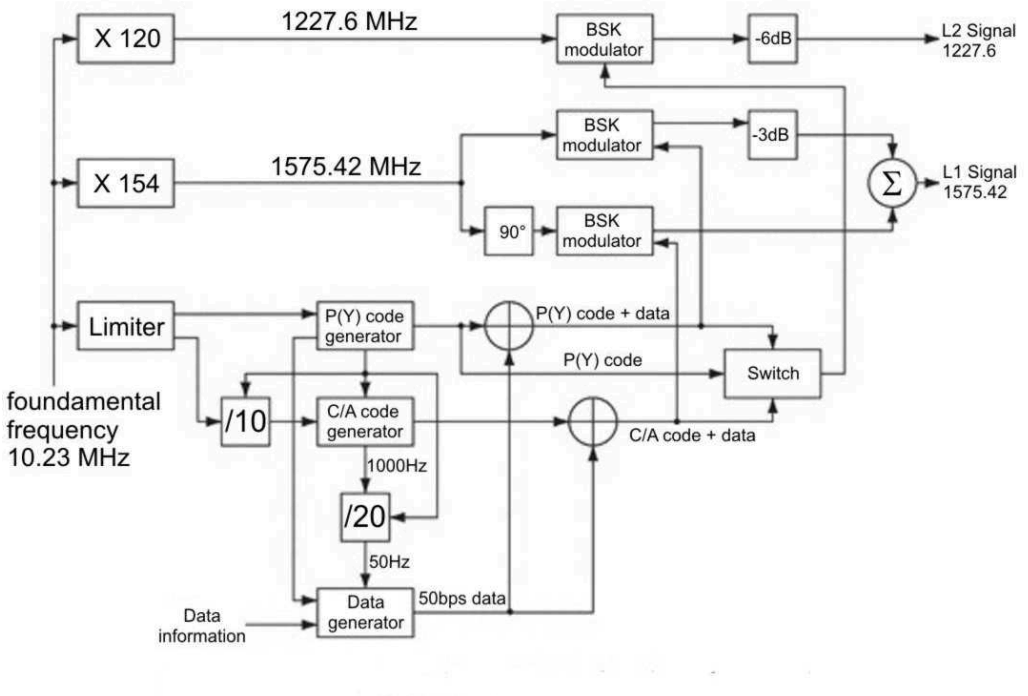


FIGURE 1.8: Block scheme concerning the generation of GPS signal at the satellite

reflected off nearby objects, a phenomenon known as multipath. It should be pointed out that the U.S. Department of Defense has reserved the right to encrypt the P-code trough anti-spoofing (AS). When AS is activated, the Y-code is transmitted instead of the P-code. The Y-code is created by combining a secret W-code with the normal P-code. Only authorized (i.e., military) users will know the W-code. It is believed that AS will only be activated in times of national emergency and for brief test purposes. fig.

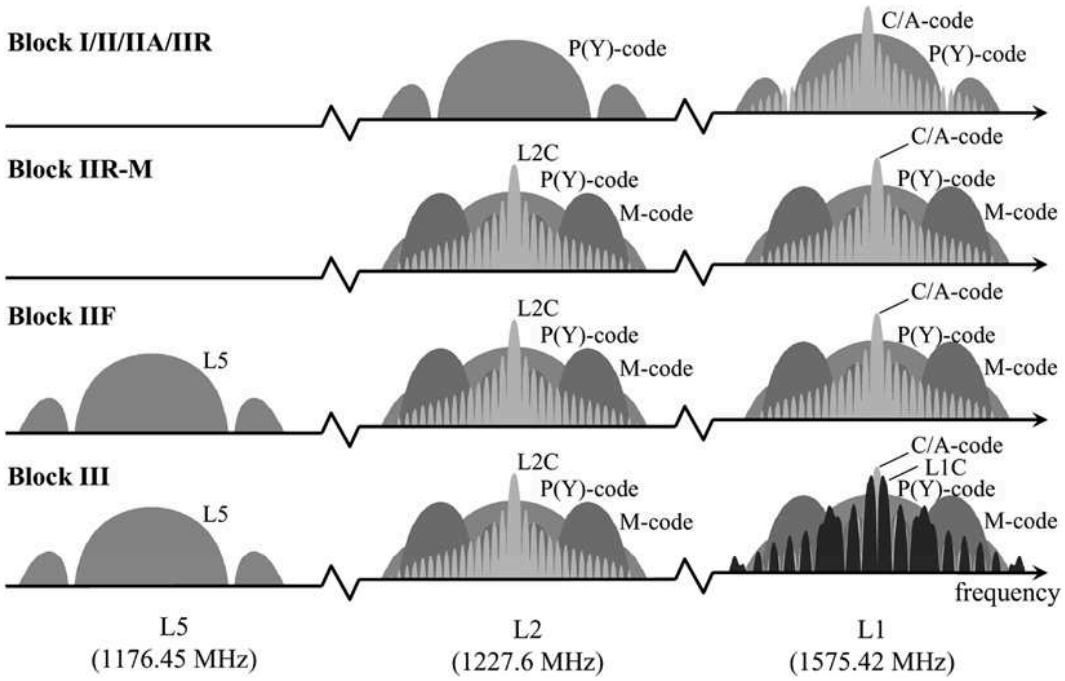


FIGURE 1.9: The normalized power spectral densities of the various GPS signals in decibels

1.9 illustrates the evolution of the GPS signals from the Block I through the Block III satellites. Shown on the figure are the normalized power spectral densities of the various GPS signals in decibels. For reference in viewing the figure, the bandwidths spanned between the first spectral nulls of the P(Y) code and L5 signals are each 20.46 MHz.

1.4 GPS Code and Phase measurements

As said there are two types of GPS observations (observables): pseudoranges and carrier phases. Pseudoranges are often used in navigation. The carrier phases are preferred in high-precision surveying. Increasingly, however, combination solutions of pseudoranges and carrier phases are becoming more common. Although the (undifferenced) phases can be used directly, it has become common practice to take advantage of various linear combinations of the original carrier phase observation, such as double differences and triple differences. Measuring pseudoranges and carrier phases involves advanced techniques in electronics. In this section the equations are developed to use pseudoranges

and carrier phases, as downloaded from the receiver, to determine geocentric positions (point positioning) or relative positions between co-observing stations (differential or relative positioning).

1.4.1 Pseudoranges

The pseudorange is a measure of the distance between the satellite and the receiver's antenna, referring to the epochs (time) of emission and reception of the codes (fig. 1.10). The transmission (travel) time of the signals is measured by correlating identical pseu-

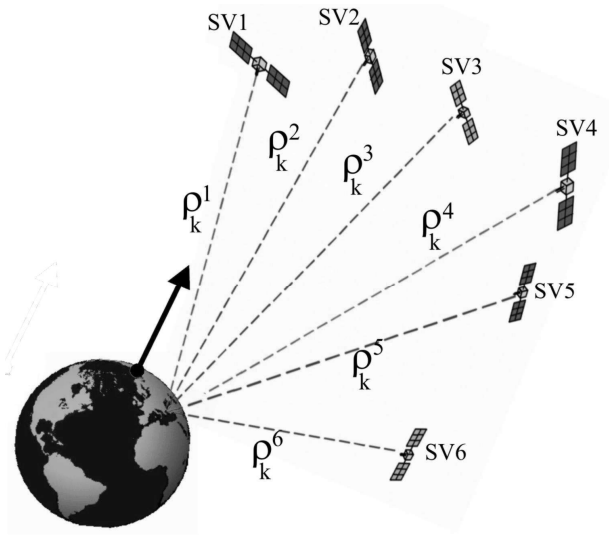


FIGURE 1.10: Pseudoranging to six satellites

dorandom noise (PRN) codes generated by the satellite with those generated internally by the receiver. The code-tracking loop within the receiver shifts the internal replica of the PRN code in time, until maximum correlation occurs. The receiver codes are derived

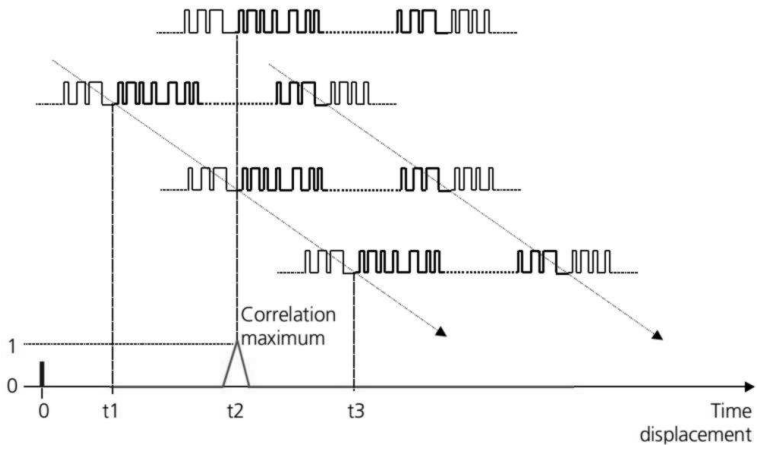


FIGURE 1.11: Measuring time difference

from the receiver's own clock and the codes of the satellite transmissions are generated by the satellite system of clocks. Unavoidable timing errors in both the satellite and the receiver clock will cause the measured pseudorange to differ from the geometric distance corresponding to the epochs of emission and reception. Pseudoranging is applicable to P-codes and/or C/A-codes. Let the symbol t_k denote the nominal time of the receiver clock k (receiver time frame) at the instant of reception of a signal and t^p denote the nominal time of the satellite clock p at emission. Nominal time refers to the time actually indicated by the respective clocks. In this terminology the "satellite clock" indicates space vehicle time as determined by on board atomic clocks which guide the code generation. The nominal times are related to true times, such as GPS time, via the clock error terms as follows:

$$t_{r,k} = t_k + dt_k \tag{1.1}$$

$$t_r^p = t^p + dt^p \tag{1.2}$$

The subscript r denotes true time. The pseudorange measurement between satellite p and receiver k, $P_k^p(t_k)$ is the difference of the nominal times such that

$$P_k^p(t_k) = (t_k - t^p)c \tag{1.3}$$

where c denotes the speed of light. The code correlation process yields the nominal time difference $t_k - t^p$ for a specific code sequence. Because t_k is known, the nominal emission time t^p is readily inferred. In this sense, measuring the reception time and the pseudorange at the receiver is conceptually identical to measuring the nominal emission time in the satellite time frame. The pseudorange (1.3) would equal the geometric distance that the signal traveled (from the instant of emission at the satellite to reception at the receiver), if the propagation medium was vacuum and if there was no clock errors and no other biases. Taking into account these errors and biases, the complete expression for the pseudorange becomes

$$\begin{aligned} P_k^p(t_k) &= (t_{r,k} - t_r^p)c - c(dt_k - dt^p) + I_{k,P}^p(t_k) + T_k^p(t_k) \\ &\quad + d_{k,P}(t_k) + d_{k,P}^p(t_k) + d_P^p(t_k) + \varepsilon_P \\ &= \rho_k^p(t_{r,k}) - c(dt_k - dt^p) + I_{k,P}^p(t_k) + T_k^p(t_k) \\ &\quad + d_{k,P}(t_k) + d_{k,P}^p(t_k) + d_P^p(t_k) + \varepsilon_P \end{aligned} \tag{1.4}$$

The signs in (1.4) are such that the left-hand side, $P_k^p(t_k)$, represents the actual measurement. The scaled difference of the true times corresponds to the vacuum distance that the signal traveled. Subtracting the clock corrections is in agreement with the understanding that the left side of the equation represents the actual measurement. Because the signal travels through the ionosphere and the troposphere, propagation delays of the codes must be taken into account. As it is explained later, the measured range is longer than the geometric distance. Thus, the numerical values of $I_{k,P}^p(t_k)$ and $T_k^p(t_k)$ are always positive and are a function of time depending on the varying conditions along the ionosphere and troposphere path of the signal. The ionospheric correction is a function of the carrier frequency while the tropospheric correction is not. However, no carrier frequency identifier is used in (1.4) in order to emphasize the general form of the pseudorange equation which is applicable to both carriers. The frequency-dependent notation will be introduced later. The symbols $d_{k,P}(t_k)$ and $d_P^p(t_k)$ denote the receiver hardware and satellite hardware code delays. To be general, these delays are marked as time-dependent, although in practice they are, hopefully, stable. The symbol $d_{k,P}^p(t_k)$ denotes the multipath of the codes which depends on the geometry of the antenna and satellite with respect to surrounding reflective surfaces. The term ε_P denotes the random measurement noise. The relativity correction is not listed explicitly because the satellite clock will be corrected to account for relativity. The subscript P identifies all terms whose numerical values are specific to pseudoranges. Note that the tropospheric term has no such subscript. The units of all terms in (1.4) are meters. To avoid overloading expressions for the subsequent development, the ionospheric and tropospheric terms, the hardware delays, the multipath term, and the measurement noise will be listed only when relevant to the discussion.

The topocentric distance $\rho_k^p(t_{r,k})$ equals the geometric distance traveled by the signal between the instants of signal emission and reception. Because distances between the receiver and the satellites differ, signals arriving simultaneously must have been emitted at different times. Since we do not know the true time $t_{r,k}$, the topocentric distance is linearized around the known nominal receiver time t_k by

$$\rho_k^p(t_{r,k}) = \rho_k^p(t_k) + \dot{\rho_k^p}(t_k)dt_k \tag{1.5}$$

where higher-order terms are neglected. Substituting (1.5) in (1.4) gives the fully developed expression for the pseudorange:

$$P_k^p(t) = \rho_k^p(t_k) - c\left[1 - \frac{\dot{\rho}_k^p(t_k)}{c}\right]dt_k + cdt^p + \ldots \tag{1.6}$$

The topocentric distance $\rho_k^p(t_k)$ at the nominal receiver epoch t_k is given by

$$\rho_k^p(t_k) = \sqrt{(u^p - u_k)^2 + (v^p - v_k)^2 + (w^p - w_k)^2} \tag{1.7}$$

The Cartesian coordinates refer to the same reference system, e.g., the WGS 84 system [DMA, 1991]. The triplet (u_k, v_k, w_k) denotes the approximate position of the receiver whose accurate position is sought. The coordinates (u_p, v_p, w_p) denote the position of the satellite at the instant of emission. Computation of the distance $\rho_k^p(t_k)$ is required for both pseudorange and carrier phase processing.

The solution is carried out iteratively, given a satellite ephemeris and approximate position of the receiver. One method begins with the nominal reception time, whereas the other starts with the nominal emission time as computed from the pseudoranges. In the first approach, the iteration begins by assuming an average value for the travel time τ_k^p. Next, the satellite position is interpolated for the epoch $(t_k - \tau_k^p)$ and the topocentric distance is computed, which is then used to recompute the travel time by dividing the distance by the velocity of light. If the discrepancy between the first and second approximation of τ_k^p is greater than a specified threshold fixed by the user,, the iteration is repeated, i.e., a new satellite position is interpolated, a new distance is computed, etc. Usually a couple of iterations are sufficient. In the second method, which requires pseudoranges, the nominal emission time is obtained from equation (1.3) as

$$t^p = t_k - \frac{P_k^p(t_k)}{c} \tag{1.8}$$

This computation is not affected by the receiver clock error dt_k and the satellite clock error dt^p. The computed nominal emission time, however, is corrupted by the ionospheric, tropospheric, hardware, and multipath delays, as seen from equation (1.4). All of these effects are negligible in the present context. For example, if all of these delays together equaled 300 m, light would transverse that distance in just 1 μsec, thus falsifying t^p by that amount. However, during 1 μsec the topocentric distance changes less than 1 mm using the estimate $|\dot{\rho}_k^p| < 800m/sec$. Even the difference between the nominal and true satellite time, which is less than 1 μsec, is negligible as far as computing the topocentric distance is concerned. Having the nominal emission time, the satellite position is interpolated from the ephemeris for that particular instant, giving (u^p, v^p, w^p). Computation of the topocentric distance requires that the Earth's rotation be taken into account during the signal travel time. The process starts by picking a good estimate for the received time, giving a first estimate of the travel time. Next, the Earth is rotated by an amount corresponding to this initial estimate of the travel time, i.e., the coordinates of the receiver change. For this new coordinate position the topocentric distance is computed and divided by the velocity of light to give a better estimate of

the travel time. The Earth is now rotated to correspond to this better estimate of the travel time. The loop is continued until the topocentric range, or the travel time, has converged. Finally, the topocentric range rate $\dot{\rho}_k^p(t)$, which is required in (1.6), can he derived from changes in the topocentric distance over time. If the station clock error is 1 μsec, the topocentric range rate term contributes less than 1 mm to the range, since the absolute value of $\dot{\rho}$ is less than 800 m/sec. Because of this small contribution, the $\dot{\rho}$ term is often neglected and is simply not listed in all equations explicitly. The notation is now simplified by omitting the subscript k to identify the nominal reception time of the station clock k. The exception id dt_k. The expression (1.6) is valid for all code observations. For pseudoranges on the L1 carrier, the complete expression is [Leick, 1994]

$$\begin{aligned} P_{k,1}^p(t) =& \rho_k^p(t) - c\left[1 - \frac{\dot{\rho}_k^p(t)}{c}\right] dt_k + cdt^p + I_{k,1,P}^p(t) + T_k^p(t) \\ &+ d_{k,1,P}(t_k) + d_{k,1,P}^p(t_k) + d_{1,P}^p(t_k) + \varepsilon_{1,P} \end{aligned} \tag{1.9}$$

All bias terms, except the tropospheric delay $T_k^p(t)$, depends on the carrier frequency and are identified accordingly with the subscript 1. Pseudoranges on L2 are identified with subscript 2.

In applications where low accuracy but instantaneous positions are required, the pseudorange is the preferred observable. Given the satellite ephemeris (that is, the position of the satellite at the epoch of emission), the receiver clock error and the receiver location are the only unknowns considered in equation (1.4). The effects of residual satellite clock errors are negligible for the typical navigation solution. This is particularly true because the time errors are indistinguishable from common ionospheric and tropospheric delays. The satellite clocks are constantly monitored with respect to GPS time as maintained by the control center. Actual offsets of the satellite clocks are approximated by polynomials in time, and are transmitted to the user as part of the navigation message. Ionospheric and tropospheric delays can be computed approximately from ionospheric and tropospheric models. Hardware delays and multipath are neglected in the case of the navigation solution. Thus, there are four unknowns left in (1.4). These can be computed using four pseudoranges measured simultaneously to four GPS satellites, giving the following system of equations:

$$P_k^1(t) = \sqrt{(u^1 - u_k)^2 + (v^1 - v_k)^2 + (w^1 - w_k)^2} - cdt_k \tag{1.10}$$

$$P_k^2(t) = \sqrt{(u^2 - u_k)^2 + (v^2 - v_k)^2 + (w^2 - w_k)^2} - cdt_k \tag{1.11}$$

$$P_k^3(t) = \sqrt{(u^3 - u_k)^2 + (v^3 - v_k)^2 + (w^3 - w_k)^2} - cdt_k \tag{1.12}$$

$$P_k^4(t) = \sqrt{(u^4 - u_k)^2 + (v^4 - v_k)^2 + (w^4 - w_k)^2} - cdt_k \tag{1.13}$$

The receiver clock error dt_k is solved together with the position of the receiver's antenna for every epoch. Thus, relatively inexpensive quartz crystal clocks can be used in the receiver rather than expensive atomic clocks. The basic requirement, however, is that there be four satellites visible at a given epoch. This visibility requirement was taken into consideration when designing the basic GPS satellite constellation for universal coverage. The pseudorange positioning technique depends on the accuracy of the satellite ephemeris, because the (u^i, v^i, w^i) coordinates are considered known in (1.10) through (1.13). The best real-time source of ephemeris data is the navigation message read by the receiver. The accuracy of the navigation solution is typically about 5-20 m if the P-code is used and less accurate if the C/A-code is used. However, the achievable position accuracy with codes is subject to changes in receiver technology, e.g. , narrow correlation for C/A-codes, and the quality of satellite transmissions which might improve in the future with replacement satellites.

1.4.2 Carrier phases

The phase observable is the difference between the received satellite carrier phase (as sensed by the receiver's antenna) and the phase of the internal receiver oscillator. The measurements are recorded at equally spaced nominal receiver clock epochs t. The measurement process cannot account for the number of whole carrier waves between the receiver and the satellite. The antenna does not "sense" the exact number of carrier waves to the satellite. The mathematical development uses the fact that the received phase was emitted at some earlier instant of time. The variation of the carrier phase observable in time relates to changes in the topocentric distance in the same way as the integrated Doppler. Sometimes the term interferometry is used to describe carrier phase techniques. Measuring waves at sub-centimeter accuracy demands that utmost attention be given to the treatment of time. After all, the time for light to travel a distance of 3 mm is only 0.01 nsec! A typical pseudorange solution allows us to synchronize the clocks, or to determine the receiver clock errors, at (only) 0.1 μsec accuracy. The remaining, unavoidable clock synchronization errors will be neutralized through appropriate differencing of the carrier phase observables.

Undifferenced Carrier Phase observables

The carrier phase observable $\varphi_k^p(t)$ for station k and satellite p is written as

$$\begin{aligned}\varphi_k^p(t) =& \varphi_k(t) - \varphi^p(t) + N_k^p(1) + I_{k,\varphi}^p + \frac{f}{c} T_k^p(t) \\ &+ d_{k,\varphi}(t) + d_{k,\varphi}^p(t) + d_\varphi^p(t) + \varepsilon_\varphi\end{aligned} \tag{1.14}$$

The symbols $\varphi_k(t)$ and $\varphi^p(t)$ denote the receiver's phase and the received satellite phase at the nominal reception time t. The subscript k is not used in connection with the nominal time t in order to simplify the notation. The symbol $N_k^p(1)$ denotes the initial integer ambiguity. It represents the arbitrary counter setting of the tracking register at the start of observations (phase lock). The terms $I_{k,\varphi}^p$ and $\frac{f}{c} T_k^p(t)$ denote the ionospheric and tropospheric effects. The terms $d_{k,\varphi}(t)$ and $d_\varphi^p(t)$ refer to the hardware delays of the receiver and the satellite, $d_{k,\varphi}^p(t)$ is the multipath, and ε_φ denotes the random carrier phase measurement noise. The carrier identification subscript is not used in (1.14) in order to keep the expression general. Terms having a subscript φ are expressed in units of cycles. The tropospheric delay is converted to cycles using the factor f/c, where f is the nominal carrier frequency.

The receiver's phase lock loop measures the time-varying function $\varphi_k(t) - \varphi^p(t)$ by shifting the receiver-generated $\varphi_k(t)$ to track (lock) the received to $\varphi^p(t)$. Thus, a phase difference is measured rather than the time difference which matches identical codes. As the phase difference increases or decreases by one cycle (2π), the carrier phase observable $\varphi_k^p(t)$ changes by one cycle accordingly (fig. 1.12). If there is a temporary blockage of the

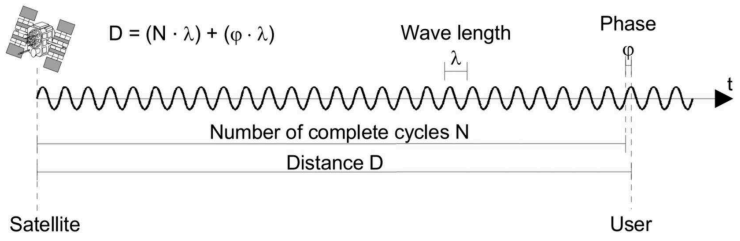


FIGURE 1.12: Measuring phase example

transmitted signal $\varphi^p(t)$, the receiver loses lock on the signal and there is the possibility that the receiver will miss some of the whole cycle change in $\varphi_k(t) - \varphi^p(t)$. Once phase lock is regained, the fractional part of $\varphi_k(t) - \varphi^p(t)$ is again measured correctly, but the counter register might show an incorrect value. In this case a cycle slip is said to have occurred. Plotting equation (1.14) as a function of time will typically show a step function. Often receivers try to resolve cycle slips internally using extrapolation. They usually write flags into the data stream to warn users that a slip might have occurred. For further understanding of $\varphi_k(t) - \varphi^p(t)$ we consider the following simplified situation: let the receiver and the satellite be located on nearby fixed locations, both operating with perfect clocks. The satellite continuously transmits carrier phases and the receiver

internally generates waves at the same frequency. For further simplification we assume that the receiver has been tuned such that at some instant, the receiver-generated phase and the received phase are both zero. Since neither the receiver nor the satellite move in this experiment, the measured difference $\varphi_k(t) - \varphi^p(t)$ will remain zero. Next, let the satellite be moved one wavelength closer to the receiver over a period of 1 sec. Over that same period the receiver will register one extra incoming wave, and the measurement of $\varphi_k(t) - \varphi^p(t)$ will then be one. Thus, the change in the phase observable reflects the change in receiver-satellite separation (fig. 1.13). The difference $\varphi_k(t) - \varphi^p(t)$ is

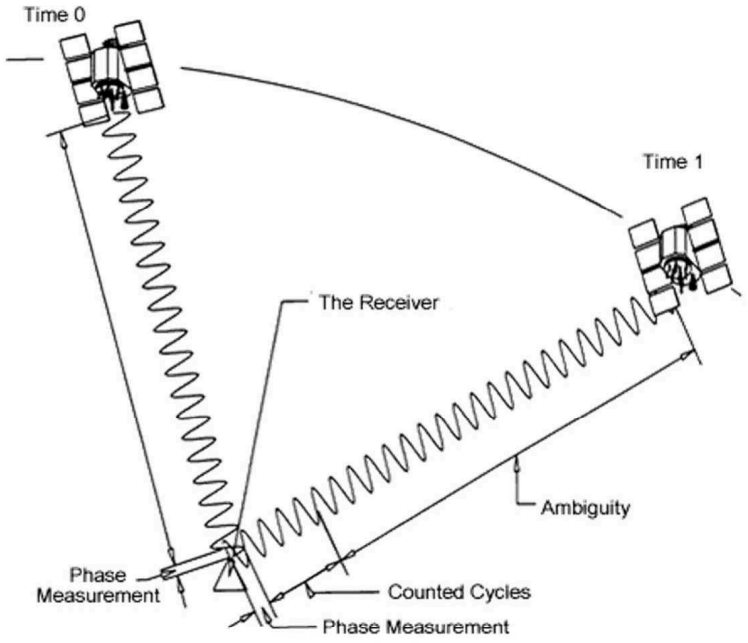


FIGURE 1.13: Satellite-receiver measuring phase example

developed for the vacuum, because the contribution from the propagation media along with other biases and delays are added separately. See equation (1.14). The idea in the development of the carrier phase equation is the equivalence of the received carrier phase and the emitted phase at the satellite, exactly τ_k^p seconds earlier. This fact is expressed by the following equation:

$$\varphi^p(t) = \varphi_T^p(t - \tau_k^p) \tag{1.15}$$

where τ_k^p is vacuum travel time. Signals received at the same time have different emission times because of the different distances to the satellites. The subscript T denotes transmission at the satellite, that is, the phase emitted by satellite p. The phases for nominal time t and true time t_r, are related to the clock errors as follows

$$\varphi_k(t_r) = \varphi_k(t) + f dt_k \tag{1.16}$$

$$\varphi_T^p(t_r - \tau_k^p) = \varphi_T^p(t - \tau_k^p) + f dt^p \tag{1.17}$$

These equations make use of the fact that the frequency of an oscillator is constant with respect to its own time system. Therefore, it is permissible to take the oscillator frequency as a constant and, by definition, to have the clock error absorb any variation. Solving (1.16) for $\varphi_k(t)$ and combining (1.17) with (1.15) gives

$$\varphi_k(t) = \varphi_k(t_r) - f dt_k \tag{1.18}$$

$$\varphi^p(t) \equiv \varphi_T^p(t - \tau_k^p) = \varphi_T^p(t_r - \tau_k^p) - f dt^p \tag{1.19}$$

Substituting these two equations into (1.14), and ignoring the propagation biases and the other delay terms, we obtain

$$\varphi_k^p(t) = \varphi_k(t_r) - \varphi_T^p(t_r - \tau_k^p) - f dt_k + f dt^p + N_k^p(1) \tag{1.20}$$

Since τ_k^p is the signal travel time, which is about 70 msec, the expansion of $\varphi_T^p(t_r - \tau_k^p)$ requires that the satellite frequency be modeled. Here we use the simple model

$$\dot{\varphi}_k^p(t_r) = f + a^p + b^p t \tag{1.21}$$

where a^p and b^p denote the satellite frequency offset and drift respectively at emission time. Thus,

$$\begin{aligned} \varphi_T^p(t_r - \tau_k^p) &= \varphi_T^p(t_r) - \int_T \dot{\varphi}_T^p(t) dt \\ &= \varphi_T^p(t_r) - [f + a^p + \frac{1}{2} b^p \tau_k^p] \tau_k^p \end{aligned} \tag{1.22}$$

Substituting (1.22) into (1.20) gives

$$\begin{aligned} \varphi_k^p(t) =& \varphi_k(t_r) - \varphi_T^p(t_r) - f dt_k + f dt^p \\ &+ [f + a^p + \frac{1}{2} b^p \tau_k^p] \tau_k^p + N_k^p(1) \end{aligned} \tag{1.23}$$

The phase terms $\varphi_k(t_r)$ and $\varphi_T^p(t_r)$ can be ignored altogether because they cannot be separated from the clock errors fdt_k and fdt^p, and because these terms will cancel in the double difference expressions which will be developed in the next chapter. Equation (1.23) shows that the satellite frequency offset and drift affect the received phase by the amount

$$\Delta\varphi^p = (a^p + \frac{1}{2}b^p\tau_k^p)\tau_k^p \tag{1.24}$$

The term $\Delta\varphi^p$ represents the impact of the change in satellite clock error during the travel time of the signal. This impact on the received is, of course, a function of the signal travel time. Since the travel time τ_k^p is short, the impact of the frequency drift term $b^p\tau_k^p$ is usually negligible. Deleting the respective terms, the expression (1.23) becomes

$$\varphi_k^p(t) = -fdt_k + fdt^p + (f + a^p)\tau_k^p + N_k^p(1) \tag{1.25}$$

The last step in the development is to relate the signal travel time to the topocentric range as

$$\tau_k^p = \frac{\rho_k^p(t) + \dot{\rho}_k^p(t)dt_k}{c} \tag{1.26}$$

Substituting (1.26) into (1.25) and including the ionospheric and tropospheric terms and the other delay terms gives

$$\boxed{\begin{aligned}\varphi_k^p(t) =& \frac{f}{c}\rho_k^p(t) - f\left[1 - \frac{\dot{\rho}_k^p(t)}{c}\right]dt_k + fdt^p + N_k^p(1) + \frac{a^p}{c}\rho_k^p(t) \\ &+ I_{k,\varphi}^p(t) + \frac{f}{c}T_k^p(t) + d_{k,\varphi}(t) + d_{k,\varphi}^p(t) + d_\varphi^p(t) + \varepsilon_\varphi\end{aligned}} \tag{1.27}$$

This expression shows that the station clock error enters in two ways: the large term consisting of the product fdt_k and the smaller term, which is a function of the topocentric range rate $\dot{\rho}_k^p$. For a dt_k of merely 1 nsec, the term fdt_k already contributes 1.5 cycles. This is about 150 times the expected carrier phase measurement accuracy. For a phase measurement accuracy of 0.01 cycles the required receiver clock accuracy should be about 0.01 nsec. A station clock error of 1 μsec contributes 0.004 cycles via Doppler term $f\dot{\rho}_k^p(t)/c$ assuming $|\dot{\rho}_k^p(t)| < 800$ m/sec. This geometry-dependent term is zero if the Doppler is zero and changes sign for approaching or departing satellites. If the station clock error does not exceed 0.1 μsec then the $\dot{\rho}$ term is negligible.

Satellite clock errors affect the phase observable through the large term fdt^p and the small frequency-offset term $a^p\rho_k^p/c$. The latter term depends on the travel time of the signal to traverse the topocentric distance. Equation (1.27) is the fully developed expression for the undifferenced carrier phase observation. Because of the linear relationship between the clock errors and the other unmodeled effects, it would be necessary to model the combined effect by one parameter per epoch per observation. Fortunately, most of these error terms are either eliminated or their impact is significantly reduced employing techniques discussed in following sections. It is instructive to compare the carrier phase expression (1.27) with that of the pseudorange (1.9). Apart from the general scaling factor f/c these two expressions differ only because of the satellite frequency offset term a^p. Therefore, the carrier phase equation is sometimes developed in the same way as the pseudorange was derived. The small a^p term is often not listed explicitly or it is simply added to account for the change in received phase due to the satellite frequency offset. Following that line of thought, there is no need to introduce terms like $\varphi_k(t_k)$ and $\varphi_T^p(t_r)$ found in (1.23).

Single Differences

If two receivers k and m observe the same satellite p at the same preset nominal receiver epoch, one can write two equations of the form (1.27). The *between-receiver single-difference* phase observable, more commonly called the *single-difference* phase observable, is defined by

$$\begin{aligned}\varphi_{km}^p(t) &\equiv \varphi_k^p(t) - \varphi_m^p(t) \\ &= \frac{f}{c}[\rho_k^p(t) - \rho_m^p(t)] + \frac{a^p}{c}[\rho_k^p(t) - \rho_m^p(t)] \\ &+ \frac{f}{c}[\dot{\rho}_k^p(t)dt_k - \dot{\rho}_m^p(t)dt_m] + N_{km}^p(1) - f(dt_k - dt_m) \\ &+ I_{km,\phi}^p(t) + \frac{f}{c}T_{km}^p(t) + d_{km,\phi}(t) + d_{km,\phi}^p(t) + \varepsilon_{km,\phi}^p\end{aligned} \tag{1.28}$$

Note how the subscripts k and m in $\varphi_{km}^p(t)$ are used to indicate the difference operation of $\varphi_k^p(t) - \varphi_m^p(t)$. Following the same convention, we have

$$N_{km}^p(1) = N_k^p(1) - N_m^p(1) \tag{1.29}$$

$$I_{km,\varphi}^p(t) = I_{k,\varphi}^p(t) - I_{m,\varphi}^p(t) \tag{1.30}$$

$$T_{km}^p(t) = T_k^p(t) - T_m^p(t) \tag{1.31}$$

$$d_{km,\varphi}(t) = d_{k,\varphi}(t) - d_{m,\varphi}(t) \tag{1.32}$$

$$d^p_{km,\varphi}(t) = d^p_{k,\varphi}(t) - d^p_{m,\varphi}(t) \tag{1.33}$$

$$\varepsilon^p_{km,\varphi}(t) = \varepsilon^p_{k,\varphi}(t) - \varepsilon^p_{m,\varphi}(t) \tag{1.34}$$

Figure 1.14 shows a conceptual view of the single-difference approach. The principal advantage of the single-difference observation is that most of the errors common to the satellite cancel. For example, the large satellite clock term dt^p has canceled. The remaining small term of the satellite frequency offset converges toward zero as the separation of the receivers decreases. The satellite hardware delay also cancels. All of these errors cancel as long as they remain constant between satellite emissions. Recall that

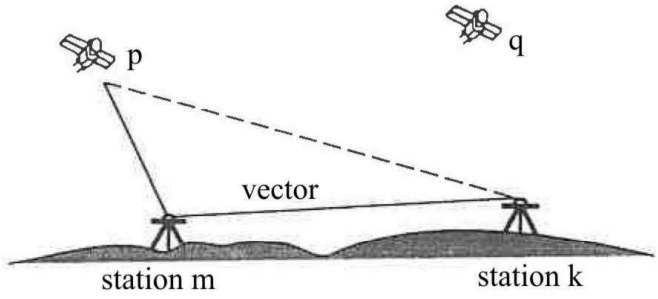


FIGURE 1.14: The single difference. Two receivers observe the same satellite at the same epoch [Leick, 1994]

even though the nominal reception time is the same, the emission times differ slightly because of different distances between the satellite and the two stations k and m. The single-difference observations, however, remain sensitive to both receiver clock errors dt_k and dt_m.

Double Differences

If two receivers k and m observe two satellites p and q at the same nominal time, the double-difference phase observable is

$$\begin{aligned}
\varphi^{pq}_{km}(t) &\equiv \varphi^p_{km}(t) - \varphi^q_{km}(t) \\
&= \frac{a^p}{c}[\rho^p_k(t) - \rho^p_m(t)] - \frac{a^p}{c}[\rho^q_k(t) - \rho^q_m(t)] \\
&+ \frac{f}{c}[\rho^p_k(t) - \rho^p_m(t)] - \frac{f}{c}[\rho^q_k(t) - \rho^q_m(t)] \\
&+ \frac{f}{c}[\dot{\rho}^p_k(t)dt_k - \dot{\rho}^p_m(t)dt_m] - \frac{f}{c}[\dot{\rho}^q_k(t)dt_k - \dot{\rho}^q_m(t)dt_m] \\
&+ N^{pq}_{km}(1) + I^{pq}_{km,\varphi}(t) + \frac{f}{c}T^{pq}_{km}(t) + d^{pq}_{km,\varphi}(t) + \varepsilon^{pq}_{km,\varphi}
\end{aligned} \tag{1.35}$$

where

$$N_{km}^{pq}(1) = N_{km}^{p}(1) - N_{km}^{q}(1) \tag{1.36}$$

$$I_{km,\varphi}^{pq}(t) = I_{km,\varphi}^{p}(t) - I_{km,\varphi}^{q}(t) \tag{1.37}$$

$$T_{km}^{pq}(t) = T_{km}^{p}(t) - N_{km}^{q}(t) \tag{1.38}$$

$$d_{km,\varphi}^{pq}(t) = d_{km,\varphi}^{p}(t) - d_{km,\varphi}^{q}(t) \tag{1.39}$$

$$\varepsilon_{km,\varphi}^{pq}(t) = \varepsilon_{km,\varphi}^{p}(t) - \varepsilon_{km,\varphi}^{p}(t) \tag{1.40}$$

See Figure 1.15 for a conceptual representation. The most important feature of the double-difference observation is the cancellation of the large receiver clock errors dt_k and dt_m (in addition to the cancellation of the large satellite clock errors). These receiver clock errors cancel completely as long as observations to satellites p and q are taken at the same time, or the receiver clock drifts between the observation epochs are negligible. The small clock terms, which are a function of the topocentric range rate, remain in (1.35). Because multipath is a function of the specific receiver-satellite-reflector geometry, it

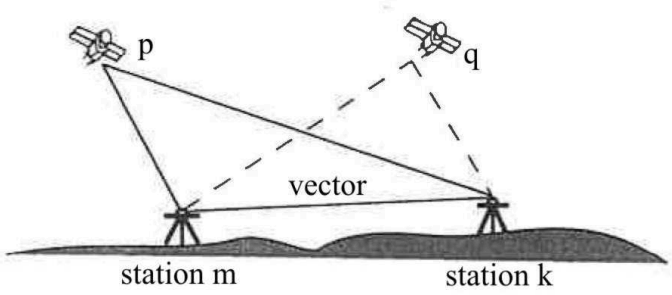


FIGURE 1.15: The double difference. Two receivers observe two satellites at the same epoch [Leick, 1994]

does not cancel in the double-difference observable. The integer ambiguity plays an important role in double differencing. If it is possible during the least-squares estimation to fix the integer, that is, to constrain the estimates $\hat{N}_{km}^{pq}$ to integers, then the fixed solution is the preferred one. Because of residual model errors the estimated ambiguities will, at best, be close to integers. Imposing the integer constraints adds strength to the solution because the number of parameters is reduced and the correlations are reduced as well. Much effort has gone into extending the baseline length over which ambiguities can be fixed. At the same time much research has been carried out in order to develop algorithms which allow the ambiguities to be fixed over short baselines using increasingly shorter observation spans.

Triple Differences

The triple difference is the difference of two double differences between different epochs:

$$\varphi_{km}^{pq}(t_2, t_1) = \varphi_{km}^{pq}(t_2) - \varphi_{km}^{pq}(t_1) \tag{1.41}$$

See Figure 1.16 for a conceptual representation. Assuming phase lock, the initial integer ambiguity $N_{km}^{pq}(t)$ has canceled in equation (1.41). Notice that the

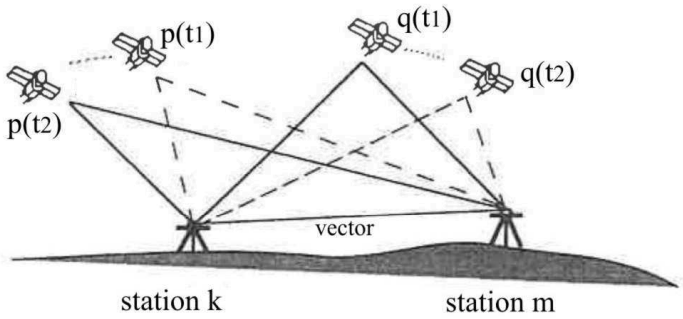


FIGURE 1.16: The triple difference. Two receivers observe two satellites from one epoch to the next [Leick, 1994]

triple and double differences have the same sensitivity with regard to clock errors and satellite frequency offset. The triple-difference observable is probably the easiest to deal with because the ambiguity cancels. The triple-difference solution is often considered a preprocessing technique to get good approximate positions for the double difference solution. Triple differences have a major advantage in that cycle slips are mapped as individual outliers in the computed residuals. Individual outliers can usually be detected and removed or corrected. The resulting cycle slip free (or nearly so) observations can be used in the double difference solution. Because of the additional differencing over time, the triple differences lose some geometric strength. The triple difference can be computed in any order. Consider the following:

$$\begin{aligned}
\varphi_{km}^{pq}(t_2, t_1) &= [\varphi_{km}^{p}(t_2) - \varphi_{km}^{q}(t_2)] - [\varphi_{km}^{p}(t_1) - \varphi_{km}^{q}(t_1)] \\
&= [\varphi_{km}^{p}(t_2) - \varphi_{km}^{p}(t_1)] - [\varphi_{km}^{q}(t_2) - \varphi_{km}^{q}(t_1)] \\
&= \varphi_{km}^{p}(t_2, t_1) - \varphi_{km}^{q}(t_2, t_1) \\
&= [\varphi_{k}^{p}(t_2) - \varphi_{m}^{p}(t_2)] - [\varphi_{k}^{p}(t_1) - \varphi_{m}^{p}(t_1)] \\
&- [\varphi_{k}^{q}(t_2) - \varphi_{m}^{q}(t_2)] + [\varphi_{k}^{q}(t_1) - \varphi_{m}^{q}(t_1)] \\
&= [\varphi_{k}^{pq}(t_2) - \varphi_{m}^{pq}(t_2)] - [\varphi_{k}^{pq}(t_1) - \varphi_{m}^{pq}(t_1)] \\
&= [\varphi_{k}^{p}(t_2) - \varphi_{k}^{p}(t_1)] - [\varphi_{m}^{p}(t_2) - \varphi_{m}^{p}(t_1)] \\
&- [\varphi_{k}^{q}(t_2) - \varphi_{k}^{q}(t_1)] - [\varphi_{m}^{q}(t_2) - \varphi_{m}^{q}(t_1)] \\
&= [\varphi_{k}^{p}(t_2, t_1) - \varphi_{m}^{p}(t_2, t_1)] - [\varphi_{k}^{q}(t_2, t_1) - \varphi_{m}^{q}(t_2, t_1)] \\
&= [\varphi_{k}^{pq}(t_2, t_1) - \varphi_{m}^{pq}(t_2, t_1)]
\end{aligned} \tag{1.42}$$

Whenever a new observable is formed by differencing over time, the epoch identifiers t_1 and t_2 become part of the symbol for the new observable. For example, the symbol $\varphi_{km}^{p}(t_2, t_1)$ denotes the difference over time of two between-receiver single differences involving the same satellite, whereas $\varphi_{k}^{pq}(t)$ is the between-satellite single difference. The symbol $\varphi_{k}^{pq}(t_2, t_1)$ denotes the difference over time of two between-satellite single differences.

1.5 GPS ranging errors

Although the GPS is clearly the most accurate worldwide navigation system yet developed, it still can exhibit significant errors. By understanding these errors, the user can to both hope to reduce them and to understand the limitations of the GPS system. This section provides only an overview of the sources of error. The precision of GPS measurements depends on the accuracy with which we can determine the satellite-receiver ranges. The errors we commit to the extent of the ranges are the bias, i.e., systematic errors in the long term, and random errors, i.e. errors in the short term.

1.5.1 The bias errors

The bias errors are grouped into the five following classes:

- Ephemeris data: errors in the transmitted location of the satellite.
- Satellite clock: errors in the transmitted clock, including SA.
- Ionosphere: errors in the corrections of pseudorange caused by ionospheric effects.
- Troposphere: errors in the corrections of pseudorange caused by tropospheric effects.
- Multipath: errors caused by reflected signals entering the receiver antenna.

The **ephemeris errors** result when the GPS message does not transmit the correct satellite location. It is typical that the radial component of this error is the smallest: the tangential and cross-track errors may be larger by an order of magnitude. Fortunately. the larger components do not affect ranging accuracy to the same degree. Because satellite errors reflect a position prediction, they tend to grow with time from the last control station upload. It is possible that a portion of the intentional SA error is added to the ephemeris as well. However, the predictions are long smooth arcs, so all errors

in the ephemeris tend to be slowly changing with time. Therefore, their utility in SA is quite limited. As reported during phase one, (Bowen, 1986) in 1984, for predictions of up to 24 hours, the RMS ranging error attributable to ephemeris was 2.1 m. These errors were closely correlated with the satellite clock, as we would expect.

The **satellite clock errors** affect both the C/A- and P-code users in the same way. This effect is also independent of satellite direction, which is important when the technique of differential corrections is used. All differential stations and users measure an identical satellite clock error. A major source of apparent clock error is SA, which is varied so as to be unpredictable over periods longer than about 10 minutes. The RMS value of SA is typically about 20 m in ranging, but this can change after providing appropriate notice, depending on need. The U.S. Air Force has guaranteed that the two dimensional RMS (2 DRMS) positioning error (approximately 90th percentile) will be kept to less than 100 m. This is now a matter of U.S. federal policy and can only be changed by order of the President of the United States. More interesting is the underlying accuracy of the system with SA off. The ability to predict clock behavior is a measure of clock quality. GPS uses atomic clocks (cesium and rubidium oscillators), which have stabilities of about 1 part in 10^{13} over a day. If a clock can be predicted to this accuracy, its error in a day ($\approx 10^5 s$) will be about $10^{-8} s$ or about 3.5 m. The experience reported in 1984 was 4.1 m for 24-hour predictions. Because the standard deviations of these errors were reported to grow quadratically with time. an average error of 1-2 m for 12-hour updates is the normal expectation.

The **ionosphere errors** are caused by the free electrons in the ionosphere. When they transit this region, the GPS signals do not travel at the vacuum speed of light. The modulation on the signal is delayed in proportion to the number of free electrons encountered and is also (to first order) proportional to the inverse of the carrier frequency squared ($1/f^2$). The phase of the radio frequency carrier is advanced by the same amount because of these effects. Carrier-smoothed receivers should take this into account in the design of their filters. The ionosphere is usually reasonably well-behaved and stable in the temperate zones; near the equator or the magnetic poles it can fluctuate considerably. An in-depth discussion of this can be found in the following. All users will correct the raw pseudoranges for the ionospheric delay. The simplest correction will use an internal diurnal model of these delays. The parameters can be updated using information in the GPS communications message (although the accuracy of these updates is not yet clearly established). The effective accuracy of this modeling is about 2-5 m in ranging for users in the temperate zones. A second technique for dual-frequency P-code receivers is to measure the signal at both frequencies and directly solve for the delay. The difference between L1 and L2 arrival times allows a direct algebraic solution. This dual-frequency

technique should provide 1-2 m of ranging accuracy, due to the ionosphere, for a well-calibrated receiver.

The **troposphere errors** are also caused by another deviation from the vacuum speed of light. Variations in temperature, pressure, and humidity all contribute to variations in the speed of light of radio waves. Both the code and carrier will have the same delays. This is deeply described in a further chapter devoted to these effects.

The **multipath errors** are caused by reflected signals entering the front end of the receiver and masking the real correlation peak. These effects tend to be more pronounced in a static receiver near large reflecting surfaces, where 15 m or more in ranging error can be found in extreme cases. Monitor or reference stations require special care in siting to avoid unacceptable errors. The first line of defense is to use the combination of antenna cut-off angle and antenna location that minimizes this problem. A second approach is to use so-called "narrow correlator" receivers which tend to minimize the impact of multipath on range tracking accuracies. With proper siting and antenna selection, the net impact to a moving user should be less than 1 m under most circumstances.

1.5.2 The random errors

The random errors are errors in the short term that may not be modeled, or deleted. These errors are inherently present in the measurement. The random error generated by the receiver, is included among the random errors. It comes from the noise of the electronics of the receiver, antenna noise, noise due to the cables, thermal noise, software accuracy, and inter-channel biases. The **receiver errors** are substantially eliminated by the technique. Initially most GPS commercial receivers were sequential in that one or two tracking channels shared the burden of locking on to four or more satellites. With modern chip technology, it is common to place three or more tracking channels on a single inexpensive chip. As the size and cost have shrunk, techniques have improved and five- or six-channel receivers are common. Most modern receivers use reconstructed carrier to aid the code tracking loops. This produces a precision of better than 0.3 m. Interchannel bias is minimized with digital sampling and all-digital designs. The limited precision of the receiver software also contributed to errors in earlier designs, which relied on 8-bit microprocessors. With ranges to the satellites of over 20 million meters, a precision of $1 : 10^{10}$ or better was required. Modem microprocessors now provide such precision along with the co-requisite calculation speeds. The net result is that the receiver should contribute less than 0.5 ms error in bias and less than 0.2 m in noise.

There are other errors induced voluntarily by the operator of the GPS system in order to degrade the positioning accuracy: this is called SA. The degradation of accuracy is obtained simultaneously with two modes:

- degradation of the accuracy of the clocks (δ-process);
- increase of the ephemeris error (ε-process).

The δ-process is obtained by "exciting" the fundamental frequency of the clock satellite. This bias has a direct impact on the observations of pseudoranges, based on the comparison between the clocks of the satellites and the receiver. The ε-process consists in truncation of some of the information submitted in navigation signal, so as not to allow an accurate calculation of the positions of the satellites. The effect of SA due to similar errors in the two receivers that acquire simultaneously, and can be eliminated by differentiating the range of measures. The effect dell'ε-process can be eliminated by using the precise ephemeris to instead of the broadcast. A more complete discussion of individual error sources can be found in next chapters.

1.6 The European navigation satellite system: Galileo

Galileo is the first global satellite positioning and navigation system to be developed by the European Union (EU) and European Space Agency (ESA) for special civil goals in the frame of a joint venture.
The EU considers important to have control of its own independent satellite configuration. It does not want to depend on systems and technologies which were originally developed for military purposes and for applications outside Europe. The Galileo program is of strategic importance for Europe, since it will result in similar revolutionary changes as happened with mobile phones in the recent past, causing the development of a new generation of services. In road and rail traffic, for example, it will enable us to predict travel times. Automatic vehicle control systems will help to decrease the number of traffic jams and accidents. Though various aerial, maritime and land traffic forms are mentioned most frequently in connection with its advantages, its use in areas such as agriculture, fishing, construction, national and civil defense, rescue and life-saving services and telecommunication cannot be neglected. The special on-board characteristics of Galileo satellites will provide new service categories such as traffic, commercial or improved research and rescue functions.
As far as the costs are concerned, they are about 3.2-3.4 billion Euro, which sum should provide for the launches of satellites and construction of a ground service system. The

investment cost is roughly comparable to other European projects and can be compared to the cost of the finished Heathrow Airport Terminal 5 or a modern motorway of 150-km length. The two primary cost bearers of the Galileo project are the EU and ESA, but investments from the private sector are also expected. The space segment of Galileo is planned to consist of 30 satellites deployed in three Mean Earth Orbits. The inclination of orbital planes to the Equator is 56°, and there will be 9 satellites and a spare one in each plane. This satellite constellation will provide better coverage on locations of higher geographical latitude – thus in the northern part of Europe – than it is realized with American GPS satellites today. The mean rotation altitude of satellites is 23,616 km; the orbital periodic time is 14 hours; their weight is 650 kg and their design life will reach 20 years. On board each satellite two atomic clocks will provide the precise frequency. One of them will be a rubidium clock, the other one a hydrogen maser clock. These clocks are made in Europe and are being tested on board the first pilot satellites. The Galileo system time will be adjusted to the International Atomic Time. The contracts for producing the first two test satellites were signed in 2003, and they have been put in orbit since then. Galileo satellites will transmit ten different signals on the following four L-band frequencies: 1176.45 MHz, 1207.14 MHz, 1278.75 MHz and 1572.42 MHz. As can be seen, the frequencies of two Galileo signals are equal to the L1 and L5 frequencies in American GPS system. The service provided by Galileo signals can be divided into two levels: basic level and selective availability. The basic level will be without charge here as well, similarly to the American GPS, but the system operators promise more reliable service for general user applications. To use the other value-added commercial and professional level, naturally, it will be necessary to pay, and it will be available only to authorized users. Consequently, these users have to be provided with excellent operation. The level of selective availability should be so high in the case of certain users that service jamming should happen not at all. The income from service will mean a significant contribution to the cost-effectiveness of this system. The ground service infrastructure of Galileo system and its sub-system of monitor stations will widen with an integrity-monitoring infrastructure separated from the sub-system. The main tasks of the subsystem of monitor stations will be to observe the satellite constellation continuously – that is, to determine each satellite's orbit – to provide time synchronization and to produce navigation messages. All of these are provided by establishing and operating 15 automatic monitor stations, a control center and 4 telemetric stations. An independent monitoring system consisting of a European integrity center and 3 radio stations will be responsible for system integrity. Each user, however, should receive a warning within a few minutes when the satellite system (or certain satellites within it) is out of order (fact and time) and their accuracy numbers. The lack of this very important service is why the American GPS system can not be

used within full limits in such critical navigation fields such as civil aviation control. Another innovation is that transceivers compatible with the Cospas–Sarsat system will be installed in Galileo satellites, which will increase present positioning accuracy by several orders of magnitude compared to current systems. In addition, the L-band transmitters would be suitable for a person in trouble to receive confirmation about that his/her signals had been received and a life-saving mission was underway. The most important element of the user system is the receiver itself.

Chapter 2

Atmospheric Radio-Wave Propagation

2.1 The Earth atmosphere

In space geodesy normally the travel time between a source in space (a satellite) to a receiver on the surface of the Earth is measured. This travel time is then converted to a distance measurement by multiplying with the speed of light in vacuum. The radio signals used by space techniques for geodetic positioning, such as the GPS, have to propagate through the Earth's atmosphere. The atmosphere will introduce an error in the distance since it will affect the propagation path of the signal and since the propagation speed of the signal in the atmosphere is lower than the speed of light in vacuum. Along their paths, the signals are significantly affected by free electrons present in the ionosphere and by the constituents of an electrically neutral atmospheric layer, which includes the lower part of the stratosphere and the troposphere. The effects on radio signals of these two media are of different origin. The ionosphere is a dispersive medium, that is, the free electrons of the ionosphere cause a frequency dependent phase advance or a group delay; the first-order effect can thus be almost completely removed by using dual-frequency observations. The neutral atmosphere causes a non-dispersive delay and the modeling of this effect requires the knowledge of the atmospheric properties in a tridimensional space.

2.1.1 The atmosphere composition

The gaseous envelope surrounding the Earth's surface, bounded to it by gravitational attraction, is by definition the Earth's atmosphere. It is composed of different constituents, which can be grouped under three main categories: dry air, water substance, and aerosols [Iribarne and Godson]. Other forms of classification can be found (e.g. [Fleagle and Businger, 1980]; [Rogers and Yau, 1989]), but the former leads more smoothly to the approach used in radio-wave propagation, which considers the atmosphere as a mixture of two ideal gases: dry air and water vapor.

Constituent	$M_i[\frac{kg}{kmol^{-1}}]$	N_i	$R_i[\frac{J}{kgK}]$	$M_iN_i[\frac{kg}{kmol^{-1}}]$	m_i
Nitrogen (N_2)	28.01348	0.78084	296.804	21.874	0.75520
Oxygen (O_2)	31.9988	0.209476	259.838	6.7030	0.231421
Argon (Ar)	39.948	0.00934	208.133	0.3731	0.0129
Carbon dioxide ($C0_2$)	44.010	0.000314	188.923	0.0138	0.000477
Neon (Ne)	20.1797	18.18 ppm	412.02	0.000367	12.67 ppm
Helium (He)	4.002602	5.24 ppm	2077.28	0.000021	7.24 ppm
Krypton (Kr)	83.80	1.14 ppm	99.22	0.000096	3.30 ppm
Hydrogen (H_2)	2.01588	0.5 ppm	4124.5	0.000001	0.03 ppm
Ozone (0_3)	47.9982	variable	173.225	-	-

TABLE 2.1: Main constituents of the Earth's dry atmosphere below 80 km.

Dry air is a mixture of gases, in which nitrogen, oxygen, and argon are the major constituents and account for about 99.95% of the total volume, as shown in Table 2.1. The first and second column represent the value of the molecular weight (M_i) and volume fraction of each constituent (N_i) [Lide, 1997]. R_i represent the computed values of the specific gas constant of each constituent, that is, $R_i = R/M_i$, where R is the universal gas constant ($R = 8314.510 \pm 0.070 Jkmol^{-1}K^{-1}$ [Lide, 1997]); M_iN_i represent the effective molecular weight of each constituent; the sum of the individual contributions yields an approximate value of 28.9644 $kgkmol^{-1}$ for the (mean effective) molecular weight of dry air; m_i represent the mass fraction of each constituent, using the computed mean molecular weight. With the exception of carbon dioxide, ozone, and other minor constituents, all the gases of this group are mixed in nearly-fixed proportions up to a height of 80-100 km.

Water vapor is almost always presence in the air and can exist in the atmosphere in any of its three physical states: water vapor, liquid droplets and ice crystals. Water in the form of vapor is a highly variable constituent of the atmosphere, both in space and time. The main source of the atmospheric water vapor is the evaporation from bodies of water and transpiration by plants. The concentration is largest near the surface and drops to very small values at higher altitudes. On average, the quantity of water vapor above

10 km is negligible [Davis et al., 1993b]. The water vapor content of the atmosphere is also a function of the local geographic conditions and meteorological phenomena; its concentration is very small in the polar regions and large desert regions, with amounts, of less than 1% of the volume of the air, but quite significant above tropical rain forests, reaching about 4% of the volume of the air (e.g. [Lutgens and Tarbuck, 1979]). The content may vary and the maximum amount possible of water vapor in dry air depends on the temperature of the air. The addition of water vapor to air (making the air humid) reduces the density of the air, which may at first appear counter-intuitive. This occurs because the molecular mass of water (18 g/mol) is less than the molecular mass of dry air (around 29 g/mol). For any gas, at a given temperature and pressure, the number of molecules present is constant for a particular volume (Avogadro's Law). So when water molecules (vapor) are added to a given volume of air, the dry air molecules must decrease by the same number, to keep the pressure or temperature from increasing. Hence the mass per unit volume of the gas (its density) decreases. The other two forms of water are water droplets and ice crystals, of which clouds are made. Aerosols are suspended particles of small size (such as smoke, dust, pollen, and organic matter). The presence of aerosols in the atmosphere is the result of a great number of activities, both human (e.g. industrial and urban pollution) and natural (e.g. volcanic activity and wind-raised dust).

2.1.2 The atmosphere vertical structure

The atmosphere can be divided into a series of layers, based on the chemical composition, vertical distribution of temperature, or degree of ionization. As regards to its chemical composition, the atmosphere is generally divided into two layers: the homosphere and the heterosphere. The **homosphere** is a layer of uniform and relatively well-mixed composition, with respect to the major constituents, extending up to about 100 km (e.g. [Wallace and Hobbs, 1977]; [Miller and Thompson, 1979]; [Iribarne and Cho, 1980]). The **heterosphere** is the layer above the homosphere, with varying composition. In this layer, molecular diffusion becomes an important process, responsible for the stratification of the gases according to their molecular weight. Positively charged particles and free electrons are a significant part of the air composition within this layer. When the temperature distribution is used as the main property in the establishment of an atmospheric segmentation, several layers are considered. The lowest layer of the atmosphere is the **troposphere**. It is characterized by a general constant decrease of the temperature with increasing height of about $6.5°C/km$, on average. The troposphere is an unstable layer, with significant atmospheric turbulence due to vertical convection

currents, particularly in the region near the Earth's surface, denominated as the boundary layer. The thermal structure of this layer, whose thickness can vary from tens of meters to one or two kilometers [Peixoto and Oort, 1992], is mainly controlled by the heating of the Earth's surface, due to solar radiation and turbulent mixing.

The troposphere contains about 80% of the total molecular mass of the atmosphere (e.g. [Wallace and Hobbs, 1977]; [Fleagle and Businger, 1980]) and, as mentioned before, nearly all the water vapor and aerosols. The upper limit of the troposphere is characterized by a, sudden change in the temperature gradient, a level called the **tropopause**, which marks the transition to the stratosphere. The tropopause height is variable and depends on time and place. It typically ranges from ~7-10 km, over the polar regions, to ~16-18 km, over the tropical and equatorial regions (e.g. [Cole et al., 1965]; [Miller and Thompson, 1979]; [Iribarne and Cho, 1980]; [Fleagle and Businger, 1980]; [Barry and Chorley, 1987]). This limit varies also from season to season (higher in summer) and with the passage of air masses. The tropopause is clearly defined in the tropical regions, less well defined at mid-latitudes and can be almost absent in the polar regions, particularly in winter[Fleagle and Businger, 1980].

The **stratosphere** is the second largest layer of the atmosphere, extending upwards to ~50 km. The temperature distribution in this layer is not uniform. In the lower part (up to ~20 km) the stratosphere is approximately isothermal. Then the temperature increases gradually with the height, due to the presence of ozone, which absorbs the ultraviolet solar radiation, reaching a maximum of about $0°C$, at ~50 km [Iribarne and Cho, 1980]. The irregular distribution of ozone over the Earth and its variability with the seasons leads to a variable temperature distribution. The level at which the maximum temperatures are reached is called the **stratopause**, and it marks the transition to the next layer, the **mesosphere**. The main characteristic of the mesosphere is the sharp decrease of temperature with increasing height; at a rate of approximately $3°C/km$ [Miller and Thompson, 1979]. The temperature reaches a minimum of about $\sim 90°C$, at the height of -85 km (e.g. [Fleagle and Businger, 1980]), the coldest point in the atmosphere. This is the level of transition to the thermosphere, called the **mesopause**.

The **thermosphere** is a layer of rapidly increasing temperature, with variations associated with solar activity. The maximum temperature, reached at ~500 km, varies between about $400°C$ and about $2000°C$, depending on the time of day, latitude and solar activity (e.g. [Iribarne and Cho, 1980]; [Fleagle and Businger, 1980]). Due to molecular diffusion, a high concentration of light gases can be found in this layer, making the air density very low. The process of dissociation of the molecular oxygen and other constituents is also significant in this layer; above ~130 km, most oxygen is in atomic form. The upper limit of this layer is not well defined (300 - 500 km) and is bounded by a nearly isothermal layer, the **thermopause**. The **exosphere** is defined as the region where the mean free path (average distance traveled by a molecule between

two collisions) is so great that the lighter particles tend to escape from the atmosphere. The main constituents of this layer are atomic oxygen, ionized oxygen, and hydrogen atoms.

The outermost layer of the atmosphere is the **magnetosphere**, characterized by a supremacy of the Earth's magnetic field over the gravitational field in the distribution of electrons and protons. One peculiarity of this layer is the existence of the Van Allen radiation belts, which are zones of a near-permanent concentration of charged particles. Beyond this layer, the Earth's atmosphere merges with that of the sun (fig. 2.1). The

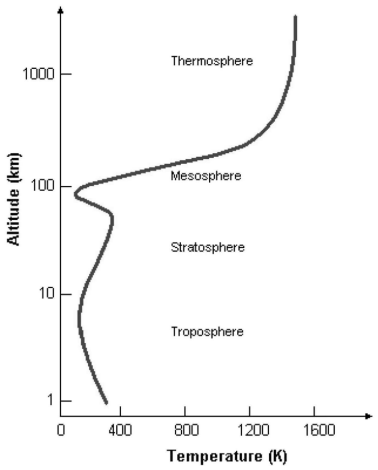


FIGURE 2.1: The first 100 Km of temperature atmospheric profile

ionization state of the atmosphere created by solar radiation has also been used to characterize the atmosphere.

The radiation emitted by the sun produces significant ionization in the upper atmosphere. The term **ionosphere** is used to designate the region of the atmosphere characterized by a strong degree of ionization, produced mainly by solar radiation and much lesser by cosmic rays. The free electrons in this layer affect significantly the propagation of the electromagnetic radiation. The ionosphere extends from ~50 km to ~1000 km, that is, it covers the region of the lower thermosphere and heterosphere. The variability of the ionization leads to a classification of different ionospheric layers, each one with boundaries not well defined. The ionospheric boundaries and structure presented here follow [Iribarne and Cho, 1980]. The **D layer** is the ionospheric region up to ~90 km characterized by the lowest concentration of electron density ($10^3 - 10^4 electrons/cm^3$ and observed during the day only. It is composed of negative ions, positive ions and free electrons. Under normal conditions, the degree of ionization is not strong enough to interfere with the propagation of the radio waves. The ionization on the **E layer**,which ranges from ~90 km to ~140 km, shows a high degree of correlation with the incident

solar radiation, making it almost absent during the night (most of the electrons recombine with positive ions). During the day-time, the electron density is about one order of magnitude greater than that observed in the D layer ($10^5 electrons/cm^3$). It is composed of positive ions and free electrons. Within this layer, a thin ionospheric layer with an augmented electron concentration is sometimes observed. Due to its transient character, it is called **sporadic E layer**, E_s, and is associated with quasi-periodic amplitude scintillations [Coco et al., 1995]. The highest ionospheric layer is the **F layer**, with a lower boundary of ~140 km, which involves two regions of different characteristics. The F1 layer is the lowest region, which exists only during daytime. The F2 layer is the region of the ionosphere with the highest electron density, reaching a maximum concentration of $10^6 electrons/cm^3$ at 250-500 km. Although it does not disappear during night-time, the electron density during this period decreases by about one order of magnitude. It is also very sensitive to the sunspot cycle. Further, sudden ionospheric disturbances due to coronal holes, solar flares, and magnetic disturbances can affect the behavior of this and other ionospheric layers, as a result of a significant increase in the electron density. Below the ionosphere, the atmosphere is electrically neutral and this region is named the **neutral atmosphere**. The neutral atmosphere includes the lower part of the stratosphere and the troposphere. Although both the ionosphere and the neutral

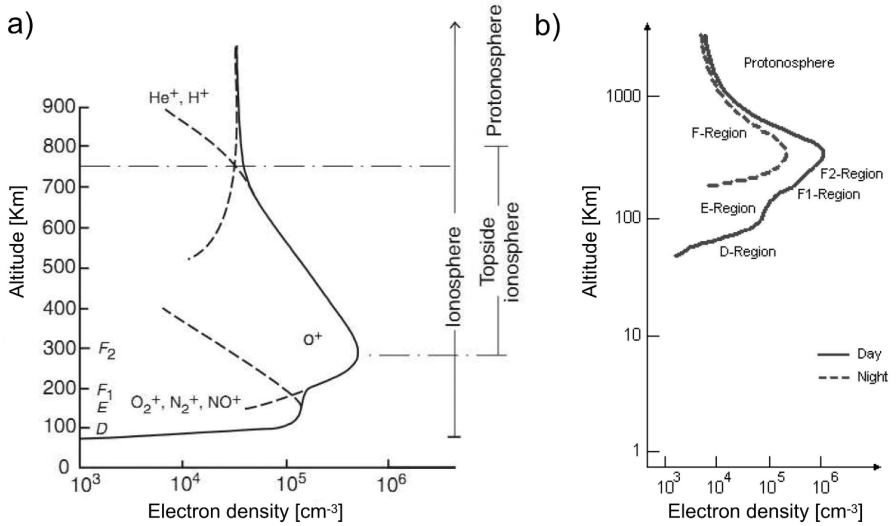


FIGURE 2.2: a) Chemical composition of ionosphere (from [Banks and Kockarts, 1973]). b) Density profiles of free electrons content in the ionosphere during the day/night)

atmosphere affect the propagation of radio waves, the study of the resulting effects have to be addressed differently. Unlike the neutral atmosphere, the ionosphere is a dispersive medium for radio waves, and its first-order effects can be removed almost completely by using two different carrier frequencies (for discussion of residual effects see, e.g., [Brunner and Gu, 1991b] and [Bassiri and Haij, 1993]). The following sections address these issues (fig.2.2). As far as the neutral atmosphere is concerned, the effect on radio signal

propagation is a function of pressure, temperature, and water vapor content. Therefore, the profiles of these variables have to be known accurately and the underlying physics understood to properly model the effect. Section 2.3 will deal with this issue.

2.2 EM waves in ionospheric plasma

The free electron population in the ionosphere is created by a complex ion production recombination mechanism via the ionization of the tenuous upper reaches of our atmosphere by ultraviolet and more energetic radiation from the sun and cosmic sources. The losses involve recombination via an intermediate attachment of free electrons to neutral atoms and molecules. The mathematics describing how electromagnetic waves, such as

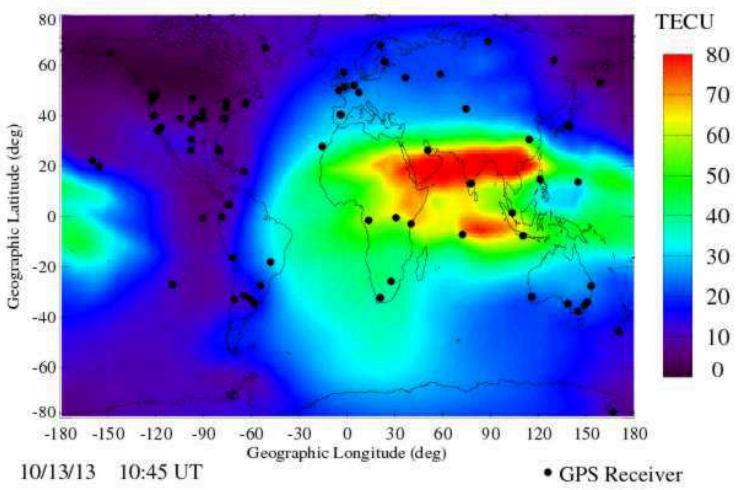


FIGURE 2.3: Global maps of ionospheric total electron content (TEC) are produced in real-time by mapping GPS observables collected from ground stations.(from iono.jpl.nasa.govl)

GPS signals, travel through a partially ionized medium in the presence of a magnetic field has been a subject of interest for decades. The theory that is typically applied for GPS frequencies is termed the magneto-ionic theory.

The refractive index of a medium defines the phase speed in the material relative to the speed in a vacuum. For a magneto-ionic medium such as the ionosphere, the refractive index depends upon several factors which include the frequency of the signal, its polarization, and the angle between the signal and the magnetic field. The applications of this theory to GPS will be reviewed below.

A magneto-ionic medium is one in which "free electrons and heavy positive ions are situated in a uniform magnetic field and are distributed with statistical uniformity, so that there is no resultant space charge [Barbarino]. A collection of charged particles in a magnetic field may also be known as a magneto-plasma, and has a characteristic

frequency of oscillation, known as the plasma frequency, f_p. Considering electrons only (as all ions are so massive in comparison that their motions are negligible), f_p is related to the electron charge, e, the electron density, N_e , the electron mass, m_e , and the permittivity of free space, e_0 , as follows:

$$\omega_p^2 = \frac{q_e^2 N_e}{m_e \varepsilon_0} \text{ or } f_p^2 = \frac{q_e^2 N_e}{4\pi^2 m_e \varepsilon_0} \tag{2.1}$$

the constants are evaluated (Table 2.2). However, the presence of a magnetic field also influences the magneto-plasma. Charged particles tend to spiral around the lines of magnetic force at a characteristic frequency known as the gyrofrequency, f_g . For electrons, the gyrofrequency depends on m_e , e, and the magnetic field vector, B, in the following relationship:

$$\omega_g = -\frac{|q_e|}{m_e} B_0 \text{ or } f_g = -\frac{q_e}{2\pi m_e} B_0 \tag{2.2}$$

When defined as above (Eq. 2.1), the plasma frequency is always a positive scalar. However, the electron gyromagnetic frequency or gyrofrequency (Eq. 2.2) is a vector, representing the direction around which charged particles will orbit in a clockwise manner (and may thus be negative for particles with a positive charge). The electron gyrofrequency is typically 1.5 MHz; the plasma frequency rarely exceeds 20 MHz.

Symbo	Quantity	Value	Unit
ε_0	Permittivity of free space	8.854×10^{-12}	farad/[m]
m_e	Electron mass	9.107×10^{-31}	[Kg]
q_e	Electron charge	-1.602×10^{-19}	Coulomb
μ_0	Permeability in a vacuum	12.57×10^{-7}	Henry/m

TABLE 2.2: Main constituents of the Earth's dry atmosphere below 80 km.

2.2.1 The refractive index in the Ionosphere

When an electromagnetic wave travels through a plasma interacts with all three components of the plasma: electrons, free ions and neutral molecules. The interaction of the wave with neutral molecules is so weak, when compared to the interaction between the wave and the charged particles, which can safely be neglected. Furthermore, since ions have a mass much greater than that of the electrons, the speed imparted to the ions by the wave is much smaller than that transferred to electrons. So when an electromagnetic wave passes through an ionized plasma only free electrons of the plasma appreciably affect the transmission of the wave, so that to model this phenomena is

sufficient to consider only the interaction between the wave and the free electrons. When a magnetostatic field $\vec{B_0}$ is applied to a plasma, the plasma becomes electrically anisotropic for electromagnetic waves. The magnetic permeability of the plasma is equal to the permeability of the vacuum μ_0, while the dielectric constant of the plasma becomes a tensor $\bar{\bar{\epsilon}}$. We can write the equation of motion of an electron, subjected to a monochromatic electromagnetic field and a uniform magnetostatic field $\vec{B_0}$. Because the electron doesn't moves with relativistic velocity, the magnetic force that acts on it, due to the magnetic field associated with the electromagnetic wave, is much smaller than that due to the electric field; this is the reason why we neglect the magnetic field.

$$m_e \frac{\partial \vec{v}}{\partial t} = q_e \vec{E} - m_e \vec{v} \omega_{eff} + q_e \vec{v} \times \vec{B_0} \tag{2.3}$$

where m_e and q_e represent the electron mass and charge, (as shown in Table 2.2), $\vec{v}$ is the average speed of each electron, $\vec{E}$ is the electric field of the incident wave, $\vec{B_0}$ is the magnetostatic field and $\omega_{eff} = 2\pi v_{eff}$ is the angular frequency of collisions (i.e. the number of collisions per unit time) between the electron and the neutral molecules of the plasma. In regime the solution is

$$v_x = \frac{q_e}{m_e} \frac{-i\Omega E_x + \omega_g E_y}{\omega_g^2 - \Omega^2} \tag{2.4}$$

$$v_y = \frac{q_e}{m_e} \frac{-i\Omega E_y + \omega_g E_x}{\omega_g^2 - \Omega^2} \tag{2.5}$$

$$v_z = \frac{q_e}{m_e} \frac{E_z}{i\Omega} \tag{2.6}$$

Where $\Omega = \omega + i\omega_{eff}$. The electric current density $\vec{J} = N_e q_e \vec{v}$:

$$J_x = \frac{N_e q_e^2}{m_e} \frac{-i\Omega}{\omega_g^2 - \Omega^2} E_x + \frac{N_e q_e^2}{m_e} \frac{\omega_g}{\omega_g^2 - \Omega^2} E_y \tag{2.7}$$

$$J_y = \frac{N_e q_e^2}{m_e} \frac{\omega_g}{\omega_g^2 - \Omega^2} E_x + \frac{N_e q_e^2}{m_e} \frac{-i\Omega}{\omega_g^2 - \Omega^2} E_y \tag{2.8}$$

$$J_z = \frac{N_e q_e^2}{m_e} \frac{1}{i\Omega} E_z \tag{2.9}$$

From (Eq. 2.7 to 2.9) we observe that in the case of a plasma which acts on an external magnetic field, the electric current density vector $\vec{J}$ is not parallel to $\vec{E}$. *The plasma behaves like anisotropic electrically medium, and therefore its complex dielectric constant is a tensor.* From the knowledge of $\vec{J}$ we can find the complex dielectric constant of the plasma starting from Maxwell's equation:

$$\vec{\nabla} \times \vec{H} = \vec{J} - i\omega\epsilon_0 \vec{E} = -i\omega \bar{\bar{\epsilon}}' \cdot \vec{E} = -i\omega \vec{D} \tag{2.10}$$

which yields

$$D_x = \epsilon_0 \left(1 - \frac{\omega_p^2(\omega + i\omega_{eff})}{\omega[(\omega + i\omega_{eff})^2 - \omega_g^2]}\right) E_x - i\epsilon_0 \left(\frac{\omega_p^2 \omega_g}{\omega[(\omega + i\omega_{eff})^2 - \omega_g^2]}\right) E_y \tag{2.11}$$

$$D_y = i\epsilon_0 \left(\frac{\omega_p^2 \omega_g}{\omega[(\omega + i\omega_{eff})^2 - \omega_g^2]}\right) E_x + \epsilon_0 \left(1 - \frac{\omega_p^2(\omega + i\omega_{eff})}{\omega[(\omega + i\omega_{eff})^2 - \omega_g^2]}\right) E_y \tag{2.12}$$

$$D_z = \epsilon_0 \left(1 - \frac{\omega_p^2}{\omega(\omega + i\omega_{eff})}\right) E_z \tag{2.13}$$

The matrix that represents the complex dielectric constant is:

$$\bar{\bar{\epsilon}}' = \begin{pmatrix} \epsilon'_{xx} & \epsilon'_{xy} & 0 \\ \epsilon'_{yx} & \epsilon'_{yy} & 0 \\ 0 & 0 & \epsilon'_{zz} \end{pmatrix} \tag{2.14}$$

We will study the propagation and the polarization properties of a monochromatic electromagnetic plane wave (such as GPS wave), which propagates in a homogeneous plasma subjected to a magnetic field (such as the Earth's magnetic field). The electric and the magnetic field of a monochromatic plane wave have the form, omitting the term $e^{-i\omega t}$.

$$\vec{E}(\vec{r}) = \vec{E_0} e^{i\vec{k}\cdot\vec{r}} \quad \text{and} \quad \vec{H}(\vec{r}) = \vec{H_0} e^{i\vec{k}\cdot\vec{r}} \tag{2.15}$$

where $\vec{E_0}$ and $\vec{H_0}$ are constant vectors, $\vec{k}$ is the wavevector, $\vec{r}$ is the position vector. We can write $\vec{k}$ as:

$$\vec{k} = \hat{n}\frac{\omega}{v} \tag{2.16}$$

where $\hat{n}$ is the unit vector in the direction of propagation and v is the speed of phase wave. The problem is to determine the vector $\vec{k}$ which describes the propagation of the wave, and the vector $\vec{E_0}$ which describes the polarization of the wave. Consider Maxwell's equations written for harmonic fields in time:

$$\vec{\nabla} \times \vec{E} = i\omega\mu_0 \vec{H} \quad \text{and} \quad \vec{\nabla} \times \vec{H} = -i\omega \bar{\bar{\epsilon}}' \cdot \vec{E} \tag{2.17}$$

Substituting the expressions 2.15

$$\vec{\nabla} \times \vec{E}_0 e^{i\vec{k}\cdot\vec{r}} = i\omega\mu_0 \vec{H} \quad \text{and} \quad \vec{\nabla} \times \vec{H}_0 e^{i\vec{k}\cdot\vec{r}} = -i\omega\overline{\overline{\epsilon}}' \cdot \vec{E} \tag{2.18}$$

we have

$$\vec{k} \times \vec{E} = \omega\mu_0 \vec{H} \quad \text{and} \quad \vec{k} \times \vec{H} = -\omega\overline{\overline{\epsilon}}' \cdot \vec{E} = -\omega\vec{D} \tag{2.19}$$

The (2.19) provides the important result that the vectors $\vec{D}$ and $\vec{H}$ are both orthogonal to the direction of propagation and between them. $\vec{H}$ is orthogonal to $\vec{E}$. Calculating the scalar product of the first of 2.19 and $\vec{D}$, we obtain:

$$\vec{k} \times \vec{E} \cdot \vec{D} = 0 \tag{2.20}$$

i.e. the vectors $\vec{k}$, $\vec{E}$ and $\vec{D}$ lie on the same plane. Because in an anisotropic medium $\vec{E}$ is not parallel to $\vec{D}$, it follows that *the vector $\vec{E}$ is not orthogonal to the direction of propagation.* Therefore the Poynting vector does not have the same direction of $\vec{k}$ and lies in the same plane of $\vec{D}$, $\vec{E}$ and $\vec{k}$ (fig. 2.4).

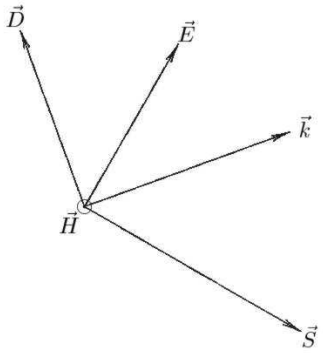


FIGURE 2.4: Representation of the vectors $\vec{k}$, $\vec{E}$ and $\vec{D}$

It can be shown that the average of the Poynting vector is parallel to the actual speed of propagation of the wave, that is the group velocity. Extracting $\vec{H}$ from the first of (2.19) and substituting in the second, the two Maxwell equations are equivalent to:

$$\boxed{\vec{k} \times (\vec{k} \times \vec{E}) = -\omega^2\mu_0\overline{\overline{\epsilon}}' \cdot \vec{E}} \tag{2.21}$$

that we can write as:

$$\vec{E}_0 - \hat{n}(\hat{n} \cdot \vec{E}_0) = \frac{1}{\epsilon_0}\frac{v^2}{c^2}\overline{\overline{\epsilon}}' \cdot \vec{E}_0 \tag{2.22}$$

taking place $\vec{k} = \hat{n}\frac{\omega}{v}$

where c is the speed of light in vacuum, v is the velocity of phase and $\hat{n}$ the unit vector of the direction of propagation. Without loss of generality we choose a Cartesian reference system so that the z axis is parallel to the $\vec{B_0}$ and the yz plane containing $\hat{n}$, as shown in fig. 2.5.

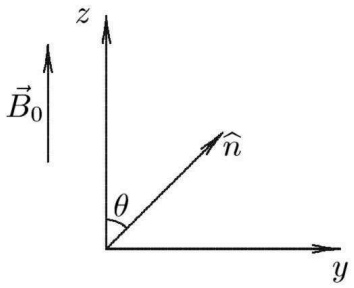


FIGURE 2.5: Representation of the vectors $\vec{B}_0$ and $\hat{n}$

where θ is the angle between $\vec{B}_0$ and $\hat{n}$. Projecting the equation (2.22) on the axes, according with the figure 2.5:

$$\begin{cases} E_{0x}\left(1 - \frac{v^2}{c^2}\frac{\epsilon'_{xx}}{\epsilon_0}\right) & -E_{0y}\frac{v^2}{c^2}\frac{\epsilon'_{xy}}{\epsilon_0} & & = 0 \\ E_{0x}\left(-\frac{v^2}{c^2}\frac{\epsilon'_{yx}}{\epsilon_0}\right) & +E_{0y}\left(\cos^2\theta - \frac{v^2}{c^2}\frac{\epsilon'_{yy}}{\epsilon_0}\right) & +E_{0z}(-\cos\theta\sin\theta) & = 0 \\ & +E_{0y}(-\cos\theta\sin\theta) & +E_{0z}\left(\sin^2\theta - \frac{v^2}{c^2}\frac{\epsilon'_{zz}}{\epsilon_0}\right) & = 0 \end{cases} \tag{2.23}$$

E_{0x}, E_{0y}, and E_{0z} represent the components of $\vec{E_0}$. Since the system (2.23) is homogeneous, it has solution different than trivial if

$$det\begin{pmatrix} \left(1 - \frac{v^2}{c^2}\frac{\epsilon'_{xx}}{\epsilon_0}\right) & -\frac{v^2}{c^2}\frac{\epsilon'_{xy}}{\epsilon_0} & 0 \\ \left(-\frac{v^2}{c^2}\frac{\epsilon'_{yx}}{\epsilon_0}\right) & \left(\cos^2\theta - \frac{v^2}{c^2}\frac{\epsilon'_{yy}}{\epsilon_0}\right) & (-\cos\theta\sin\theta) \\ 0 & (-\cos\theta\sin\theta) & \left(\sin^2\theta - \frac{v^2}{c^2}\frac{\epsilon'_{zz}}{\epsilon_0}\right) \end{pmatrix} = 0 \tag{2.24}$$

Setting

$$\epsilon_1 = \frac{\epsilon'_{xx}}{\epsilon_0} - i\frac{\epsilon'_{xy}}{\epsilon_0};\ \epsilon_2 = \frac{\epsilon'_{xx}}{\epsilon_0} + i\frac{\epsilon'_{xy}}{\epsilon_0};\ \epsilon_3 = \frac{\epsilon'_{zz}}{\epsilon_0} \tag{2.25}$$

we obtain:

$$-\frac{v^2}{c^2}\epsilon_3\cos^2\theta+\frac{v^4}{c^4}\frac{\epsilon_1+\epsilon_2}{2}\epsilon_3-\frac{v^2}{c^2}\frac{\epsilon_1+\epsilon_2}{2}\sin^2\theta+\frac{v^4}{c^4}\frac{\epsilon_1+\epsilon_2}{2}\epsilon_3\cos^2\theta \\ -\frac{v^6}{c^6}\epsilon_1\epsilon_2\epsilon_3+\frac{v^4}{c^4}\epsilon_1\epsilon_2\sin^2\theta=0 \tag{2.26}$$

Which is a quadratic equation in v^2/c^2. The solution provides, in function of the direction of propagation of the wave, the values of the propagation constant for which it is possible the existence of the wave in the plasma. We write the (2.26) in a different form:

$$-\tan^2\theta=\frac{\left(\frac{v^2}{c^2}-\frac{1}{\epsilon_1}\right)\left(\frac{v^2}{c^2}-\frac{1}{\epsilon_2}\right)}{\left(\frac{v^2}{c^2}-\frac{1}{\epsilon_3}\right)\left[\frac{v^2}{c^2}-\frac{1}{2}\left(\frac{1}{\epsilon_1}+\frac{1}{\epsilon_2}\right)\right]} \tag{2.27}$$

The formula (2.27) is useful for immediate analysis of the cases of wave propagation along the magnetic field direction and along the direction orthogonal to it. We neglect the terms dependent on the ion motion. In the case in which the propagation is parallel to $\vec{B_0}$ that is, according with figure 2.5 $\theta = 0°$, the equation (2.27) leads to the two solutions:

$$\left(\frac{v^2}{c^2}\right)_1=\frac{1}{\epsilon_1}\quad\text{and}\quad\left(\frac{v^2}{c^2}\right)_2=\frac{1}{\epsilon_2} \tag{2.28}$$

Consequently, the electromagnetic wave that passes through the plasma will be composed of two waves with different propagation constants k_0' and k_0'' :

$$k_0'=\frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega(\omega+i\omega_{eff}-\omega_g)}} \tag{2.29}$$

$$k_0''=\frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega(\omega+i\omega_{eff}+\omega_g)}} \tag{2.30}$$

If propagation occurs along the y axis, that is perpendicular to $\vec{B_0}$, θ is equal to $\frac{\pi}{2}$ and the solutions are:

$$k_{\frac{\pi}{2}}''=\frac{\omega}{c}\sqrt{1-\frac{\omega_p^2}{\omega(\omega+i\omega_{eff})}} \tag{2.31}$$

$$k''_{\frac{\pi}{2}} = \frac{\omega}{c}\sqrt{1 - \frac{\frac{\omega_p^2}{\omega^2}}{1 + i\frac{\omega_{eff}}{\omega} - \frac{\omega_g^2}{(\omega^2 + i\omega_{eff}\omega - \omega_p^2)}}} \tag{2.32}$$

The solution of equation 2.26 expressed in the general form is

$$\boxed{k'_\theta/k''_\theta = \frac{\omega}{c}\left[1 - \frac{X}{1 + iZ - \frac{1}{2}\frac{Y_T^2}{(1-iZ-X)} \pm \sqrt{\frac{1}{4}\frac{Y_T^4}{(1+iZ-X)^2} + Y_L^2}}\right]^{\frac{1}{2}}} \tag{2.33}$$

or for the phase refractive index is:

$$\boxed{n_\pm^2 = 1 - \frac{X}{1 + iZ - \frac{1}{2}\frac{Y_T^2}{(1-iZ-X)} \pm \sqrt{\frac{1}{4}\frac{Y_T^4}{(1+iZ-X)^2} + Y_L^2}}} \tag{2.34}$$

where $X = (\omega_p/\omega)^2$, $Y = (-\omega_g/\omega)$, $Z = (\omega_{eff}/\omega)$, $Y_L = Y\sin\theta)$ and $Y_T = Y\cos\theta)$. The equation 2.34 is the famous Appleton [1] equation for magneto-ionic theory of the ionosphere. We have thus found two waves traveling in arbitrary direction θ with different propagation constants. The wave with k'_θ is called "ordinary wave", while the wave with k''_θ is called extraordinary wave. For $\theta = 0$, i.e. for propagation along the magnetic field direction (fig. 2.5), the two waves move along the same direction, and the resulting wave is their composition. For $\theta \neq 0$, the two waves are separated (because they have different refractive indexes) thus highlighting the phenomenon of birefringence. As regards the GPS signal, the waves emitted in a far region ($\approx 22000Km$) where there are not free electrons, they will not suffer the phenomenon of birefringence but they, when enter in the ionosphere ($\approx 500Km$), are submitted to the Faraday rotation that we will explain below. In fact, starting from the special case in which the propagation is parallel to $\vec{B_0}$, being $\theta = 0°$, the equations 2.23 become:

$$\begin{cases} E_{0x}\left(1 - \frac{v^2}{c^2}\frac{\epsilon'_{xx}}{\epsilon_0}\right) & -E_{0y}\frac{v^2}{c^2}\frac{\epsilon'_{xy}}{\epsilon_0} & = 0 \\ E_{0x}\left(-\frac{v^2}{c^2}\frac{\epsilon'_{yx}}{\epsilon_0}\right) & +E_{0y}\left(1 - \frac{v^2}{c^2}\frac{\epsilon'_{yy}}{\epsilon_0}\right) & = 0 \\ & E_{0z}\frac{v^2}{c^2}\frac{\epsilon'_{zz}}{\epsilon_0} & = 0 \end{cases} \tag{2.35}$$

[1] Sir Edward Victor Appleton (1892-1965): English physicist, who was awarded in 1947 the Nobel Prize for physics for his work on the physical properties of the upper atmosphere and for discovering the Appleton layer.

It follows immediately that the component along the z axis of the electric field is zero. Replacing in equation (2.35) the eigenvalues of equations (2.28) we have the following solutions of the equations (2.29 and 2.30):

$$\frac{E_{0x}}{E_{0y}} = i \quad \text{and} \quad \frac{E_{0x}}{E_{0y}} = -i \tag{2.36}$$

That is, the waves are *right hand circular polarization (RHCP)* and *left hand circular polarization (LHCP)*, an example is shown in figure 2.6. Then the electrical field vector of the two waves can be written as:

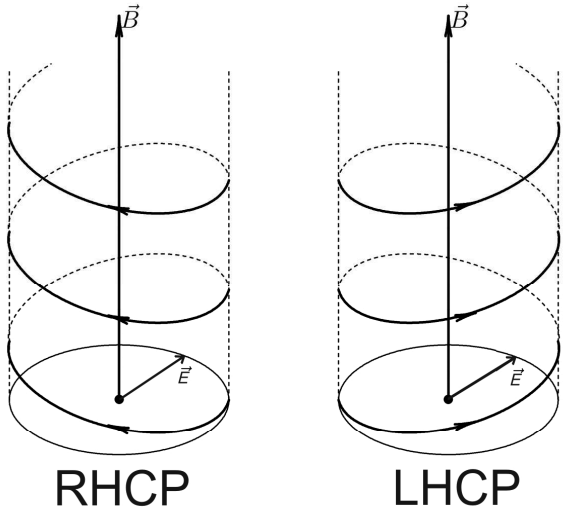


FIGURE 2.6: On left an example of right hand circular polarization (RHCP) wave; on right an example of left hand circular polarization (LHCP) wave.

$$\vec{E'} = (\hat{x} - i\hat{y})Ae^{ik_0'z} \quad \text{and} \quad \vec{E''} = (\hat{x} + i\hat{y})Ce^{ik_0''z} \tag{2.37}$$

Clearly E' is a right circularly polarized wave and E'' is a left circularly polarized wave. The sum of these two waves can provides the composed wave:

$$\vec{E} = \vec{E'} + \vec{E''} = \hat{x}(Ae^{ik_0'z} + Ce^{ik_0''z}) + \hat{y}(-iAe^{ik_0'z} + iCe^{ik_0''z}) \tag{2.38}$$

To study the polarization of this wave, we consider the ratio of the components along the axes $\frac{E_x}{E_y}$:

$$\frac{E_x}{E_y} = i\frac{1 + \frac{C}{A}e^{[i(k_0''-k_0')z]}}{1 - \frac{C}{A}e^{[i(k_0''-k_0')z]}} \tag{2.39}$$

If the waves $\vec{E}'$ and $\vec{E}$" have the same module, then the constants A and C become equal in magnitude. This is equivalent to say that the EM wave entering the plasma is linearly polarized. This assumption is true if

$$\left|\frac{C}{A}\right| = 1 \quad \text{that is,} \quad \frac{C}{A} = e^{i\phi_0} \tag{2.40}$$

Substituting the solution of 2.40 into equation 2.39, the solution is:

$$\frac{E_x}{E_y} = \cot\left[\frac{k_0' - k_0''}{2}z - \frac{\phi_0}{2}\right] \tag{2.41}$$

The phase angle ϕ_0 represents the initial direction (for z = 0) of the electric field, without loss of generality can be put equal to zero. **Since this relationship is real, the ratio E_x/E_y at each point z is linearly polarized, but the angle of orientation of its polarization plane (the plane containing $\vec{E}$ and $\vec{k}$) depends on z.** Then components E_x/E_y wheel if z increases or decreases. In other words, the resulting wave is subjected to **Faraday rotation**. The angle τ of which the resulting vector $\vec{E}$ rotates when the wave has traversed a unit distance is given by:

$$\tau = \frac{k_0' - k_0''}{2} \tag{2.42}$$

The rotation is in the clockwise direction if $k_0' > k_0''$. With the help of the equations 2.29 and 2.30 τ, supposing that $\omega_{eff} = 0$ can be written:

$$\tau = \frac{1}{2}\frac{\omega}{c}\left[\sqrt{1 - \frac{\omega^2}{\omega(\omega - \omega_g)}} - \sqrt{1 - \frac{\omega^2}{\omega(\omega + \omega_g)}}\right] \tag{2.43}$$

which reproduces the dependence of the Faraday rotation τ with the frequency. We observe that if a wave traveling parallel to $\vec{B}_0$, it is subject to a Faraday rotation in a clockwise direction (if $k_0' > k_0''$); whether traveling antiparallel to $\vec{B}_0$, it is subject to a Faraday rotation in a counterclockwise direction (see figure 2.7). *This is the reason why GPS signal is circular polarized and it implies that the RHCP GPS signals can travel with unchanged polarization for almost all values of h.*

2.2.2 Simplified Refractive Index for GPS Signals

To evaluate the ionospheric effects on the characteristic wave/GPS signal as it travels the ionosphere, we need the refractive index of the medium. The expression for the

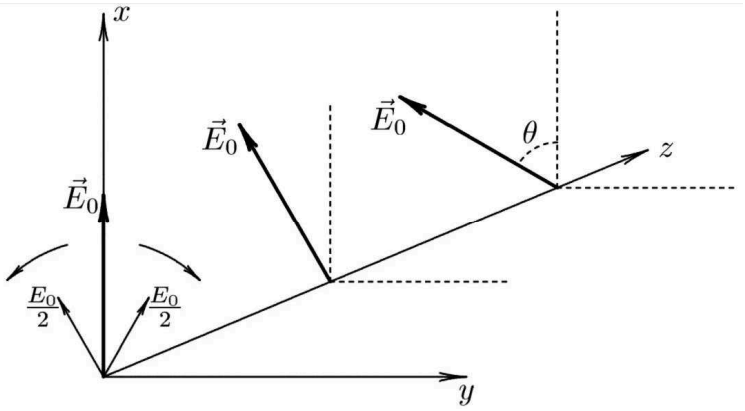


FIGURE 2.7

refractive index in eq. 2.34 is fairly complex to evaluate, and is usually simplified. For GPS frequencies, X is generally less than 2.7×10^{-4} and Y is less than 1.3×10^{-3} (using an approximate maximum value for plasma frequency and typical value for gyrofrequency of 20 and 1.5 MHz respectively [Klobuchar, 1996], and the L2 frequency, 1227.6 MHz). On rearranging eq. 2.34, then expanding it as a binomial series, discarding terms smaller than 1 part in 10^{-9} as suggested by [Brunner and Gu, 1991a], taking the square root and again expanding and discarding, the following simplified expression for the Appleton–Hartree equation is obtained:

$$n_{\pm} = 1 - \frac{1}{2}X \pm \frac{X|Y\cos\theta|}{2} - \frac{1}{8}X^2 \tag{2.44}$$

Now, for the formula to be correct for the GPS RHCP polarization the lower sign must be used when $0 \leq \theta \leq \pi/2$, and the upper sign when $\pi/2 \leq \theta \leq \pi$. Equation 2.44 is then equivalent to:

$$n = 1 - \frac{1}{2}X - \frac{XY\cos\theta}{2} - \frac{1}{8}X^2 \tag{2.45}$$

or equivalent to:

$$n = 1 - \frac{1}{2}\frac{A_p N_e}{f^2} - \frac{1}{2}\frac{A_p A_g N_e B\cos\theta}{f^3} - \frac{1}{8}\frac{A_p^2 N_e^2}{f^4} \tag{2.46}$$

where $A_p = \frac{e^2}{4\pi^2\epsilon_0 m_e}$ and $A_g = \frac{e}{2\pi m_e}$.

In the above expanded and simplified equation for the refractive index of the ionosphere, the term in f^2 becomes the first order term. On evaluation, replacing the constants, it reduces to the more familiar $40.3N_e/f^2$. The second and third order refractive index terms then arise from those in f^3 and f^4 respectively.

2.2.3 Refractive index of modulated GPS code

A signal, or modulated carrier wave, can be considered to result from the superposition of a group of waves of different frequencies centered on the carrier frequency (fig.2.8. If the medium is dispersive, the modulation of the signal will propagate with a different speed from that of the carrier; this is called the group velocity. Corresponding to the phase refractive index, n, we can define a group refractive index, n_g. The equivalent formula for the refractive index for the modulated code can be derived from the formula 2.46 :

$$n_g(f) = n + f\frac{dn}{df} = 1 - \frac{1}{2}\frac{A_p N_e}{f^2} - 2\left(\frac{1}{2}\frac{A_p A_g N_e B\cos\theta}{f^3}\right) - 3\left(\frac{1}{8}\frac{A_p^2 N_e^2}{f^4}\right) \tag{2.47}$$

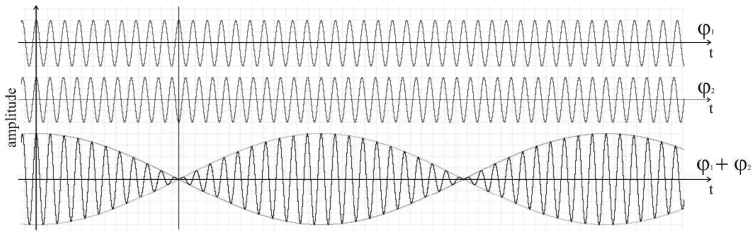


FIGURE 2.8: The two superimposed sine waves with slightly different frequencies (top and center) move at the phase velocity, their envelope (bottom) moves at the group velocity.

2.2.4 Correction to GPS Signal from Ionospheric Refractive Index Expansion Terms

Integrating the effect of the refractive index of Eq. 2.45 along the curved path leads to the following equation for the ionospheric effects on the carrier phase in units of length:

$$\Phi_1 = \rho + N_1\lambda_1 - \frac{1}{2}\frac{A_p\int N_e dL}{f^2} - \frac{1}{2}\frac{A_p A_g\int N_e B\cos\theta dL}{f^3} - \frac{1}{8}\frac{A_p^2\int N_e^2 dL}{f^4} \tag{2.48}$$

where ρ is the geometric range between the satellite and the receiver, λ the wavelength and dL denotes integrals along the path. In Eq. 2.48 and subsequently we denote N_1 to be the integer ambiguity. The subscript here refers to the GPS signal with frequency f_1 . An analogous formula holds for the GPS frequency f_2 . For the sake of simplicity, equation 2.48 does not show corrections for non-dispersive errors, such as tropospheric delay. Once the values for the constants in A_p and A_g are inserted, the effect on the carrier phase becomes (with units of metres):

$$\Phi_1 = \rho + N_1\lambda_1 - \frac{40.3\int N_e dL}{f^2} - \frac{1.1284\times 10^{12}\int N_e B\cos\theta dL}{f^3} - \frac{812.0\int N_e^2 dL}{f^4} \quad (2.49)$$

and we can define the first, second and third order refractive index correction terms as:

$$I1 = \frac{40.3\int N_e dL}{f^2};\ I2 = -\frac{1.1284\times 10^{12}\int N_e B\cos\theta dL}{f^3};\ I3 = -\frac{812.0\int N_e^2 dL}{f^4} \quad (2.50)$$

These corrections are for the GPS L1 carrier phase. The appropriate expressions for L2 are formed by replacing f_1 by f_2 in Eq. 2.50. The first-order ionospheric error depends on the electron density, the signal frequency and the path length, this effect can introduce tens of meters to the signal path. The second-order ionospheric error is a function of both the ionospheric electron content, the geomagnetic field and frequency. The third-order ionospheric error is much smaller in magnitude than the previous two terms and has the same independent variables of the first-order term. However, the second- and third-order terms can add up to 15 mm of the path delay under high ionospheric conditions Fritsche et al. (2005). To implement the corrections in a practical GPS analysis, approximations of the integrals are usually performed to reduce computing time and avoid the need for a profile of N_e with height.

2.2.5 Ionospheric-free linear combination

The traditional approach to removing the effect of the ionosphere is the 'ionosphere-free' linear combination, often known as the LC or L3 combination:

$$\Phi_L C = \frac{f_1^2}{f_1^2 - f_2^2}\Phi_1 - \frac{f_2^2}{f_1^2 - f_2^2}\Phi_2 \quad (2.51)$$

where Φ_i is the expression for the phase from Eq. 2.49, and again the units are meters. While this combination cancels the first order term (assuming that $\int N_e dL$ is identical for both the L1 and L2 frequencies), the second and third order terms do not cancel completely. The small (cm level) second- and third-order ionospheric effects are modeled using global maps of total electronic content and the IGRF model of the magnetic field [Petrie et al., 2010] Alternative linear combinations of the signals are often made for other purposes [Hofmann-Wellenhof et al., 2001] and should not be confused with the combination described above.

2.3 EM waves in neutral atmosphere

After the signals of the GPS have passed through the ionosphere they also need to pass through the neutral atmosphere (primarily the troposphere) before they are observed at the surface of the Earth. For microwaves like GPS there is practically no frequency dependence; thus it is impossible to remove the tropospheric delay with a multi-frequency combination like it is in the ionospheric case. Thus the tropospheric delays need to be corrected for by other means.

2.3.1 The refractive index in neutral atmosphere

For a neutral, non homogeneous medium and in absence of free charges, like the troposphere, the propagation of electromagnetic waves are described by Maxwell's equations, in the form:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{2.52}$$

$$\vec{\nabla} \times \vec{B} = \mu\varepsilon(\vec{r})\frac{\partial \vec{E}}{\partial t} \tag{2.53}$$

$$\vec{\nabla} \cdot (\varepsilon(\vec{r})\vec{E}) = 0 \tag{2.54}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{2.55}$$

where $\vec{E}$ and $\vec{B}$ are the electric field and magnetic field vectors, respectively, $\varepsilon(\vec{r})$ the electric permittivity dependent by the position of the vector $\vec{r}$, μ the constant magnetic permeability. Assuming that the spatial variations in ε is small, the equations can be combined into forming a wave equation for the electric field:

$$\nabla^2\vec{E} + \omega^2\mu_0\varepsilon(\vec{r})\vec{E} = -\vec{\nabla}\left(\frac{1}{\varepsilon(\vec{r})}\vec{E} \cdot \vec{\nabla}\varepsilon(\vec{r})\right) \tag{2.56}$$

The right-hand term of the equation 2.56 does not allow the exact solution of the equation that is not easy to solve because ε depends on r. In order to solve the wave equation for inhomogeneous media, we must use the approximate methods. Suppose that the frequency of the electromagnetic wave is high. In this case, the variations of the physical properties of the medium occur on distance large compared to the wavelength. We can say that locally the field propagates in a homogeneous medium. So at high frequencies, such that $|\vec{\nabla}\varepsilon_r|\lambda \ll 1$ (condition certainly reasonable for the GPS), we can represent the electromagnetic field as:

$$\vec{E}(\vec{r}) = \vec{E}_0(\vec{r})e^{i(k_0 L(\vec{r})} \quad \text{and} \quad \vec{H}(\vec{r}) = \vec{H}_0(\vec{r})e^{i(k_0 L(\vec{r})} \tag{2.57}$$

where $\vec{E}_0(\vec{r})$, $\vec{H}_0(\vec{r})$, $L(\vec{r})$ are functions of position and frequency. The wave number for vacuum (k_0) is defined as $k_0 = \omega/c$, with $c = 1/\sqrt{\mu_0 \varepsilon_0})$ is the speed of light in vacuum. $L(\vec{r})$ is a scalar function called "optical path", $E_0(\vec{r})$ and $\vec{H}_0(\vec{r})$ are complex functions.

By inserting these equations into Maxwell's equations (Eqs. 2.54–2.53), we get after some calculations (Barbarino Cap 8)

$$\vec{\nabla} \times E_0(\vec{r}) + ik_0 \vec{\nabla} L(\vec{r}) \times \vec{E}_0(\vec{r}) = i\omega\mu_0 \vec{H}_0(\vec{r}) \tag{2.58}$$

$$\vec{\nabla} \times H_0(\vec{r}) + ik_0 \vec{\nabla} L(\vec{r}) \times \vec{H}_0(\vec{r}) = -i\omega\varepsilon(\vec{r})\vec{E}_0(\vec{r}) \tag{2.59}$$

assuming ω (i.e. k_0) sufficiently large (this is the case of microwaves), the first term in the left hand of equations 2.58 and 2.59 can be neglected.

$$\vec{\nabla} L(\vec{r}) \times \vec{E}_0(\vec{r}) = \mu_0 c \vec{H}_0(\vec{r}) \quad \text{and} \quad \vec{\nabla} L(\vec{r}) \times \vec{H}_0(\vec{r}) = -\varepsilon(\vec{r}) c \vec{E}_0(\vec{r}) \tag{2.60}$$

Then, after some rearrangements, leading to:

$$(\nabla L(\vec{r}))^2 = \varepsilon(\vec{r})\mu_0 c^2 \tag{2.61}$$

and being $n = c\sqrt{\varepsilon(\vec{r})\mu_0}$:

$$(\nabla L(\vec{r}))^2 = n(\vec{r})^2 \tag{2.62}$$

where $\vec{\nabla} L$ comprises the components of the ray directions, L is the optical path length, n is the refractive index of the medium, and $\vec{r}$ is the position vector. This equation is the well-known as "Eikonal equation". The surfaces $L(\vec{r})$= constant are called geometrical wave surfaces or the geometrical wave-fronts. The above mentioned Eikonal equation is a partial differential equation of the first order for $n(\vec{r})$ and it is possible to express it in many alternative forms. In general, the Eikonal equation can be written in the Hamiltonian canonical formalism as follows ([Born and Wolf, 1999]; [Cerveny, 2005]; [Nafisi et al., 2012a].

$$\hat{\mathcal{H}}(\mathbf{r}, \nabla L) = \frac{1}{\alpha}\left((\nabla L \cdot \nabla L)^{\alpha/2} - n(\mathbf{r})^{\alpha}\right) = 0 \tag{2.63}$$

$$\frac{d\mathbf{r}}{du} = \frac{\partial \hat{\mathcal{H}}}{\partial \nabla L} \tag{2.64}$$

$$\frac{d\nabla L}{du} = -\frac{\partial \hat{\mathcal{H}}}{\partial \mathbf{r}} \tag{2.65}$$

$$\frac{dL}{du} = \nabla L \cdot \frac{\partial \hat{\mathcal{H}}}{\partial \nabla L} \tag{2.66}$$

Here α is a scalar value related to the parameter of interest u (see Table 2.3). Equations (2.63)–(2.65) can be used for constructing a ray-tracing system for any specific application by simply selecting a correct value for the scalar α [Nafisi et al., 2012a].

α		Parameter of interest
0	$du = dT$	Travel time T along the ray
1	$du = ds$	Arc-length s along the ray
2	$du = d\sigma = \frac{dT}{n^2}$	Natural variables along the ray

TABLE 2.3: Different cases of α parameter for Hamiltonian formalism

For $\alpha = 0$, the parameter u represents travel time t along the ray. If $\alpha = 1$, the parameter u represents the arc-length s along the ray. In case of $\alpha = 2$, the parameter u is equal to dt/dn, which represents the natural variables along the ray [Cerveny, 2005].

Selecting $\alpha = 1$ and solving the equation Eq. (2.63), we can obtain the equation to calculate the optical path length L

$$L = \int_S n(r, \theta, \lambda) ds \tag{2.67}$$

The refractivity index n and the optical path length L can be calculated using Eq. (2.67). Implementation of the above ray-tracing system can vary substantially, with different degrees of complexity and accuracy. Examples of ray-tracing algorithms in atmospheric studies are given by [Bean and Thayer, 1959], [Thayer, 1967], [Budden., 1985], [Davis, 1986], [Mendes, 1999], [Pany, 2002], [Böhm and Schuh, 2003], [Thessin, 2005], [Hulley and Pavlis, 2007], [Hobiger et al., 2008], [Nievinski, 2009], [Wijaya, 2010], [Nafisi et al., 2012a]. Several ray-tracing algorithms were compared by [Nafisi et al., 2012b]. Also it is possible to express the Eikonal equation and the ray-tracing system in curvilinear non-orthogonal coordinates systems. For details see [Cerveny et al., 1988].

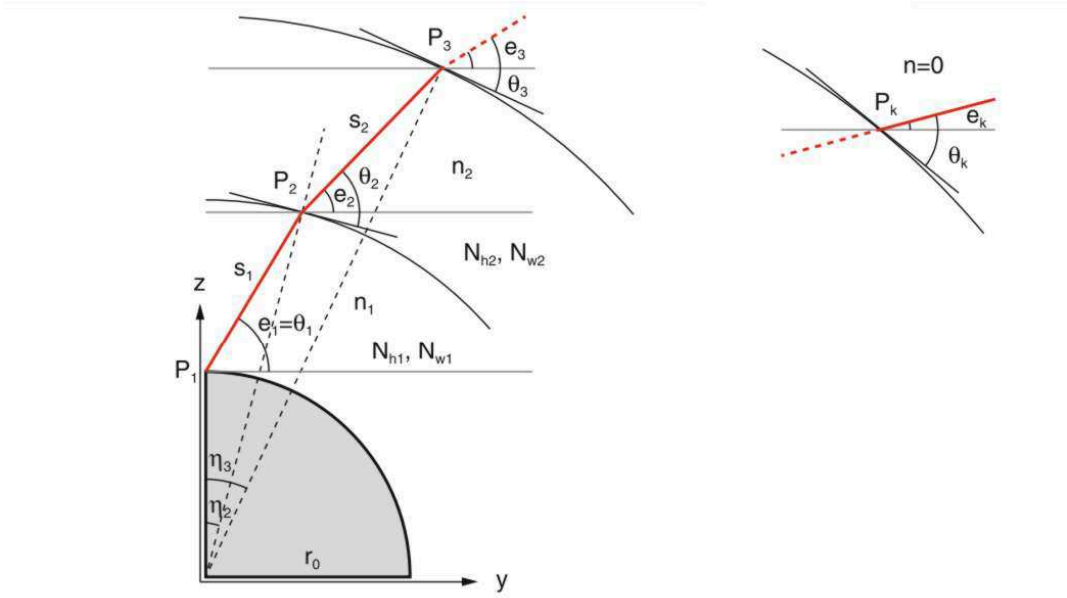


FIGURE 2.9: Geometry of a 1D ray-tracing method, for a receiver located at P_1 and the upper limit of the troposphere at P_k.

2.3.2 The refractivity

In the neutral atmosphere of the Earth $n(\vec{r})$ is very close to one, thus it is more convenient to use the so called refractivity. The refractivity N is related to the refractive index by

$$N = (n - 1) \cdot 10^6 \tag{2.68}$$

In general the refractivity is a complex number. It can be divided into three parts

$$N = N_0 + N'(\nu) - iN''(\nu) \tag{2.69}$$

When the spatial and temporal variations of N are small, the effect on the propagation of electromagnetic waves caused by the real and the imaginary parts of the refractivity can be considered separately. For the signals of space geodetic techniques traveling through the atmosphere, this separation is a reasonable assumption since the wavelengths are shorter than a few decimeters. The real part of the refractivity $(N_0 + N'(\nu))$ causes refraction and propagation delay of signals traveling through the atmosphere. It consists of a frequency-independent (non-dispersive) part N_0 and a frequency dependent (dispersive) part $N'(\nu)$. The imaginary part of the refractivity, $N''(\nu)$, causes absorption. Since the observables of space geodetic techniques (e.g. GNSS, VLBI, and SLR) typically are measurements of the travel time of the signals, the absorption is typically not important since it does not affect the propagation delay. Of course, absorption will

affect the delay measurements by increasing the noise; higher attenuation will cause the signal-to-noise-ratio to be lower, and thus the accuracy of the measured delay will be worse (in the worst case the signal cannot be detected). However, there is typically no need for modeling this effect in the space geodetic data analysis. Thus, in the following we will concentrate on the real part of the refractive index and the effects caused by it. The real part of refractivity can be expressed as a function of the densities of the different atmospheric gases and the temperature T ([Debye, 1929])

$$N = \sum_i \left(A_i(\nu)\rho_i + B_i(\nu)\frac{\rho_i}{T} \right) \tag{2.70}$$

where ρ_i is the density of the i-th gas, and A_i and B_i are constants. The $B_i \frac{\rho_i}{T}$ term is caused by the permanent dipole moment of the molecules. Since water vapor is the only major atmospheric gas having a permanent dipole moment, we can ignore this term for all other gases. The relative concentrations of the dry atmospheric gases are approximately constant (except carbon dioxide). Thus we can assume that $\rho_i = x_i \rho_d$, where x_i is constant and ρ_d is the density of dry air. This makes it possible to express the refractivity as a function of pressure, temperature, and humidity ([Essen and Froome, 1951])

$$\begin{aligned} N &= \sum_i A_i(\nu) x_i \rho_d + A_w(\nu)\rho_w + B_w(\nu)\frac{\rho_w}{T} + A_{lw}(\nu)\rho_{lw} = \\ &= k_1(\nu)\frac{p_d}{T} Z_d^{-1} + k_2(\nu)\frac{p_w}{T} Z_w^{-1} + k_3(\nu)\frac{p_w}{T^2} Z_w^{-1} + k_4(\nu)\rho_{lw} \end{aligned} \tag{2.71}$$

Where ρ_{lw} is the density of liquid water. It is here assumed that the liquid water droplets are small compared to the wavelength ($< 1mm$ for microwave techniques), the expression of Path Delays in the Neutral Atmosphere for larger droplets becomes more complicated ([Solheim et al., 1999]). However, normally the liquid water contribution to the refractivity ($k_4(\nu)\rho_{lw}$) is neglected since it is small [Solheim et al., 1999]. The variables Z_d and Z_w are compressibility factors for dry air and water vapor, respectively. These describe the deviation of the atmospheric constituents from an ideal gas. The compressibility factor for the i-th constituent of air is given by

$$Z_i = (pM_i)/(\rho_i RT) \tag{2.72}$$

where M_i is the molar mass (Table 2.1) and R is the universal gas constant. For an ideal gas we have $Z = 1$. [Owens, 1967] obtained expressions for Z_d^{-1} and Z_w^{-1}:

$$Z_d^{-1} = 1 + p_d \left[57.97 \cdot 10^{-8} \left(1 + \frac{0.52}{T} \right) - 9.4611 \cdot 10^{-4} \frac{T - 273,15}{T^2} \right] \tag{2.73}$$

$$\begin{aligned} Z_w^{-1} = 1 + 1650 \frac{p_w}{T^3} [1 - 0,01317(T - 273,15) \\ + 1,75 \cdot 10^{-4}(T - 273,15)^2 + 1,44 \cdot 10^{-6}(T - 273,15^3)] \end{aligned} \tag{2.74}$$

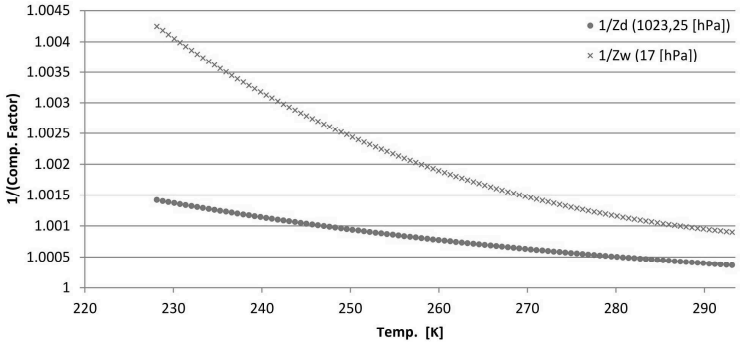


FIGURE 2.10: The compressibility factor behavior in function of temperature. A pressure value of 1013,25 and 17 [hPa] for dry and wet compressibility factor respectively, were fixed.

where p_d and p_w are expressed in hPa and T in K. By using the equations 2.73 and 2.74, figure 2.10 shows the inverse of compressibility factor behavior in function of temperature. A pressure value of 1013,25 and 17 [hPa] for dry and wet compressibility factor respectively, were used. Figure 2.11 shows the total refractivity for frequencies between 0 and 100 GHz for the case when the total pressure is 1013hPa, the temperature is 300K, and the relative humidity is 100 % (and for three different values of liquid water concentration). The refractivity was calculated using the Millimeter-wave Propagation Model (MPM) ([Liebe, 1985]; [Liebe, 1989]; [Liebe et al., 1993]). As can be seen, the variations in the refractivity as function of frequency are relatively small. The biggest variations are in the range 50–70GHz, a region where several strong absorption lines exist for oxygen. Below 40GHz the refractivity is more or less constant. There are small variations around the 22.235GHz water vapor absorption line, however these can typically be neglected. Since all space geodetic techniques that use microwaves operate at frequencies well below 40GHz, we can consider the refractivity to be frequency independent for microwaves. Thus the phase ($c_{phase} = c_0/n$) and group velocities ($c_{group} = c_0/(nf\partial n/\partial f)$) in the troposphere will be equal. In figure 2.11 three different cases are shown corresponding to different concentrations of liquid water: 0 g/m^3 0.05 g/m^3 (e.g. fog), and 1 g/m^3 (e.g. inside a cloud). The impact of liquid water on the refractivity is typically neglected since it is relatively small, although the effect should be considered in order to achieve highest accuracy in the presence of dense clouds. The difference between the

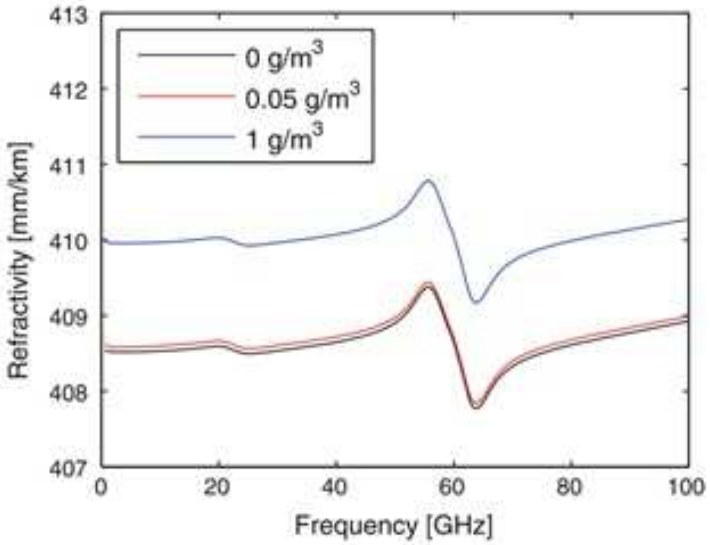


FIGURE 2.11: The total refractivity as a function of frequency. The total pressure is 1013 hPa, the temperatre 300 K, and the relative humidity is 100% Three different cases are shown corresponding to different concentrations of liquid water: 0, 0.05, and 1 g/m^3.[Nilsson, 2013]

case with 1 g/m^3 liquid water and the case with no liquid water is about 1.44 mm/km for the frequencies below 10 GHz, and then it decreases slightly with frequency to about 1.35 mm/km at 100 GHz.

This agrees with the published values for k_4 in the microwave range, which is generally about 1.45 ([Liebe et al., 1993]; [Solheim et al., 1999]). By ignoring the liquid water term and assuming that the refractivity is frequency independent, Eq 2.71 becomes

$$N = k_1 \frac{p_d}{T} Z_d^{-1} + k_2 \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1} \tag{2.75}$$

Several different laboratory measurements have been performed in order to determine the constants k_1, k_2, and k_3 (e.g. [Bevis et al., 1994]), Thayer [Thayer, 1974], [Hill et al., 1982]. Rüeger ([Rüeger, 2002a] ; [Rüeger, 2002b]) summarized and discussed many of these investigations, and calculated the "best average" values for the constants. These values are given in Table 2.4. The accuracy of k_1 is 0.015% and the accuracy of the water vapor part of the refractivity is 0.15%. Strictly speaking the constant k_1 is dependent on the relative concentrations of the different dry atmospheric gases. Thus if these change, k_1 will also change. Most dry atmospheric gases have stable concentrations. Of the major gases only the concentration of carbon dioxide show a significant variation (it is increasing with 1.5–2 ppm/year). Rüeger ([Rüeger, 2002a] ; [Rüeger, 2002b]) gives a formula for calculating k_1 for different carbon dioxide concentrations. In total the carbon

dioxide makes k_1 about 0.03 % larger compared to a carbon dioxide free atmosphere. The k_1 value given by [Rüeger, 2002a] [Rüeger, 2002b]

	$k_1(K/hPa)$	$k_2(K/hPa)$	$k_3(K^2/hPa)$
375 ppm CO_2	77.6890	71.2952	375463
392 ppm CO_2	77.6900	71.2952	375463

TABLE 2.4: The "best average" values of the k_1, k_2, andk_3 coefficients in Eq. 2.75, as presented by ([Rüeger, 2002a];[Rüeger, 2002b])

assumed a carbon dioxide concentration of 375 ppm (2004 level). Table 2.4 also shows k_1 for a carbon dioxide concentration of 392 ppm (2012 level). The concentration of carbon dioxide also shows an annual variation of about 5 ppm, meaning that k_1 will have an annual variation of about $2.8 \cdot 10^{-4}$ K/hPa. This variation is negligible for all practical purposes. Using Eq. 2.72 it is possible to rewrite Eq. 2.75 as

$$N = k_1 \frac{R}{M_d}\rho + k_2' \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1} = N_h + N_w \tag{2.76}$$

where $k_2' = k_2 - k_1 \frac{M_w}{M_d}$ and:

$$N_h = k_1 \frac{R}{M_d}\rho \tag{2.77}$$

$$N_w = k_2' \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1} \tag{2.78}$$

N_h is called the hydrostatic refractivity and N_w the wet (or non-hydrostatic) refractivity. The hydrostatic refractivity depends only on the total density of air, while the wet part depends only on the partial pressure of water vapor and the temperature. Figure 2.12 shows examples of vertical profiles of N_h and N_w. While the hydrostatic part is larger than the wet part, the wet refractivity is much more variable and difficult to model. We will see in sec.2.3.4 that the effect of the hydrostatic refractivity on the propagation of microwaves can be accurately estimated from just the surface pressure, while the modeling of the wet part is more complicated. It should be noted that in the literature sometimes a division of the refractivity into a dry and a wet part is used (e.g. [Perler et al., 2011]). The dry refractivity will be the part caused only by the first term of the righthand side of Eq. 2.75, while the other two terms are designated as the wet part. It is important to remember that the wet refractivity obtained in this case is not the same as the wet (i.e. non-hydrostatic) refractivity obtained when dividing the refractivity into a hydrostatic and wet part (Eq. 2.78). The division into dry and wet

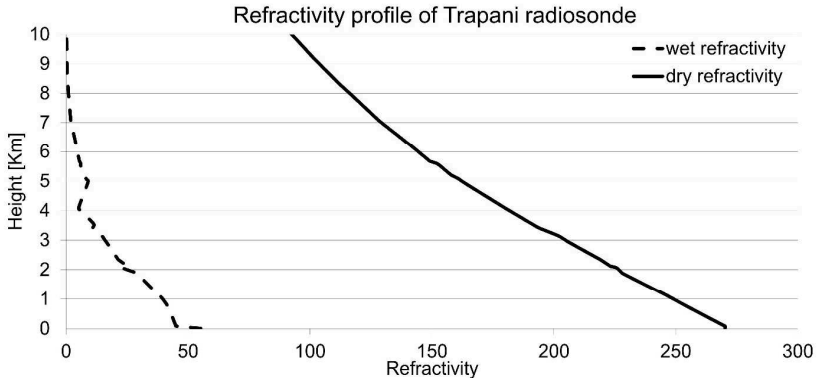


FIGURE 2.12: Examples of vertical profiles of the hydrostatic and wet refractivity. The profiles are calulated using radiosonde data from Trapani, Italy.

parts makes sense in that it clearly separates the contributions from the dry gases and water vapor (part of the hydrostatic refractivity is caused by water vapor). However, there are practical advantages of using the division into hydrostatic and wet parts, making it more commonly used. As shown in sec.2.3.4 the propagation delay caused by the hydrostatic refractivity can easily be inferred from surface pressure measurements.

2.3.3 Definition of Path Delay in the Neutral Atmosphere

The electric path length L (propagation time divided by the speed of light in vacuum) of a ray propagating along the path S through the atmosphere will be

$$L = \int_S n(s)ds \tag{2.79}$$

The electric path will be longer than the geometric length G of a straight line between the endpoints of the path for two reasons (see figure 2.13). Firstly, the propagation velocity is lower in the atmosphere than in vacuum. Secondly, the path S taken by the ray is, according to Fermat's principle, the path which minimizes L. The atmospheric delay, ΔL, is defined as the excess electric path length caused by the atmosphere

$$\Delta L = L - G = \int_S n(s)ds - G = \int_S [n(s) - 1]ds - \int_S ds - G = 10^{-6}\int_S N(s)ds + S - G \tag{2.80}$$

where S is the geometric length of the actual propagation path of the ray. By dividing the refractivity into hydrostatic and wet parts using Eq. (2.76) we get

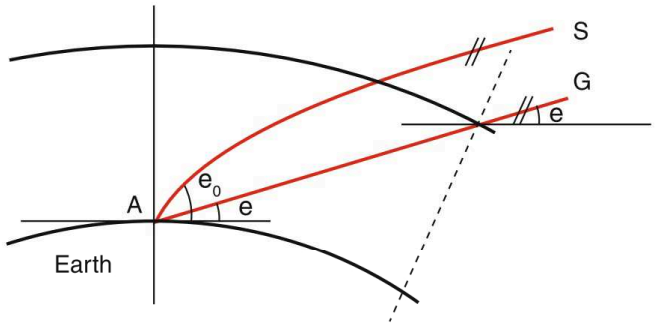


FIGURE 2.13: Path taken by a signal through the atmosphere. The signal will take the path with the shortest propagation time (S). Since the signal propagates slower in the atmosphere than in vacuum, the geometrical length of S will be larger than the straight path G. [Nilsson, 2013]

$$\Delta L = 10^{-6}\int_S N_h(s)ds + 10^{-6}\int_S N_w(s)ds + S - G = \Delta L_h + \Delta L_w + S - G \qquad (2.81)$$

where ΔL_h and ΔL_w are called the hydrostatic and wet delay, respectively. Commonly, the effect of bending, $S - G$, is by convention considered to be part of the hydrostatic delay, i.e. the hydrostatic mapping function (see sec.2.4.1) includes the bending effect. In space geodesy it is common to refer the slant delays to the delays in the zenith direction (using mapping functions, see sec.2.4.1). The zenith hydrostatic delay ΔL_h^z and the zenith wet delay ΔL_h^z are given by

$$\Delta L_h^z = 10^{-6}\int_{h_0}^{\infty} N_h(s)ds \qquad (2.82)$$

$$\Delta L_w^z = 10^{-6}\int_{h_0}^{\infty} N_w(s)ds \qquad (2.83)$$

where h_0 is the altitude of the site.

2.3.4 Hydrostatic Delay

From Eqs. (2.77) and (2.82) we see that the

only depends on the total density and not on the mixing ratio of wet and dry parts. Following [Davis et al., 1985], the hydrostatic delay can be determined starting from the hydrostatic equation

$$\frac{dp}{dz} = -\rho(z)g(z) \tag{2.84}$$

where g (z) is the gravity along the vertical coordinate z, and integration of Eq. (2.84) yields the pressure p_0 at the height h_0

$$p_0 = \int_{h_0}^{\infty} \rho(z)g(z)dz = g_{eff}\int_{h_0}^{\infty} \rho(z)dz \tag{2.85}$$

Instead of the height-dependent gravity g (z) , we introduce the mean effective gravity g_{eff}

$$g_{eff} = \frac{\int_{h_0}^{\infty} \rho(z)g(z)dz}{\int_{h_0}^{\infty} \rho(z)dz} \tag{2.86}$$

and the inversion yields the height h_{eff} which is the height of the center of mass of the atmosphere above the site and can be determined with

$$h_{eff} = \frac{\int_{h_0}^{\infty} \rho(z)zdz}{\int_{h_0}^{\infty} \rho(z)dz} \tag{2.87}$$

[Saastamoinen, 1972b] used the approximation for the effective height

$$h_{eff} = (0.9h_0 + 7300) \pm 400[m] \tag{2.88}$$

which holds for all latitudes and all seasons. By combining 2.77 and 2.82 is thus possible to determine the the zenith hydrostatic delay from the pressure p_0 at the Earth surface.

$$\Delta L_h^z = 10^{-6}k_1\frac{Rp_0}{M_d g_{eff}} \tag{2.89}$$

We follow [Saastamoinen, 1972b] and [Davis et al., 1985] to find the appropriate coefficients in Eq. (2.89). At first, the gravity g_{eff} at the effective height h_{eff} is determined with

$$g_{eff} = 9.8062[1 - 0.00265cos(2\theta) - 0.31 \cdot 10^{-6}h_{eff}] \tag{2.90}$$

which combined with Eq. (2.88) can be written as

$$g_{eff} = g_m \cdot f(\theta, h_0) \tag{2.91}$$

with $g_m = 9.7840$ and

$$f(\theta, h_0) = [1 - 0.00266cos(2\theta) - 0.28 \cdot 10^{-6} h_0] \tag{2.92}$$

where θ and h_0 are latitude and orthometric (or ellipsoidal) height of the station. Thus, the zenith hydrostatic delay is

$$\Delta L_h^z = 10^{-6} k_1 \frac{R p_0}{M_d g_m f(\theta, h_0)} \tag{2.93}$$

and after substitution of all values we get for the zenith hydrostatic delay in meters

$$\Delta L_h^z = 0.0022768 \frac{p_0}{f(\theta, h_0)} \tag{2.94}$$

where p_0 is in hPa. The molar masses M_d and M_w are constant up to heights of about 100 km ([Davis, 1986]), which is essential for all troposphere delay models. The errors in the zenith hydrostatic delays are mainly caused by errors in k_1 and in the surface pressure measurements. Assuming meteorological conditions at medium latitudes the zenith hydrostatic delay is about 2.3 m at sea level. An error in the surface pressure of 1 hPa causes an error of about 2.3 mm. In order to reach an accuracy of 0.1 mm, the pressure has to be measured with an accuracy of 0.05 hPa. The error due to the assumption of hydrostatic equilibrium depends on the wind and is about 0.01 % (0.2 mm path delay). Under severe weather conditions vertical accelerations can reach 1% of the gravity acceleration, which can cause errors in the zenith hydrostatic delays of about 20 mm ([Davis et al., 1985]). Usually for geodetic purposes, there are three possibilities to obtain pressure values at the stations: local pressure recordings at the sites, pressure values from numerical global weather models, or empirical models for the pressure ([Böhm et al., 2009a]).

Unfortunately, local pressure measurements are usually not available, in particular at GNSS stations. Thus, to get consistent values of a priori zenith hydrostatic delays for global GNSS networks it is preferable to take these values from numerical weather models. For example, the Department of Geodesy and Geoinformation (GEO) at the Vienna University of Technology provides zenith hydrostatic delays calculated from European Centre for Medium-Range Weather Forecasts data (ECMWF). These are provided on global grids ($2.5° \times 2.0°$) and with a temporal resolution of 6 h starting in 1994 ([Böhm

et al., 2009a]). For scientific purposes also forecast values are made available so that they can be used for real-time applications without significant loss of accuracy ([Böhm et al., 2009b]). Empirical models are then proposed, for instance by [Hopfield, 1969]

$$p = 1013.25 \left(\frac{T_k - \alpha h}{T_k} \right)^{\frac{M_d g}{R\alpha}} \tag{2.95}$$

where h represents the altitude above sea level, $T_k = 293.15K$ is the atmospheric temperature at sea level, $\alpha = 6.5K/km$ is the normal lapse rate of temperature with elevation , gravity g at the surface of the Earth ($9.7867m/s^2$), the gas constant $R = 8.31447J//molK$ and the molar mass of dry air $M_d = 0.0289644kg/mol$). for dry air. More sophisticated empirical models are UNB3m ([Leandro et al., 2006]) or Global Pressure and Temperature (GPT; [Böhm et al., 2007]). UNB3m is based on meteorological parameters (pressure, temperature, humidity, temperature lapse rate, and water vapor pressure height factor) at five latitude bands which are symmetric with respect to the equator (similar to the Niell Mapping Functions ([Niell, 2006])). Input parameters for GPT are station latitude, longitude, height and the day of the year, which is similar to the Global Mapping Functions (GMF; [Böhm et al., 2006a]) as both, GPT and GMF, are based on spherical harmonics up to degree and order 9 2.5.

Pressure	Local recordings	Grid values	GPT
Availability	At sites	All (interp.)	All
Time span	Per observ.	Since 1994	Unlimited
Spatial resolution	Per site	$(2.5° \times 2.0°)$	Spherical harmonics (9/9)
Time resolution	Per observ.	6h	Annual

TABLE 2.5: Availability of pressure values from local recordings at the sites, from numerical weather models (e.g. the hydrostatic zenith delays from ECMWF data as provided by the Vienna University of Technology), and from the empirical model GPT [Nilsson, 2013]

Empirical models like GPT are always available for all time epochs, but the spatial resolution is limited as it is represented by spherical harmonics up to degree and order 9 ($\sim 20°$ in latitude/longitude), respectively. The model only includes an annual variation with the zero phase set to 28 January, so it cannot capture short-term and sub-annual weather phenomena. The error in estimating the pressure will result in an error in the position, especially the vertical component, estimated with a space geodetic technique. In Sect.2.4.1 a rule of thumb relating the error in the delay to the error in the vertical coordinate is presented, from this we find that 3 hPa (7 mm zenith delay error) correspond to 1 mm station height difference [Nilsson, 2013]. GPT only accounts for an annual variation of the pressure with rather small amplitudes compared to the other (e.g. weekly) variations of the pressure. Thus, the standard deviations of the differences

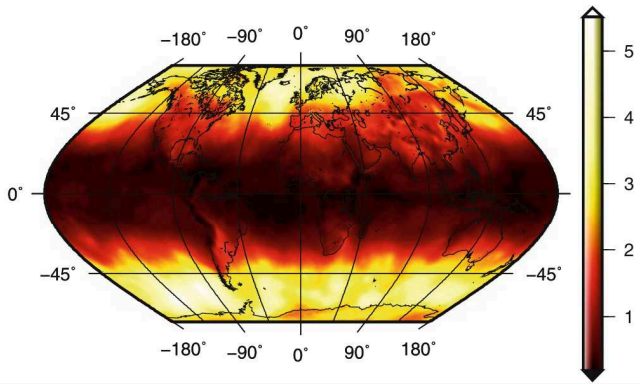


FIGURE 2.14: Simulated station height standard deviations in mm for GPT with respect to the pressure values provided by the ECMWF (based on 36 epochs in 2005 for a 2° latitude times 2.5° longitude grid (modified from [Böhm et al., 2009b])

to the grid values from the ECMWF are almost the same for the GPT model. In fig. 2.14 the simulated station height standard deviations are plotted for the case of using GPT compared pressure values from the ECMWF. There is an increase of the standard deviations towards higher latitudes (with maximum values at around ±60° latitude), which is due to the larger pressure variations in these regions compared to equatorial regions [Böhm et al., 2009a].

2.3.5 Wet Delay

From equations (2.83) and (2.78) the zenith wet delay is

$$\Delta L_w^z = 10^{-6}\left[\int_{h_0}^{\infty}\left(k_2'\frac{p_w}{T}Z_w^{-1}\right)dz + \int_{h_0}^{\infty}\left(k_3\frac{p_w}{T^2}Z_w^{-1}\right)dz\right] \tag{2.96}$$

The first term in Eq. (2.96) is about 1.6 % of the second term. The derivation of a model to account for the zenith wet delay (ΔL_w^z) is by far more challenging than the one for the hydrostatic delay. This is due to high spatial and temporal variability and unpredictability of the amount of water vapor. Thus, the temperature and the water vapor content at the Earth surface are not representative for the air masses above. This is the reason why numerous models have been developed over the past few decades for the wet delay, while preserving Saastamoinen's model (with slight modifications) for determining the hydrostatic delay. The zenith wet delay varies between a few mm at the poles and about 40 cm above the equatorial regions. In order to keep millimeter accuracy in space geodetic techniques, the (ΔL_w^z) is nowadays estimated as an additional parameter within the data analysis. Nevertheless, some models are listed below and can be used as an initial value in the data analysis or for applications not requiring

high accuracy. [Saastamoinen, 1972a.,b]proposes the calculation of the zenith wet delay (ΔL_w^z) based on ideal gas laws using a simple relation

$$\Delta L_w^z = 0.0022768(1255 + 0.05T_0)\frac{p_{w0}}{T_0} \tag{2.97}$$

where p_{w0} is the water vapor pressure and T_0 is the temperature at the surface. Likewise to the hydrostatic delay, [Hopfield, 1969] proposes an expression for (ΔL_w^z) as follows

$$\Delta L_w^z = \frac{10^{-6}}{5} N_w(h_0) h_w \tag{2.98}$$

with $N_w(h_0)$ the refractivity of wet air at the surface (located at height h_0) and a mean value $h_w = 11000$ m for the height of the troposphere up to which water vapor exists.[Ifadis, 1986] proposes to model the zenith wet delay as a function of surface pressure, partial water vapor pressure and temperature.[Mendes and Langley, 1998] derived a linear relation between ΔL_w^z and partial water vapor pressure. Some other models are being described by [Mendes, 1999]. An approximate relation between water vapor pressure and delay reads

$$\Delta L_w^z \approx \frac{0.217 p_w}{T} \tag{2.99}$$

Assuming an isothermal atmosphere with exponential decrease of water vapor pressure p_w, and assuming that water vapor exists until a height of 2 km, we get an approximation for the wet delay as a function of water vapor pressure at the Earth's surface p_{w0}

$$\Delta L_w^z \approx \frac{748 p_{w0}}{T_0^2} \tag{2.100}$$

An even simpler way is a rule of thumb that suggests that the wet zenith delay in cm equals the water vapor pressure in hPa at the Earth's surface

$$\Delta L_w^z[cm] \approx p_{w0}[hPa] \tag{2.101}$$

In any case, information of water vapor pressure and/or temperature at the surface has to be known. If no surface meteorological observation is available, we can use the simple model of the standard atmosphere where p_w can be calculated as a function of the relative humidity f , i.e.

$$p_w = \frac{f}{100} e^{(-37.2465+0.213166T-0.000256908T^2)} \tag{2.102}$$

2.3.5.1 Conversion of Zenith Wet Delays to Precipitable Water

The zenith wet delay can be related to the amount of integrated water vapor above the station. Following Eq. (2.96) and using the expression for integrated mean temperature T_m ([Bevis et al., 1992])

$$T_m = \frac{\int_S \left(\frac{e}{T} Z_w^{-1}\right) ds}{\int_S \left(\frac{e}{T^2} Z_w^{-1}\right) ds} \approx \frac{\int_{h_0}^{\infty} \left(\frac{e}{T} Z_w^{-1}\right) dz}{\int_{h_0}^{\infty} \left(\frac{e}{T^2} Z_w^{-1}\right) dz} \tag{2.103}$$

we can write

$$\Delta L_w = 10^{-6} \left[k_2' + \frac{k_3}{T_m}\right] \int_S \frac{e}{T} Z_w^{-1} ds \tag{2.104}$$

Applying the ideal gas laws, Eq. 2.104 can be reformulated as

$$\Delta L_w = 10^{-6} \left[k_2' + \frac{k_3}{T_m}\right] \frac{R}{M_w} \int_S \rho_w ds \tag{2.105}$$

To be able to calculate the mean temperature T_m the vertical profiles of the water vapor and temperature have to be known. Such data can be obtained from radiosonde measurements or calculated (and predicted) from operational meteorological models ([Wang et al., 2005]). In absence of this data, the empirically derived model by e.g. [Bevis et al., 1992] and [Emardson and Derks, 2000] can be used. The determination of the mean temperature Tm from Eq. (2.103) is based on the weighting with water vapor pressure in the atmosphere. Since water vapor is mainly located near the Earth surface the mean temperature T_m will be highly correlated with the temperature at the Earth surface T_0. Using 8718 profiles of radiosonde launches at 13 stations in the United States between 27 and 65 ° northern latitude, between 0 and 1600 m height, and over a time span of 2 years [Bevis et al., 1992] found

$$T_m \approx 70.2 + 0.72 T_0 \tag{2.106}$$

with a standard deviation of ± 4.74 K. It is clear from Eq. (2.105) that the wet delay is proportional to the integrated water vapor content IWV ($IWV = \int_0^{\infty} \rho_w dz$). Since IWV is a variable that can be easily obtained from numerical weather prediction models

or measured by other techniques, it is of great interest to have a simple expression for calculating the wet delay from IWV, and vice versa. Thus we define the proportionality constant $\prod$ such that

$$IWV = \prod \Delta L_w^z \tag{2.107}$$

where ΔL_w^z is the wet tropospheric delay in the zenith direction. By comparing Eqs. (2.105) and (2.107), we find that $\prod$ can be related to T_m by

$$\prod = \frac{10^6 M_w}{\left[k_2' + \frac{k_3}{T_m}\right] R} \tag{2.108}$$

The integrated water vapor in zenith direction can also be provided as precipitable water (PW) which corresponds to the height of the equivalent water column above the station

$$PW = \frac{IWV}{\rho_{w,fl}} \tag{2.109}$$

where $\rho_{w,fl}$ is the density of liquid water in kg/m^3. With a dimensionless quantity κ we can relate the ΔL_w^z and PW

$$PW = \kappa \Delta L_w^z \tag{2.110}$$

with κ defined as

$$\kappa = \frac{\prod}{\rho_{w,fl}} \tag{2.111}$$

The precipitable water is roughly 0.16 of ΔL_w^z. This value can vary by more than 15 %, mainly as a function of latitude and season ([Fölsche, 1999]). Errors in the factor $\prod$ are mainly caused by errors in T_m and the constants in Eq. (2.108). According to [Fölsche, 1999] the influence of errors in the mean temperature is at least one order of magnitude larger than errors in the constants. It is shown in the following how accurate T_m has to be determined to get the amount of water vapor with a certain accuracy. The partial derivative of $\prod$ with respect to the mean temperature yields

$$\frac{\partial \prod}{\partial T_m} = \frac{10^6 M_w k_3}{R \left[k_2' + \frac{k_3}{T_m}\right]^2 T_m^2} \tag{2.112}$$

This means that $\prod$ is changed by about 20 kg/m^3 if T_m (= 270 K) is changed by 4 K. Assuming a zenith wet delay of 200 mm this corresponds to an error in the precipitable water of about 4 mm. Requirements for better accuracies of PW (1 mm or better) indicate that the real weather data should be used to derive the mean temperature instead of approximations such as provided in Eq. (2.106).

2.4 Modeling Delays in the Neutral Atmosphere

Basically, there are two ways to handle the atmospheric delays when analyzing space geodetic data; either external measurements of the atmospheric delays are used to correct the measurements, or the atmospheric delays are parameterized and estimated in the data analysis. As seen in Sect.2.3.4 the hydrostatic delay can be accurately determined from surface pressure measurements. However, the wet delay cannot be estimated accurately from meteorological measurements at the surface. Thus it is common when analyzing GPS data to use surface pressure measurements to model the hydrostatic delay, while the wet delay is estimated in the data analysis. In the data analysis the tropospheric delays are modeled using mapping functions and gradients (see Sect. 2.4.1). An alternative strategy is to use external estimates of the hydrostatic and wet delays. Such estimates could for example be obtained from ray-tracing though numerical weather models (Sect.2.3), a technique also commonly applied for deriving expression for the tropospheric mapping functions.

2.4.1 Mapping Functions and Gradients

In the analysis of space geodetic data the troposphere path delay $\Delta L(e)$ at the elevation angle e is usually represented as the product of the zenith delay ΔL^z and an elevation-dependent mapping function $mf(e)$ with

$$\Delta L(e) = \Delta L^z \cdot mf(e) \tag{2.113}$$

This concept is not only used to determine a priori slant delays for the observations, but the mapping function is also the partial derivative to estimate residual zenith delays. Typically, the zenith delay is estimated with a temporal resolution of 20–60 min in GPS analysis. In the analysis of space geodetic observations the zenith delays are estimated, together with other parameters like the station clocks and the stations heights (fig.2.15).

The partial derivatives of the observed delays with respect to the station heights are dependent on the elevation angle only, but whereas the partial derivatives with respect

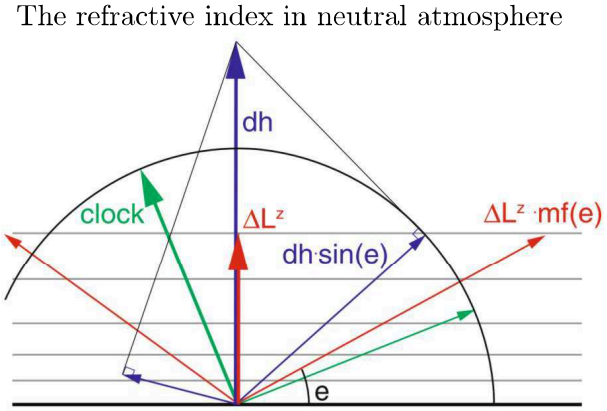


FIGURE 2.15: Different elevation-dependence of the tropospheric delays (red), clock values (green), and height components (blue). Whereas the partial derivatives for the estimation of clocks and height components are strictly 1 and $sin(e)$, respectively, the partial derivative for the estimation of the zenith delay is the mapping function which is approximately $1/sin(e)$, but is not perfectly known (modified from [Rothacher et al., 1998])

to the clocks (=1) and the station heights (=$sin(e)$) are exactly known, the partial derivatives for the zenith delays (i.e. the mapping functions) are only known with a limited accuracy. Consequently, due to the correlations between zenith delays, station heights, and clocks, any imperfection of the mapping function will also result in errors in the station height estimates (and clock errors). Considering Eq. (2.113) we find the following relationship: if the mapping function is in error (too large), then the estimated zenith delay ΔL^z is too small, because the observed tropospheric delay $\Delta L(e)$ does not change. Consequently, the estimated station height goes up. [Niell, 2001] set up a rule of thumb specifying that the error in the station height is approximately one third of the delay error at a cutoff elevation angle of 7°.The corresponding decrease of the zenith delay is about one half of the station height increase.

2.4.1.1 Azimuthal Symmetry: Mapping Functions

Assuming azimuthal symmetry of the neutral atmosphere around the station (i.e. at a constant elevation angle the delay is not dependent on the azimuth angle of the observation), the approach as described in Eq. (2.114) is used (e.g. [Davis et al., 1985])

$$\Delta L(e) = \Delta L_h^z \cdot mf_h(e) + \Delta L_w^z \cdot mf_w(e) \tag{2.114}$$

$\Delta L(e)$ is the total path delay of the microwaves in the neutral atmosphere and e is the elevation angle of the observation to the satellite (vacuum or geometric elevation angle). ΔL_h^z and ΔL_w^z are the a priori zenith hydrostatic and the estimated wet delays, and $mf_h(e)$ and $mf_w(e)$ are the so-called mapping functions which provide the ratio

of the delay to the delay in zenith direction. The input to both mapping functions is the vacuum elevation angle e, because the bending effect is accounted for by the hydrostatic mapping function. Errors in the zenith hydrostatic delays or the mapping functions have an influence on station height errors as described with the rules of thumb by [Davis et al., 1985] or mentioned above. The following two examples illustrate this rule of thumb, which depends on the actual distribution of elevation angles and on whether elevation angle-dependent weighting is used. The zenith hydrostatic and wet delays shall be 2000 and 200 mm, respectively, the minimum elevation angle is 5°, and the corresponding values for the true hydrostatic and wet mapping functions are 10.15 ($mf_h(5°)$) and 10.75 ($mf_w(5°)$) [Böhm et al., 2006b].

- We assume an error in the total pressure at the station of +10 hPa, e.g. when using the "mean" pressure from GPT during a time of low pressure. +10 hPa correspond to ≈ +20 mm zenith hydrostatic delay [Saastamoinen, 1972b], which is then mapped down to 5° elevation angle using the wrong mapping function (wet instead of hydrostatic, factor -0.6 = 10.15–10.75). At 5° elevation angle the mapping function error is - 12 mm, and one fifth of it, i.e. - 2.4 mm, would be the resulting station height error. This results in a kind of atmosphere loading correction (see Part 4, [Wijaya, 2010]), because during a pressure low the station heights go up [Tregoning and Herring, 2006].

- We consider an error in the wet mapping function of 0.1 ($mf_w(5°)$ =10.85 instead of 10.75) or in the hydrostatic mapping function of 0.01 ($mf_h(5°)$mfh) =10.16 instead of 10.15). The error at 5° elevation angle is in both cases 20 mm, i.e. the resulting error in the station height would be approximately +4 mm.

The scale height of the wet part in the troposphere is about 2 km, whereas the scale height of the hydrostatic part is about 8 km (cf. fig.2.16). The mapping functions describe the ratio $(AB)/(B_0B)$ (wet) and $(AC)/(C_0C)$ (hydrostatic). Due to the curvature of the Earth and the smaller scale height of the wet part, the hydrostatic mapping function is smaller than the wet mapping function. Exceptions are mapping functions for observations at very low elevation angles where the geometric bending effect, which is attributed to the hydrostatic mapping function, is increasing considerably. Thus, the mapping functions are a measure for the thickness of the atmosphere compared to the Earth radius [Niell, 2001]. If the thickness of the atmosphere gets smaller, it appears to be flatter, and the mapping function approaches $1/\sin(e)$. Assuming a flat and evenly stratified atmosphere the mapping function is $1/\sin(e)$. For higher elevation angles (> 20° elevation) these mapping functions are sufficiently accurate. [Marini, 1972] showed that the dependence on the elevation angle of the mapping functions for

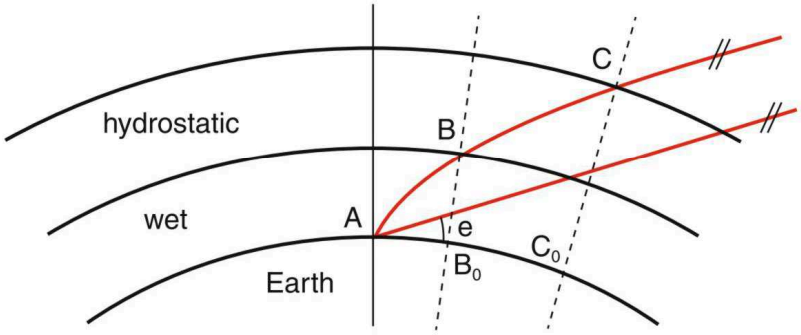


FIGURE 2.16: The scale height of the wet part of the troposphere is about 2 km, the scale height of the hydrostatic part is about 8 km. The mapping functions describe the ratio of the paths $(AB)/(B_0B)$ (wet) and $(AC)/(C_0C)$ (hydrostatic). The wet mapping function is larger than the hydrostatic mapping function

any horizontally stratified atmosphere can be described with continued fractions, when a, b, c, etc. are constants (Eq. 2.115). For verification [Marini, 1972] used standard atmosphere data but no real weather data.

$$mf(e) = \frac{1}{\sin(e) + \frac{a}{\sin(e) + \frac{b}{\sin(e) + \frac{c}{\sin(e) + \cdots}}}} \tag{2.115}$$

This concept was first used in a model for the refraction of the hydrostatic atmosphere [Marini and Murray, 1973] which has then been applied in the analysis of geodetic and astrometric VLBI observations for a long time. The zenith delay corresponds to that by [Saastamoinen, 1972b], and the mapping functions are represented by a continued fraction form with two coefficients a and b. The first mapping functions for space geodetic applications with different coefficients for the hydrostatic and wet parts were published by [Chao, 1974] who replaced the second $\sin(e)$ by $\tan(e)$ to get unity in zenith direction. [Davis et al., 1985] developed the mapping function CfA2.2 for the hydrostatic delays down to 5° elevation; it is based on the approach by [Chao, 1974] but extended by an additional constant c. Based on ray-tracing through various standard atmospheres with elevation angles between 5 and 90°, the coefficients a, b, and c were determined as functions of pressure, water vapor pressure, and temperature at the Earth surface, and from vertical temperature gradients and the height of the troposphere. [Herring, 1992] developed coefficients for the mapping function MTT (MIT Temperature) as functions of latitude, height, and the temperature at the site. Unlike [Davis et al., 1985] he did not use standard atmospheres but radiosonde data. The MTT mapping functions are based on a slightly changed continued fraction form which is widely accepted nowadays

$$mf(e) = \frac{1 + \frac{a}{1+\frac{b}{1+c}}}{\sin(e) + \frac{a}{\sin(e)+\frac{b}{\sin(e)+c}}} \tag{2.116}$$

The factor in the denominator ensures that the mapping function is equal to one in the zenith direction. The strong dependence of the MTT mapping function [Herring, 1992] on surface temperature induced [Niell, 1996] to develop the New Mapping Functions (NMF, now often called Niell Mapping Functions). The NMF do not use meteorological parameters at the sites, but only the day of the year (doy), station latitude, and station height as input parameters. Thus, they can be easily applied at stations without meteorological sensors, which often is the case for GNSS stations. [Niell, 1996] used standard atmosphere data at various latitudes to determine hydrostatic and wet mapping functions down to 3° elevation. Similar to [Davis et al., 1985] and [Herring, 1992], he used ray-tracing methods to determine the coefficients a, b, and c of the continued fraction form in Eq. (2.116). NMF is based on sine functions to describe the temporal variation of the coefficients. The period is 365.25 days and the maximum/minimum is set to January 28 (doy 28). There is also a height correction for the hydrostatic NMF (NMFh) which describes that mapping functions increase with increasing height, i.e. the atmosphere above the site becomes flatter. [Niell, 2000] was the first to determine mapping functions from numerical weather models which are often available with a time resolution of 6 h. However, some practical and conceptual limitations in the computation of the NMF induced [Böhm and Schuh, 2004] to develop the Vienna Mapping Functions (VMF). The VMF1 are valid (tuned) for elevation angles above 3°, and the largest deviations from ray-traces at other elevations show up at about 5° elevation angle. The Vienna Mapping Function 1 is realized as discrete time series (resolution 6h) of coefficients a, either on a global grid or at certain geodetic sites (see http://ggosatm.hg.tuwien.ac.at/). Mind that with the gridded version of the VMF1, the height correction of [Niell, 1996] has to be applied. [Niell, 2006] compared the mapping functions VMF1, IMF, GMF, and NMF with mapping functions derived from ray-tracing of radiosonde data in 1992, which were assumed to be the most accurate reference possible. The standard deviation was converted to station height scatter with the rule of thumb by [Niell, 2001] (one third of the delay at 7° elevation angle). The best agreement was found for the VMF1, both for the hydrostatic and the wet mapping function. All hydrostatic mapping functions show the lowest scatter at the equator, because there are only small pressure and temperature variations. The scatter increases with station latitude. The situation is different for the wet mapping functions where the scatter is largest at the equator. This is due to the fact that the zenith wet delays are largest over the equator (up to 40 cm). At the poles, there is hardly any humidity; thus, errors in the wet mapping functions are not critical

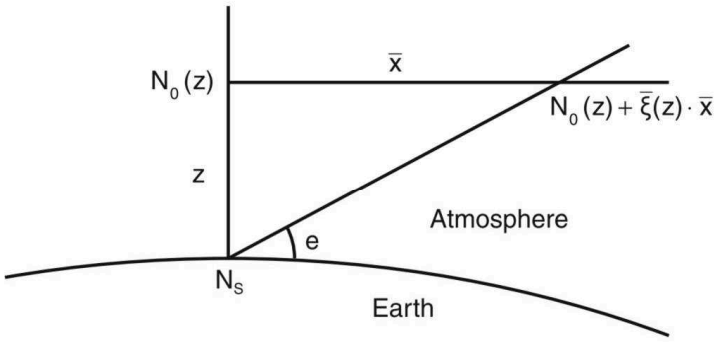


FIGURE 2.17: The refractivity in the vicinity of the vertical profile above the site can be determined with the linear horizontal gradients of refractivity ξ

for the estimation of station heights (and zenith wet delays). Again, the performance of the wet VMF1 is best.

2.4.1.2 Azimuthal Asymmetry: Gradients

Mapping functions as described above allow for the modeling of path delays under the assumption of azimuthal symmetry of the neutral atmosphere around the station. Consequently, vertical refractivity profiles above the sites are sufficient to determine the path delays at arbitrary elevation angles or the mapping functions, respectively, because the refractivity is always taken from the vertical profile as it is the case for the VMF1. However, due to certain climatic and weather phenomena path delays will not be constant at varying azimuths. For example, at sites at northern latitudes the path delay towards south will be systematically larger than towards north, because the height of the troposphere above the equator is larger than above the poles. In the following the derivation of linear horizontal gradients is shown following [Davis et al., 1993a]. The Taylor series up to degree one for the refractivity at a station is (fig.2.17)

$$N(\mathbf{x}, \mathbf{z}) = \mathbf{N_0(z)} + \xi(\mathbf{z}) \cdot \mathbf{x} \tag{2.117}$$

$$\xi_i(z) = \left. \frac{\partial N(\mathbf{x}, z)}{\partial x_i} \right|_{x=0} \tag{2.118}$$

$N_0(z)$ is the refractivity above the site, $\mathbf{x}$ is the horizontal position vector (origin is placed at the site), and $\xi(z)$ is the linear horizontal gradient vector of refractivity at height z. The index i refers to the i-th component of $\mathbf{x}$: 1 towards north and 2 towards east. The path delay (hydrostatic or wet) at an arbitrary direction can be found by

integration of Eq. (2.117) along the path s. If expressed with elevation angle e and azimuth angle a, we get

$$\Delta(a,e) = 10^{-6}\int_0^\infty N(s)ds = 10^{-6}\int_0^\infty N_0(z)ds + 10^{-6}\int_0^\infty \xi(z)\cdot\mathbf{x}ds \tag{2.119}$$

$$\Delta(a,e) = \Delta L_0(e) + 10^{-6}\int_0^\infty \xi(z)\cdot\mathbf{x}ds \tag{2.120}$$

where ΔL_0 is the delay without gradients. In Eq. (2.120) the difference between the paths with and without gradients has been neglected. [Davis et al., 1993a] state that this difference is as large as 1 mm for N = 300, $\partial N/\partial x_i = 1km^{-1}$, and $e = 20°$. If the concept of an azimuth-dependent mapping function is used, Eq. (2.120) can be written as

$$\Delta(a,e) = \Delta L^z \cdot mf(a,e) \tag{2.121}$$

with

$$mf(a,e) = mf_0(e) + \delta mf(a,e) = mf_0(e) + 10{-}6\int_0^\infty \zeta(z)\cdot\mathbf{x}ds \tag{2.122}$$

where

$$\zeta(z) = \frac{\xi(z)}{\Delta L^z} \tag{2.123}$$

and mf_0 and ΔL^z are the mapping function and the path delay in zenith direction for the symmetric case. Thus, the gradients cause a change in the mapping function which can be described by an additional term δmf . With

$$\mathbf{x}(a,e) \approx z\cdot\cot(e')(\cos(a)\ \hat{\mathbf{n}} + \sin(a)\ \hat{\mathbf{e}}) \tag{2.124}$$

$$ds \approx dz\cdot mf_0(e) \tag{2.125}$$

and normalized gradients of refractivity

$$\zeta(z) = \zeta_n\cdot\hat{\mathbf{n}} + \zeta_e\cdot\hat{\mathbf{e}} \tag{2.126}$$

and when $\hat{\mathbf{n}}$ and $\hat{\mathbf{e}}$ refer to the unit vectors in north and east direction and e' is the refracted elevation angle (which only differs from the geometric elevation angle at low elevations), we get

$$\delta(a,e) \approx 10^{-6} m f_0(e) \cot(e') \left(\cos(a) \int_0^\infty z \cdot \zeta_n(z) \cdot dz + \sin(a) \int_0^\infty z \cdot \zeta_e(z) \cdot dz \right) \tag{2.127}$$

and

$$\delta mf(a,e) = mf_0(e) + \cot(e')[Z_n \cos(a) + Z_e \sin(a)] \tag{2.128}$$

when

$$\mathbf{Z = 10{-}6} \int_{\mathbf{0}}^{\infty} \mathbf{z} \cdot \zeta(\mathbf{z}) \mathbf{dz} \tag{2.129}$$

Equation (2.127) shows that the elevation-dependence of the azimuth-dependent mapping function $\delta mf(a,e)$ consists of two parts: the dependence on mf_0 and on the factor $\cot(e')$. As already mentioned the mapping functions are dependent on the geometric elevation angle e, whereas the cotangent depends on the refracted elevation angle e', because close to the site this angle determines the refraction. It holds that

$$e' = e + \delta e(e) \tag{2.130}$$

where δe can be described with [Davis et al., 1993a]

$$\delta e \approx 10^{-6} N_s \cot(e) \tag{2.131}$$

N_S is the refractivity at the Earth surface, and for $N_S = 300$, $e = 5^\circ$ we get $\delta e \approx 0.2^\circ$. Since δe is small, $\cot(e')$ can be expanded into a series, and we get for the deviation from the symmetric mapping function

$$\delta(a,e) = mf_0(e) \cot(e')[1 - 10^{-6} N_s \csc^2(e)][(Z_n \cos(a) + Z_e \sin(a)] \tag{2.132}$$

With the delay gradients (or just gradients) **G**

$$\mathbf{G} = \mathbf{Z} \cdot \mathbf{\Delta L^z} \tag{2.133}$$

we get for the path delay

$$\Delta(a,e) = \Delta L_0(e) + m f_(e) \cot(e)[1 - 10^{-6} N_s \csc^2(e)][(G_n \cos(a) + G_e \sin(a)] \tag{2.134}$$

The equations above can be used to determine gradients by integrating over the horizontal gradients of refractivity along the site vertical [Böhm et al., 2007]

$$G_a = 10^{-6} \int_0^\infty \xi_a z dz \tag{2.135}$$

where a denotes the azimuth direction (e.g. E or N). It is important to estimate gradients in the analysis of space geodetic observations, in particular when observing at low elevation angles. On the other hand, there is no need to apply a priori gradients if no constraints are applied on the estimation of gradients.

2.5 Applications of space geodetic techniques for atmospheric studies

As discussed earlier, it is important to have a good model of the delay in the neutral atmosphere in order to obtain the highest accuracy in the space geodetic results (e.g. station positions). Since external estimates of the wet delay with high enough accuracy are typically not available (at least for microwave techniques), the common way of handling the wet delay in the data analysis is to estimate it, i.e. by modeling it using mapping functions and gradients as described in Sect.2.4.1. Thus the results of the data analysis will also contain information about the tropospheric delay, which is closely related to the IWV (see Sect.2.3.5.1). Several studies investigated the accuracy of the zenith wet delays and IWV estimated from VLBI and GNSS (e.g. [Herring et al., 1990], [Bevis et al., 1992]), and these show that it is on the same level or better then that of other techniques. Thus there exists a large interest in applying space geodetic techniques, especially GNSS, for atmospheric studies. For example, zenith wet delays can be used to study climate trends (Sect. 2.5.1), or assimilated into numerical weather prediction models to improve weather forecasts (Sect.2.5.2). With wet delays estimated

from a local GNSS network one can even attempt to estimate the 3D structure of the atmospheric water vapor by applying tomographic methods (Sect.).

2.5.1 Long-Term Water Vapor Trends

Since the zenith wet delay is closely related to the integrated water vapor content (see Sect.2.3.5.1), we can analyze ΔL_w^z estimated from space geodetic techniques to study the variations of the atmospheric water vapor content in time. For example, it is possible to study diurnal and seasonal variations as well as long term trends in the water vapor content. Such information is of great interest in climatology since the water vapor content is closely related to the temperature. Climate models typically predict that the average relative humidity remains constant as the temperature changes [Trenberth et al., 2003]. Since the saturation water vapor pressure depends approximately exponentially on the temperature, this means that a a change in the temperature will cause a corresponding change in the water vapor content. It is predicted that an increase in temperature of 1 K will increase the water vapor content by 6–7 % [Trenberth et al., 2003]. It is important to monitor the water vapor content since water vapor is a greenhouse gas, in fact the most important one. Additionally, higher water vapor content can also indicate an intensified hydrological cycle, including increased precipitation. Several studies have calculated long-term trends in ΔL_w^z (or IWV) estimated from GNSS and VLBI, e.g. [Gradinarsky et al., 2002]; [Ning and Elgered., 2012].

2.5.2 GNSS Meteorology

Water vapor is a very important parameter in meteorology and in order to get accurate weather forecasts it is very important to have accurate measurements of the water vapor content. The water vapor content indeed is highly variable in both space and time, and traditional instruments (e.g. radiosondes) do not pro-vide the water vapor content with high enough spatial and temporal resolution. With the establishment of relative dense GNSS networks the meteorological community has started to be interested in assimilating zenith total delays or zenith wet delays estimated from these GNSS networks in the numerical weather prediction models [Poli et al., 2007]. Several investigations of assimilating GNSS tropospheric delays in numerical weather prediction models have been performed. For example, in Europe this has been investigated in the projects COST-716 [Elgered et al., 2005]. It has been shown that the quality of the forecasts improve if GNSS data are assimilated, especially in cases of high precipitation [Karabatic et al., 2011].

2.5.3 GNSS Tropospheric Tomography

Tomography is a method which can be used to estimate the 2D or 3D structure of a quantity from measurements of the integral of the quantity along different paths. It is a method commonly used in medicine, seismology, material science, and a number of other fields. Tomography can also be applied to atmospheric delay measurements in order to reconstruct the 3D structure of the wet (or total) refractivity. This requires that the slant wet delays are measured by several stations in a local (inter-station distance maximum a few km) network. The only space geodetic technique for which such dense networks are available is GNSS. A picture of a GNSS tomography scenario is shown in fig.2.18. In order to estimate the wet refractivity field from the observed wet delays, the atmosphere above the GNSS network is parameterized. The most commonly used parameterization is voxels, although other parameterizations are also possible [Perler et al., 2011]. Voxel parameterization means that the atmosphere is divided into a number of boxes (called voxels, volume pixels) in which the refractivity is assumed constant. Thus the wet tropospheric delays along the rays of the observed GNSS signals can be described by a linear combination of the voxel refractivities

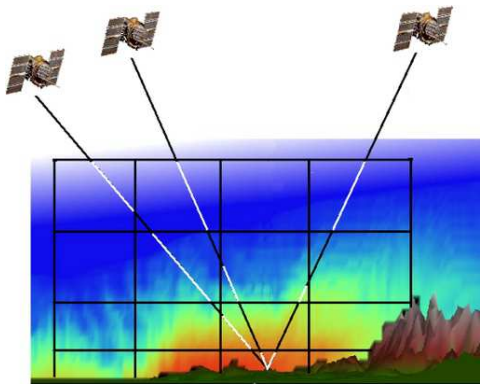


FIGURE 2.18: A picture of a GNSS tomography scenario [Nilsson, 2013]

$$d_i = \sum_{j=1}^{n_{vox}} A_{ij} m_j \tag{2.136}$$

d_i is the wet tropospheric delay along the i-th ray, n_{vox} is the number of voxels, m_j is the refractivity of the j-th voxel, and A_{ij} is the distance traveled by ray i in voxel j. Since the station and satellite coordinates are normally known, A_{ij} can be calculated. Having observations of the tropospheric delays along several different rays, a linear system of equations is obtained

$$\mathbf{d} = \mathbf{Am} \tag{2.137}$$

d and **m** are vectors containing the wet delays and the refractivities, respectively, and **A** is a matrix containing the A_{ij} values. By inverting the system, the voxel refractivities are estimated. There are however a few problems with this method. One is how to estimate the slant wet delays along the GNSS signal rays. One way is to estimate the zenith wet delay and gradients in a normal GNSS analysis, then use these to calculate the slant wet delays. This is for example done by [Champollion et al., 2005]. However, this method will limit the accuracy since it assumes that all horizontal variations in the refractivity above a station are linear, something which is not always realistic. In order to improve the slant wet delays, it is often assumed that the post-fit residuals of the GNSS analysis will contain the unmodeled parts of the slant wet delays, and thus adding these to the calculated slant wet delays will give the true delays [Troller et al., 2006]. Another approach is to model the slant delays by using Eq. (2.136) in the GNSS data analysis instead of zenith delays and gradients. First results using this approach are presented by [Nilsson and Gradinarsky, 2006] and [Nilsson et al., 2007]. Another problem is the normally weak geometry since tomography ideally requires that there are rays crossing the investigated volume in all possible directions. In GNSS tomography, however, all rays are going between the top of the troposphere to the stations on the surface of the Earth, while there are no rays entering and/or leaving the voxel grid at the sides. This makes the sensitivity to the vertical refractivity profile very low and as a result the equation system (2.137) will be ill-conditioned. This problem can be solved by either constraining the refractivity field to some a priori field obtained either by models or external measurements like radiosondes. The problem is not as big if the GNSS stations are placed at very different altitudes Furthermore, since the satellite geometry will change during the day, some voxels may at times have no or only a few rays passing through it. Thus, in order to avoid singularity problems, constraints need to be applied.

Chapter 3

The etnean GPS network: data, check and elaboration

3.1 Introduction

The main task of this work is the analysis of the potentiality of GPS to detect the amount of wet delays in the troposphere, to improve the DInSAR technique through an atmospheric water vapor delay filtering. Since the zenith wet delay is closely related to the integrated water vapor content which is the main driver of meteorological events, these studies should have also meteorological applications. To obtain output data, the definition of the best processing strategy was the main starting point. We analyzed all of the available GPS data using the GAMIT/GLOBK software [Herring et al., 2010]. The delay that the EM waves emitted by GPS satellites accumulate through the troposphere, were calculated starting from the Etna network "Etn@net" data, owned by Istituto Nazionale di Geofisica e Vulcanologia - Osservatorio Etneo (INGV-OE). Because of its height (3300 m from sea level) and the quite variable weather conditions, as regards the content of water vapor, the estimation of atmospheric delays using GPS measurements have exceptional importance to calibrate the SAR interferograms and to determine the "effective" deformation of the volcanic measured with this technique.
Considering the availability of a GPS network on a volcanic area, a "by product" of the study concerns the detection of volcanic plume in the atmosphere, which will be dealt on Chapter 4

3.2 The Etn@net characteristics

Routine use of GPS for monitoring ground deformation started at Mt. Etna in 1988, when a network of 18 benchmarks was surveyed by both GPS and EDM (Electronic Distance Measurement) techniques [Briole et al., 1990]. Since that first survey, the network has been improved and measured by GPS in survey mode almost yearly. In late 2000, the installation of a continuous GPS permanent network at Mt. Etna began. The network geometry was gradually upgraded in the following years [Palano et al., 2010] to reach the current configuration of 40 permanent stations that provide a dense cover of most areas of the volcano edifice (Figure 3.1).

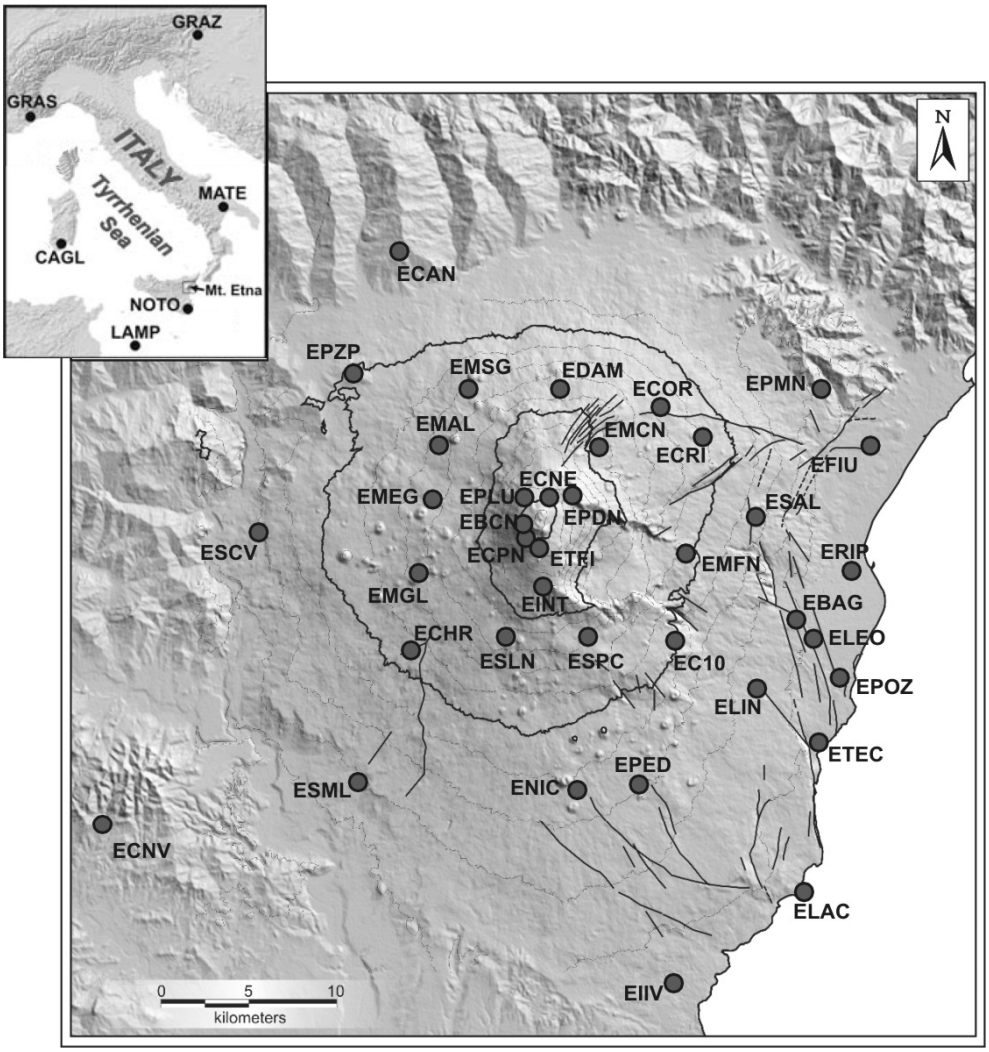


FIGURE 3.1: Sketch map of Mt. Etna. The permanent GPS network stations are indicated as circles. Inset, the IGS stations used in this study [Bruno et al., 2012].

The Mt. Etna volcano GPS network (Etn@net) allows monitoring of the volcanic deformation at different time scales. We can measure slow (days or months) or fast (minutes

or hours) changes in site positions using different techniques in relation to data acquisition and processing. In the first case, we process GPS data collected at 30-s sampling rates on a daily basis, while in the second we perform high-frequency (1 Hz) instantaneous GPS positioning. The increasing number of GPS stations has allowed to deeper the knowledge of the volcano dynamics [Aloisi et al., 2011] [Bruno et al., 2012] [Palano et al., 2010] and to highlight some important kinematic features. Table 3.1 shows the list of GPS stations composing the Etn@net together with information about receiver, antenna, sample frequency and transmission type. The GPS receivers are firstly used as reference station applications. For this reason the GPS receivers have to incorporates a very high precise GNSS measurement engine. The essential characteristics for a GPS reference receiver are: the acquisition within seconds, an excellent signal strength, the reliable tracking to low elevations, an algorithm for phase and code multipath suppression, an high precision code and phase measurements and, for remote stations, a low power consumption. GPS and GLONASS are continually improving and the European Galileo and Chinese Compass systems are emerging to provide additional benefits for world-wide users. For this reason, the receiver has to be able to track the new multi-frequency signals from modernized GPS. During the last years, the IGNV improved its network updating the GPS hardware. There is a particular care in choosing the sites of the station. The host site affects the quality of the measurement mainly due to the multipath. As already explained, the multipath depends on the geometry of the antenna and satellite with respect to surrounding reflective surfaces.

Choke ring antennas are recognized as having the best accuracy and tracking performance compared to all other antenna types. They are widely used in the INGV infrastructure network. The choke ring antenna benefits of the multipath rejection and high signal to noise ratio, whilst at the same time improving tracking of low elevation satellites. An example of a Leica choke ring antenna owned INGV is reported in figure 3.2 A choke ring antenna consists of a number of conductive concentric cylinders around a central antenna. Due to its delicate construction, it is often enclosed in a protective cover or "radome" when placed outside and exposed to the elements. Choke ring antennas are notable for their performance to reject multipath signals from a source. Since the path that a signal takes from transmitter to receiver can be used to measure the distance between the two, this makes it highly suited for GPS and radar applications. In a GPS ground based receiver, a choke ring antenna can provide millimeter precision measurements for surveying and geological measurement applications. Obviously today's all the GPS receivers of Etn@net are equipped with choke ring antennas.

Abbrev.	Site	GPS Receiver	GPS Antenna	Rate[sec]
EBAG	Baglio	GMX902GG	AT504GG	1
EBCN	Bocca Nuova	GX1220	AT504GG	1
ECAN	Masscria Cannata	GRX1200GG Pro	AT504	1
ECHR	Santa Chiara	GMX902	AT504	30
ECNE	Cratere Nord Est	GRX1200GG Pro	AX1203+GNSS	1
ECNV	Catenanuova	GMX902	AT504	30
ECOR	M. Corruccio	GMX902GG	AT504GG	1
ECPN	Cratere del Piano	GRX1200GG Pro	AX1202GG	1
ECRI	M. Crisimo	GPR1200	Ax1202	1
EDAM	Etna Dammusi	GX1220	AT504	1
EFAR	Contrada Farella	GRX1200GG Pro	AT504	1
EFIU	Fiumefreddo	GMX902GG	AT504GG	30
EIIV	Catania	GRX1200GG Pro	AX1202GG	1
EINT	Intermedia	GX1220	AT504	5
ELAC	Isola Lachea	GRX1200GG Pro	AT504GG	1
ELEO	San Leonardello	GX1220	AT504	1
ELIN	Linera	GRX1200	Ax1202	1
EMAL	Monte Maletto	GX1220	AT504	30
EMCN	Monte Conca	GX1220	AT504	1
EMEG	Monte Egitto	GX1220	AT504	30
EMFN	Monte Fontane	GRX1200GG Pro	AT504	1
EMGL	Monte Gallo	GX1220	TRM29605	30
EMSG	Monte Spagnolo	GRX1200GG Pro	AT504	1
ENIC	Nicolosi	GRX1200GG Pro	AT504	1
EPDN	Pizzi Deneri	GX1220	AX1202GG	1
EPED	Pedara	GX1220	AT504GG	30
EPLU	Punta Lucia	GX1220	AX1202GG	1
EPMN	Piedimonte	GRX1200	AT504	1
EPOZ	Pozzillo	SR502	AT504	30
EPZF	Pizzo Felice	GMX902GG	AT504GG	30
ERIP	Riposto	GMX902GG	AT504GG	30
ESAL	Sant Alfio	GRX1200GG Pro	AT504	1
ESCV	Scorciavacca	GMX902	AT504	30
ESLN	Serra La Nave	GRX1200GG Pro	AT504	1
ESML	Santa Maria di Licodia	SR502	AT504	30
ESPC	Serra Pizzuta	GX1220	AT504	1
ETEC	Santa Tecla	GMX902GG	AT504GG	1
ETFI	Torre del Filosofo	1230Plus	AT504GG	30

TABLE 3.1: Summary of today's GPS station hardware. The abbreviation of the station and its site, the receiver and antenna type, together with the sample frequency and the height are shown

FIGURE 3.2: An example of a Leica choke ring antenna.

3.3 GAMIT Processing Strategy

The elaboration of the geodetic GPS data allows to obtain a high-precision output of position, velocity and atmospheric parameters. In the present work the GPS data coming from the Etna network were processed using the GAMIT package. GAMIT is a GPS analysis package developed at MIT Massachusetts Institute of Technology [Herring et al., 2010], the Harvard-Smithsonian Center for Astrophysics (CfA), and the Scripps Institution of Oceanography (SIO) for estimating station coordinates and velocities, stochastic or functional representations of post-seismic deformation, atmospheric delays, satellite orbits, and Earth orientation parameters. The software is designed to run under any UNIX operating system like Linux for example.
High-precision geodetic measurements with GPS are performed using the carrier beat phase. This observable is the difference between the phase of the carrier wave implicit in the signal received from the satellite, and the phase of a local oscillator within the receiver. The phase can be measured with sufficient precision that the instrumental resolution is a millimeter or less in equivalent path length. For the highest relative-positioning accuracies, observations must be obtained simultaneously at each epoch from several stations (at least two), for several satellites (at least two), and at both the L1 and L2 GPS frequencies, as explained in Chapter 1)

3.3.1 GAMIT Processing Algorithms

GAMIT incorporates difference-operator algorithms that map the carrier beat phases into single and double differenced phases. These algorithms extract the maximum relative positioning information from the phase data regardless of the number of data outages, and take into account the correlations that are introduced in the differencing process. In the presence of cycle slips, initial processing of phase data is often performed using triple difference or Doppler observations in order to obtain a preliminary estimates of station or orbital parameters. GAMIT software uses triple differences in editing but not parameter estimation.

GAMIT is composed of distinct program modules which perform the functions of preparing the data for processing (MAKEXP, MAKEJ, MAKEX and FIXDRV).

The first MAKEXP creates driver file for the elaboration, session information file and batch input files from the RINEX observation files and the station information file. Secondly, MAKEJ using the information in global broadcast navigation file and the partial derivative in the observation file creates the satellite clock information file. MAKEX, using the session information, satellite clock information and predetermined coordinate information of the station, creates values of receiver clock offset during observation span from pseudorange and input observation file. Lastly FIXDRV, running with different modules, creates the output files. The main requirement of any GPS geodetic measurement session is an accurate model of the motion of satellites. The motion of a satellite can be described, in general, by a set of six initial conditions (Cartesian position and velocity, or osculating Keplerian elements, for example) and a model for the forces acting on the satellite over the span of its trajectory. To model accurately the motion, we require knowledge of the acceleration induced by the gravitational attraction of the sun, moon, and higher order terms in the Earth's gravity field, and some means to account for the action of non-gravitational forces due to solar radiation pressure and gas emission by the spacecraft's batteries and attitude control system. In GAMIT the reference orbits for the satellites integration is performed by program (ARC). Others GAMIT programs preparing for the computing residual observations (*obs.* − *calc.*) and partial derivatives from a geometrical model (MODEL), detecting outliers or breaks in the data (AUTCLN), and performing a least-squares analysis (SOLVE).

Although the modules can be run individually, they are tied together through the data flow (fig. 3.3). The most processing is best done with shell scripts and a sequence of batch files set up a driver module (FIXDRV) for modeling, editing, and estimation.

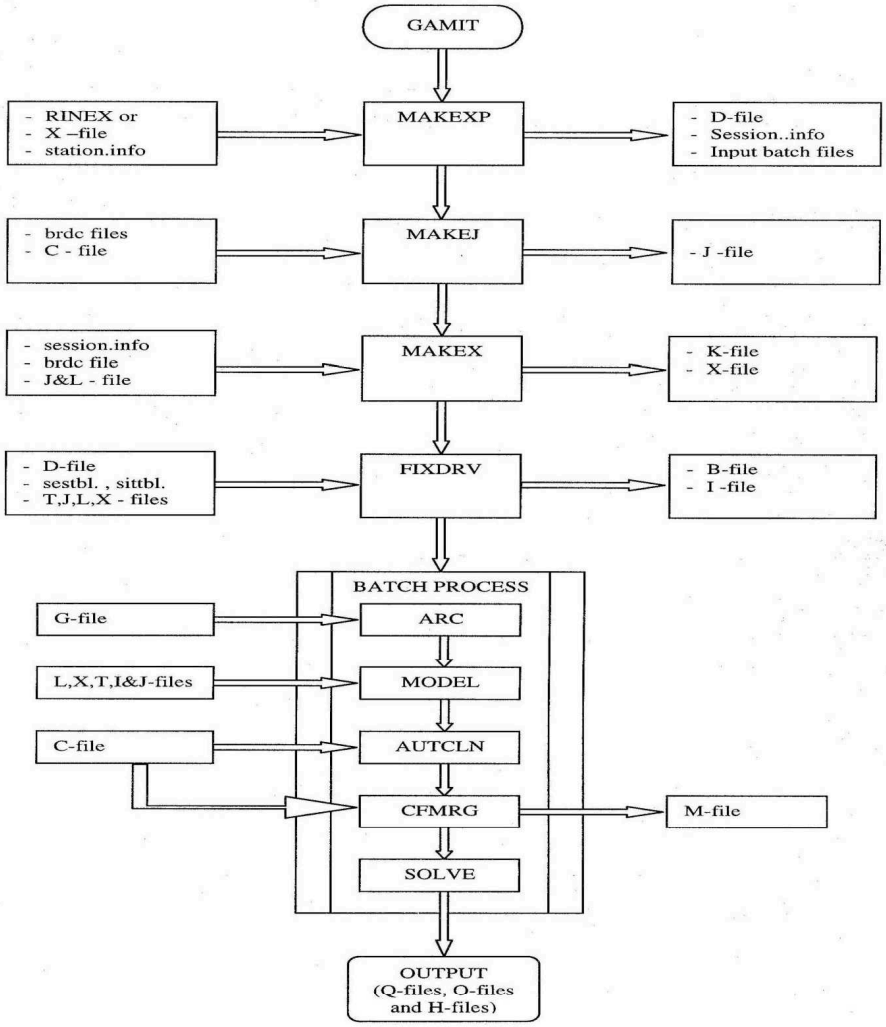


FIGURE 3.3: General layout of GAMIT processing (see appendix A for help about the abbreviations.

3.3.2 Parameter Estimation

GAMIT incorporates a weighted least-squares algorithm to estimate the relative positions of a set of stations, orbital and Earth-rotation parameters, zenith total delays (dry and wet), gradients (N-S and E-W) rescaled at 10 degrees elevation and phase ambiguities by fitting to doubly differenced phase observations. The last output of processing are the LC undifferenced residuals. Since the functional (mathematical) model relating the observations and parameters is non-linear, GAMIT produces two solutions, the first to obtain coordinates within a few decimeters, and the second to obtain the final estimates. How to get the output GAMIT data begins, first and foremost, from the navigation and observation data files for each day of the year (doy) at the receiver in compressed RINEX (The Receiver Independent EXchange) format. The navigation data is discarded and precise ephemeris data is downloaded from Scripps Orbit and Permanent Array Center (SOPAC) data archives. Six IGS stations (CAGL, MATE, NOT1,

GRAS, GRAZ, ZIMM) to tie to ITRF are used together with Etna network in all the elaborations. In the subsequent text, we will describe all input files required for GAMIT elaboration.
Before processing GPS data using GAMIT, the information should be organized into sessions, defined as the spans during which all the stations tracks simultaneously the phase of two or more satellites. A brief description how a single session is processed is reported. After having create a working directory, we have to copy or link all the necessary files. The main files needed for a single session processing are:

- RINEX observation and navigation files;
- station coordinates in the form of an *L-file*;
- receiver and antenna information for each site file (*station.info*);
- satellite list and scenario file (*session.info*);
- control files for the analysis (*sittbl.* and *sestbl.*);
- G- and T-files from external ephemerides;
- links to the following global files:
  - nutations (*nutabl.*);
  - lunar and solar ephemerides (*soltab.* and *luntab.*);
  - geodetic datums (*gdetic.dat*);
  - leap seconds (*leap.sec*);
  - spacecraft, receiver, and antenna characteristics (*svnav.dat, antmod.dat, rcvant.dat*)
  - Earth rotation (*pole., ut1.*);
  - ocean tides (*stations.oct* and *grid.oct*)
- the control files for sh_gamit.

RINEX observation and navigation files
The GPS data acquired by Etn@net GPS stations in RAW format, are sent, through a wireless network surrounding Mount Etna, in the headquarters of the INGV-OE. Here are stored in a database available to the scientists. Data rate is 30 seconds, and all files contain observations for a 24 hour period, from 00 : 00 : 00 till 23 : 59 : 30 GPS time. File naming follows the convention "NNNNddds.yy" where NNNN is the site name (e.g EIIV), ddd refers to the day of the year (e.g 024 means January 24th), s is the session number (e.g 0), and yy is the two last digit of the year. Examples of RINEX header is

```
1     2.11           OBSERVATION DATA    G (GPS)             RINEX VERSION / TYPE
2  teqc  20100ct21                        20110708 09:45:04UTCPGM / RUN BY / DATE
3  MSXP|IAx86-PII|bcc32 5.0|MSWin95->XP|486/DX+               COMMENT
4  teqc  20100ct21                        20110708 09:45:04UTCCOMMENT
5  BIT 2 OF LLI FLAGS DATA COLLECTED UNDER A/S CONDITION       COMMENT
6  EPLU                                                        MARKER NAME
7  -Unknown-           -Unknown-                               OBSERVER / AGENCY
8  ----------          LEICA GX1220        2.14/2.121          REC # / TYPE / VERS
9                      LEIAX1202                               ANT # / TYPE
10   4879005.8155  1306023.8700  3886682.4815                  APPROX POSITION XYZ
11        0.0000        0.0000        0.0000                  ANTENNA: DELTA H/E/N
12     1     1                                                WAVELENGTH FACT L1/2
13     5    L1    C1    L2    P2    D1                        # / TYPES OF OBSERV
14    30.0000                                                  INTERVAL
15   2007     8    17     0     0    0.0000000     GPS         TIME OF FIRST OBS
16                                                            END OF HEADER
17  07  8 17  0  0  0.0000000  0  6G14G03G22G11G19G01
18  115827317.621 7  22041203.159    90255052.28346  22041202.6274       1365.203
19  108950960.424 8  20732673.199    84896862.51047  20732673.0144      -1819.676
20  115268805.249 8  21934919.219    89819861.51846  21934916.6004      -2520.625
21  120611492.440 7  22951603.260    93982982.89446  22951601.9914       2605.179
22  107381926.289 8  20434095.770    83674227.86048  20434092.9394        778.223
23  120541180.395 7  22938223.209    93928203.21346  22938223.7834       3047.710
24  07  8 17  0  0 30.0000000  0  6G14G03G22G11G19G01
25  115786617.055 7  22033458.194    90223337.55746  22033457.4534       1348.098
26  109005776.944 8  20743104.341    84939576.68547  20743104.2754      -1834.921
27  115344545.824 8  21949331.982    89878880.14846  21949329.6134      -2528.944
28  120533383.586 7  22936739.716    93922118.89046  22936738.3494       2601.956
29  107358806.796 8  20429696.146    83656212.66748  20429693.5294        762.921
30  120449854.898 7  22920844.676    93857040.49546  22920845.1284       3040.465
31  07  8 17  0  1  0.0000000  0  6G14G03G22G11G19G01
32  115746439.090 7  22025812.605    90192030.05446  22025811.9704       1330.438
33  109061051.844 8  20753622.798    84982648.03247  20753622.9744      -1850.195
34  115420537.328 8  21963792.761    89938094.30846  21963790.2704      -2537.121
35  120455372.554 7  22921894.657    93861331.09546  22921893.2724       2598.695
36  107336152.122 8  20425385.212    83638559.67048  20425382.4724        747.324
```

FIGURE 3.4: Rinex file example containing L1 and Lc data for each epoch

given in figure 3.4. As summarize in line 13, the RINEX observation file contains the L1 and L2 carrier beat phases, C/A code, P code of L2 and shift doppler of L1. The receiver and antenna type, together with initial station coordinates and antenna offsets and start and stop times are also stored in the file. According with the RINEX standard, beginning from line 17, the time of the acquisition and the identification of the satellites tracked in each receiver channel are recorded. The lines from 18 to 23, represent the values of L1, C/A, L2, P2, D1 for each linked satellite, ordered according to the line 17. The line 24 repeats the same explained scheme, but referred to the next epoch.

Preparing the L-file

The L-file contains the coordinates of all the stations to be used in the data processing. Two forms of coordinate formats are supported: geocentric (spherical) coordinates at the epoch of the observation, or Cartesian coordinates and velocities at a specified epoch.

Creating the station information file

All of the receiver and antenna information, specific for each particular site, are recorded in the station.info file. The entered values correspond to a single occupation, of either one day or a series of days. In this file format, the number of entries (columns) is variable and determined by a list beginning with the keyword *SITE. The most important among the entries in station.info are the antenna type (AntCod) and specification of how the height-of-instrument was measured (HtCod), since this directly affects the estimated heights from the analysis. This information is entered into the file in form of keywords

and later converted by GAMIT to L1 and L2 phase-center value offsets (PCV). Entries for horizontal offsets (Ant N, Ant E) of the antenna from the monument can also be entered. A GPS monument is a concrete structure whose philosophy is to create a highly rigid frame which is in solidarity with the movements of the ground. The station.info values are added to the coordinates of the monument in computing the antenna phase-center position.

Creating a scenario file

The scenario file is also known as the session.info and it contains the start time, sampling interval, number of observations, and satellites (PRNs) to be used in generating X-files for each day. The session.info file can be specific to a given measurement session or contain all of the scenarios used for all the measurement sessions processed. It can also be generated automatically to a specific measurement session by the program makexp using the input start/stop time and the satellites available on the navigation file.

Control files for the analysis (sittbl. and sestbl.)

The site control table *sittbl.* contains the number and name of stations to be used in the measurement session. The session control table (*sestbl.*) contains all the entries useful to define the most important aspect of the elaboration, the a priori constraints in units of meters, the elevation cutoff, the model for variations in the phase center of the receiving antenna, the way the receiver clock can be controlled and the atmospheric model used in the computation. More details about this file can be found in [Herring et al., 2010]. In summary, the *sestbl.* entries can be put into some categories: analysis controls, data weighting, ambiguity resolution, atmospheric, orbit, MODEL and SOLVE parameters.

G- and T-files from external ephemeris

Orbital information is input to GAMIT as a tabular ephemeris (T-) file, which contains the positions of all the satellites at 15-minute intervals throughout the observation span, or a G-file of initial conditions which are integrated by arc to create T-file. The most accurate and reliable method of obtaining a starting T-file is to first download SP3 file(s) from IGS analysis center.

Global Files

These files are called global because they can be used for many measurement sessions over the time interval for which they are valid (usually at least a year). These tables are linked to each working directory in order to minimize the storage space that will be needed if we store them in each working directory. The global files needed for a single session processing are explained below.

(nutabl.) This is the nutation table and contains nutation parameters in tabular form for transforming between an inertial and Earth-fixed system.

(soltab. and luntab.) soltab. is the solar tabular ephemeris, which is a tabulation of

the position of the Earth with respect to the sun. luntab. is the lunar tabular ephemeris which contains the Moon's position x, y, and z for exact Julian date.

(gdetic.dat) This is the table of parameters of geodetic datums which are specified by the standard ellipsoid parameters, semi-major axis (in meters), Cartesian offsets (in meters) from the geocenter.

(leap.sec) This a table of jumps (leap seconds) in TAI-UTC since 1 January 1982.

(svnav.dat, antmod.dat, rcvant.dat) svnav.dat gives the correspondence between spacecraft numbers and PRN numbers for each GPS satellite, its mass, and it's yaw parameters. The table is updated after each launch or change in yaw status. antmod.dat contains the table of antenna phase center offsets and variations as a function of elevation and azimuth. rcvant.dat is a table of correspondences between 6-character codes and the full names of receivers and antennas used in the RINEX files.

(pole., ut1.) pole. is pole table and contains polar motion values in tabular form for interpolation in different GAMIT modules. ut1 is a table which contains TAI-UT1 values.

(stations.oct and grid.oct) These tables contain ocean tidal loading components of stations and global grid respectively.

The control files for sh_gamit

The script sh_gamit take us, with a single command, from raw or RINEX data over a range of days, to a solution. The only preparation required is setting up the control files:

(process.defaults) This file is edited to specify the computational environment, sources for internal and external data and orbit files, start time and sampling interval, and instruction for archiving results.

(sites.defaults) We edit this file to specify the International GPS Service (IGS) stations which are to be used in the measurement session and how the station log data are to be handled. Even though sh_gamit can automatically download RINEX and ephemeris files from SOPAC and precise orbits from IGS ftp-servers, it is much simple when all the data are in the local system in each directory.

3.3.3 The GAMIT output files

The analysis of the GPS Rinex files coming from the Etna network (figure 3.4 and table 3.1) was performed using the GAMIT software. For these observations, we used a variety of geodetic-quality (low-multipath) antennas, and included in our analysis elevation-dependent models for the antenna phase centers [Schmid et al., 2005]. Standard models for precession, nutation, Earth rotation, and solid Earth and ocean tides are applied (IERS Conventions 2010). The positions of the GPS satellites are taken from

the Final Orbits of the International GNSS Service (http://igscb.jpl.nasa.gov), which typically have an accuracy of 1–2 cm. For our ETNA analysis, we combined the data from our local network (fig. 4.16a) with those from six IGS stations in the region (fig. 4.16b), included in the International Terrestrial Reference Frame (ITRF2005) realization and used as reference stations to tie the network. We used a 2-h spacing for the zenith delays and gradients estimations. The delay along the line between the station and the satellite is computed by separate "mapping functions" for the hydrostatic ("dry") and water vapor ("wet") components. The dry mapping function is applied to that part of the modeled delay defined by the a priori pressure value, and the wet mapping function is used to compute the partial derivatives for the estimated correction to the zenith delay. For our analysis, we used a standard constant model for surface pressure and the mapping functions developed by [Niell, 1996]. We included only data above 15 degrees elevation angle in order to minimize multipath effects and errors in the mapping functions. The first-order ionospheric delay is removed by forming the "ionosphere-free" combination of the L1 and L2 phases. The small (order of cm) second- and third-order ionospheric effects are modeled using global maps of total electronic content and the IGRF model of the magnetic field [Petrie et al., 2010]. After estimating station positions and atmospheric parameters from the doubly differenced phase, GAMIT produces residuals for the undifferenced phases by estimating clock corrections which cancel in the double-differences. We use these undifferenced post-fit phase residuals as input in the test phase in chapters 4 and 5.

3.4 The Quality Control Check

The GPS network Etn@net, today's, is made up of 40 GPS receivers that collect data at various bit rate. All these data must be checked and analyzed. The geodetic GPS data, as well known, allow to obtain high-precision outputs, like position, velocity and atmospheric parameters. To achieve this level of precision, it is mandatory to perform a careful maintenance and operation of the site. In addition, the importance of the reliable data becomes more evident if they are applied to obtain informations about the homogeneity of the atmosphere or the presence of plume. During this studies, we analyzed a great amount of data and almost immediately, we realized that it was necessary to check the data quality; the easiest way was to use the program TEQC.
TEQC was designed and developed by the University Navstar Consortium (UNAVCO) Facility, the program is named after its three main functions: Translation, Editing, and Quality Checking [Estey and Meertens, 1999]. With TEQC, we can extract a variety of information from a RINEX file. Some of the information we get are: the receiver clock slips, receiver cycle slips, site multipath, satellite elevation and azimuth angle, receiver

clock drift, and receiver signal-to-noise ratios. In our studies we use the software to produce a file with a full report that shows the signal quality with time for each SV detected.

An extract of the quality check file generated by TEQC, starting from the Rinex file EMGL2470.07o is reported in figure 3.5. The first three lines represent the observation

```
Observations start   : 2007 Sep  4 00:00:00.000
Observations  end    : 2007 Sep  4 23:59:30.000
Observation interval : 30.0000 second(s)

 SV  #+hor <ele> #+mask <ele> #reprt #compl    L1     L2     P1     P2     CA    L2C
--- ------ ----- ------ ----- ------ ------ ------ ------ ------ ------ ------ ------
G 1   1154 25.25    897 31.04    821      0    821    820      0      0    821      0
G 2   1184 22.08    932 26.72    861      0    861    861      0      0    861      0
G 3    859 41.24    740 47.08    739      0    739    739      0      0    739      0
G 4   1156 23.65    891 29.21    822      0    822    819      0      0    822      0
[blurred]
G29    860 41.49    740 47.41    739      0    739    739      0      0    739      0
G30    928 39.87    813 44.80    813      0    813    813      0      0    813      0
G31   1181 23.03    920 28.17    856      0    856    856      0      0    856      0
```

FIGURE 3.5: An example of the quality check file generated by TEQC. The time limits and the observables sample time interval, together with the table summarizing what was seen by each SV are reported.

file's time limits and the observables sample time interval. The table below, summarizes what was seen by each SV, in order the columns in each line are:

SV: the PRN number (G for GPS and R for GLONASS);
#+hor: the total number of all types of observations above the horizon for this SV;
<ele>: the mean elevation of SVs above the horizon for epochs with observations;
#+mask: the number of observations above the elevation mask for this SV;
<ele>: the mean elevation of SVs above the elevation mask for epochs with observations;
#reprt: the number of reported observations with any data reported for this SV;
#compl: the number of complete observations reported for this SV;
Next are six columns with the number of **L1, L2, P1, P2, C/A,** and **L2C** observations for this SV.

Along with this, other tables and information are returned by TEQC software during the analysis of RINEX file. Although we used TEQC a lot, expecially during the conversion from binary format to RINEX format, the quality check instrument is not the right tool for our analysis. The quality check file, in fact, returns information about a single file; for the analysis that we will discuss in the next chapters, it is necessary to build a synoptic view of the station's behavior.

In this regard, we used Gamit to elaborate data from January 1 to June 30, 2012, in order to observe the characteristics of each GPS receiver of the network. By writing a code developed in MATLAB, we load the LC undifferenced residuals for each couple of antenna and receiver. Gamit write one residual file, called DPH file, for each antenna and satellite pair, consequently our session of elaboration have been generate about 230,000 DPH files. We developed a MATLAB code to load all the LC undifferenced phase residual, we tried to load one year of DPH data but the personal computer resources are insufficient to handle such amount of data. We plotted the LC residual so as to deduce the anomalies for each station.

From the 38 GPS receiver analyzed, the following figures are just some of them and report four representative different examples of the station behavior.

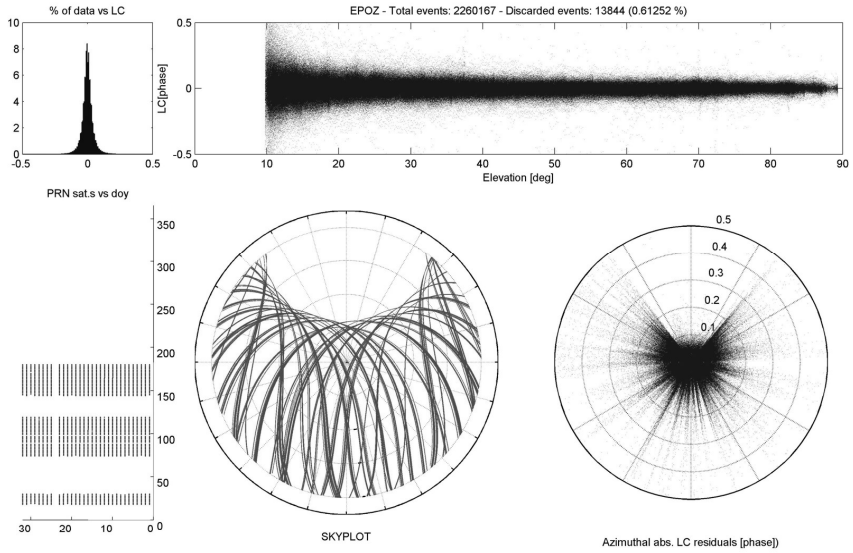


FIGURE 3.6: A synoptic view of the undifferenced LC residuals in the GPS station EPOZ, from 01 January to 30 June 2012. The top left histogram represents the distribution of the LC postfit residuals expressed in percent; the top right figure shows the LC postfit residuals versus the elevation angle expressed in degrees; the bottom left represents a rolled graph containing the PRNs versus the GPS doys of 2012; the bottom center is the post elaboration skyplot, reported to the considered GPS receiver; the bottom right is a polar graph containing the absolute value of the LC postfit residuals versus the azimuthal angle expressed in degrees.

Figure 3.6 shows an example of the residuals during the first half of 2012 for the GPS station EPOZ. The LC postfit undifferenced residuals behavior can be understood only plotting a lot amount of data in a single figure.

The top left histogram represents the distribution of the LC postfit residuals expressed in percent of data. With the help of this plot we establish immediately if the residuals follow a normal distribution. The χ^2 test was performed also, confirming the assumption. The top right plot shows the LC postfit residuals versus the elevation angle expressed in degrees. This is a very important plot because allow to obtain information about the

multipath of the station, the presence of obstacles, information about the PCV modelling and in general the elevation dependent noise of the station.

The Figure 3.7 allows us to understand what we have previously stated. It must be

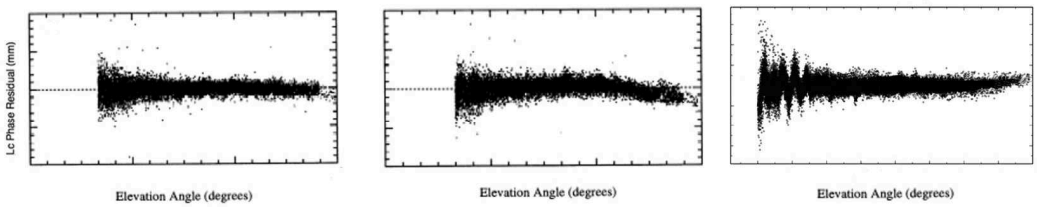


FIGURE 3.7: Three typical envelopes of the LC phase residuals versus elevation angle.

read in a qualitative manner (absolutely not quantitative), three typical trends of LC GPS undifferenced residuals versus the elevation angle are reported, this is the reason why we omitted the values in the axes. The first (figure 3.7 left) is a normal LC residual behavior, an hard dependency from the elevation angle can be revealed. The center plot shows the results of a generic station in which a microwave absorber was placed. The phase center variation model is no longer able to correct the observations, an anomalous negative slope appears at high elevation angles.

The plot on the right of figure 3.7 shows the phase residual of a GPS receiver placed

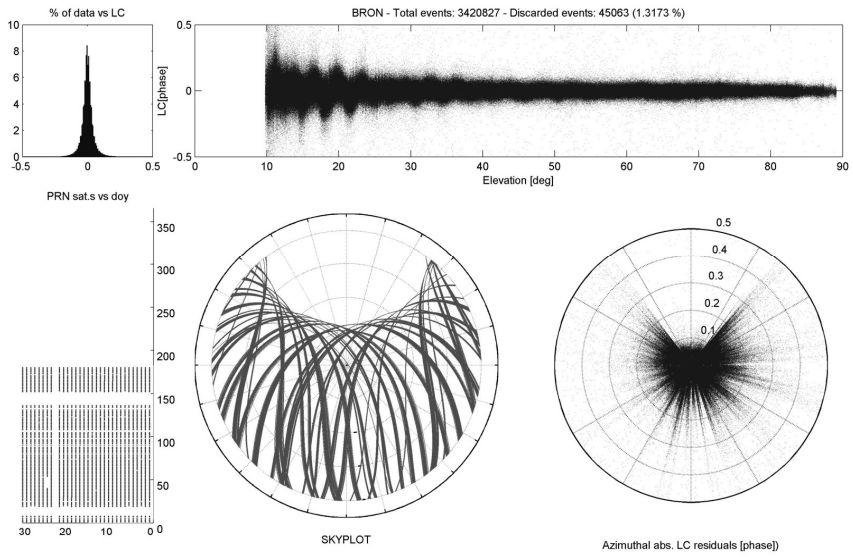


FIGURE 3.8: A synoptic view of the undifferenced LC residuals in the GPS station BRON, from 01 January to 30 June 2012. The top left histogram represents the distribution of the LC postfit residuals expressed in percent; the top right figure shows the LC postfit residuals versus the elevation angle expressed in degrees; the bottom left represents a rolled graph containing the PRNs versus the GPS doys of 2012; the bottom center is the post elaboration skyplot, reported to the considered GPS receiver; the bottom right is a polar graph containing the absolute value of the LC postfit residuals versus the azimuthal angle expressed in degrees.

over a reflective surface (a flat, smooth ground), a typical wave form appears at low elevation angles.

From these considerations, the behavior of figure 3.6 top right is as expected. The bottom left plot (figure 3.6) represents a rolled graph containing the satellite PRN numbers versus the GPS day of year (doy), it is useful because allows immediately to estimate the data loss; in that case for example there are large ranges of missing data. The bottom center plot represent the post elaboration skyplot reported by the Gamit software. We use this plot to evaluate the presence of obstacles or anomalies in the satellite and receiver link . Finally, the bottom right is a polar graph containing the absolute value of the LC postfit residuals versus the azimuthal angle expressed in degrees. This plot should be read together with the elevation dependence residuals plot, they give us a 4π information of the state of the residuals. As said above, the figure 3.6, shows no particular anomalies, except for different intervals of missing data. Figure 3.8 is referred to the station BRON, it shows two intervals of missing data of about ten days during may 2012. A typical wave form appears at low elevation angles, suggesting that the GPS station has been placed over a reflective surface. There are no relevant anisotropies in the space concerning the LC phase residual. Example 3 is reported in figure 3.9, that

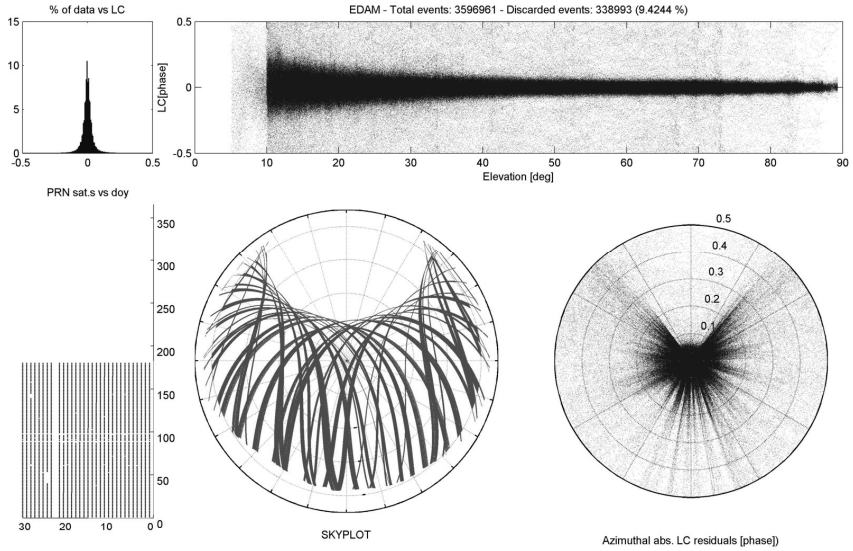


FIGURE 3.9: A synoptic view of the undifferenced LC residuals in the GPS station EDAM, from 01 January to 30 June 2012. The top left histogram represents the distribution of the LC postfit residuals expressed in percent: the top right figure shows the LC postfit residuals versus the elevation angle expressed in degrees; the bottom left represents a rolled graph containing the PRNs versus the GPS doys of 2012; the bottom center is the post elaboration skyplot, reported to the considered GPS receiver; the bottom right is a polar graph containing the absolute value of the LC postfit residuals versus the azimuthal angle expressed in degrees.

refers to the station EDAM. A clear evidence of an obstacle in the south-west side is present. There are no anisotropies in the space LC phase residual but they are greater

then the previous figures (i.e. figure 3.8). Figure 3.10, referred to the station EBCN,

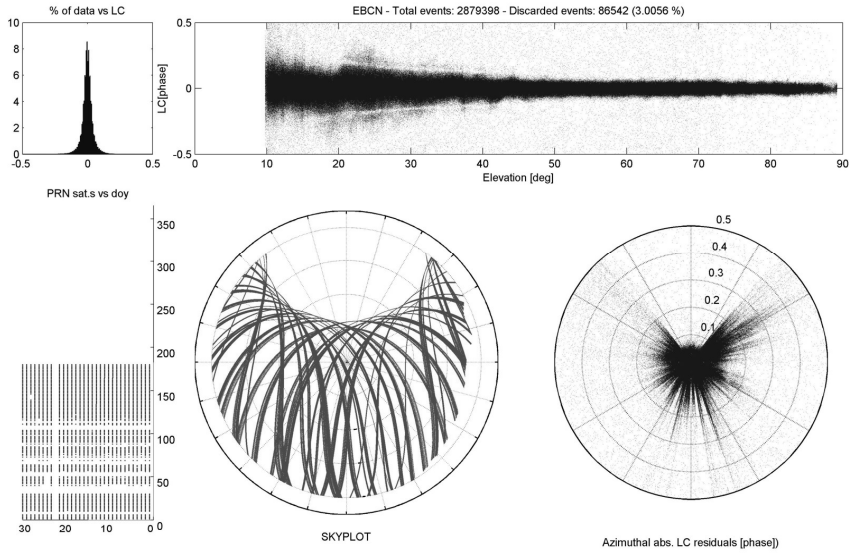


FIGURE 3.10: A synoptic view of the undifferenced LC residuals in the GPS station EBCN, from 01 January to 30 June 2012. The top left histogram represents the distribution of the LC postfit residuals expressed in percent; the top right figure shows the LC postfit residuals versus the elevation angle expressed in degrees; the bottom left represents a rolled graph containing the PRNs versus the GPS doys of 2012; the bottom center is the post elaboration skyplot, reported to the considered GPS receiver; the bottom right is a polar graph containing the absolute value of the LC postfit residuals versus the azimuthal angle expressed in degrees.

shows many problems: discontinuous data together with the presence of obstacles in the east side of station are revealed. There is also an abnormal behavior of GPS LC phase residual between 10 to 40 degrees. The map view of the residuals shows also larger values of the residues in the north-east side of station.

We produce the same figures for all stations processed. Although this is a starting point to extract additional important information from the noise data, this type of analysis was conducted for the first time during this study. The information obtained can be used in many aspects, one of which, to estimate the horizontal gradients and the atmospheric anomalies. The developed codes used in this section have shown that there are limits on the amount of data to be analyzed simultaneously. For this reason, future studies should take into account these problems, trying to overcome them. The reason that led us to move from TEQC to a MATLAB code is justified because the aggregated data were analyzed successfully.

In the future studies, analysis like those shown in this chapter, will be computed in a time varying way for each station of the network, so as to identify and improve the inadequate receivers.

Chapter 4

Volcanic plume detection by GPS signal

4.1 Introduction

Mt. Etna, Italy, is one of the most active volcanoes in the world. In the last years explosive activity usually forms weak plumes that rise some kilometers above the vent and may cause difficulties for aviation operations and the population living on the volcano's flanks. Due to the persistent activity, the Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Etneo (INGV-OE) has improved the monitoring of Etna volcanic plumes over recent years with the use of new instruments and volcanic ash dispersal models. Observations of volcanic plume dispersal and fallout are carried out by analyzing multispectral infrared measurements from the Spin Enhanced Visible and Infrared Imager on board the Meteosat Second Generation geosynchronous satellite, visual and thermal images from the video surveillance system, and data collected from a network of radar disdrometers. Furthermore, mapping of tephra deposits and prompt analysis of samples collected shortly after the end of the explosive event are also routinely carried out. Every day, an automatic procedure downloads wind data of two high-resolution meteorological models, runs these models and transfers hazard maps to the INGV-OE control room and to the Italian Civil Protection, which manages emergencies during volcanic activity in Italy. The results of the simulations are always compared with data collected from the monitoring system for validation purposes. After the Eyjafjallajokull eruption in 2010, there was much debate on the type of information that remote sensing and ground-based observations can provide during an eruption (e.g., plume height, erupted mass, particle size distribution). The main problem is that volcanic ash dispersal models needed well-constrained eruption source parameters to improve the reliability

of the forecast. A recent workshop on Ash Dispersal Forecast and Civil Aviation held in Geneva on October 18–20, 2010, produced a list of data acquisition techniques able to detect volcanic ash and furnish data on the main eruption source parameters. The list contains AIRS, ASTER, AVHRR, GOES-11, GOES- 12,13,14,15, Grimm EDM 107, Grimm Sky OPC, IASI, IMO-radar, Infrasonic Array, IR-SO2, LIDAR, MISR, MODIS, MTSAT, OMI, PLUDIX, SEVIRI, Thermal Camera, UV Camera and VOLDORAD. A Data-Acquisition Summary Document was also drawn up with the aim of describing the features for each of these methodologies. However, the capability of Global Positioning System to detect properties of a volcanic plume was not considered among these techniques, most likely because there are only a few cases using this technology with respect to such plumes.

The Etna volcano is particularly suited for an in-depth investigation into the potential of GPS observations to detect volcanic plumes, owing to both the high frequency of explosive episodes over recent years and also the well-developed GPS network.

In the following sections we will discuss two different approaches in order to examine the capability of the GPS network to detect volcanic plumes at Etna. The first approach is applied on the signal strength of the GPS carrier phase data. The GPS signals strength also called signal-to-noise ratio (SNR) data are used to detect the plume. The second approach, instead, is statistical: we analyze the single difference post fit residual of elaboration signals to assert the hypothesis that the plume affects the GPS data. The main results of the second approach have been published in a paper on a specialized journal [Aranzulla et al., 2013]

4.2 The September 4, 2007 Etna eruption test case

A spectacular lava fountain episode occurred in the afternoon of September 4, 2007, from the South East Crater, one of Mt. Etna's central craters. Although a marked increase in seismic tremor began from 14:00 GMT, the volcanic plume was visible only from 16:00 GMT by the INGV-OE video surveillance system. An eruption column rose up to 2 km above the crater rim and was dispersed by the westerly wind until 03:00 GMT on 5 September (Fig. 4.1).

This event was almost steady for the entire duration and formed a thin tephra deposit on the eastern flank of the volcano. Volcanic particles fell on the east flank and formed a clastogenic lava flow which continued for 4 km from the volcanic vent [Andronico et al., 2008]. A detailed analysis of the tephra deposit showed that the total erupted volume was 3.9–4.9 $\times$ $10^5 m^3$ and the particle size distribution had a peak at 1 mm [Andronico et al., 2008]. The volcanic plume dispersal was also observed by satellite.

FIGURE 4.1: Volcanic plume produced during the 4-5 September, 2007 lava fountaining activity at Mt. Etna, in Italy. Photo courtesy of D. Andronico; SEVIRI images of the volcanic plume at 18:00 GMT on the topleft corner.

Images showed the formation of a hot spot at about 16:00 GMT and the volcanic plume spreading eastwards which was well visible beginning from 16:45 GMT. It should be noted that during this event, forecasts of the PUFF and TEPHRA models concurred perfectly with the observations [Coltelli, 2007].
This eruption was considered as a good test case to evaluate the potential of the GPS in detecting Etna volcanic plumes mainly due to the stationary nature and prolonged duration of the event.
In the present study, we analyzed data coming from nineteen GPS receivers available during the investigated period of September 2007 (table 4.1). A map of the GPS location is shown in (Fig. 4.2).

For these observations, we used a variety of geodetic-quality (low-multipath) antennas (Table 4.1).

4.3 The Signal to Noise Ratio approach

Signal to Noise Ratio data provide no information about the distance between the satellite transmitting the signal and the receiver, and thus make no direct contribution to positioning solutions. For this reason, SNR data are generally ignored by geophysicists and geodesists. However, SNR data are important because they can directly measure

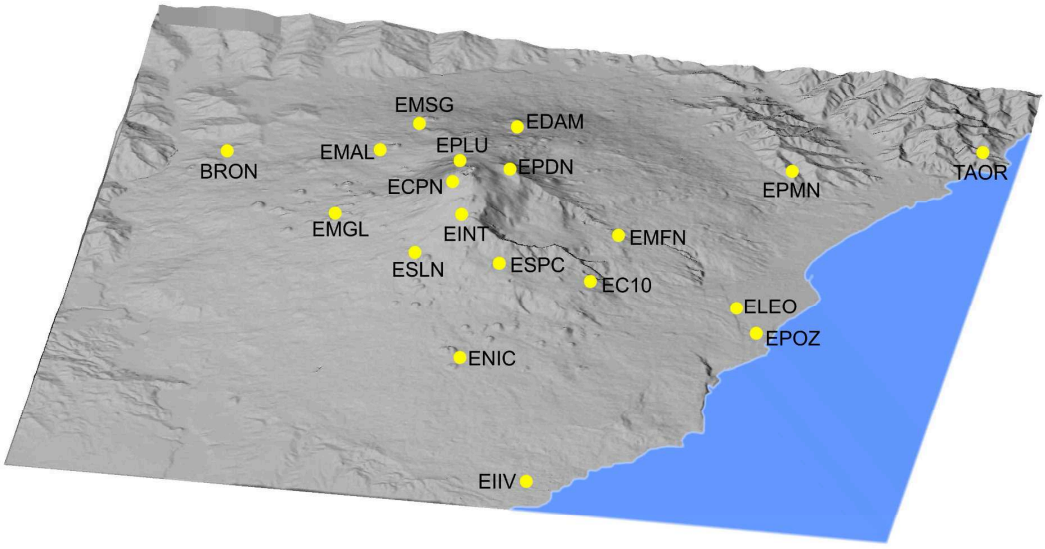


FIGURE 4.2: A map of the GPS receivers position over the Volcano on September 2007

signal blockages. If, for example, an ash laden volcanic plume crosses a GPS signal, parts of that signal are attenuated and scattered. This means that the signal that does arrive at the GPS receiver has less power than it would ordinarily. In contrast, water in the atmosphere does not significantly reduce GPS signal power. This lack of sensitivity to water is the same reason that L-band radars are being developed to observe ash clouds [Donnadieu, 2012].

GPS instruments have been rarely used to detect effects of the eruption above ground. [Houlie' et al., 2005a], [Houlie' et al., 2005b], for example, use the GPS data obtained during the eruption of Miyakejima volcano (Japan), occurred in 2000, and found anomalous values in the ionosphere-free linear combination phase measurements. It is important to note that the eruptions that commonly occur at the Etna volcano, are one tenth in magnitude, with respect to the investigated activity from [Houlie' et al., 2005a] during August 18, 2000 of Miyakejima volcano. In [Larson, 2013], instead, a method is proposed to detect volcanic plumes using GPS signal strength data. Limitations of the method are assessed using GPS data collected during the 2008 and 2009 eruptions of the Okmok and Mt. Redoubt volcanoes [Larson, 2013]. The algorithm proposed was used to detect plumes for two of the four largest explosive events from the 2009 Mt. Redoubt eruption sequence [Larson, 2013], [et al., 2012]. During the analyzed events, ash was rose up to heights of 19 and 15 km, respectively. Typically the Mount Etna's plume height reaches heights of 2 to 5 Km above vent of volcano. We analyze the opportunity to apply the signal strength method [Larson, 2013] for detecting the smaller volcanic plumes on Mount Etna. The GPS signal strength (signal-to-noise ratio, SNR) data are used. SNR data are very sensitive to large ash particles and can be modeled without estimating position, thus providing a clearer picture of a volcanic plume.

Station	Receiver type	Antenna type	Dome
BRON	LEICA GMX902	LEIAX1202GG	NONE
EC10	TRIMBLE 4700	TRM29659.00	NONE
ECPN	LEICA GX1220	LEIAX1202	NONE
EDAM	LEICA GX1220	LEIAT504	NONE
EIIV	LEICA GRX1200PRO	TRM29659.00	NONE
EINT	LEICA GX1230	LEIAX1202	NONE
ELEO	TRIMBLE 4700	TRM33429.00-GP	NONE
EMAL	LEICA GX1220	TRM29659.00	NONE
EMFN	TRIMBLE 4700	TRM29659.00	NONE
EMGL	LEICA GX1220	TRM29659.00	NONE
EMSG	LEICA GRX1200GGPRO	LEIAT504	SCIT
ENIC	LEICA GRX1200PRO	TRM29659.00	NONE
EPDN	LEICA GX122	LEIAX1202	NONE
EPLU	LEICA GX122	LEIAX1202	NONE
EPMN	LEICA GRX1200PRO	LEIAT504	SCIT
EPOZ	LEICA SR520	LEIAT504	SCIT
ESLN	LEICA GRX1200	LEIAT504	NONE
ESPC	LEICA GX1220	TRM29659.00	NONE
TAOR	LEICA GMX902	LEIAX1202	NONE

TABLE 4.1: Summary of the GPS station hardware. The name of the station, the receiver type and the antenna/radome type are shown

4.3.1 SNR and C/N_0

In addition to carrier phase and code observables, SNR data are routinely recorded by GPS receivers but are less frequently reported in RINEX data files [Gurtner and Mader, 1990]. When available in RINEX, these data are reported as observable types S1 and S2, and recorded as raw signal strengths or SNR values by the receiver, for L1 and L2 phase observations. The RINEX translation software TEQC [Estey and Meertens, 1999] now reports SNR in dB (power or amplitude) for the majority of common GPS geodetic receivers. Signal-to-noise ratio is a measure used in science and engineering that compares the level of a desired signal to the level of background noise. GPS receiver manufacturers primarily report this quantity assuming a 1 Hz bandwidth, or dB-Hz. It is defined as the ratio of signal power to the noise power in a given bandwidth, usually expressed in decibels. A ratio higher than 1:1 indicates more signal than noise.

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{4.1}$$

where P_{signal} is the average signal power and P_{noise} is the average noise power in a given bandwidth. Because many signals have a very wide dynamic range, SNRs are often

expressed using the logarithmic decibel scale. In decibels, the SNR is defined as

$$SNR[dB] = 10log_{10}\frac{P_{signal}}{P_{noise}} \tag{4.2}$$

According to the equation 4.2, value of SNR greater then zero indicates more signal than noise. Power can be expressed in decibels by forming the ratio of the power to a reference power level. Typical reference levels are one watt and one milliwatt. A power level in decibel-milliwatts can be computed from a power expressed in milliwatts as:

$$\begin{aligned} P[dBW] = 10log_{10}\left(\frac{P[W]}{1[W]}\right) = 10log_{10}\left(\frac{P[W]}{1000[mW]}\right) = 10log_{10}\left(\frac{1}{1000}\frac{P[W]}{1[mW]}\right) = \\ = 10log_{10}\left(\frac{1}{1000}\right) + 10log_{10}\left(\frac{P[W]}{1[mW]}\right) = -30 + 10log_{10}\left(\frac{P[W]}{1[mW]}\right) = P[dBm] - 30 \end{aligned} \tag{4.3}$$

If the unit of power are already in common decibel units, such as dBW or dBm, then according to eq.s 4.1 and 4.3

$$\begin{aligned} SNR = 10log_{10}\frac{P_{signal}}{P_{noise}} = 10log_{10}P_{signal} - 10log_{10}P_{noise} = \\ = S[dBm] - N[dBm] = S[dBW] - N[dBW] \end{aligned} \tag{4.4}$$

where S and N are expressed in units of decibel/milliwatt (dBm) or decibel/watts (dBW). The carrier-to-noise density (C/N_0), on the other hand, is usually expressed in decibel-Hertz (dB-Hz). For the GPS L1 C/A signal it is defined as the power of the original unmodulated carrier power divided by the noise power spectral density. Since noise sources like thermal noise generate power in proportion to the bandwidth of the system in question, a method of describing the power level independent of the bandwidth is desirable. Power spectral density is a measure of power in each unit of bandwidth. Thermal noise has a constant power spectral density. The power of thermal noise generated is a function of the temperature and the noise bandwidth. It is independent of the center frequency of that bandwidth. The noise power spectral density for noise is kT, where k is Boltzmann's constant and T is the absolute temperature. Boltzmann's constant is the ratio of the energy in a molecule to its temperature and is equal to $k = 1.38 \times 10^{-23} J/K$, expressed in the units of joules per Kelvin degrees.

The sources of white noise in a GNSS receiver are usually described by the antenna noise temperature and the receiver noise temperature. The antenna temperature models the

noise entering the antenna from the sky whereas the receiver noise temperature models the thermal noise due to the motion of charges within a device such as the GPS receiver front-end. Ambient thermal noise is typically calculated at 290K, this is a reference generally taken as the effective noise temperature of the Earth:

$$N_T = kT = 1.38 \times 10^{-23} \frac{[J]}{[K]} \cdot 290[K] = 4.00 \times 10^{-21}[J] \tag{4.5}$$

Since a watt is one joule-per-second, we can also express the units for power spectral density as watts-per-hertz. Ambient thermal noise power spectral density is then:

$$N_T = 10log_{10}\left(4.00 \times 10^{-21} \frac{[W]}{[Hz]}\right) = -204\frac{[dBW]}{[Hz]} = -174\frac{[dBmW]}{[Hz]} \tag{4.6}$$

To calculate the L1 C/A carrier-to-noise density for a GPS receiver operating at the thermal noise floor ($N_0 = N_T$) we need the carrier power. The C/A code GPS signal specification give us a nominal value for the carrier power received at the surface of the Earth, specifying this to be at a power level of –160 dBW or above.
Using the eq. 4.4, the carrier power and the thermal noise floor we compute:

$$\begin{aligned} C/N_0[(dB - Hz] = C - (N - BW) = C - N_0 = SNR + BW = \\ = -160[dBW] - (-204[dBW]/[Hz]) = 44[dB][Hz] \end{aligned} \tag{4.7}$$

Where C is the carrier power in dBm or dBW, N is the noise power in dBm or dBW, N_0 is the noise power density in dBm-Hz or dBW-Hz, BW is the bandwidth of observation, which is usually the noise equivalent bandwidth of the last filter stage in a receiver's RF front-end. Converting the value of 4.7 to a signal-to-noise ratio for the same conditions, assuming a C/A code receiver bandwidth of 4 MHz ($BW = 10log_{10}(4 \cdot 10^6)[Hz+] = 66[dB][Hz]$).

$$SNR[dB] = \frac{C}{N_0} - BW = 44[dB][Hz] - 66[dB][Hz] = -22[dB][Hz] \tag{4.8}$$

The signal power is 22 dB below the noise power coming into the receiver. It is only through despreading the C/A code that the signal can be detected at all. After the signal is despread, it filtered to a narrower bandwidth. The decrease in bandwidth eliminates most of the noise, which is spread over the entire bandwidth, but leaves the signal. This is the mechanism by which process gain is achieved in a spread-spectrum receiver. A more complete description of process gain can be found in [Dixon, 1984].

4.3.2 GPS signal strength

The description of the GPS signal structure is explained in chapter 1. Figure 4.3 shows a track of SNR for two satellites (PRN13 and PRN29) of block II-R and II-RM (Table 1.2 cap1). The top panel highlights two satellite tracks for L1 (1.5754 GHz) SNR using the public C/A code. The bottom panel shows L2 (1.2276 GHz) SNR data using two different codes, L2C (which is also public but present only in the blocks II-RM and II-F) and the encrypted L2P code.

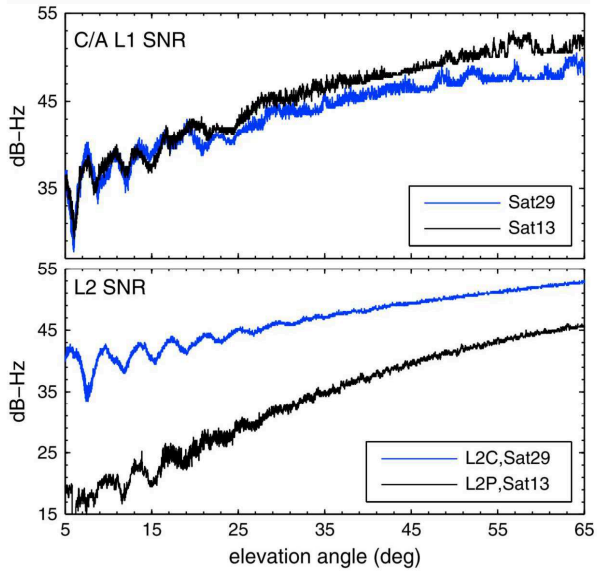


FIGURE 4.3: (top) L1 SNR data derived from the C/A code. Satellite 29 is offset by 2 dB-Hz to highlight the data for large elevation angles. (bottom) L2 SNR data using L2C (satellite 29) and L2P (satellite 13) [Larson, 2013].

Both panels show SNR values slowly increasing as the satellite rises from 5° to 65° in elevation angle. This is known as the direct signal effect. The slow increase in SNR is primarily due to the antenna gain pattern. The L1 SNR data above 25° have high-frequency noise that is both systematic and random. In contrast, the high-elevation L2 SNR data have much smaller levels of high-frequency noise. This difference is due to the fact that the C/A code is much shorter than L2P or L2C and thus suffers from cross-channel interference. However, L2P has much lower SNR values than C/A or L2C because the receiver cannot use the encrypted code in its retrieval. As a final comment, L2C is only available on satellites launched after 2005 Sec. 1.2.1. Furthermore, it is often not tracked unless the user requests it. The oscillations you see in SNR data at elevation angles of 25° are caused by ground reflections. A reflected signal travels a longer distance than the direct signal and interferes with the direct signal; this causes the observed modulations. The frequencies in the SNR data below 25° can be related to soil moisture content, snow depth, and sea level height [Larson et al., 2008], [Larson et al., 2009], [Larson, 2013]. Here only the higher elevation angle SNR data, of interest

for plume sensing, will be used. Because it has better precision at higher elevation angles, the L2P SNR data are used to isolate the effect of plumes. Figure 4.5 shows an example of L2 SNR data for EIIV station at Mt.Etna.

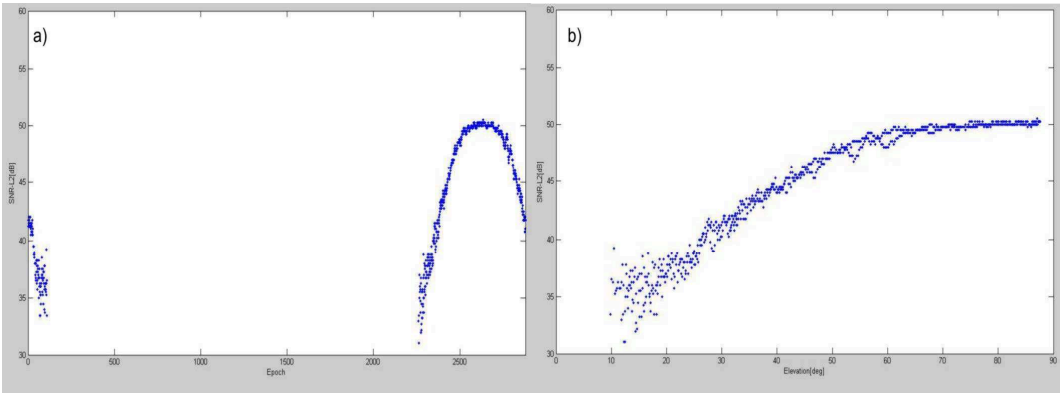


FIGURE 4.4: (An example of L2-SNR data for EIIV station at Mt.Etna. a) The L2-SNR values expressed in [dB-Hz] vs. the time expressed in epochs (1 epoch is equal to 30 seconds). b) The L2-SNR values expressed in [dB-Hz] vs. satellite elevation angle expressed in degrees

As expected, the SNR data vary smoothly by $\approx 30dB - Hz$ over the satellite arc.

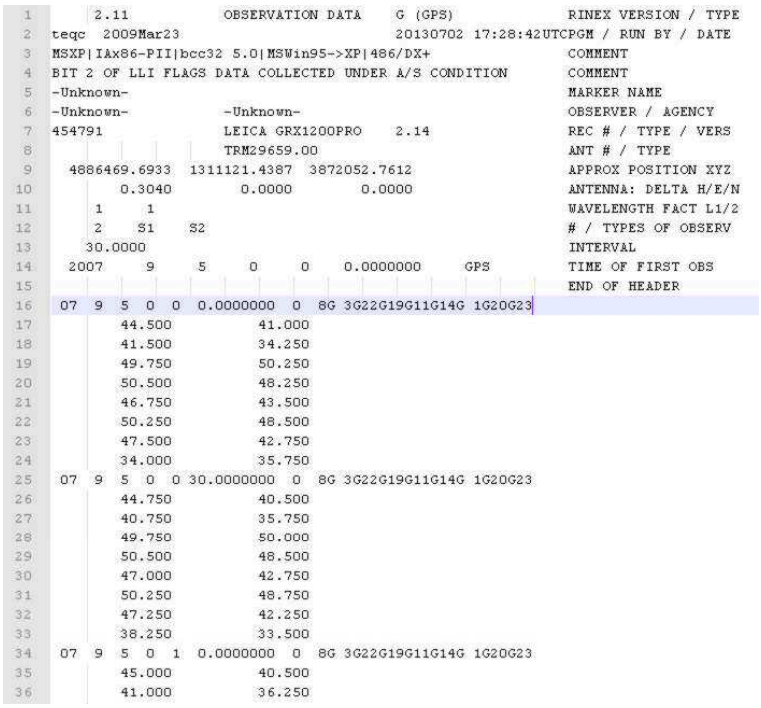
4.3.3 SNR data extraction

As said previously, the SNR is recorded in the binary data (RAW) by the GPS receiver. By using the TEQC software [Estey and Meertens, 1999], developed by UNAVCO, the RAW data from native binary format to Rinex format (for details see chapter 3) is translated. Starting from the RAW data recorded by the Mount Etna network during the 4-5 September (fig. 4.2), trough a code developed in MATLAB, we read raw files and generate three files for each GPS receiver and day of year (doy): a Rinex files containing SNR data, an azimuth file and an elevation file. The core of this code is the TEQC command, useful to obtain the three data file (table 4.2). For example, if we want to obtain data from the receiver called "EIIV", the command is:

teqc -O.dec 30 -O.obs S1S2 -R binary_input_EIIV > RINEX_observation_EIIV

Where we give the desired decimation interval in seconds (-O.dec 30), with the explicit list of the desired observables and order (-O.obs S1S2). The option -R allow us to eliminate the GLONASS data. The figure 4.5 shows an example of the created Rinex file containing the SNR1 and SNR2 data.

As explain in sec. 3.3.1 each file consists of a header section and a data section. Just to remember, the data that have to be ingested in the MATLAB workspace begin from line 16 of figure 4.10. The line 16 represents the time of the acquisition and the satellites



```
     2.11           OBSERVATION DATA    G (GPS)             RINEX VERSION / TYPE
teqc  2009Mar23                        20130702 17:28:42UTCPGM / RUN BY / DATE
MSXP|IAx86-PII|bcc32 5.0|MSWin95->XP|486/DX+                COMMENT
BIT 2 OF LLI FLAGS DATA COLLECTED UNDER A/S CONDITION        COMMENT
-Unknown-                                                    MARKER NAME
-Unknown-           -Unknown-                                OBSERVER / AGENCY
454791              LEICA GRX1200PRO    2.14                 REC # / TYPE / VERS
                    TRM29659.00                              ANT # / TYPE
  4886469.6933  1311121.4387  3872052.7612                   APPROX POSITION XYZ
        0.3040        0.0000        0.0000                   ANTENNA: DELTA H/E/N
     1     1                                                 WAVELENGTH FACT L1/2
     2    S1    S2                                          # / TYPES OF OBSERV
    30.0000                                                  INTERVAL
  2007     9     5     0     0    0.0000000     GPS          TIME OF FIRST OBS
                                                             END OF HEADER
 07  9  5  0  0  0.0000000  0  8G 3G22G19G11G14G 1G20G23
        44.500          41.000
        41.500          34.250
        49.750          50.250
        50.500          48.250
        46.750          43.500
        50.250          48.500
        47.500          42.750
        34.000          35.750
 07  9  5  0  0 30.0000000  0  8G 3G22G19G11G14G 1G20G23
        44.750          40.500
        40.750          35.750
        49.750          50.000
        50.500          48.500
        47.000          42.750
        50.250          48.750
        47.250          42.250
        38.250          33.500
 07  9  5  0  1  0.0000000  0  8G 3G22G19G11G14G 1G20G23
        45.000          40.500
        41.000          36.250
```

FIGURE 4.5: Rinex file example containing the SNR data for each epoch

to which the receiver is linked. The lines from 17 to 24, represent the values of SNR1 and SNR2 for each linked satellite, ordered according to line 16. The line 25 repeats the same explained scheme, but referred to the next epoch. As each Rinex file contains 2880 epochs per day and each receiver can simultaneously link 7-9 GPS satellites per epoch, the ascii file consists of about 24000 lines per day. So it was necessary to implement the MATLAB routine to read all files. Since other files, as the azimuth and elevation files, have different standard of tabulation but the same excessive samples problem it was mandatory automating all the processes by implementing a MATLAB codes. This allow to study all the cases where an evidence of plume dispersal in atmosphere in the future is present.

4.3.4 Methods

The following example reports the method used to isolate the plume effect on the signal. No data below elevation angles of 30° are used, as these data are both less precise and more likely to be impacted by ground reflections as shown in figure 4.6. Data on the 3 days before and after the eruptive event are shifted by 4 min/d to account for the repeating satellite geometry, to taking in account the sideral day. Figure 4.6 shows the SNR values referred to the station EDAM and satellite 11, for all the 7 studied days. The repeatability of the SNR data during the seven days of the plume event relative to the pair EDAM-PRN11 is evident. After the cut-off filtering of 30°, the data are averaged.

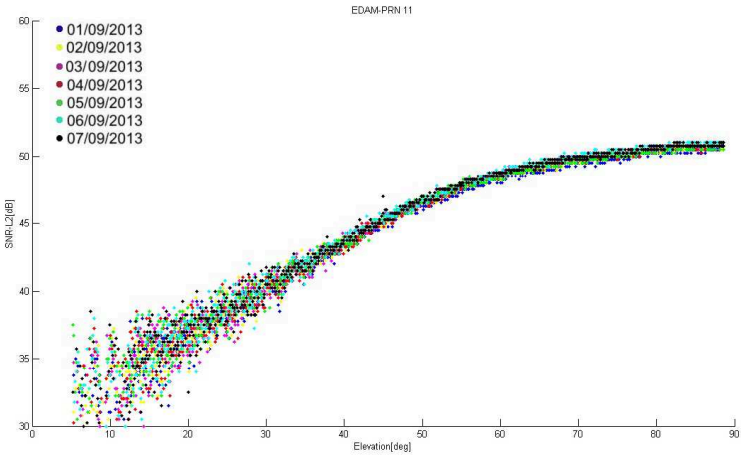


FIGURE 4.6: SNR data relative to EDAM-PRN11 during the days in the neighborhood of the plume emission.

The average is calculated as the mean value of the six days (three before and three after) the plume event. The obtained values are then smoothed over 10 points. In figure 4.7 for the pair EDAM-PRN11, the black line represents the averaged and smoothed data. Examples reported in figures 4.6 and 4.7 show that there is a high repeatability of the

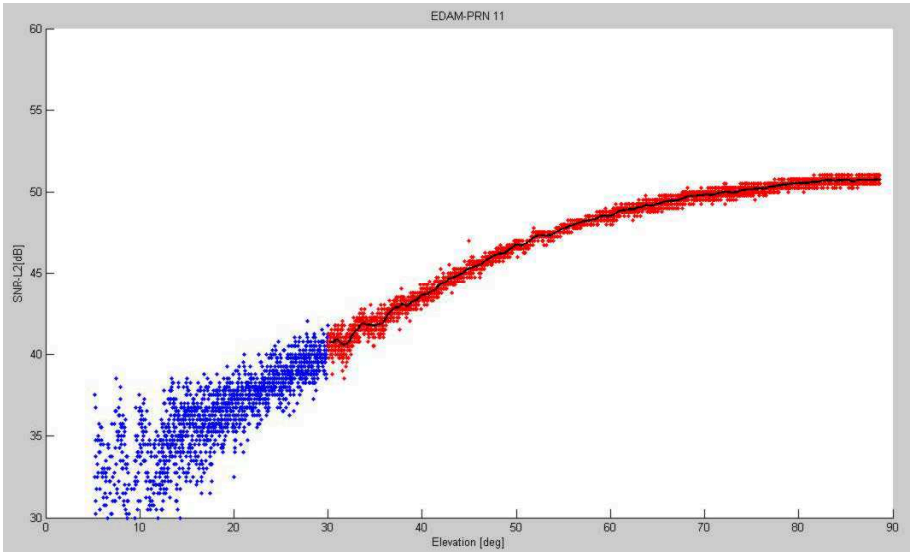


FIGURE 4.7: An elevation cut off angle of 30° is applied. The example of SNR data referred to the EDAM station and satellite 11, in blue the data before the cut off, in red the filtered data. The black curve represents the smoothed values.

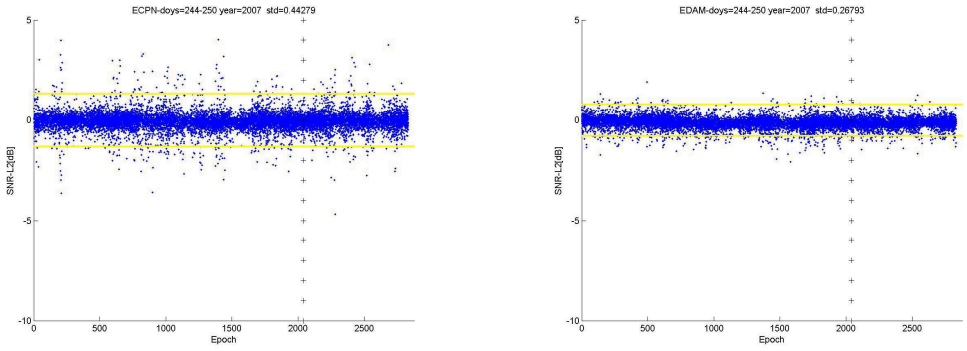
SNR vs. elevation angle data.

This observation suggests that, if a perturbation like a plume occurs, it should be revealed trough an SNR analysis.

However the figure 4.6 doesn't give information about the epoch in which an anomaly occurs. For this and other reasons that will be explained in the next section, in the following figures epoch vs. SNR data will be plotted.

4.3.5 Results

Table 4.2 shows the results of the RAW files conversion. The elevation and azimuth files referred to BRON, EC10, ELEO and TAOR stations are missed, preventing the use of these stations in the following analysis. The smoothed SNR model, which represents the expected values for a given station and satellite at a given time, is then subtracted from the SNR data collected on the day of the eruption. This SNR change is what it is used to detect the presence of plumes. The standard deviation of the SNR data for the 30 min before each eruption is calculated to provide a quality control measure, and the detections threshold is fixed to be 2.5 times this standard deviation. If no SNR data are available before the eruption, the satellite track is discarded. Finally, no detection is reported (regardless of its standard deviation), unless its value exceeds 2 dB-Hz, in order to avoid reporting false detections. The figures from 4.8 to 4.15 show the obtained results for all the 15 available stations.



(a) SNR-ECPN (b) SNR-EDAM

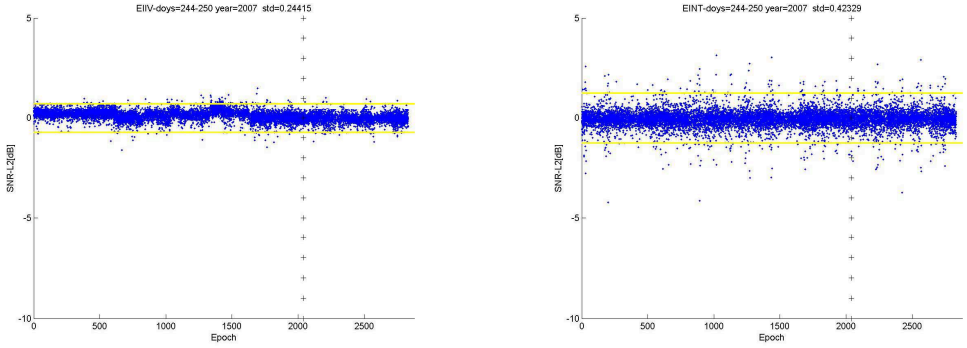
FIGURE 4.8: SNR data collected on the day of the eruption minus the smoothed SNR model referred to ECPN and EDAM stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.

4.3.6 Discussion

Despite this method has been applied to detect effects of volcanic plumes in eruptions of greater magnitude than of Mt. Etna [Larson, 2013], the figures from 4.8 to 4.15 to do not show clear evidence of the plume presence during the 4 September 2007 event.
For EMFN and ENIC stations respectively, figures 4.10b and 4.12a, indeed, show high discrepancy from the common behavior SNR value. However the EMFN anomalies occurs from epoch 1000 to 1500, and the ENIC anomalies occurs between 750 and 1700, and between 2000-2200, while the eruption starts after the epoch 2000. Thus these anomalies are not related to the volcanic plume, although its presence in the figure 4.1

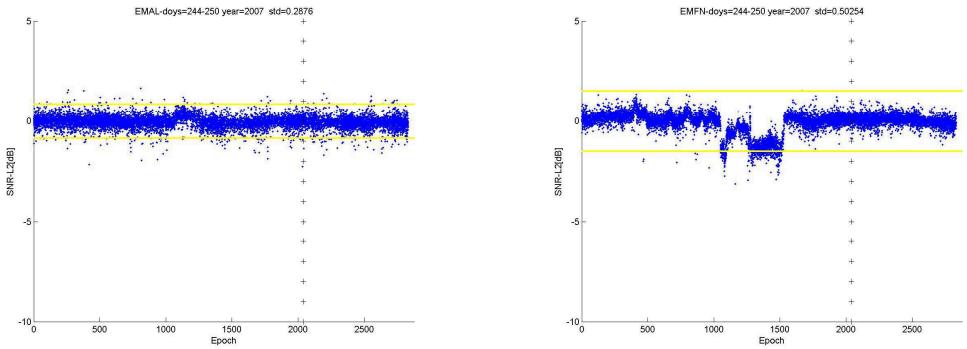
	244 (01/09/2007)			245 (02/09/2007)			246 (03/09/2007)			247 (04/09/2007)			248 (05/09/2007)			249 (06/09/2007)			250 (07/09/2007)		
	S1S2	ele	azi	S1S2	ele	azi	S1S2	ele	azi	S1S2	ele	azi	S1S2	ele	azi	S1S2	ele	azi	S1S2	ele	azi
BRON	.			.			.			.			.			.			.		
EC10	.			.			.			.			.			.			.		
ECPN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EDAM	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EIIV	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EINT	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ELEO	.			.			.			.			.			.			.		
EMAL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EMFN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EMGL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EMSG	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ENIC	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EPDN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EPLU	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EPMN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EPOZ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ESLN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ESPC																					
TAOR	.			.			.			.			.			.			.		

TABLE 4.2: Results of the RAW files conversion into Rinex, azimuth and elevation files for 4 Sept. 2007 event.



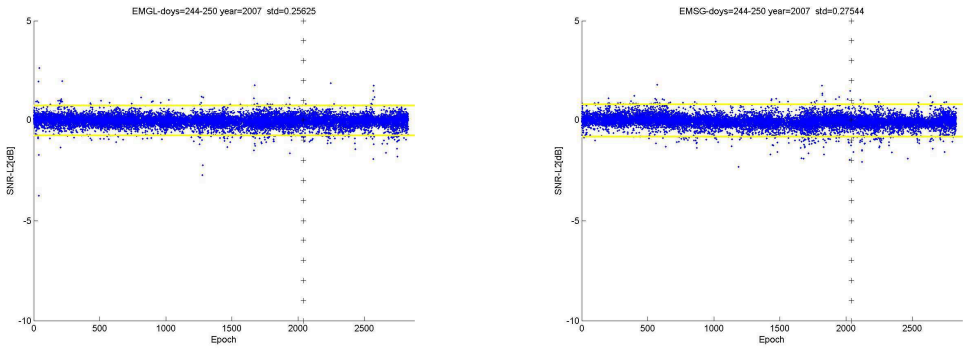
(a) SNR-EIIV (b) SNR-EINT

FIGURE 4.9: SNR data collected on the day of the eruption minus the smoothed SNR model referred to EIIV and EINT stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.



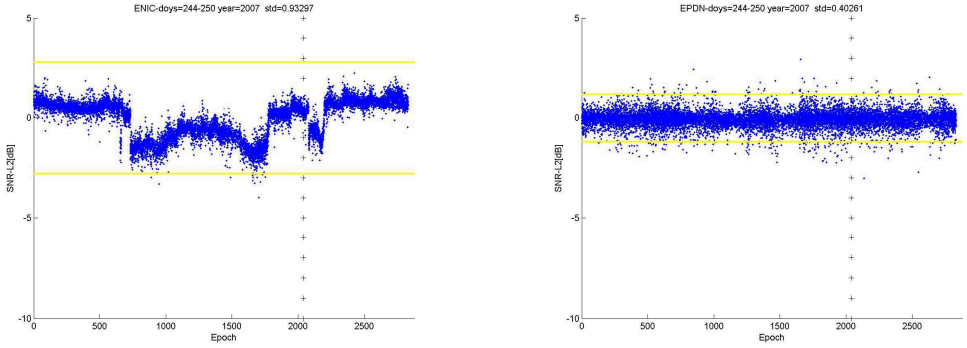
(a) SNR-EMAL (b) SNR-EMFN

FIGURE 4.10: SNR data collected on the day of the eruption minus the smoothed SNR model referred to EMAL and EMFN stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.



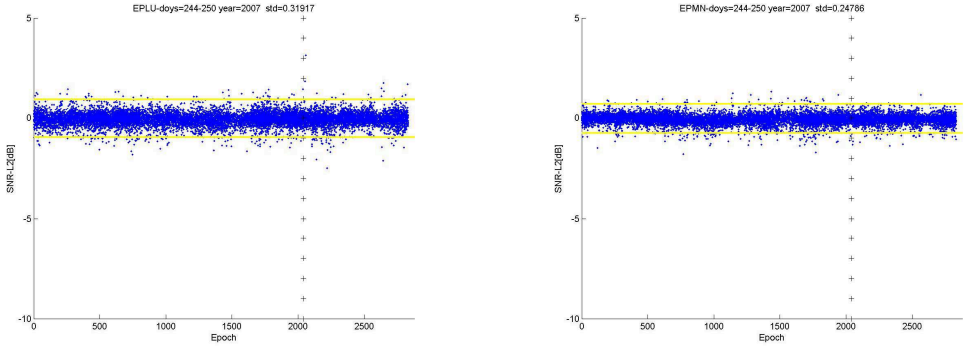
(a) SNR-EMGL (b) SNR-EMSG

FIGURE 4.11: SNR data collected on the day of the eruption minus the smoothed SNR model referred to EMGL and EMSG stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.



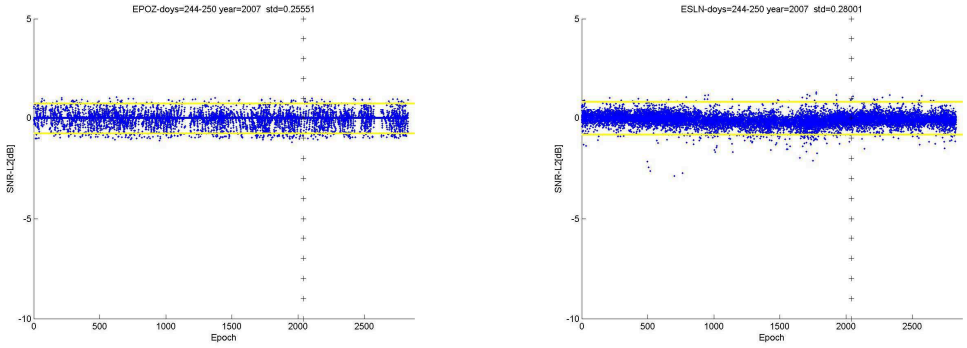
(a) SNR-ENIC (b) SNR-EPDN

FIGURE 4.12: SNR data collected on the day of the eruption minus the smoothed SNR model referred to ENIC and EPDN stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.



(a) SNR-EPLU (b) SNR-EPMN

FIGURE 4.13: SNR data collected on the day of the eruption minus the smoothed SNR model referred to EPLU and EPMN stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.



(a) SNR-EPOZ (b) SNR-ESLN

FIGURE 4.14: SNR data collected on the day of the eruption minus the smoothed SNR model referred to EPOZ and ESLN stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.

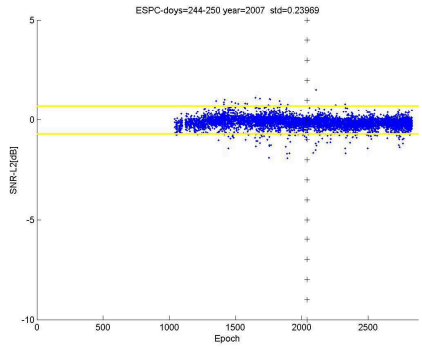


FIGURE 4.15: SNR data collected on the day of the eruption minus the smoothed SNR model referred to ESPC and stations. The horizontal yellow lines represent the value of $\pm 2.5\sigma$, the vertical black crosses represent the begin of the plume emission.

is well evident.

At this point, we can't say anything about these anomalies, only underline their existence. In the future we would like to apply this algorithm to all the events after 2007, since the number of GPS receivers increased from 19 to 35 during the last 6 years; the higher number of data might improve the quality of the results. Because the Mount Etna plume of 4 September was smaller than the eruptions studied by Larson in [Larson, 2013], perhaps a finer understanding of the interactions and a lower noise level of stations , will help us to achieve better results on Mount Etna.

4.4 Statistical approach

Using the results of the previous method, we pass to present the results of a different approach to the detection of the volcanic plume. The same September 4, 2007 activity was used as a test case. As we explained in Chapter 2 , water vapor, hydrometeors and particulates induces propagation delays in the signal path. Volcanic plumes are largely made up of volcanic ash, water vapor, volcanic gas (CO_2 , NO_2 , SO_2) and hydrometeors [Oppenheimer et al., 2003]. As explined in previous chapters, the main problem with modeling the water vapor is that it is not in hydrostatic equilibrium and therefore is poor modeled by surface measurements. When delays cannot be modeled or measured, they are considered as errors in the GPS position solution [Meertens et al., 1997]. After estimating station positions and atmospheric parameters from the doubly differenced phase, GAMIT can produce residuals for the undifferenced phases. We use these undifferenced post-fit phase residuals as input in a testing procedure for the presence of a volcanic plume.

The LC signal, which is combined from the two GPS signal carriers, is more strongly affected than L1 and L2 carriers [Solheim et al., 1999]. This is the reason why we choice to investigate the LC signal. The first approach was similar of previous section, soon superseded by a statistical method, more reliable because it is not affected by errors introduced by the interpretation.

In order to highlight the presence of volcanic plume in the atmosphere from the GPS signals, we follow two steps: firstly, we analyze the GPS data collected by the Etna network in order to obtain the GPS post-fit phase residuals ("Instrumentation and GPS analysis" section); secondly, we use a statistical approach to investigate whether the plume can be detected in post-fit phase residuals ("'The statistical approach" section).

4.4.1 Methods

Instrumentation and GPS analysis

In the present study we analyze the same September 4, 2007 activity. The difference is that, with this approach, we process 30 days of data, between August 5 (day-of-year 217) and September 4 (day-of-year 247), 2007. The RINEX files containing the observables were created by using TEQC translator.

For these observations, we used a variety of geodetic-quality (low-multipath) antennas (table 4.1), and so in our processing elevation-dependent models for the antenna phase centers have been considered [Schmid et al., 2005].

The GPS data coming from the Etna network (figure 4.2 and table 4.2) were processed using the GAMIT package developed by Massachusetts Institute of Technology [Herring

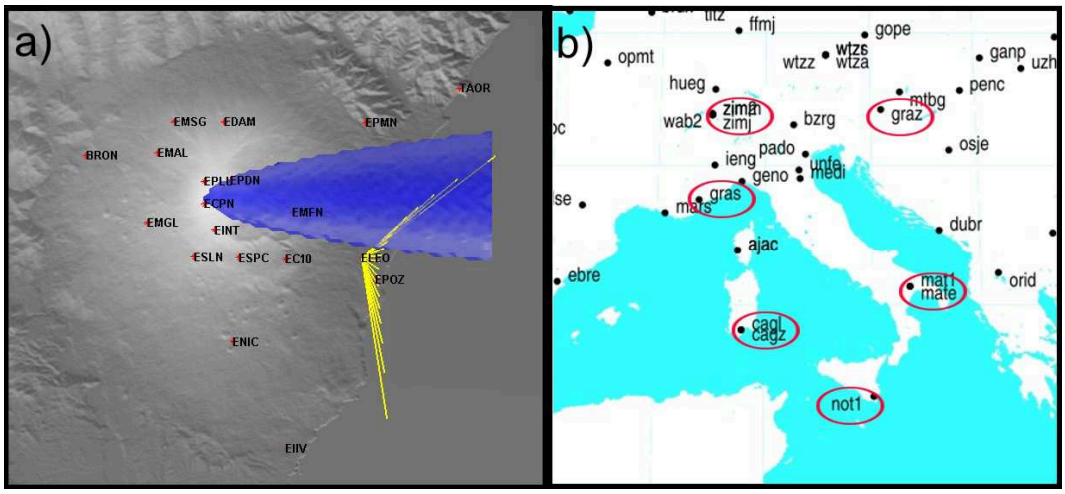


FIGURE 4.16: a) Map of the GPS receivers on the volcano together with the simplified model of the volcanic plume for the September 4-5, 2007 event (in blue). Red Stars represent the GPS receivers. The yellow lines join the satellite and ELEO antenna. b) Map of the IGS stations used as stable sites in this work

et al., 2010] (see chapter 3). More details of the parameters used during the elaboration are reported in section 3.3.3. In order to evaluate the quality of the GPS data analysis, the repeatability of the daily station positions and comparison with ITRF2005 were carried out. For the same IGS station (CAGL), the daily position computed by GAMIT is compared with values downloaded from the SOPAC website (http://sopac.ucsd.edu) (fig. 4.17). The figure shows the good agreement between the two different datasets. Table 4.3 shows instead the RMS of the position repeatability of all the stations in the analyzed network. Figure 4.18 shows instead data from three GPS stations located at different heights (EIIV at 88 m, EMAL at 1551 m and EPLU at 2964m) that clearly highlights the good repeatability of the results of the daily INGV-OE monitoring elaboration.

The statistical approach

The analysis of the GPS residuals was focused on understanding if the plume may affect the GPS elaboration in the residual component of the Etna network and to quantify, if it is the case, the relation between the presence of a volcanic plume and its effect on the GPS residual. In our analysis we considered the residual signals as coming from a stochastic process. Figure 4.19 shows the steps followed in our statistical analysis.

For a better understanding of the algorithm, we divided the process in 4 steps.

Step 1 is the construction of a simplified model of the volcanic plume for September 4, 2007 event. Supported by the satellite images of the volcanic plume (fig. 4.1), we approximate the plume geometry with a paraboloid (fig. 4.16a) represented by the equation:

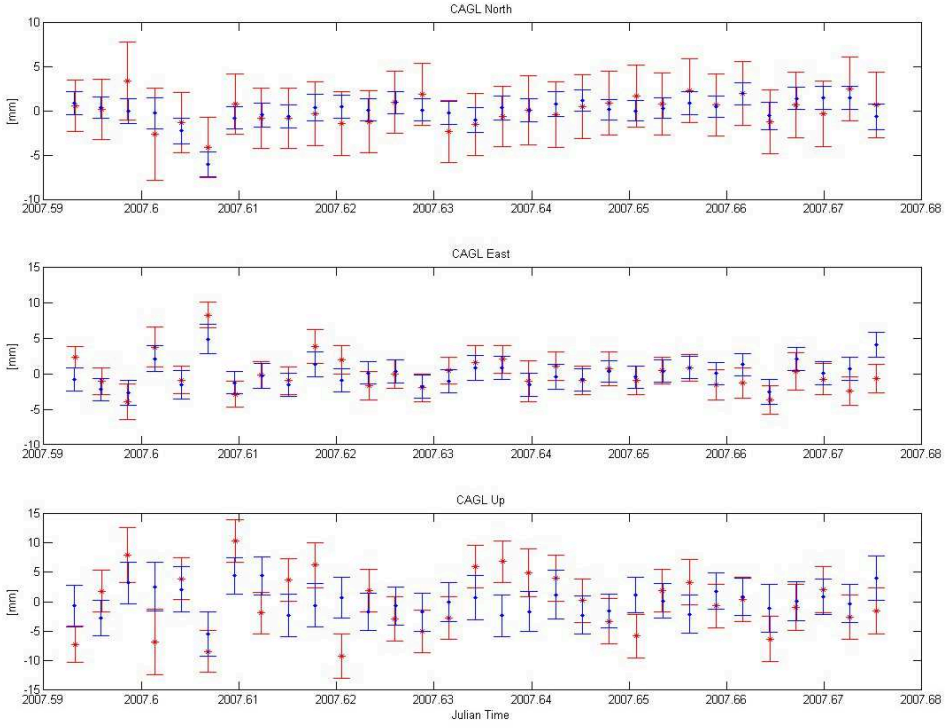


FIGURE 4.17: Comparison of the calculated points from INGV-OE (blue) with data available from SOPAC (red) (http://sopac.ucsd.edu), expressed in the International Terrestrial Reference Frame ITRF2005.

$$x = \frac{y^2}{a^2} + \frac{z^2}{b^2} \tag{4.9}$$

where the origin is the volcanic vent, x axis (m) is toward E in accordance with the main wind direction during the eruption, y axis (m) is toward N as the crosswind direction from the emission plume centerline and z axis (m) is normal to the $x - y$ plane. The a and b coefficients were chosen empirically in order to fit the available satellite images of the volcanic plume. In order to take into account the plume dispersion along the plume axis, we adopted an exponential density decay using a similar law found in the tephra deposit [Andronico et al., 2008].

In **Step 2** we integrate the positions of satellites and receivers with the plume model. For each epoch, we calculate the intersection points P_1 and P_2 between the link Antenna-Satellite and the computed volcanic plume region (fig. 4.16a).

In **Step 3**, we calculated the total amount of plume (normalized to the emission rate) for the segment P_1P_2 by the following equation:

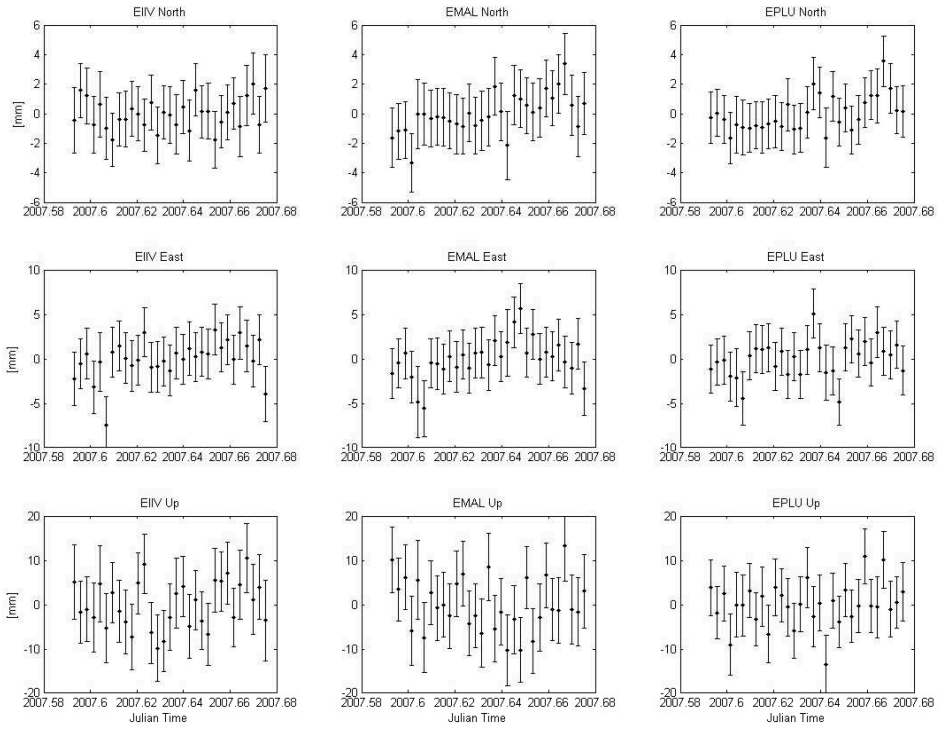


FIGURE 4.18: Daily position of three INGV-OE stations: EIIV at 88 m, EMAL at 1551 m and EPLU at 2964 m. x axis is in Julian date.

$$C_{AS}(T) = \int_{p_1}^{p_2} e^{-\frac{x}{\tau}} ds \tag{4.10}$$

where $C_{AS}(T)$ is the total relative amount plume for the pair antenna A and satellite S at the epoch T, τ_x is the decay constant of the plume concentration along the downwind axis. According to the deposit and satellite images, the constant τ_x was empirically set to 5 km. The computation of volcanic plume dispersion was useful in order to find the most likely epochs in which the link between Antenna and Satellite crossed the plume with high probability. Indeed, we chose a threshold on $C_{AS}(T)$ for selecting the crossing-epochs. The threshold was set empirically to 0.1 to discard ambiguities in the crossing of the plume due to the model approximation. For each SAP, we found the Plume Crossing Period (PCP) on the day of the eruption.

In **Step 4** we evaluated the statistical tests. In order to stress the effect of the plume on the residual, we applied a high-pass filter on the time series of residual r(t) which approximates the first derivative as:

$$dr(t) = r(t) - r(t-1) \tag{4.11}$$

Station	North	East	Up
BRON	1.3	1.9	5.2
CAGL	1.4	1.7	2.4
EC10	1.7	3.8	10.4
ECPN	1.3	1.8	3.8
EDAM	1.5	2.5	4.1
EIIV	1.0	2.2	5.3
EINT	1.9	2.4	4.8
ELEO	1.8	3.2	7.2
EMAL	1.3	2.2	6.1
EMFN	1.8	6.5	10.7
EMGL	1.8	1.8	6.6
EMSG	1.2	1.3	5.6
ENIC	1.2	1.7	5.8
EPDN	1.5	2.4	4.1
EPLU	1.2	2.1	4.7
EPMN	1.7	3.0	6.6
EPOZ	1.6	2.7	6.9
ESLN	1.4	2.5	4.9
ESPC	2.4	3.3	7.9
GRAS	1.4	1.0	2.9
GRAZ	0.9	1.3	2.4
MATE	1.1	1.2	2.3
NOT1	1.1	1.6	7.1
TAOR	2.0	2.2	5.5
ZIMM	1.2	1.0	2.2

TABLE 4.3: The millimeter weighted RMS of the position repeatability of all the stations in the network analyzed. The name of the GPS Station and the North, East and Up components are shown.

If PCP was long enough (more than 80 epochs), in order to detect signal feature variations, we estimated the energy of the extracted signal α of the derivative time series $dr(t)$ by integrating the squared series. we then calculated α for the same periods in the 30 days before the eruption, taking into account the sidereal day to be sure that the antennas pointed at the same satellite constellation. we found that the distribution of α could not be assimilated to a normal distribution using the Shapiro-Wilk test [Shapiro and Wilk, 1965]. The probability distribution of α, was hence empirically approximated by a Chi-squared (χ^2) distribution. Later, for the set of all the energies of each SAP, we found the best fit of χ^2 using the maximum Likelihood approach described in [Hahn and Shapiro, 1994]. For each SAP, the estimated χ^2 characterizes the distribution of the energy of the GPS residual noise in the considered PCP on days without the volcanic plume. Under the hypothesis of χ^2 distribution on the calculated energies, we evaluated the probability p that the energy α at the eruption day exceeds a threshold β by

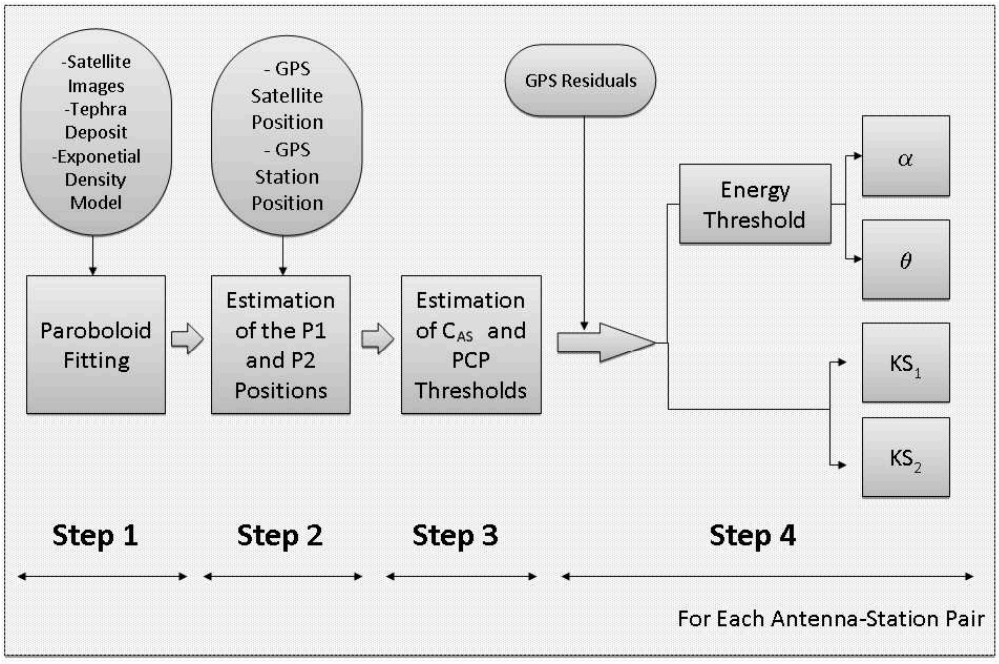


FIGURE 4.19: Flow chart of the plume detection statistical approach

calculating:

$$p = 1 - \int_0^{\beta} \chi^2 dx \tag{4.12}$$

We were interested in finding the threshold β above which there is a probability p to find values of α. For our purposes, taking into account the possible uncertainties on the parameters, we chose to fix a strict value of $p = 1\%$. In this case, the threshold was calculated for each estimation of χ^2. We consider that if α belongs to the no-plume distribution χ^2, there is a probability of 1% to find $\alpha > \beta$. In our study We considered the coefficient $\rho = \alpha/\beta$. In this way, We could expect that for SAP with $\rho > 1$, the plume affected the residues with a probability of 99%. We performed the same analysis on different periods of the same length in duration of PCP, for the same day but at times that did not coincide with the eruption. Therefore, having a set of $\{\rho_T\}$, We were able to calculate another index of deviation:

$$Y_t = \frac{\rho_T - \mu_p}{\sigma_p} \tag{4.13}$$

where ρ_T is the ρ coefficient calculated during the period T in the day of the eruption,μ_p and σ_p are, respectively, the mean value and the standard deviation of the set of ρ_T

indexes on the same day. We evaluated this statistic Y_T only if the number of useful periods was > 6. Given the few samples of the population, we chose to use a robust estimation of standard deviation calculating the midspread Δ and,using the definition of midspread, approximating $\sigma = 0.7413\Delta$. Under the hypothesis of Y_T normally distributed $N(0;1)$ [McPherson, 1990], we can assert that there is a probability of 1 % that

$$Y_{PCP} > \left\{ x : \int_{-\inf}^{x} N(0;1)dx = 0.99 \right\} = 2.32 \tag{4.14}$$

Also, in this case, we defined the coefficient $\theta \cong Y_{PCP}/2.32$, thus for with $\theta > 1$, the plume affects the residuals with a probability of 99 %. Furthermore, in order to assess the anomalies in the time series of residuals during the plume event, we carried out another statistical test. we used the two-sample Kolmogorov-Smirnov (KS) test [Massey, 1951] to compare the distributions of the plain residuals during the PCP at the eruption day and at previous quiet days. This test checks, as null hypothesis, whether the two data samples come from the same distribution by comparing their empirical distribution functions. The test is suitable for our case because it does not specify what the common distribution is (e.g., normal or not normal). we chose a significance level for the tests of 1 %. For each SAP, we performed the K–S test in two different ways:

- The distribution of the residuals during the PCP at the eruption day against the distribution of the residuals during the PCP in the previous 30 days (KS-1).
- The distribution of the residuals during the PCP at the eruption day against the distribution of the residuals during the 48 h before the eruption (KS-2).

The rejection of both tests suggests that the distribution of the residuals during the plume is different from the same distribution during the quiet state, highlighting a possible relationship between the residual and the volcanic plume. Finally, we combined all the performed tests and asserted that the GPS signal crossed the volcanic plume reasonably if simultaneously:

$$\boxed{\rho > 1; \qquad \theta > 1; \qquad KS-1 = 1; \qquad KS-2 = 1.}$$

4.4.2 Results

Table 4.4 and Table 4.5 show the results of the statistical tests performed in our analysis for those stations in which it is possible to compute all the parameters ($\rho, \theta, KS -$

$1, KS-2$) at least for one satellite. For each SAP, we consider that the plume affects the measurements when all four tests were passed (meaning that $\rho > 1, \quad \theta > 1, \quad KS-1 = 1, \quad KS-2 = 1$).

	EC10		ELEO		EMFN		EPDN		EPMN		EPOZ		TAOR	
PRN	ρ	θ	ρ	θ	ρ	θ	ρ	θ	ρ	θ	ρ	θ	ρ	θ
1							0.1	-1.2	0.4	-1.3			0.5	0.1
2														
3			0.8	-0.3			0.8	0.2			0.5	-0.3		
4														
5							1.2	0.2	0.9	0.3			0.8	0
6			0.9	0.1	1.2	0.5	1	-0.2						
7														
8														
9														
10														
11					0.1	-1.5								
12														
13														
14					0.6	0	0.3	-0.5	0.5	-0.4				
15														
16	0.9	0	0.8	-0.1	1.6	0.6	1	0	***2***	***3.8***	0.9	-0.2	***1.2***	***1.5***
17														
18	***2***	***1.7***	0.8	0	1.3	0.9	1	0.9	0.7	-0.2	1	0.2	0.7	0.3
19														
20							0.8	-0.1	***1.5***	***3.7***				
21	1.1	0.5	***1.3***	***1.1***			1.3	0.6			0.9	0.6		
22	1.9	0.4	***5.8***	***2.8***	***1.9***	***2.7***			1	0.2	***4.1***	***2.7***	0.8	-0.3
23														
24	1.1	0.9	***1.1***	***1***	***1.3***	***1***	1.1	0.1			1	0.4		
25														
26														
27														
28														
29														
30					1.1	1	1.2	0.1	0.7	-0.4			1.1	0.3
31														
32														

TABLE 4.4: ρ and θ values of the statistical analysis of the interaction between Plume and GPS undifferenced residual. Antenna–satellite pairs passing all the tests are bolded

We found that the pairs: EC10-PRN18, ELEO-PRN21, ELEO-PRN22, ELEO-PRN24, EMFN-PRN22, EMFN-PRN24,EPMN-PRN16, EPMN-PRN20, EPOZ-PRN22 and TAOR-PRN16 detected the volcanic plume. It is notable that only 10 among the 51 SAPs passed the four tests above described. Hence, this means that 20 % of SAPs detected the volcanic plume (Table 4.6).

	EC10		ELEO		EMFN		EPDN		EPMN		EPOZ		TAOR	
PRN	$T1$	$T2$	$T1$	$T2$	$T1$	$T2$	$T1$	$T2$	$T1$	$T2$	$T1$	$T2$	$T1$	$T2$
1							0	1	0	1			0	1
2														
3			0	1			1	0			0	1		
4														
5							1	1	1	0			0	0
6			0	0	1	0	1	0						
7														
8														
9														
10														
11					0	1								
12														
13														
14					1	0	0	0	0	0				
15														
16	0	1	1	0	1	1	0	0	***1***	***1***	1	0	***1***	***1***
17														
18	***1***	***1***	0	0	1	0	0	0	0	0	0	1	1	1
19														
20							1	0	***1***	***1***				
21	0	0	***1***	***1***			1	1			1	1		
22	1	0	***1***	***1***	***1***	***1***			1	1	***1***	***1***	0	1
23														
24	1	1	***1***	***1***	***1***	***1***	1	1			1	1		
25														
26														
27														
28														
29														
30					0	0	0	0	0	1			0	1
31														
32														

TABLE 4.5: $T1$ and $T2$ values of the statistical analysis of the interaction between Plume and GPS undifferenced residual. The tests T1 and T2 assume value 0 if the null hypothesis is accepted, 1 if it is rejected. Antenna–satellite pairs passing all the tests are bolded

Among all the stations, ELEO showed a greater residual. Figure 4.20 shows the GPS residual of ELEO station (in the length unit) linked to PRN satellites. In the figure 4.20 the green time-series represent the undifferenced LC residuals of the 30 days before the activity, the blue time-series represents the LC residual of the day of eruption and red points denote the plume crossing period.

The anomalies in the LC signal (blue line) between 2680 and 2800 epochs (about 22:00 and 23:00 GMT) for PRN22 are evident. Among all SAPs, this was the clearest change

PRN	EC10	ELEO	EMFN	EPDN	EPMN	EPOZ	TAOR
1				0	0		0
2							
3		0		0		0	
4							
5				0	0		0
6		0	0	0			
7							
8							
9							
10							
11			0				
12							
13							
14			0	0	0		
15							
16	0	0	0	0	***1***	0	***1***
17							
18	***1***	0	0	0	0	0	0
19							
20				0	***1***		
21	0	***1***		0		0	
22	0	***1***	***1***		0	***1***	0
23							
24	0	***1***	***1***	0		0	
25							
26							
27							
28							
29							
30			0	0	0		0
31							
32							

TABLE 4.6: Statistical evidence (true[1]/false[0]) of the interaction between Plume and GPS undifferenced residual.

in the LC signal with respect to the LC signal of the previous 30 days (green lines), resulting in a higher value of ρ (Table 4.4). This could be due to the model of antenna which does not have a multipath filtering system, being a Compaq L1/L2 antenna. However, we note that both ELEO and EPOZ show higher values of anomalies with satellite 22 which is surely due to the plume crossing. These two stations are in fact located at low altitude above sea level (i.e., dozens of meters) and are very near each other (fig. 4.16a). Consequently, the volcanic plume probably affected the two stations similarly, in agreement with our results. Figure 4.21 shows instead the GPS residuals of EMFN station.

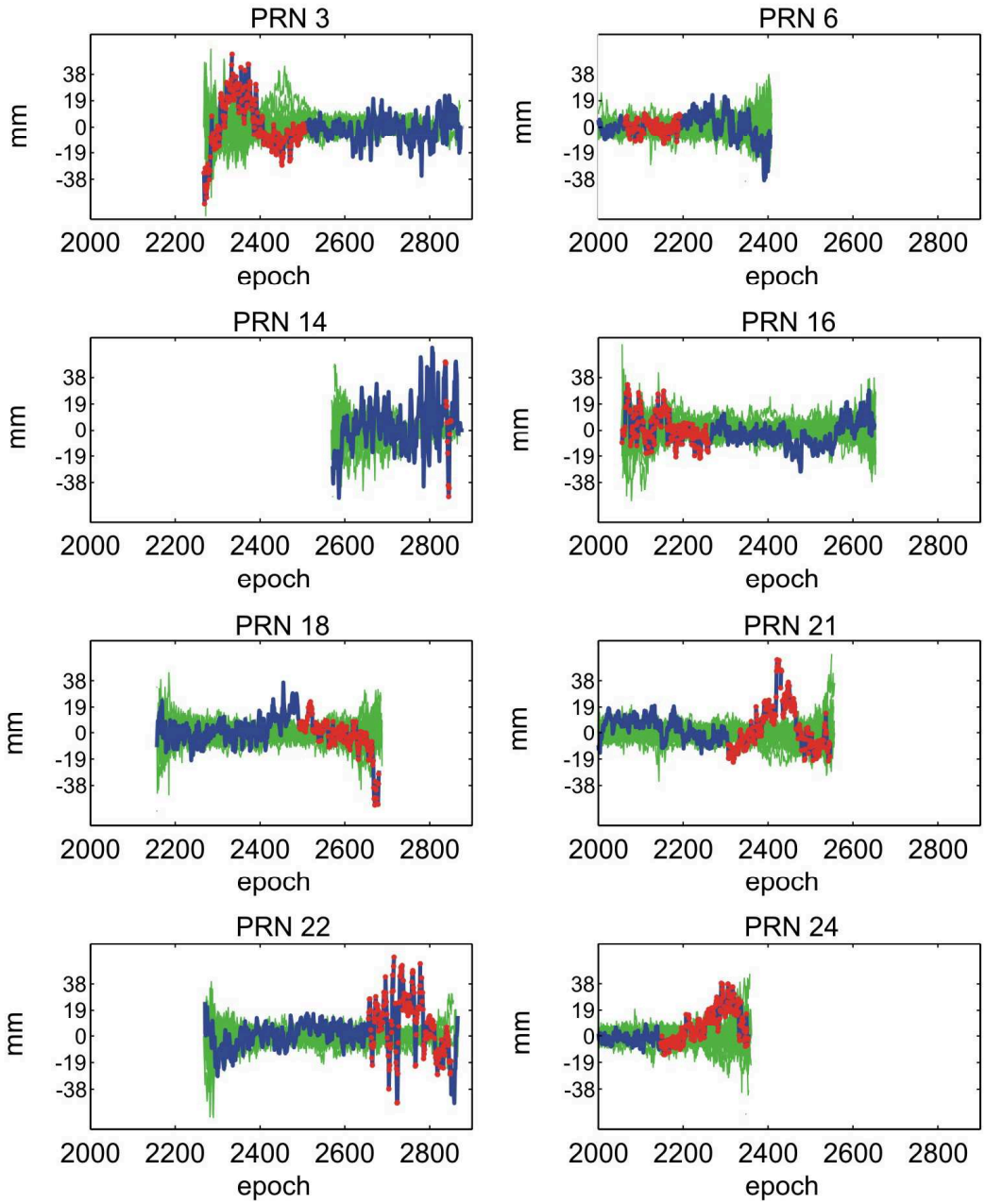


FIGURE 4.20: GPS residuals of the ELEO station with PRN3, PRN6,PRN14, PRN16, PRN18, PRN21, PRN22 and PRN24 during the eruption day (blue line), the previous 30 days (green line) and the covered period (red points). The x and y axes represent the GPS epoch and residuals in length units, respectively

Although there is no clear evidence of the anomaly, we observe that the signal is treated with a high-pass filter for two tests, pointing out anomalies which otherwise would be not visible. Figure 4.22 instead shows a greater LC residual value with respect to the previous days for the EPMN.

However, this particular SAP did not pass the test. we believe that this could be due to the LC signal to noise ratio (SNR), which is low in this particular SAP. The

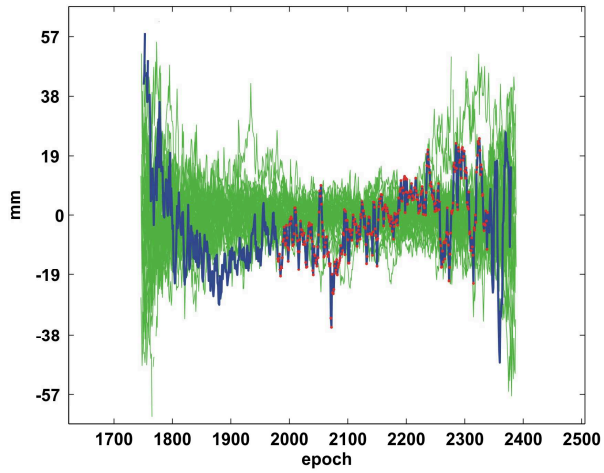


FIGURE 4.21: GPS residuals of the EMFN station with satellite PRN 24 during the eruption day (blue line), the previous 30 days (green line) and the covered period (red points). The x axis represents the GPS epoch, and the y axis represents the GPS residuals in length units

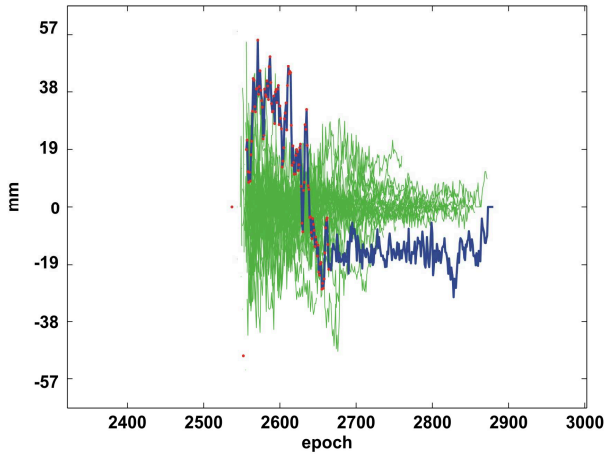


FIGURE 4.22: GPS residuals of the EPMN station with satellite PRN 14 during the eruption day (blue line), the previous 30 days (green line) and the covered period (red points). The x axis represents the GPS epoch, and the y axis represents the GPS residuals in length units

obvious evidence of anomalies is therefore not a sufficient condition to clearly identify the presence of the plume.

4.4.3 Discussion

We demonstrate that the presence of a volcanic plume in the atmosphere affects the GPS signal during the September 4, 2007 eruption of Mt. Etna. GPS was able to detect the volcanic plume generated from this mid-intensity explosive activity. This was tested through a statistical algorithm that indicates which stations and satellites were unambiguously affected by the GPS signal. The choice to investigate the LC signal is

related to the plume composition. Indeed, volcanic plumes are largely made up of volcanic ash, water vapor, volcanic gas (CO_2, NO_2, SO_2) and hydrometeors [Oppenheimer et al., 2003]. While the delay from water vapor is due to the polar nature of the water molecules, the phase delay induced by droplets, hydrometeors or aerosol particles is related to permittivity. The LC signal, which is verify of the two GPS signal carriers, is more strongly affected than L1 and L2 carriers [Solheim et al., 1999]. Their results show that the components that mainly affect the GPS signal are in order: water vapor, hydrometeors, aerosol and volcanic ash. We showed that a number of SAPs presented high anomalies during the activity of the September 4, 2007. The choice of this eruption as a test case in our study was largely guided by the long-lasting activity and the presence of quasi-stationary eruptive phenomena that remained constant in intensity and dispersal for almost 12 h. The relationship of the volcanic plume presence with LC GPS residual has been proved by [Houlie' et al., 2005a], [Houlie' et al., 2005b] studying the plianian eruption of Miyakejima. However, it is notable that the explosive activity of the September 4, 2007 Etna eruption produced a total volume of 3.9–4.9 $\times$ $10^5 m^3$ for pyroclastics [Andronico et al., 2008], which is ten times less than the estimated value for the August 18, 2000 activity of Miyakejima volcano ($4 \times 10^6 m^3$ see [Nakada et al., 2005]; [Saito et al., 2005]). In agreement, the residual values on LC frequency in millimeters found in our work are ten times less than residual values found by [Houlie' et al., 2005a], [Houlie' et al., 2005b], highlighting a possible relationship between residuals and volcano energy. In our study, We need to represent the volcanic plume geometrically. As a reasonable approximation, We decided to use data collected during monitoring activities because they are the nearest representation of the real volcanic plume dispersal. Consequently, We considered the plume region similar to a paraboloid fitting the real tephra deposit [Andronico et al., 2008] and satellite images (fig. 4.1). In case of missing data, results from volcanic ash dispersal models, such as those available in the control room of INGV-OE [Scollo et al., 2009], could be used to estimate the actual plume geometry. Furthermore, the activity of September 4, 2007, produced a weak plume. These plumes are characterized by lower vertical velocities than the wind and typically follow bent-over trajectories as a result of the strong wind advection [Bonadonna et al., 2005]. For this reason, our approach may be improved by including a nearest representation of the volcanic dispersal, for example by using more sophisticated volcanic ash dispersal models [Costa et al., 2006]. We found that while the presence of the volcanic plume is necessary in order to affect the residual, this is insufficient because not all the SAPs intersecting the plume show an anomaly that passes all the tests, as can be seen in table 4.6. We retain that this estimation could be improved through a better representation of the volcanic plume dispersion, a further analysis of GPS signal features and with improvements in the reliability of the statistical test. In order to test the reliability of the statistical approach, the two different tests were developed to assess the anomalies in

the residual signals during the plume crossing, compared with the same signal measured in the previous 30 days for the same satellite constellation (i.e., the same epochs) and the same signal measured on the day of eruption (to assess that the anomalies did not affect the signal before the beginning of the explosive activity). We stated that no SAP was able to pass the four tests. However, it should be noted that results must only be seen in a probabilistic manner. In the future, new statistical approaches could be used to improve the capability of GPS signals to detect volcanic plumes. Furthermore, according to the different types of antenna which each receiver is equipped, we note that the comparison of the LC residual data periods is always carried out among data of the same station and never among different stations. Consequently, a different type of antenna can only influence its sensitivity to detect volcanic plumes but not the reliability of this test. Finally, we believe that, due to the high number of GPS stations, a more accurate analysis of these data during explosive activity could furnish new complementary information and be complementary to the already existing monitoring system of Etna volcanic plumes at INGV-OE, thereby contributing to reduce the risk to air traffic operation.

Chapter 5

The GPS tomography

5.1 Introduction

The satellite techniques are today frequently used in geodesy. Measures of the size and shape of the Earth play an important role, especially in the volcanic areas. One technique capable of such measures is the differential SAR interferometric (DInSAR) technique. The measurement of ground deformations with DInSAR technique is based on the comparison between two SAR images acquired at different times (called interferograms). If a ground deformation occurs between two satellite passes in the same area, the path differences between the sensor and the ground (through the two passes) lead to a difference in the phase of the back-scattered radar wave [Massonnet et al., 1993], [Ulaby et al., 1981]. Among various perturbations affecting interferograms, atmospheric artifacts are one of the most significant and, probably, the most difficult to identify and reduce [Zebker et al., 1997], [Williams et al., 1998], [Delacourt et al., 1998]. The electromagnetic (EM) waves, emitted from GPS and SAR satellites, indeed propagate in the neutral atmosphere at velocities lower than the vacuum depending on the local characteristics of the medium. Variations of pressure, temperature and water content cause changes in the effective refractive index along the ray path and therefore may corrupt any phase change measurement due to a deformation and degrade any geophysical parameters estimated from such measurements. The neutral atmosphere (stratosphere and troposphere) is the portion of the atmosphere where the variations that significantly affect the interferograms occur. Due to the prominent topography of the volcanoes, the use of DInSAR technique may significantly suffer from atmospheric heterogeneities. As a consequence, several studies have been carried out on Mt. Etna since 1996 to better understand the influence of neutral atmosphere on measurements of ground deformation

phenomenon viewed from DInSAR. These studies investigate the possibility of evaluating and correcting the tropospheric effects of SAR images through different approaches: by using the integrated tropospheric delay obtained from the GPS analysis [Beauducel et al., 2000], starting from empirical data used to predict atmospheric conditions trough numerical model [Delacourt et al., 1998], [Wadge et al., 2002], [Webley and Wadge, 2004]. All these studies are based on the use of GPS output data analysis to sound the atmosphere, to estimate its temporal variability and then integrate the results in SAR interferograms. The need to use an atmospheric tomography for evaluating the effects of atmospheric anomalies on DInSAR measurements is due to the fact that the SAR satellites are a side looking radar (SLR), that mean to observe the scene trough an off-nadir angle ranging from 20° to 40°, depending on the radar satellite used. In the following we present a method to obtain a 3D electromagnetic waves velocity field, starting from the GPS output analysis data, in order to project along any line of sight (LOS) the effects of the tropospheric anomalies eventually existing on the atmosphere crossed by the actual radar signal. Starting from the previous studies [Bruno et al., 2007], [Bonforte et al., 2001], we propose some advances aimed to assessing a method to obtain a suitable tomography of EM waves propagation velocity of the lower atmosphere, starting from delays measured by using an existing GPS network.

5.2 The inverse problem

The inverse problem consists of using the actual result of some measurements to infer the values of the parameters that characterize the system, figure 5.1. While the forward problem has (in deterministic physics) a unique solution, the inverse problem does not [Tarantola, 2005]. In the ideal case, an exact theory that gives a complete description

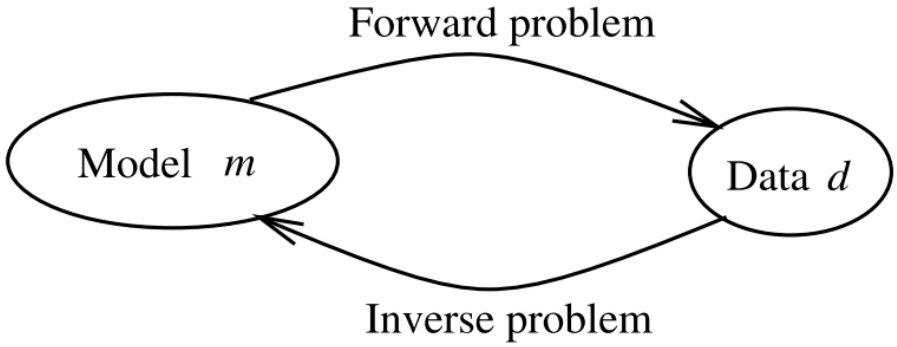


FIGURE 5.1: The traditional definition of the forward and inverse problems [Snieder and Trampert, 2000].

of a physical system exists. For some selected examples such a theory exists assuming that the required infinite and noise-free data sets would be available. This means that

the model has infinitely degrees of freedom. However, in actual experiments the amount of data that can be used for the definition of the model is usually finite and affected by noises and bias. A simple count of variables shows that the data cannot carry sufficient information to determine the model uniquely. In the context of linear inverse problems this point has been raised by [Backus and Gilbert, 1967] and [Backus and Gilbert, 1968] more recently by [Parker, 1994] and [Tarantola, 2005]. The fact that in actual experiments a finite amount of data is available to reconstruct a model with infinitely degrees of freedom implies that the inverse problem is not unique in the sense that there are many set of parameters of the model that explain the data equally well. The model obtained from the inversion of the data is therefore not necessarily equal to the true model that it is looking for. This implies that the view of inverse problems as shown in figure 5.1 is too simplistic. For actual problems, inversion really consists of two steps. Let the true model be denoted by m and the data by d. From the data d one reconstructs an estimated model $\widetilde{m}$, this is called the estimation problem, see figure 5.2. Apart from estimating a model $\widetilde{m}$ that is consistent with the data, it is

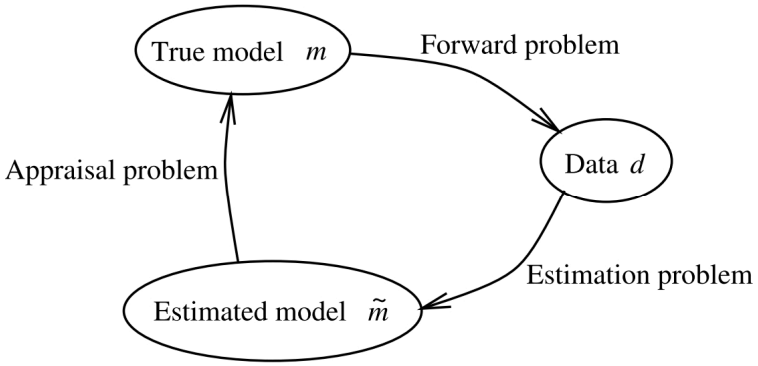


FIGURE 5.2: The inverse problem viewed as a combination of an estimation problem plus an appraisal problem [Snieder and Trampert, 2000].

also necessary to investigate the relations between the estimated model $\widetilde{m}$, and the true model, m. The "appraisal problem" consists in determining what parameters of the true model are recovered by the estimated model and what are the errors associated to them. The essence of this discussion is that *inversion = estimation + appraisal.* It does not make much sense to make a physical interpretation of a model without the knowledge of the errors of parameter and the limits of the model [Trampert, 1998] and [Toomey and Foulger, 1989]. In general there are two reasons why the estimated model differs from the true model. The first reason is the non-uniqueness of the inverse problem that causes several equivalent models fitting the data. Technically, this model null-space exits due to inadequate sampling of the model space. The second reason is that actual data (and physical theories more often than we would like) are always contaminated by errors and so the estimated model is therefore affected by these errors as well.

5.2.1 Solving finite linear system of equations

As discussed in the previous section, the inverse problem maps a finite number of data onto a model. In most practical applications like geophysics or physic of atmosphere, the model is a continuous function of the space coordinates and therefore has infinitely many degrees of freedom. For the moment we will ignore this and will assume that the model can be characterized by a finite number of parameters. For a finite-dimensional model, the model parameters can be ordered in a vector $\mathbf{m}$, and similarly the data can be ordered in a vector $\mathbf{d}$. The matrix $\mathbf{A}$ relates the data to the model through the product $\mathbf{Am}$. This matrix is often referred to as the "theory operator". Indeed, it contains all the information on physics and mathematics we have chosen to model in the given problem. In practice, the data are contaminated with errors $\mathbf{e}$, so that the recorded data and the model are related by:

$$\mathbf{d} = \mathbf{Am} + \mathbf{e} \tag{5.1}$$

It should be noted that often there is an certain arbitrariness in the choice of the model parameters that are contained in the model vector $\mathbf{m}$. For example, if one wants to describe the density in the Earth one could choose a model where the Earth's mantle and the core have a uniform density, in that case there are two model parameters. Alternatively, the density in the Earth can be expanded in a large amount of eigenfunctions defined on the sphere such as spherical harmonics for lateral variations and polynomials for depth variations; in that case there are much more model parameters. These two different parameterizations of the same model correspond to different model parameters $\mathbf{m}$ and to a different matrix $\mathbf{A}$. This example illustrates that the model $\mathbf{m}$ is not necessarily the true model, but that the choice of the model parameters usually contains a restriction on the class of models that can be constructed. Below we will refer to $\mathbf{m}$ as the true model regardless of the difficulties in its definition. Since the estimated parameters, by using the inverse approach, will be different from the true model, the estimated model is denoted by $\widetilde{\mathbf{m}}$. There are many ways for designing an inverse operator that maps the data on the estimated modele.g. [Menke, 1984]; [Tarantola, 2005]; [Parker, 1994]. Whatever estimator one may choose, the most general linear mapping from data to the estimated model can be written as:

$$\widetilde{\mathbf{m}} = \mathbf{A}^{-\mathbf{g}}\mathbf{d} \tag{5.2}$$

The operator $\mathbf{A}^{-\mathbf{g}}$ is called the generalized inverse of the matrix $\mathbf{A}$. In general, the number of data is different from the number of model parameters. For this reason $\mathbf{A}$ is

usually a non-square matrix, and hence its formal inverse does not exist. Later we will show how the generalized inverse $\mathbf{A}^{-\mathbf{g}}$ may be chosen, but for the moment $\mathbf{A}^{-\mathbf{g}}$ does not need to be specified. The relation between the estimated model $\widetilde{\mathbf{m}}$ and the true model $\mathbf{m}$ follows by inserting (5.1) in expression (5.2):

$$\widetilde{\mathbf{m}} = \mathbf{A}^{-\mathbf{g}}\mathbf{A}\mathbf{m} + \mathbf{A}^{-\mathbf{g}}\mathbf{e} \tag{5.3}$$

The matrix $\mathbf{A}^{-\mathbf{g}}\mathbf{A}$ is called the *resolution kernel*, this operator is given by:

$$\mathbf{R} = \mathbf{A}^{-\mathbf{g}}\mathbf{A} \tag{5.4}$$

Better known as "resolution matrix". Expression (5.3) can be interpreted by rewriting it in the following form:

$$\widetilde{\mathbf{m}} = \mathbf{m} + \underbrace{(\mathbf{A}^{-\mathbf{g}}\mathbf{A} - \mathbf{I})\mathbf{m}}_{\text{Limited Resolution}} + \underbrace{\mathbf{A}^{-\mathbf{g}}\mathbf{e}}_{\text{Error propagation}} \tag{5.5}$$

In the ideal case, the estimated model equals the true model vector: $\widetilde{\mathbf{m}} = \mathbf{m}$ meaning that our chosen parameters, ordered in vector $\mathbf{m}$, may be estimated independently from each other. The last two terms in equation (5.5) account for blurring and artifacts in the estimated model. The term $(\mathbf{A}^{-\mathbf{g}}\mathbf{A} - \mathbf{I})\mathbf{m}$ describes the fact that components of the estimated model vector are linear combinations of different components of the true model vector. We only retrieve averages of our parameters and "blurring" occurs in the model estimation as we are not able to map out the finest details [Snieder and Trampert, 2000]. In the ideal case this term vanishes; this happens when $\mathbf{A}^{-\mathbf{g}}\mathbf{A}$ is equal to the identity matrix. With (5.4) this means that for perfectly resolved model parameters the resolution matrix is the identity matrix:

$$\text{Perfect resolution: } \mathbf{R} = \mathbf{I} \tag{5.6}$$

As noted earlier, usually there is a certain ambiguity in the definition of the model parameters that define the vector $\mathbf{m}$. The resolution operator tells us to what extend we can retrieve the model parameters independently from the estimation process. However, the resolution matrix does not tell us completely what the relation between the estimated model and the real underlying physical model is, because it does not take into account to what extent the choice of the model parameters has restricted the model that can be obtained from the estimation process. The last term in (5.5) describes how the errors

e are mapped onto the estimated model. These errors are not known deterministically, otherwise they could be subtracted from the data. A statistical analysis is needed to describe the errors in the estimated model due to the errors in the data. When the data d_j are uncorrelated and have standard deviation σ_{d_j}, the standard deviation σ_{m_i} in the model estimate $\widetilde{m_i}$, resulting from the propagation of data errors only, is given by:

$$\sigma_{m_j}^2 = \sum (A_{ij}^{-1}\sigma_{d_j})^2 \tag{5.7}$$

Ideally, one would like to obtain both: a perfect resolution and no errors in the estimated model. Unfortunately this cannot be achieved in practice.

5.2.2 Least squares estimation

Let us for the moment consider the case where the number of independent data is larger than the number of unknowns. In that case, the system $\mathbf{d} = \mathbf{Am}$ cannot always be satisfied for any given model $\mathbf{m}$ because of possible errors contained in the data vector making the equations inconsistent. A common way to estimate a model is to seek the model $\widetilde{\mathbf{m}}$ that gives the best fit to the data in the sense that the difference, between the data vector $\mathbf{d}$ and the recalculated data $\mathbf{A}\widetilde{\mathbf{m}}$ is made as small as possible. This means that the least-squares solution is given by the model that minimizes the following cost function:

$$S = ||\mathbf{d} - \mathbf{Am}||^2. \tag{5.8}$$

As shown in detail by [Strang, 1988] this quantity is minimized by the following model estimate:

$$\widetilde{\mathbf{m}} = (\mathbf{A^T A})^{-1}\mathbf{A^T d} \tag{5.9}$$

5.2.3 Minimum norm estimation

In some problems the number of unknowns is greater than the number of parameters. For a under determined system of equations the minimum norm solution is defined as the model that fits the data exactly, $\mathbf{Am} = \mathbf{d}$, and that minimizes $||\mathbf{m}||^2$. Using Lagrange multipliers one can show that the minimum-norm solution is given by:

$$\widetilde{\mathbf{m}} = \mathbf{A^T}(\mathbf{AA^T})^{-1}\mathbf{d} \tag{5.10}$$

a detailed derivation is given by [Menke, 1984].

5.2.4 Mixed determined problems

In the least-squares estimation, we assume that we had enough information to evaluate all model parameters, even though contradictions occurred due to measurement errors. The problem is then purely over-determined and as a consequence $\mathbf{A^TA}$ is regular. In the minimum norm solution, we assume no contradictions in the available information, but we don't have enough equations to evaluate all model parameters. This is the case of a purely under-determined problem and here $\mathbf{AA^T}$ is regular.
The most common case, however, is that we have contradictory information on some model parameters, while others cannot be assessed due to a lack of information. Then neither $\mathbf{A^TA}$ nor $\mathbf{AA^T}$ can be inverted and the problem is ill-posed.
Even if the inverse matrices formally exist, they are often ill-conditioned meaning that small changes in the data vector lead to large changes in the model estimation. This means that errors in the data will be magnified in the model estimation. Clearly a trick is needed to find a set of model parameters that is not too sensitive to small changes in the data. To this aim, [Levenberg, 1944] introduced a damped least-squares solution. From a mathematical point of view, ill-posedness and ill-conditioning result from zero or close to zero singular values of $\mathbf{A}$. Suppose one has a matrix $\mathbf{M}$ with eigenvalues λ_n and eigenvectors $\widehat{\mathbf{v}}_\mathbf{n}$:

$$\mathbf{M}\widehat{\mathbf{v}}_\mathbf{n} = \lambda_n\widehat{\mathbf{v}}_\mathbf{n} \tag{5.11}$$

One readily finds that the matrix $(\mathbf{M} + \gamma\mathbf{I})$ has eigenvalues $(\lambda_n + \gamma)$:

$$(\mathbf{M} + \gamma\mathbf{I})\widehat{\mathbf{v}}_\mathbf{n} = (\lambda_n + \gamma)\widehat{\mathbf{v}}_\mathbf{n} \tag{5.12}$$

This means that the eigenvalues of a matrix can be raised by adding the scaled identity matrix to the original matrix. This property can be used to define the *damped least-squares solution*:

$$\widetilde{\mathbf{m}} = (\mathbf{A^TA} + \gamma\mathbf{I})^{-1}\mathbf{A^Td} \tag{5.13}$$

Since the matrix $\mathbf{A^T A}$ has positive eigenvalues, they are moved away from zero when the constant γ is positive. Alternatively, the solution (5.13) can be found by minimizing the following cost function:

$$S = ||\mathbf{d} - \mathbf{Am}||^2 + \gamma||\mathbf{m}||^2 \tag{5.14}$$

This expression clearly shows what the effect of the damping is. Minimizing the first term of (5.14) amounts to finding the model that gives the best fit to the data. Minimizing the last term of (5.14) amounts to finding the model with the smallest norm. In general we cannot minimize both terms simultaneously, but in minimizing (5.14) we try to find a model that both fits the data reasonably well and whose model size is not too large. The parameter γ controls the emphasis we put on these conflicting requirements and for this reason it is called the *damping-factor* or *trade-off parameter*.

5.3 The SIMULPS approach

Although GPS is also used to estimate the delays of EM waves through the atmosphere, relatively few studies exploit this characteristic of GPS to obtain tomographies of the atmosphere [Hoeven et al., 2002]. Our research starts from previous achievements both on seismic tomography [Aloisi et al., 2002] and GPS atmospheric sounding [Bonforte et al., 1999], [Bonforte et al., 2001] for assessing a method to produce a tomography of lower atmosphere (neutral atmosphere) over the Mt. Etna area, by suitably modifying a software originally implemented for seismic tomography SIMULPS12 [Thurber and Ellsworth, 1980], [Thurber, 1983], [Thurber and Aki, 1987], [Thurber et al., 1993]. As in the seismic case, where it is necessary to fix the seismometers and the Earthquake positions and the a-priori wave velocity fields to evaluate the tomography of underground wave velocity, in our case we have to fix corresponding parameters to compute the atmospheric EM wave velocity tomography. We assume the satellites as seismic stations and the GPS receivers as Earthquakes [Bruno et al., 2007]; in particular, since the position of GPS stations is well known, they are considered as points of shot in seismology, namely Earthquakes of well known position and time. To study the tomography problem, in this work we use an interpolative function defined by values specified at nodal points (knots) within a three-dimensional grid [Aki, 1993], [Lee and Pereyra, 1993]; a complete discussion of the problem is examined in [Thurber and Ellsworth, 1980], [Thurber, 1983], [Thurber and Aki, 1987] and [Thurber et al., 1993]. Some processing parameters were

analyzed and several spatial configurations were tested to determine the best knots position. In order to assess whether benchmarks are suitable to obtain the desired results, a few tests were carried out.

5.3.1 GPS Satellites: position and selection

As in local Earthquake tomography (LET), we suppose to ideally enclose inside a parallelepiped the satellites, the receivers and the portion of atmosphere we want to analyze. The base of the parallelepiped of the investigated area should be comparable with the Etna area. Since the GPS satellites orbit at altitudes between 18000 and 22000 km, assuming a cut-off angle of 15° for GPS measurements, as usual for geodetic applications of GPS, the base of the parallelepiped results extremely large (see fig.5.3). For instance,

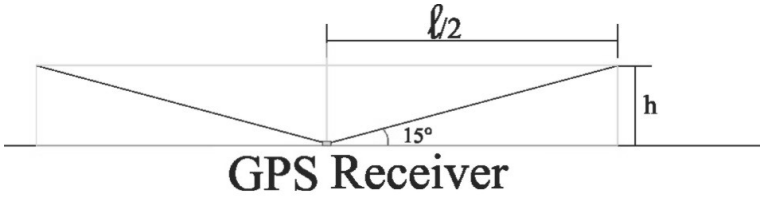


FIGURE 5.3: Sketch of cone representing the maximum angular aperture of the GPS antenna assuming a cut-off angle of 15°.

if we consider a satellite altitude of 18000 km, we obtain a base for parallelepiped of $1.8 \times 10^{10} km^2$ that is oversized if compared to that of Mt. Etna area, which is about $7.5 \times 10^3 km^2$. Since we are only interested in the neutral atmosphere and also because the ionospheric GPS delay is eliminated by a "iono-free" linear combination, we reduced the dimension of our system, "shifting down" the satellites ideally along the directions connecting them with the GPS receivers. The new satellite coordinates (here called synthetic coordinates) result from the intersection between the straight line joining satellite and receiver with the ellipsoid 10 km away from the Earth ellipsoid (fig.5.4). According to this model, we "reduced" the height of satellites to 10 km above the Earth ellipsoid inside a parallelepiped with a base of about $70 \times 70 (4.9 \times 10^3) km^2$, which is close to the dimension of Mt. Etna area. Therefore the geometrical distance between the satellite S and the receivers R will not be the real one, but depends on the new coordinates of satellite. The reduction of the height of GPS satellites led to suppose that tropospheric delay is due only to the first 10 km of atmosphere (troposphere) and not to all neutral atmosphere (troposphere and stratosphere), which extends as far as about 50 km above sea level. This seems a reasonable assumption, because the troposphere contains about 90% of the air mass and 99% of water vapour of atmosphere, which is the main source of the atmospheric delays for the GPS EM waves [Essen and Froome, 1951]. After a preliminary processing of GPS data, performed by using GAMIT software [Herring

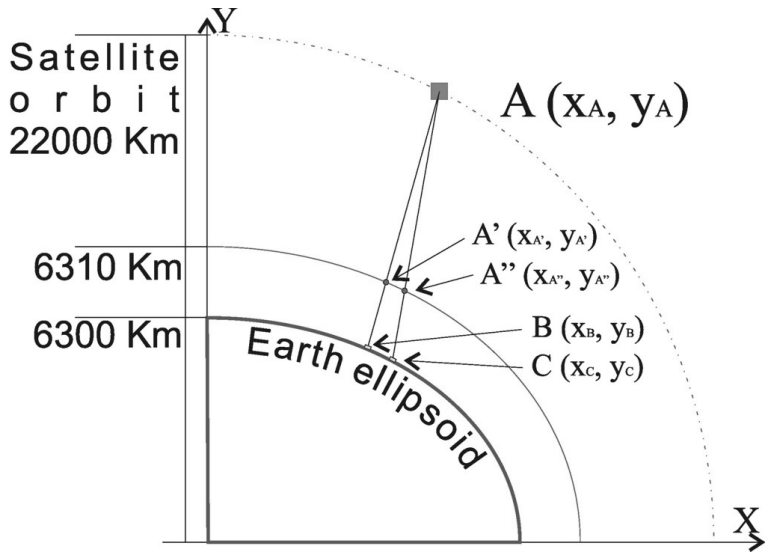


FIGURE 5.4: Scheme of satellite projection from the actual position $A(x_A, y_A)$ to the positions $A'(x_{A'}, y_{A'})$ and $A''(x''_A, y''_A)$ relatively to the stations $B(x_B, y_B)$ and $C(x_C, y_C)$ respectively.

et al., 2010], we carried out the statistical study of the residuals. In order to discard bad quality data, a selection of datasets is performed by inspecting the undifferenced LC phase residuals for each satellite-receiver pair. The statistical analysis of residuals is a good approach to evaluate the quality of data and guides their selection [Bruno et al., 2007]. According to the results of this analysis we filter the data set assuming a threshold value for residuals of two standard deviations of the data set. The filtered data have a Gaussian distribution of the residuals centred in zero and with standard deviations of about 6 mm; based on previous studies [Bruno et al., 2007] we consider that the data set now has a good quality.

5.3.2 The adopted atmospheric model

In SIMULPS12, the distribution of wave velocity is computed on knots of a three-dimensional grid, in which velocity varies continuously in all directions, with linear interpolation among knots [Evans et al., 1994]. Knots are defined by the intersections of a set of planes in three orthogonal directions, one horizontal and two vertical. The a priori model we used is characterized by a grid in which velocity of EM waves in the atmosphere varies only with respect to the altitude. According to the equations 2.61 and 2.62, the EM waves velocity depends on refraction index n, excluding the imaginary part of the refractive index that indicates the amount of absorption loss when the EM wave propagates through the material. Since the refractive index is related to the refractivity N by the eq. 2.68, that led to eq. 2.75, we can express EM waves velocity through the following expression:

$$v = \frac{c}{n} = \frac{10^6}{N + 10^6} \tag{5.15}$$

Rewriting the equation 2.75:

$$N = k_1 \frac{p_d}{T} Z_d^{-1} + k_2 \frac{p_w}{T} Z_w^{-1} + k_3 \frac{p_w}{T^2} Z_w^{-1} \tag{5.16}$$

and as already mentioned earlier, p_d is the partial pressure of dry air (expressed in [hPa]), T is the atmospheric temperature (expresed in [K]), p_w is the partial pressure of water vapour (expressed in [hPa]),while the empirical constants k_1, k_2, k_3, have the numerical values of table 2.4.

$$N = k_1 \frac{p_d}{T} + k_2 \frac{p_w}{T} + k_3 \frac{p_w}{T^2} \tag{5.17}$$

By applying the 5.17 we may compute the refractivity at an arbitrary altitude. Unfortunately, we do not have regular atmospheric sounding on Mt. Etna and the closest radiosonde launches are at Trapani (i.e. about 200 km west of Mt. Etna). Thus, we have to use models to predict p_d, T and p_w, to apply the 5.17 and to determine the a priori velocity model that we use for the tomography. In the following part of this chapter we assess the models to predict p_d, T and p_w and compare these estimations with actual atmospheric sounding to evaluate the accuracy of these models and, eventually, calibrate them.

Temperature distribution

As mentioned in chapter 2 section 2.3.4, the variation of temperature with height h defined as [Saastamoinen, 1972b], [Saastamoinen, 1972a.]:

$$T = T_0 - \alpha h \tag{5.18}$$

where

$$\alpha = \frac{\partial T}{\partial h} \tag{5.19}$$

In the figure 5.5, a comparison between the temperature model computed through the eq. 5.19 (black line) assuming $T_0 = 10$, compared with the values measured from the radiosonde launched from Trapani is plotted. We can see that, especially between 0 to 3 Km and 5.5 to 10 Km, the measured values have a slope rate close to the model values.

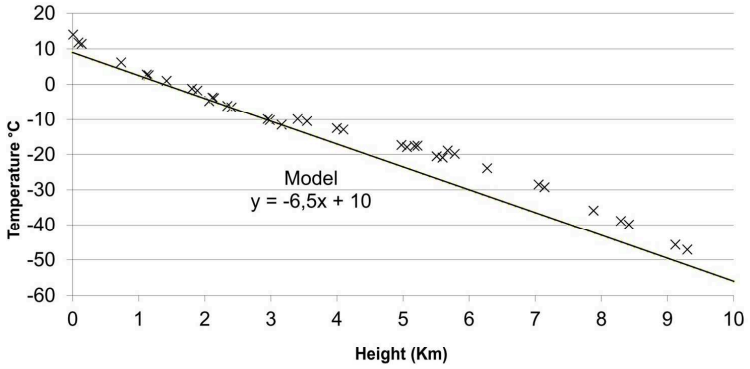


FIGURE 5.5: A comparison between the temperature values computed through the model and the measured data from the radiosonde launched from Trapani.

Between 3 to 5.5 Km, instead, a lower slope rate is measured. This example shows that it is necessary to have information about the vertical profile of the atmosphere in order to obtain a "good starting model" for refractivity. However todays we do not have this informations on Etna, although INGV research team started recently the launches of some atmospheric balloons test. Future cooperation will allow to have the profiles of temperature, pressure and moisture needed. For the present study in the SIMULPS approach, we used the "Normal Temperature and Pressure" (NTP) reference conditions (20 [°C], 293,15 [K] and 1013.25 [mbar]) to compute the values of variables at each altitude.

Pressure distribution

To calculate the pressure distribution within the troposphere, the following equation (the Barometric formula) relates atmospheric pressure p to altitude h, according to the eq. 2.95 [Saastamoinen, 1972b], [Saastamoinen, 1972a.]:

$$p = p_0 \left(\frac{T_0 - \alpha h}{T_0} \right)^{\frac{M_d g}{R\alpha}} \tag{5.20}$$

where p_0 is the pressure value in $[Pa]$ and T_0 is the atmospheric temperature in $[K]$ recored at sea level, α is the normal lapse rate of temperature with elevation (6.5 $[K/m]$), h is the height on which we want to calculate the pressure value $[m]$, g is the gravity at the surface of the Earth (9.7867 $[m/s^2]$), R is the universal gas constant (8.31447 $[J/(mol \cdot K])$ and M_d is the molar mass of dry air in (0.0289644 $[kg/mol]$). For the same reasons as above, the value of pressure was set to NTP reference conditions equal to 1013.25 [hPa]. We compare the results with those sampled from the Trapani radiosonde (Fig. 5.6). As expected, for the reasons discussed in chapter 2, the comparison between modeled and measured data shows a good agreement. For our purposes, it isn't necessary that the water vapor pressure at eight h is well known, while it is more important to

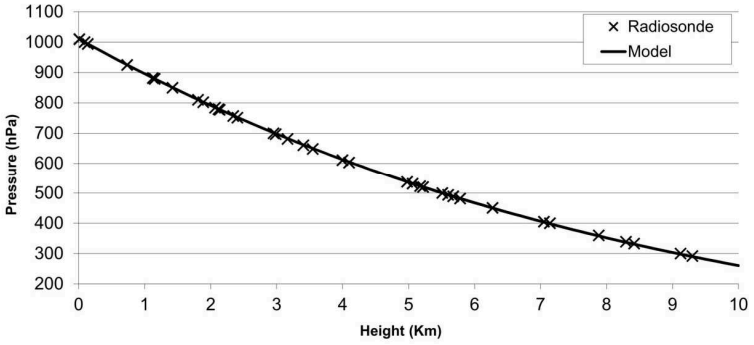


FIGURE 5.6: Plot of the atmospheric pressure versus altitudes: the line represents the theoretical atmospheric pressure, the crosses represent the values measured by the radiosonde lunched at Trapani.

estimate a possible variation range for that pressure values. Since the highest value of water vapor pressure at each height h corresponds to the saturated water vapor pressure, the water vapor pressure e is calculated starting from a relative humidity value of 100 %, using the Clausius-Clapeyron expression:

$$e = \frac{H_0}{100} \cdot e_{sat}(t) = \frac{H_0}{100} 6.1094 \cdot e^{\frac{17.625t}{t+243.04}} \tag{5.21}$$

where: H is the relative humidity (%), t is temperature in [°C] and e_{sat} is the saturated water vapor pressure in $[hPa]$ at the temperature t. From the equation 5.17 we are able, now, to evaluate if the refractivity field (and consequently the wave velocity) reconstructed through the tomography is realistic. We may consider the two limit situations: the first called *d*ry condition (which estimate the refractivity by using only the first term of the equation 5.17) and the second called *s*aturated condition (which estimate the refractivity by using all the three terms of the equation 5.17). Thus, it is possible to calculate the refractivity and consequently the propagation velocity of radio waves corresponding to different heights from sea level, for the two extreme conditions: dry and saturated. Figure 5.7 shows the two refractivity profile (dry and wet), compared with the samples of radiosonde.

All the values of refractivity estimated by using the radiosonde data are inside the computed refractivity range. We use the values of dash line of figure 5.7 as a priori atmospheric model in the inversion. These values in fact have been obtained assuming a dry atmosphere. Table 5.1 shows a detail of values computed trough the model described above.

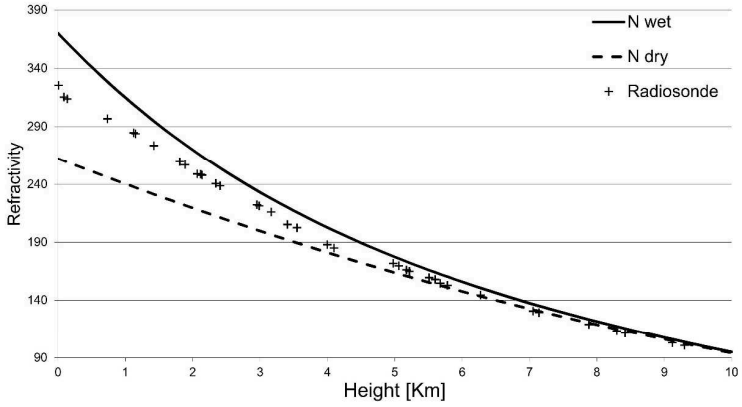


FIGURE 5.7: Refractivity values vs. height. The crosses represent the refractivity values of radiosonde. The dash and continuous lines represent the refractivity limit in dry and saturated conditions.

5.3.3 Grid fixing

In SIMULPS knots where the velocity is computed are defined by the intersections of a set of planes in three orthogonal directions, one horizontal and two vertical. Typically the planes are closely spaced in the center of the target volume, where the most rays should be, and more widely spaced in the periphery. A critical aspect of this parameterization is the placement and the spatial density of knots within the volume to be imaged (in our assumption a parallelepiped). To fix the position of the parallelepiped in the space we set a layer (π_1) tangent to the ellipsoid in the point of coordinates 37.75° N 14.99° E (the projection of the central crater of Mt. Etna on the ellipsoid). Then we calculate the parallel plane 10 Km from π_1 (called π_2). These two planes define respectively the base and the top of the parallelepiped containing the portion of atmosphere we want to study. Many tests were made to fix the position of the planes, both horizontal and vertical, within the parallelepiped because the problem of choosing an acceptable nodal distribution is not unique [Evans et al., 1994]. To fix the knots we use four parameters: the KHIT, defined below, the derivative weight sum (DWS), the resolution (RS) and the most important spread function (SF). The KHIT represents the number of ray paths running near the considered knot, it gives information about the density of rays near the knot but it is an un-weighted count; this parameter is important to evaluate immediately the goodness of the grid and to fix a cut off value below which the knot must be excluded from elaboration. The DWS provides an average relative measure of the density of rays near a given velocity knot, this measure of ray distribution is superior to an unweighted count of the total number of rays but it is influenced by a model parameter, since it is sensitive to the spatial separation of a ray from the nodal location [Toomey and Foulger, 1989]. The magnitude of DWS depends on the selected step length of the arc connecting the start and end points, smaller step lengths yield larger DWS, therefore DWS provides

Height[m]	Temp.[K]	Temp[°C]	Press.[hPa]	e sat[hPa]	N(max)	N(min)
0	293.15	20.00	1013.25	23.33	369.97	262.35
400	290.55	17.40	967.56	19.83	346.49	253.41
800	287.95	14.80	923.54	16.80	324.89	244.64
1200	285.35	12.20	881.15	14.19	305.00	236.04
1600	282.75	9.60	840.35	11.94	286.69	227.62
2000	280.15	7.00	801.09	10.01	269.80	219.38
2400	277.55	4.40	763.32	8.36	254.21	211.32
2800	274.95	1.80	727.00	6.95	239.80	203.46
3200	272.35	-0.80	692.09	5.76	226.46	195.78
3600	269.75	-3.40	658.54	4.76	214.10	188.29
4000	267.15	-6.00	626.32	3.91	202.62	181.00
4400	264.55	-8.60	595.38	3.20	191.94	173.90
4800	261.95	-11.20	565.69	2.61	181.98	167.00
5200	259.35	-13.80	537.20	2.11	172.67	160.29
5600	256.75	-16.40	509.89	1.71	163.96	153.77
6000	254.15	-19.00	483.70	1.37	155.79	147.44
6400	251.55	-21.60	458.61	1.09	148.11	141.30
6800	248.95	-24.20	434.58	0.87	140.87	135.35
7200	246.35	-26.80	411.58	0.69	134.03	129.58
7600	243.75	-29.40	389.57	0.54	127.57	124.00
8000	241.15	-32.00	368.52	0.42	121.44	118.59
8400	238.55	-34.60	348.40	0.33	115.62	113.36
8800	235.95	-37.20	329.18	0.25	110.08	108.30
9200	233.35	-39.80	310.82	0.19	104.81	103.42
9600	230.75	-42.40	293.29	0.15	99.78	98.70
10000	228.15	-45.00	276.57	0.11	94.98	94.14

TABLE 5.1: The refractivity values computed trough the model. Height, temperature, pressure and saturated water pressure are used to compute the refractivity values in dry (N min) and saturated (N min) conditions.

only a relative measure of the ray distribution. The RS provides the weights utilized in the averaging process of model parameters to estimate. An instructive way to examine the RS matrix is to plot the elements of the matrix in the three space of the study volume. For simultaneous inversion with hundreds of knots, it is impractical to examine pictorially these elements for each knot. The definition of a spread function (eq. 5.22) provides a synoptic view of RS [Toomey and Foulger, 1989]:

$$S_j = \log\left[\frac{1}{s_{jj}}\sum_{k=1}^{N}\left(\frac{s_{kj}}{s_{jj}}\right)^2 D_{jk}\right] \tag{5.22}$$

where s_{kj} is an element of the RS matrix, N is the number of the parameters, D_{jk} (eq. 5.23) is the distance from j-th to k-th knot.

$$D_{jk} = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2 + (z_k - z_j)^2} \tag{5.23}$$

Several configurations were tested to determine the best knots position (table 5.2), analyzing for each of them the values of KHIT, DWS and SF. In a series of test inversions, the horizontal and vertical separation of knots were varied from 2 to 7 Km. We fix a

	X layer [Km]	Y layer [Km]	Z layer [Km]	Knots
Test 1	-20,-15,-13,-11,-9, -7,-5,-3,-1,1,3,5,7, 9,11, 13,15,20	-30,-19,-15,-11,-9,-7,-5,-3,-1,1,3,5,7,9,11,15, 19,30	0,2,4,6,8,10	1944
Test 2	-20,-15,-12,-9,-6, -3,0 ,3,6,9,12,15,20	-20,-15,-12,-9,-6,-3,0,3,6,9,12,15,20	0,2,4,6,8,10	1014
Test 3	-50,-25,-16,-12,-8,-4,0,4,8,12,16,25,50	-50,-25,-16,-12,-8, -4,0,4,8,12,16,25 ,50	0,2,4,6,8,10	1014
Test 4	-20,-15,-10,-5,0,5,10,15,20	-20,-15,-10,-5,0,5,10,15,20	0,2,4,6,8,10	486
Test 5	-40,-20,-12,-6,0,6,12,20,40	-30,-16,-8,0,8,16,30	0,2,4,6,8,10	378
Test 5b	-40,-20,-12,-6,0,6,12,20,40	-40,-20,-12,-6,0,6,12,20,40	1,4,7,10	324
Test 6	-30,-14,-7,0,7,14,30	-30,-14,-7,0,7,14,30	1,3,5,7,9	245
Test 6b	-30,-14,-7,0,7,14,30	-30,-14,-7,0,7,14,30	2,3,5,7,9	245

TABLE 5.2: Knots positions performed to fix the optimal grid spacing. In the first column the number of test, in the second, third and fourth columns the values (in km) of layer positions with respect to the origin of the system are reported. In the last column the number of the knots for each test.

KHIT threshold value of 9, i.e. the knots touched less then 9 rays, are excluded from the inversion. The results of simultaneous inversion for all the nodal configurations are better for larger grid spacing. We chose a configuration grid having 245 knots (Test 6b in table 5.2). Concerning to the horizontal layers, we fix them at 2, 3, 5, 7, 9 Km above the ellipsoid level. The layers better configured for the inversion are the upper layers, specifically the layers at 5, 7 and 9 km from the ellipsoid level. Figure 5.8 shows a plan view of the knots produced by the vertical planes intersection. Moreover, the values of DWS vary substantially from a layer to another with the biggest values in the upper layers. As the value of DWS depends on the step length of the ray-path, in these tests we fix the value of DWS equal to 0.1 $[Km]$. Fig. 5.9 shows the value of DWS versus SF. Because well-sampled knots have a large value of DWS versus a smaller values of

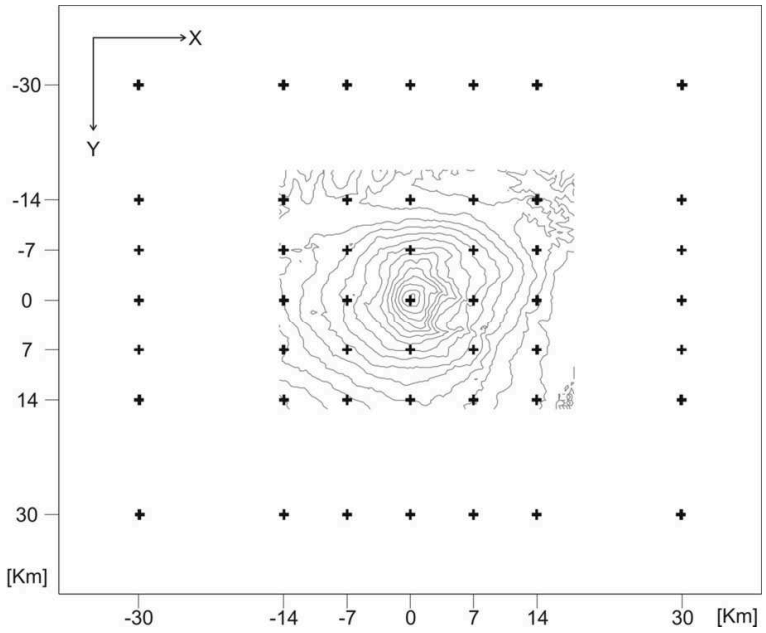


FIGURE 5.8: Plan view of the knots fixed for tomographic inversion.

SF, as a threshold of adequate resolution [Toomey and Foulger, 1989] we chose a spread function value of 2 as an upper limit of well-resolved knots

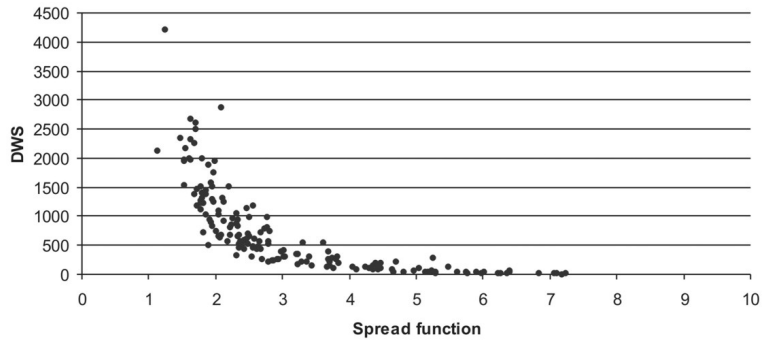


FIGURE 5.9: Plot of DWS versus Spread Function; this highlights a clear change of trend around the SF value of 2.

5.3.4 Tests

One of the aims of this work is to evaluate if the actual configuration (i.e. number and position of satellites, stations and knots) is suitable to obtain the "corrections" with respect to the starting velocity model, to obtain the tomography. To this end, we carried out a few tests, by assuming a set of "structures" (i.e. volume in which the velocity of waves varies from the surrounding space) of well-defined shapes (e.g. points, cube, checkerboard). The tests aimed to verify if the software is able to identify the structures by inverting the theoretical travel time obtained as direct computation by

SIMULPS12 software (i.e. the "synthetic travel times"). The "synthetic travel time" is computed taking into account the presence of the anomalies. We tested three different GPS network configurations, with 15, 30 and 90 GPS receivers. The configuration with

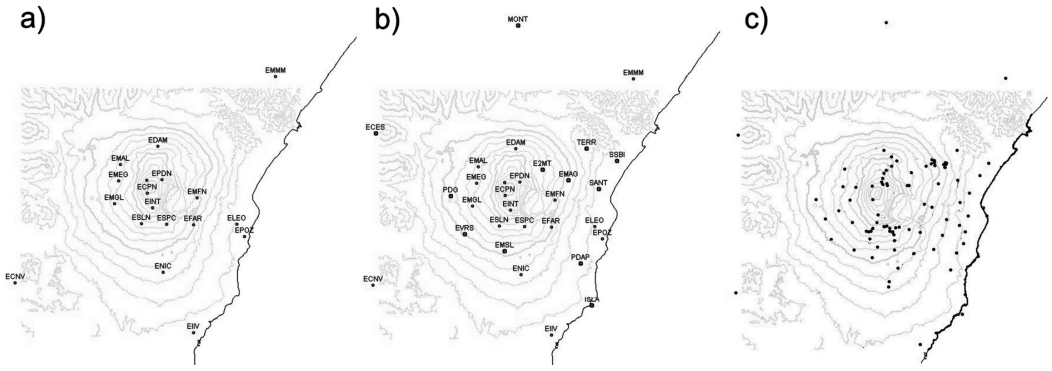


FIGURE 5.10: Plan view of the three GPS network configurations studied: a) GPS permanent network until the first quarter of 2007; (b) GPS permanent network until 2012; (c) configuration with 90 stations built of permanent stations and the entire benchmarks now existing on Mt. Etna.

15 stations corresponds to the permanent GPS network until the first quarter of 2007 (Fig.5.10a), while that with 30 stations corresponds to the permanent GPS network until 2012 (Fig.5.10b). The configuration with 90 stations is an ideal network that includes both permanent stations and the whole benchmarks now existing on Mt. Etna (Fig.5.10c). In the tests with 15 and 30 GPS receivers, two GPS satellite spatial config-

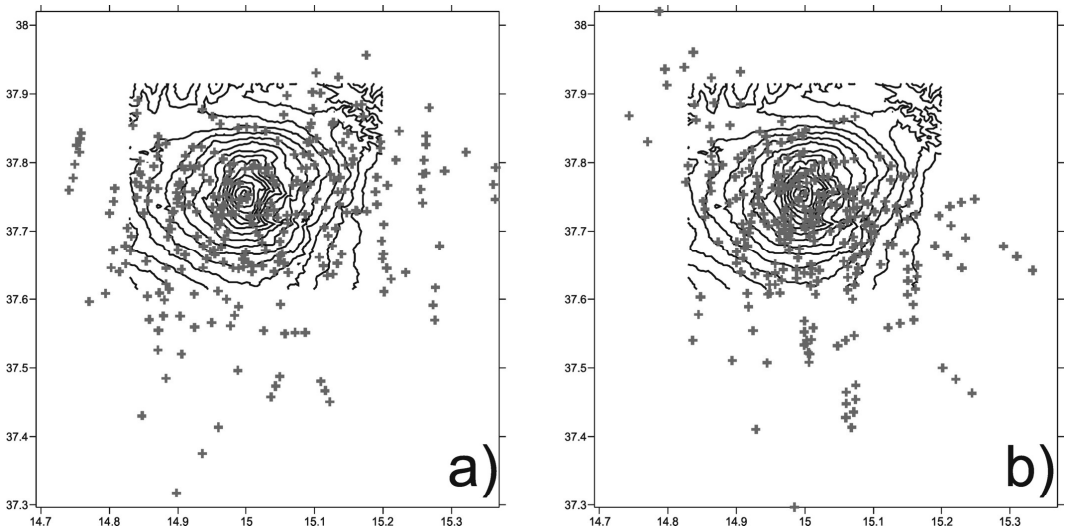


FIGURE 5.11: Plan view of GPS satellites shifted down positioning relevant to the two random days analyzed in the tests.

urations were analyzed in order to investigate the variability of the well-resolved knots (and consequently the well-resolved area) with satellite positions. Results are reported in two different ways: the first is numeric, the second is graphic; in the latter the obtained values are overlapped with the spread function. Before starting elaborations, it

is necessary to calculate the damping value for each elaboration. This value is selected empirically, by running a series of single iteration inversions aimed at exploring a wide range of damping values, and plotting data misfit versus model variance. We obtain a concave roughly hyperbolic trade off curve and the value of damping is selected on the maximum variation of slope curve (see fig. 5.9).

Since we want to investigate the atmosphere anomalies, we assume that these anomalies don't change for a time shorter than 30 minutes; this allows us to increase the number of observables of three times because we consider the satellite position 15 minutes before and after the fixed time. The position of satellites is obtained by reading the IGS precise ephemeris, which are provided with a sample rate of 15 minutes and an accuracy of about 5 [cm]. The satellite positions are then projected according to the procedure explained above; the results of this procedure, for the two selected days are shown in figure 5.11.

The first analyzed configuration includes 15 stations and 8 satellites, which imply 360 observables. figure 5.12a and 5.12b show the contour plots of the SF values of 2 and 2,5 (solid and dash line respectively) in the investigated area for the two configuration of satellite positions. We visualize data using a "natural neighbor" griding method. Differences between the spread function shapes are due to a different GPS satellites placement of the two investigated situations. The graphs reported on figure 5.13 and 5.14 show the values of the velocity, for the models relevant to the three well-defined shapes (e.g. points, cube, checkerboard) and the tomographic inversions, respectively. Obviously, as the inversion value to the model one is closest, as the result of the tomography to describe the actual structure is more suitable. fig. 5.14 reports the results as maps of gridded values for different layers, at different elevations. As visible in Fig. 5.13 and 5.14, the tomography provides information about the presence of atmospheric anomalies in the studied area, although the quality of the information is different by considering the different structures. In the inversion of the "point" and "bubble" structure scheme, for instance, the tomography produces a sort of "reverberation" of the anomalies in the adjoining layers but the maximum magnitude of variations are revealed in the right places. The test on the "checkerboard" structure clearly shows that inside places where SF is less than 2 the anomalies were well resolved, while outside the anomalies becomes blurred and unrecognizable. Furthermore, this last test shows that the image quality increases with the altitude of the considered layer.

The second analyzed configuration includes 30 stations and 8 satellites. In this case we may use 720 observables to compute the tomographic inversion. Fig. 5.15a and 5.15b show a plan view contour plots of the SF in the investigated area for the two configurations of satellites position. In accordance with the higher number of observables with respect to the case of 15 stations, the Fig. 5.16 and 5.17 suggest a better reconstruction of the anomalies. Although the reverberation of the anomalies in the adjoining layers is

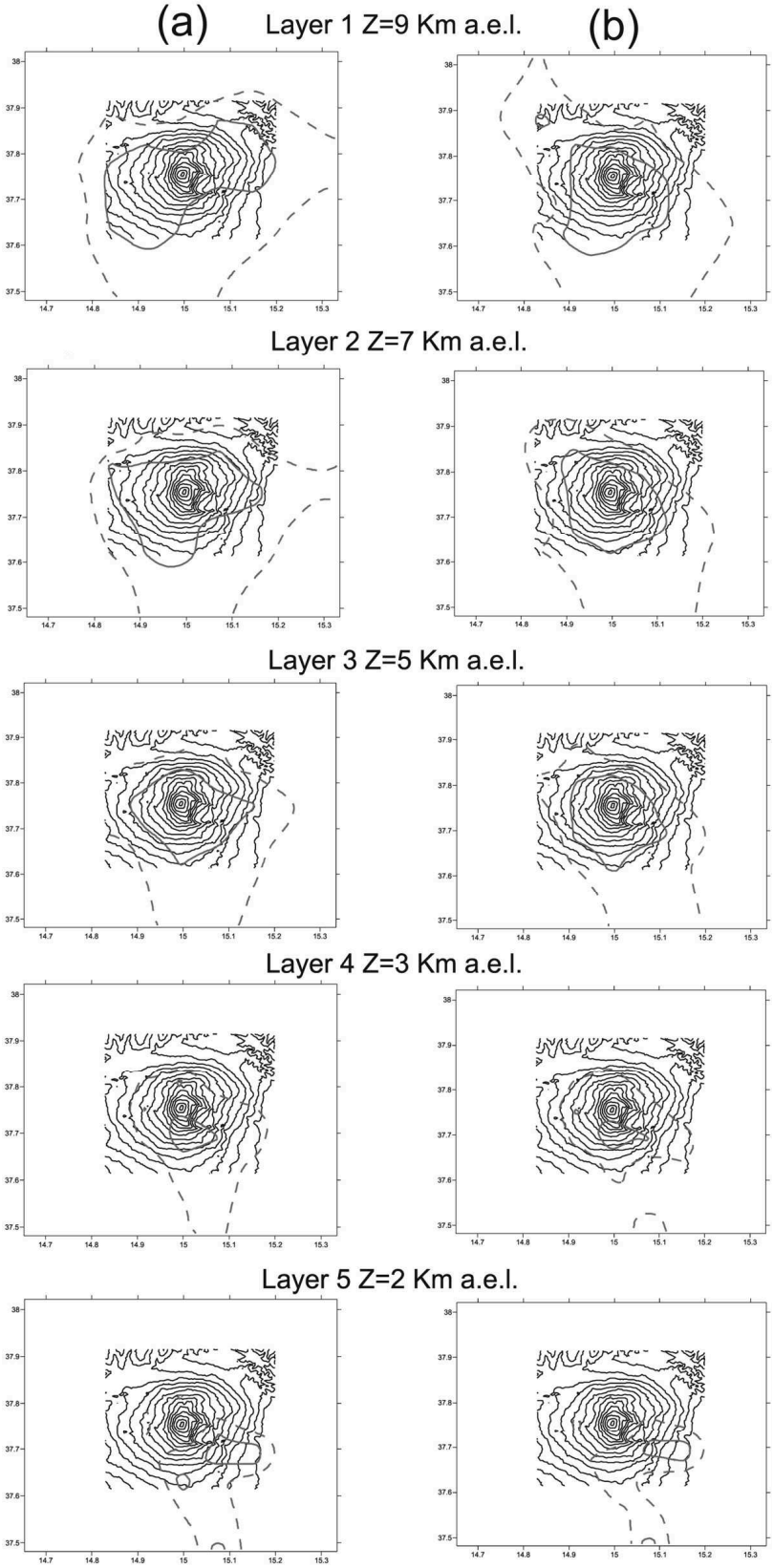


FIGURE 5.12: Plan view contour plot of the SF values obtained with 360 observables (15 stations), using the satellites spatial configuration reported: (a) in figure 5.11a, (b) in figure 5.11b. The solid lines represent the contour plot of the SF values 2, the dash lines, instead, SF values equal to 2,5.

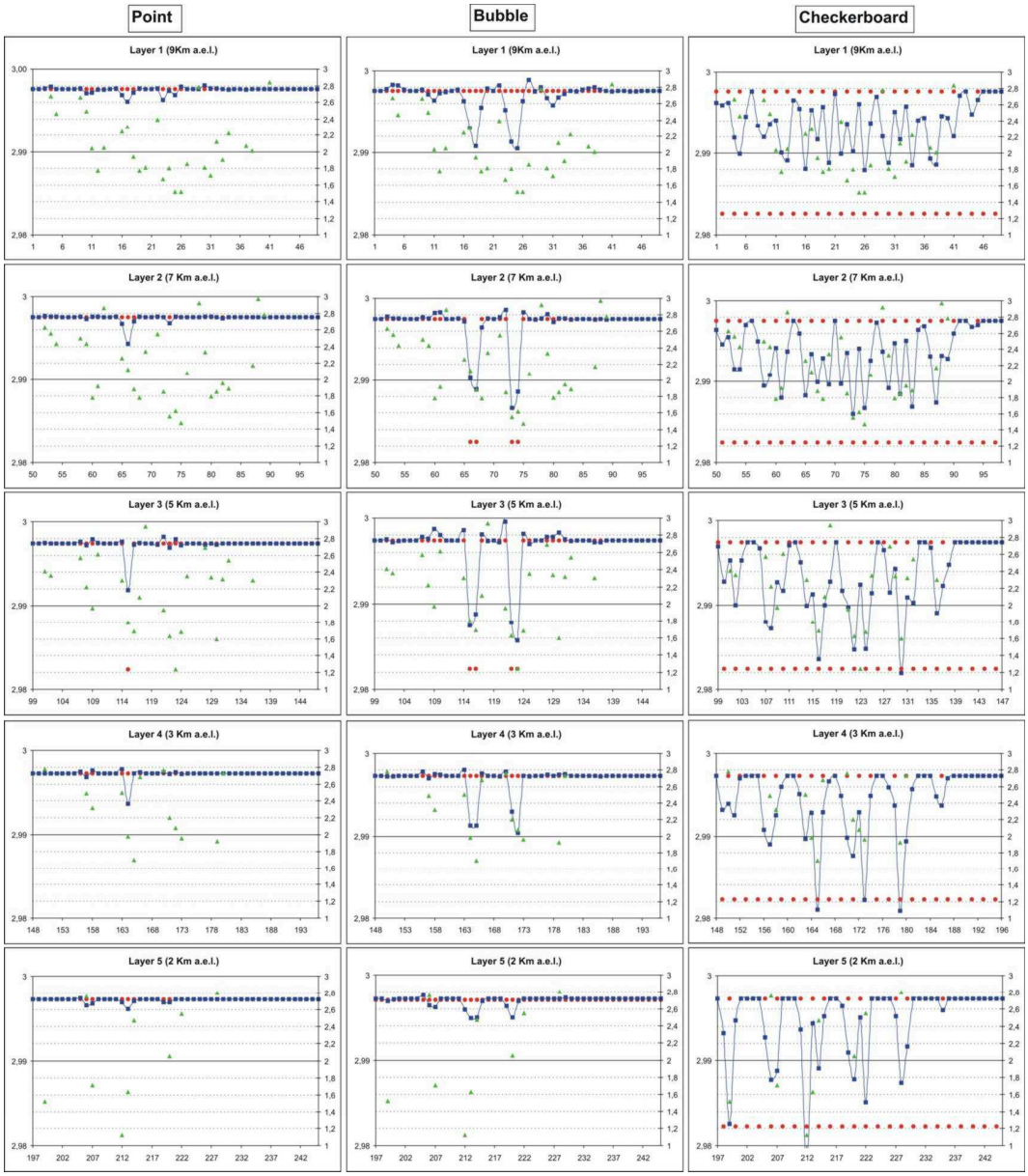


FIGURE 5.13: Results of the tests used to asses the tomographic inversions. On the abscissa axes are the i.d. number of the knots, according to the software criteria. On the primary axes the velocity values and on secondary axes the SF values (green points). The plots show the comparison between expected EM wave velocity (red points) and tomographically inverted velocity (blue points) obtained by using 360 observables (15 stations) and the spatial configuration reported in Figure 9.a. for the three shapes (e.g. points, cube, checkerboard).

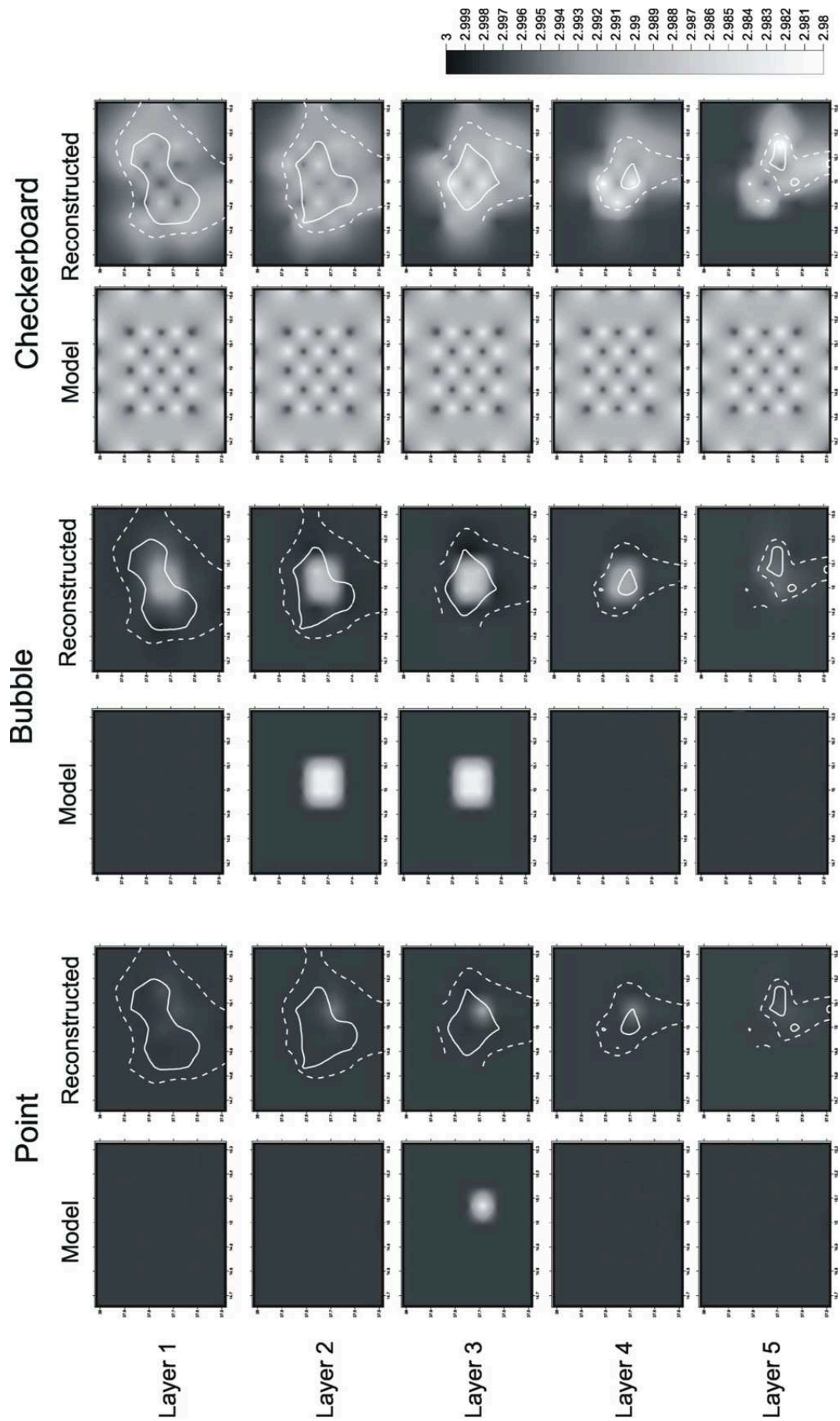


FIGURE 5.14: Tomographic images of the same results reported in figure 5.13. Each reconstructed image shows moreover the spread function contour plot. Dashed lines surround area with SF values less than 3, while continuous lines surround area with SF values less than 2.

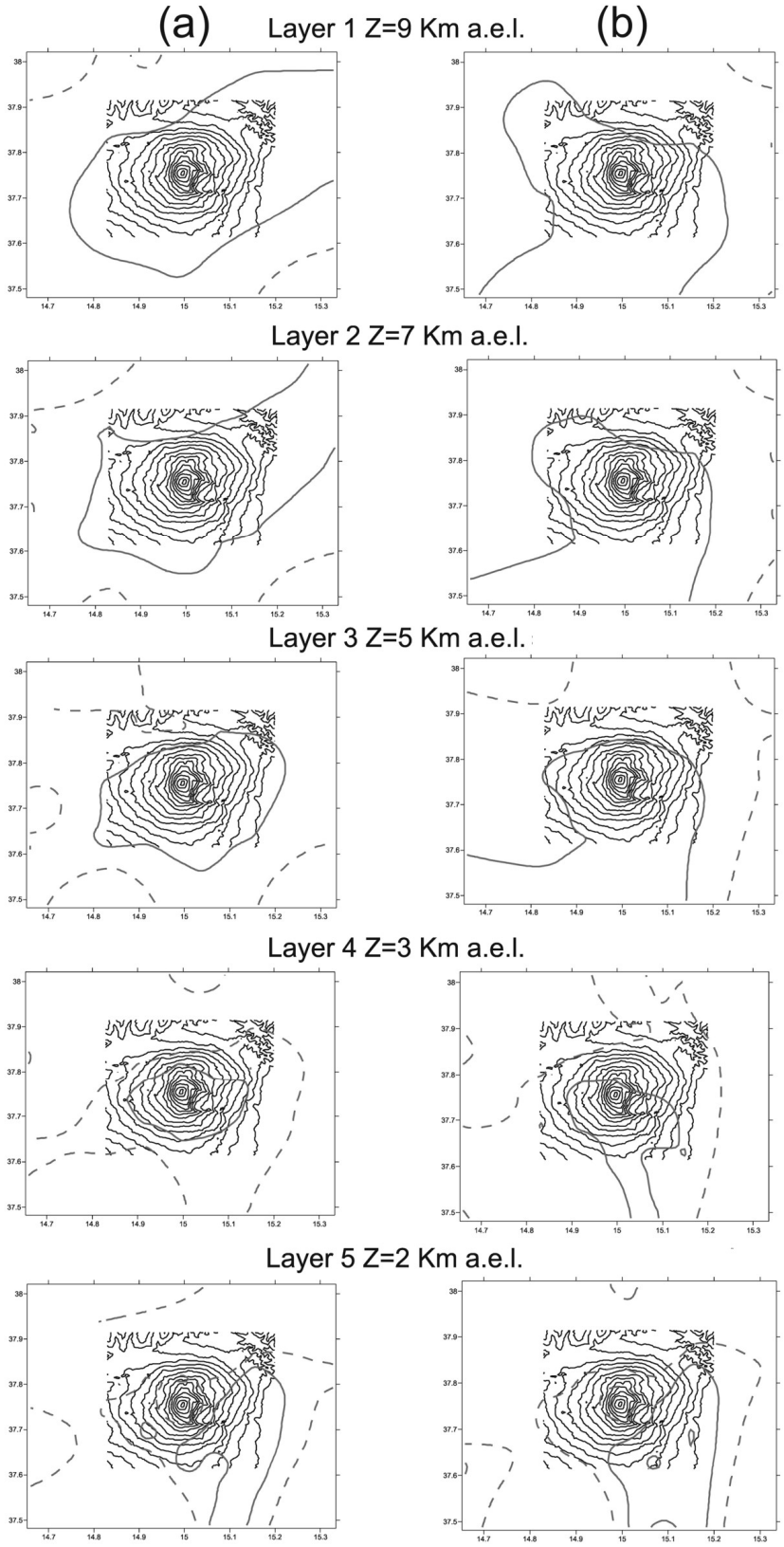


FIGURE 5.15: Plan view contour plot of the SF values obtained with 720 observables (30 stations), using the satellites spatial configuration reported: (a) in figure 5.11a, (b) in figure 5.11b. The solid lines represent the contour plot of the SF values 2, the dash lines, instead, SF values equal to 2,5.

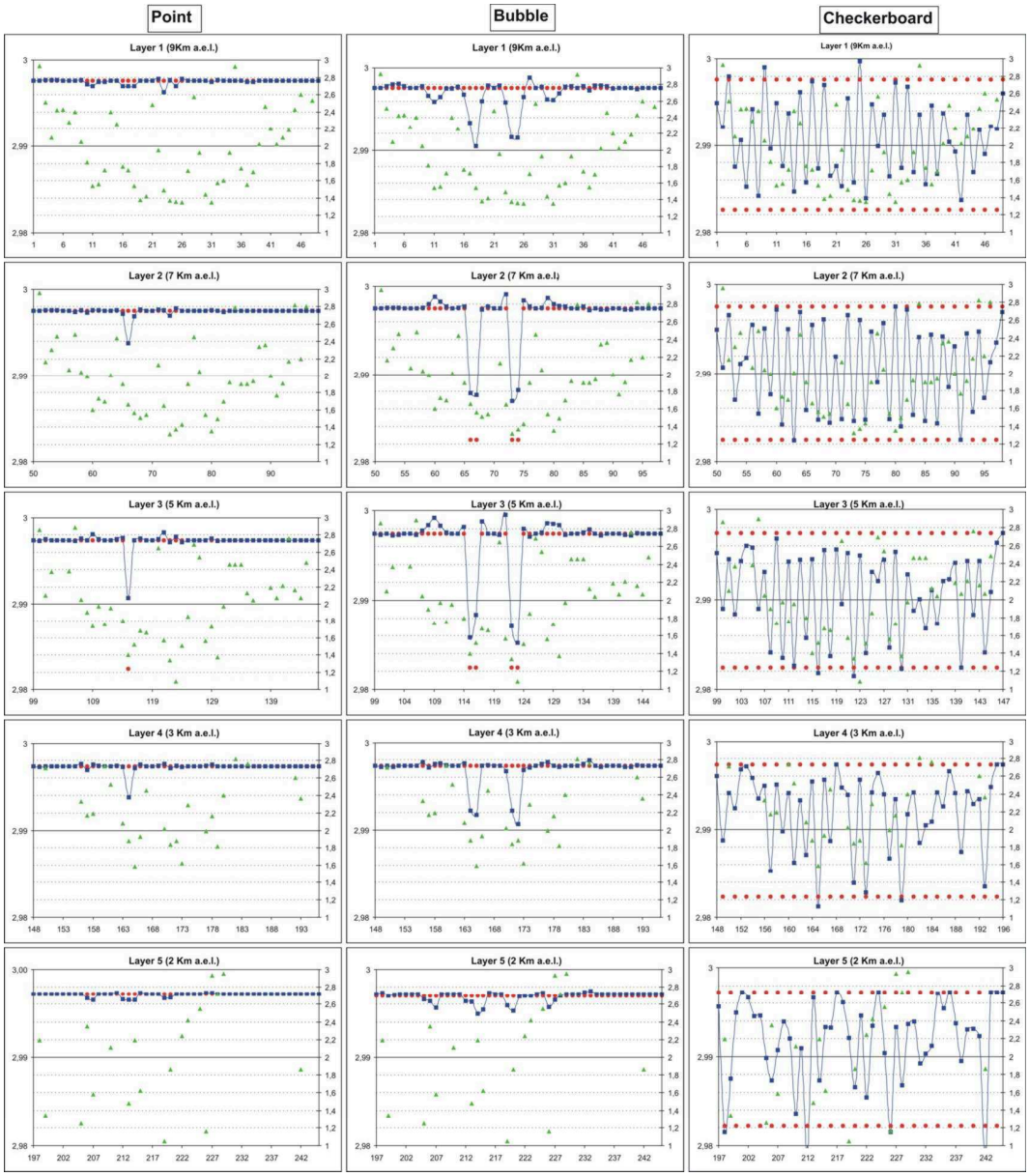


FIGURE 5.16: Numeric results of the test used to asses the tomographic inversion. On the abscissa axes are the i.d. number of the knots, according to the software criteria. On the primary axes the velocity values and on secondary axes the SF values (green points). The plots show the comparison between expected EM wave velocity (red points) and tomographically inverted velocity (blue points) obtained by using 720 observables (30 stations) and the spatial configuration reported in Figure 5.11.a. for the three shapes (e.g. points, cube, checkerboard).

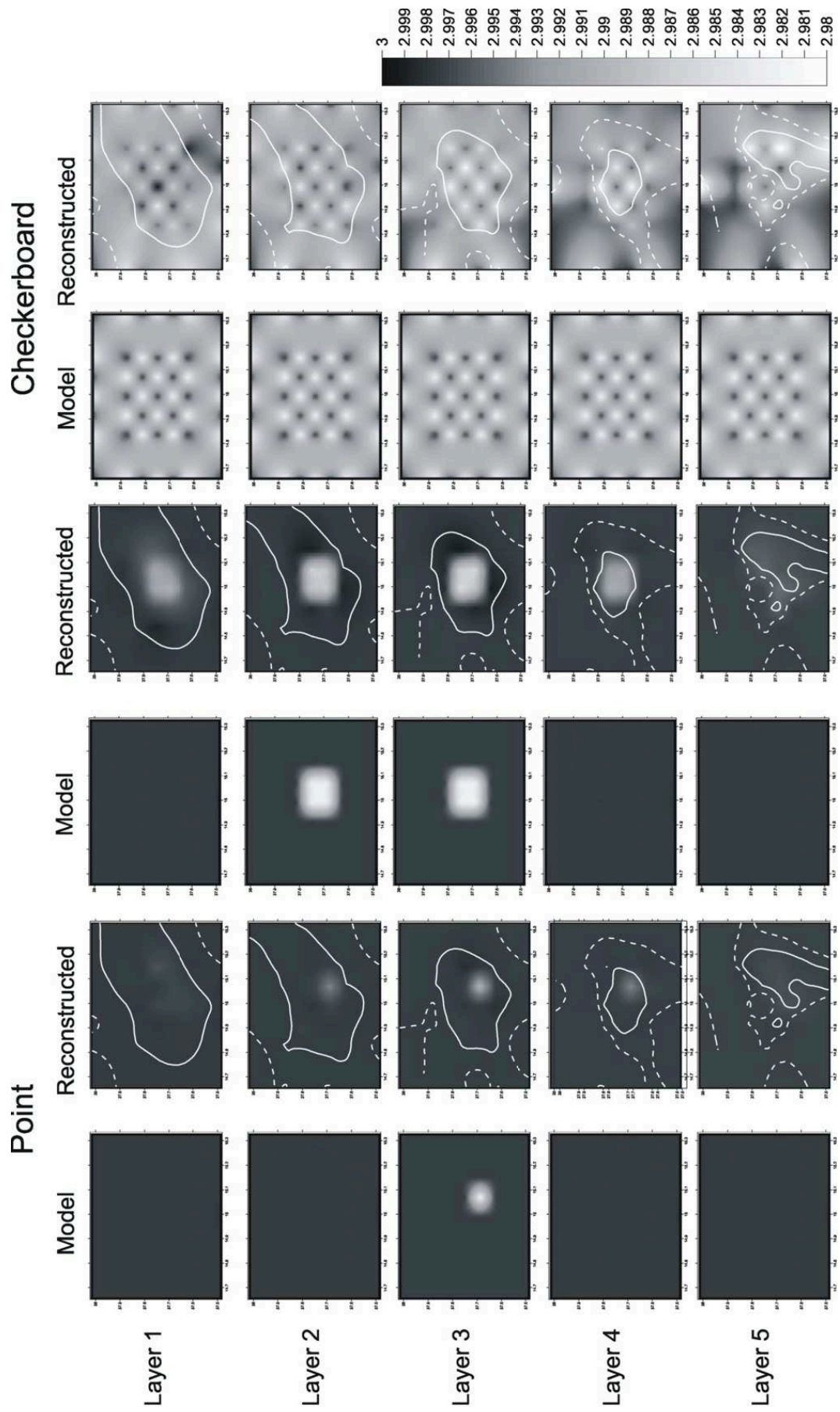


FIGURE 5.17: Tomographic images of the same results reported in Figure 5.16. Each reconstructed image shows moreover the spread function contour plot. Dashed lines surround area with SF values less than 3, while continuous lines surround area with SF values less than 2.

still present with the same magnitude of the previous case, the reconstructed values at different knots are generally closer to the model values than the previous corresponding tests. Furthermore, the contours of the spread function (Fig. 5.15a and 5.15b) are wider than those reported in Fig. 5.12a and 5.12b respectively.
As concerning the "checkerboard inversion" in Fig.5.17, we can confirm the right choice for cut off value of spread function adopted. For 720 observables case, we have a substantial improvement of the reconstructed anomalies. In fact, comparing Fig. 5.17 and Fig. 5.14, we can see a better overlapping of the reconstructed knots value with respect to the model knots value.
As a last test, we check the maximum input data that SIMULPS12 may process. In this case, we assume to process information provided by a configuration that includes 90 (Fig. 5.10c) receivers and 8 satellites. Although this test may be considered an "unrealizable test case" this is useful to understand the limits of the adopted software. The number of obtaining observables is about 2160 but SIMULPS12 can process a maximum of 1800 observables, so we delete data having greater value of residual until 1800 observables remain. Fig. 5.18 shows a plan view contour plots of the spread function in the investigated area for the processed configuration. The reconstructed values from this simulation are shown in Fig. 5.19 and 5.20. The reverberation of the anomalies in the adjoining layers is still present with the same magnitude of the previous cases. The value of reconstructed knots is still different from the real value and equal to the values found in the previous case. Even if the number of observables is much higher than the previous simulations, the growth of the area covered by the contour plot of the SF is not comparable with the increase of number of the receivers. This test suggests that, if we extend the number of receivers up to a lot of dozen, the increase in RS is lower compared with the cost of the survey and the computational efforts. The tests above described led to the conclusion that the value of 2 can be assumed as upper limit of the quality of the results, as the test on the checkerboard structure, reported on Fig. 5.17, shows.

5.3.5 Discussion

The main problem with the SIMULPS approach has been to adapt the seismic software to a atmospheric problem. In particular the main critical points are the GPS satellite shifting down, which was necessary to reduce the base of the investigated area, and the definition of 3D grid.
We have tested three different network configurations with 15, 30 and 90 GPS receivers respectively. Results of the tests show that the dimension of the well resolved area of tomographic images increases with the increasing number of GPS receivers but, in

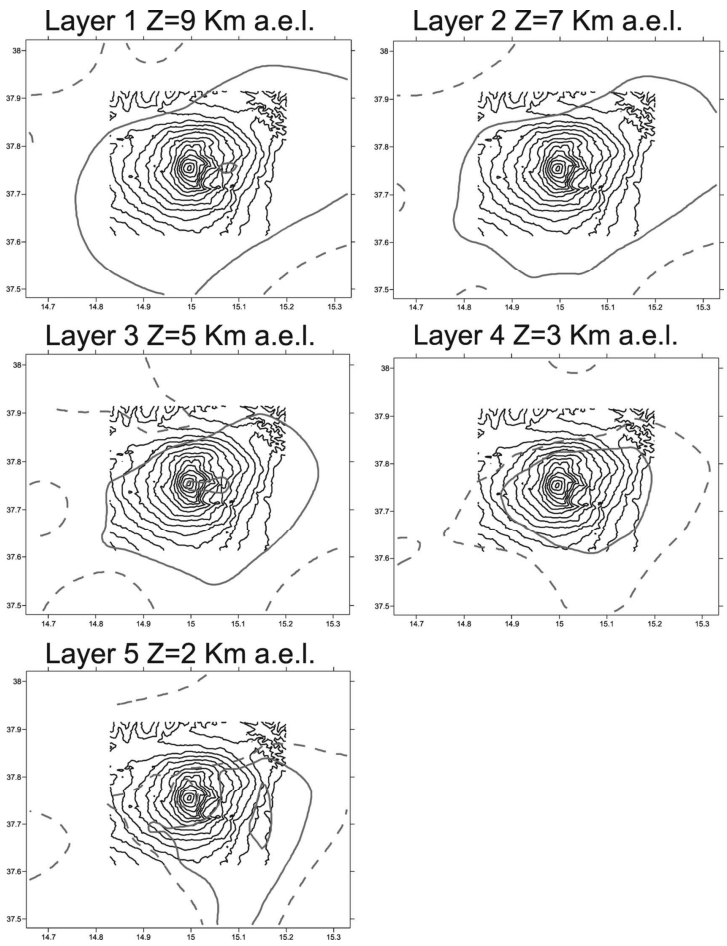


FIGURE 5.18: Plan view contour plot of the SF values obtained with 1800 observables (90 stations) using the first satellite spatial configuration reported in Figure 5.11a. The solid lines represent the contour plot of the SF values 2, the dash lines, instead, SF values equal to 2,5.

the last configuration with 90 GPS receivers (where the number of observables is much higher than the previous simulations) the growth of the area covered by the contour plot of the SF is not comparable with the increase of number of the receivers.

The test was also aimed to fix the a priori atmospheric model and the critical values of main parameters involved in the tomographic inversion as SF. Unfortunately, we realized that the use of SIMULPS did not give good results for these purposes. About the sensitivity of SIMULPS, indeed, although it is considered a milestone for seismic tomography, this software doesn't seem suitable for studies of the atmosphere . The changes that we may expect in the atmosphere, consistent with the studied physical model, can not be revealed by SIMULPS because they are too small of three order of magnitude. Nevertheless, the work carried out to retrieve the three anomalies and the results obtained applying SMULPS at atmosphere are important for many aspects. In any case results of the test on the number of stations and their disposal, and results of the test on the number and location of voxels, are still valid. Moreover we could demonstrate

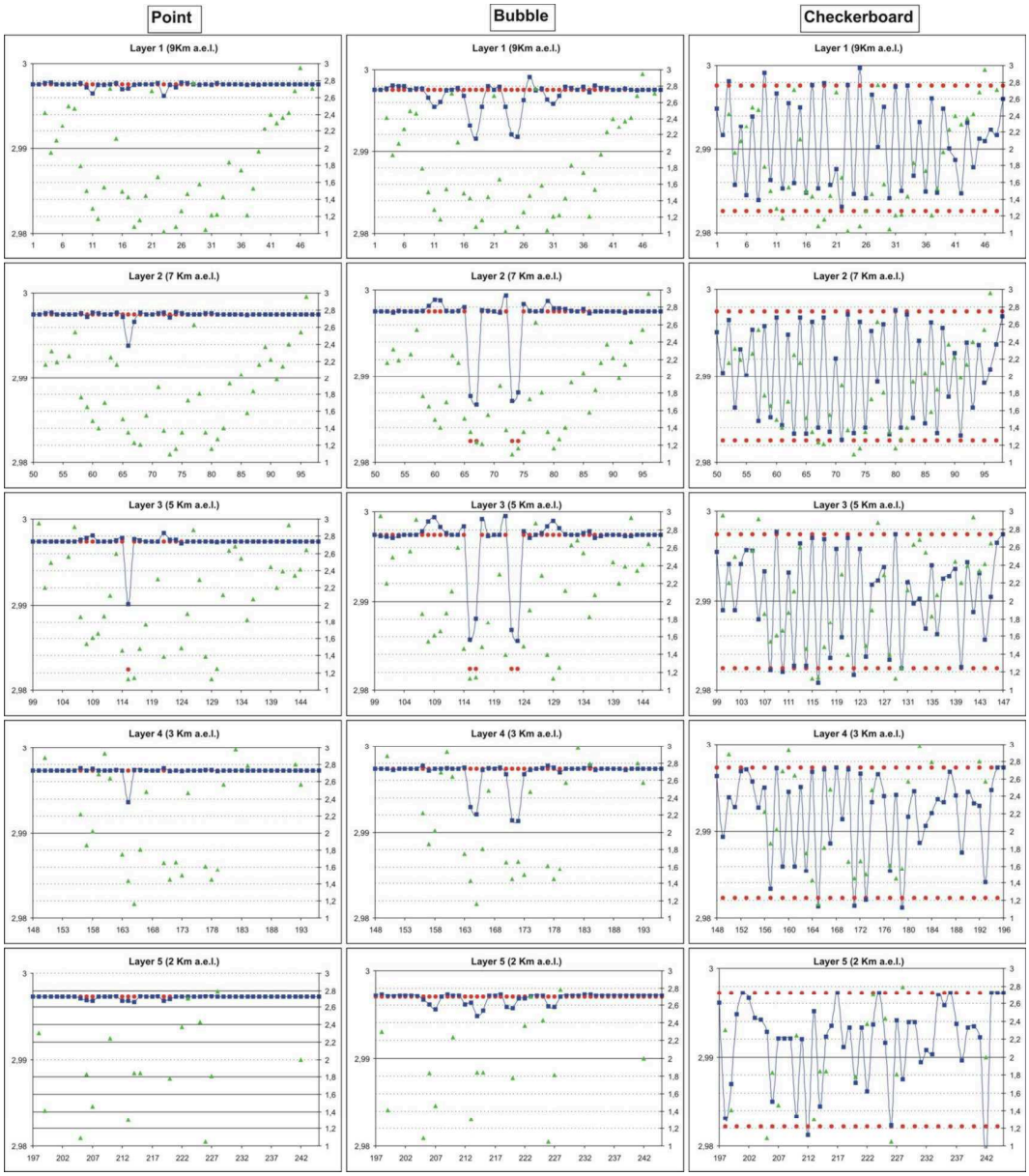


FIGURE 5.19: Numeric results of the test used to asses the tomographic inversion. On the abscissa axes are the i.d. number of the knots, according to the software criteria. On the primary axes the velocity values and on secondary axes the SF values (green points). The plots show the comparison between expected EM wave velocity (red points) and tomographically inverted velocity (blue points) obtained by using 1800 observables (90 stations) and the spatial configuration reported in Figure 5.11.a. for the three shapes (e.g. points, cube, checkerboard).

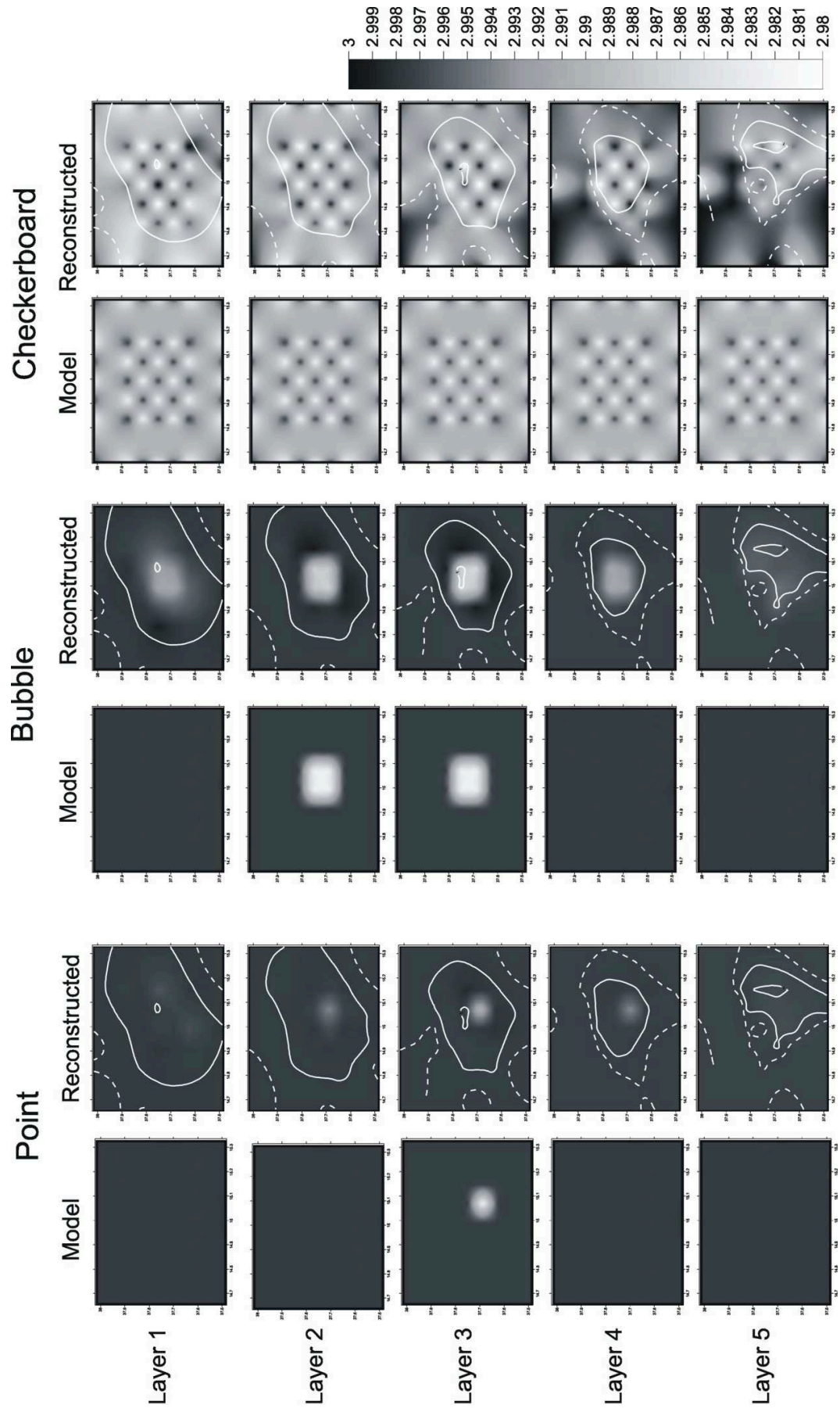


FIGURE 5.20: Tomographic images of the same results reported in Figure 5.19. Each reconstructed image shows moreover the spread function contour plot. Dashed lines surround area with SF values less than 3, while continuous lines surround area with SF values less than 2.

that the spread function is an irreplaceable tool for recognizing "good" or "bad" areas. Finally we have become familiar with the sensitive parameters for proper tomography and we have built an atmosphere model that can be used for more advanced studies, as reported in ... where a new code is presented, written for tropospheric tomography on the basis of the acquired experience and results.

5.4 MATLAB approach

The experiences and knowledge gained in the previous section were used for developing the tropospheric tomography software designed and implemented entirely in the context of this PhD. The development environment chosen for this purpose is MATLAB.
MATLAB, which name stands for MATrix LABoratory, is a technical computing environment for high-performance numeric computation and visualization. MATLAB integrates numerical analysis, matrix computation, signal processing, and graphics into an easy-to-use environment where problems and solutions are expressed just as they are written mathematically, without much traditional programming. We used MATLAB in order to acquire, computing and illustrate the GPS atmospheric data to obtain a tomography results.

5.4.1 The code

We suppose that the best way to explain how the software works is to show a simplified flow chart of the code. We want to emphasize again that the scheme has been traced starting from the SIMULPS experience. The need to rewrite the code allowed us to better understand the algorithms and problems related to the computation of tomography. In the future we hope to deeper understand this technique and apply more efficient resolution algorithms . The figure 5.21 is a summary, by sectors, of the code wrote during this PhD. We compose a conceptual flow chart of the software and added some numerical references to explain the blocks separately.

1) The geometry
The first few lines of code, not shown in the figure, are to specify the day and time that we want to elaborate. In general, to solve any inverse problem, it is necessary to know the geometry of the investigated system. As explained in chapter 3, GAMIT elaboration produces different output data, most of them are used in the tomography. Unlike other output data that are at higher frequency, GAMIT output for GPS receiver position consists in a daily solution. Other methods exist for estimating the position of a GPS station with different frequencies, however, for our purposes, this data frequency

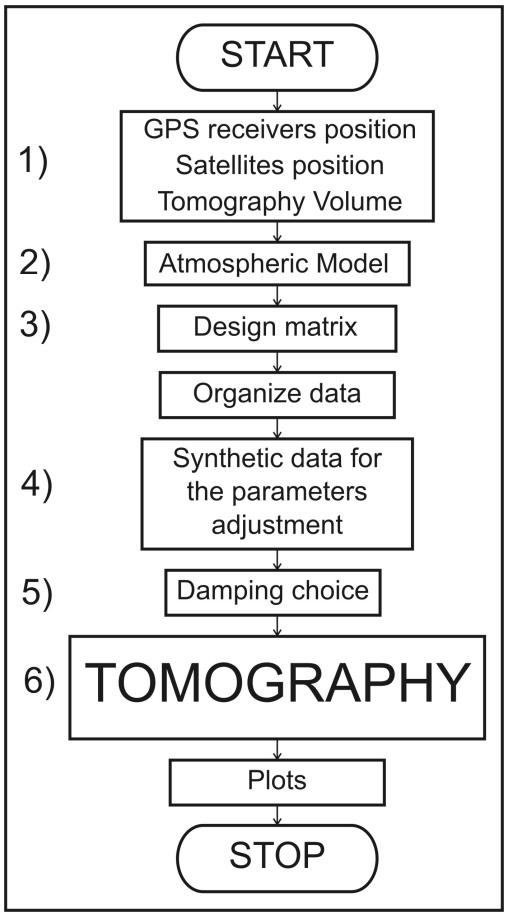


FIGURE 5.21: The flow chart of the software developed during this PhD

is sufficient. The GPS receiver positions, together with their errors, were read automatically from GAMIT output data file and located in the workspace of MATLAB. A typical errors in the precise daily position for each GPS receiver are shown in table 4.3. Together with the receivers, the satellite positions must be imported too. GPS satellite positions can be obtained in two ways: the first is to download an ASCII file (*.SP3) containing the data calculated from IGS site, the second way consists to read the *.DPH files coming from the output of GAMIT. Although the seconds are computed from *.SP3 files, the *.DPH files contain also the LC-undifferenced phase residuals that have been used in chapter 4, sec.4.4. As regards the position of the satellite, the main difference between the SP3 and a DPH file is that: the SP3 file contains the precise ephemeris, that is the coordinates of all GPS satellite with a frequency of 15 minutes and an error of 5[cm]; instead the DPH file contains the position of the linked satellite at each epoch (30 s). GAMIT interpolate data, according with Keplerian parameters, giving back the azimuth and elevation of satellite. Figure 5.22 shows an example of a *.DPH file.

The script to load the *.DPH files was wrote for the purposes of chapters 3 and 4. The satellite positions were read, converted in cartesian coordinate with respect to the GPS

```
* Clock information for site BRON receiver LEI . PRN 12
* Epoch  L1 cyc   L2 cyc   P1 cyc   P2 cyc   LC cyc   LG cyc   PC cyc   WL cyc   N cyc  LSV   Azimuth    Elev  PF
     1   -33.13   -40.43     2.84     0.16   -4.127   -29.00     6.91    -6.93   -95.14   11  102.3088  32.1206 1
     2   -30.15   -38.49     2.25     1.23   -0.418   -29.73     3.29    -7.90   -76.86   11  102.5090  31.9583 1
     3    -1.13    -1.47     2.24     1.99    0.043    -1.17     1.75     0.18    -4.59   11  102.7085  31.7958 0
     4    -1.17    -1.52     1.78     2.16    0.031    -1.20     0.26     0.14     0.31   11  102.9075  31.6332 0
     5    -1.22    -1.55     1.77     1.01   -0.035    -1.18     2.51     0.02    -8.91   11  103.1058  31.4705 0
     6    -1.25    -1.59     1.86     0.46   -0.024    -1.22     3.83    -0.05   -14.13   11  103.3036  31.3076 0
     7    -1.26    -1.62     0.63     0.77    0.001    -1.26     0.08    -0.18    -1.76   11  103.5008  31.1446 0
     8    -1.29    -1.65    -0.55     0.40   -0.021    -1.27    -2.20    -0.37     4.75   11  103.6974  30.9815 0
     9    -1.34    -1.72    -1.07     1.74   -0.004    -1.34    -6.18    -0.30    19.58   11  103.8934  30.8182 0
    10    -1.39    -1.76    -0.03     0.57   -0.028    -1.36    -1.19    -0.31     1.63   11  104.0888  30.6549 0
```

FIGURE 5.22: An example of *.DPH file generated by GAMIT. The columns 12 and 13 illustrate the azimuth and elevation data referred to the epoch of column 1

receiver and finally conveted in UTM. At the end of this process they are located in the workspace of MATLAB.

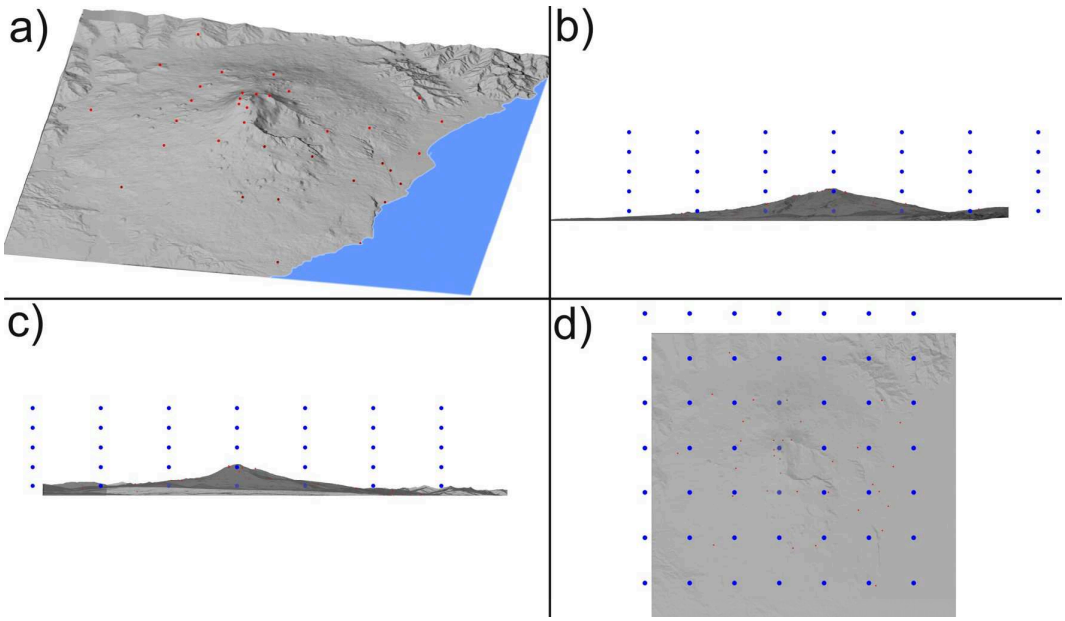


FIGURE 5.23: An example of the investigated volume. Measurements of the box are $49 \times 49 \times 10$ [Km^3], for a total of 245 voxels ($7 \times 7 \times 5$). a) A synoptic view of the GPS receivers position in the test case; b) view of the East flank of Mount Etna, the blue dots represents the centre of the voxels; c) view of the Sud flank of Mount Etna, the blue dots represents the centre of the voxels; d) a Plan view of the Mount Etna, the blue dots represents the centre of the voxels.

To fix the geometry, it is now necessary to specify the characteristic of the tomographic volume. In particular, the center and the dimensions (length, width and height) are inputs of the operator. According to the tests carried out by using SIMULPS, we will use the same grid above. Thus, the geometric approach is the same compared with the SIMULPS tests: the dimension of the box are $49 \times 49 \times 10$ [Km^3], for a total of 245 voxels ($7 \times 7 \times 5$). Figure 5.23 shows the investigated tomographic volume. The voxels are placed in the investigated geographical of Mount Etna area and numbered in progressive way from south to north and from west to est, starting from the lower layer.

2) Atmospheric model

There are not difference in the determination of the atmospheric model with respect to the tests with SIMULPS but the way to initialize the parameters is different. We created a database by using the values of pressure, temperature and humidity collected from the weather station placed at the Physic Department, Laboratory of environmental physics, University of Catania. The humidity data are collected with a precision of $\pm$ 1 %, the atmospheric pressure data of $\pm$ 0.1 [hPa]%, and the temperature of $\pm$ 0.1 [$^{\circ}C$]. The database is organized in columns containing: hour, minute, second, day, month,

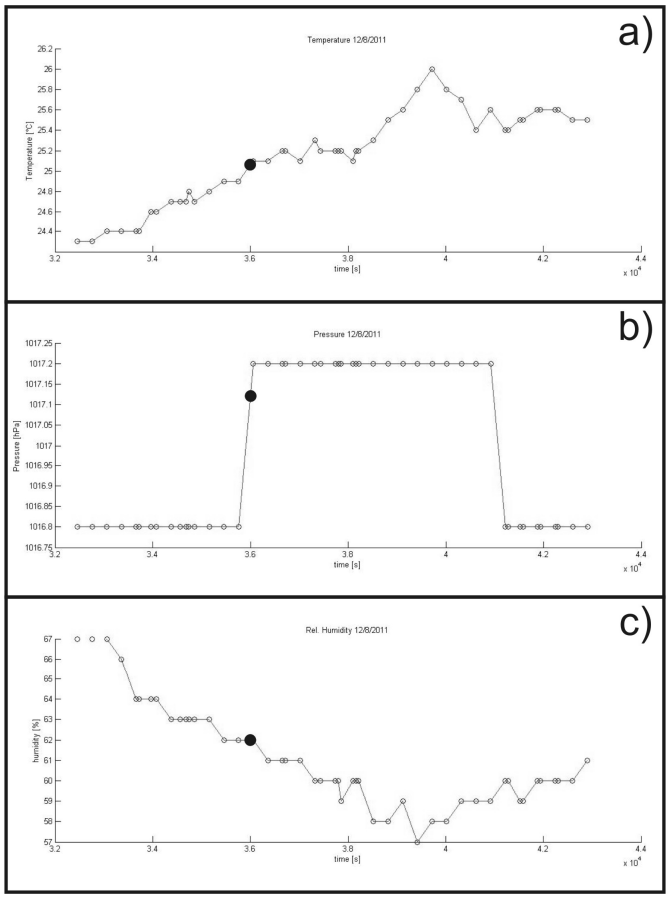


FIGURE 5.24: The 12 August 2011 atmospheric data values from 9 to 11 a.m. The humidity (a), pressure (b) and temperature (c) data have been collected from the weather station placed at the Physic Department, Laboratory of Geophysics and physical environment, University of Catania

year, pressure / humidity / temperature value, time (in [s]) from the first of January of the chosen year. Each year consists of 3 files loaded in the MATLAB workspace. As the epoch in which we want to obtain the tomography is set, the software extracts the atmospheric data from the database. If the intersection of the choice epoch and the meteo data point returns null value, the values of pressure, temperature and humidity are obtained by linear interpolation of the two closest values. Figure 5.24 shows an example of the extracted data.

The extracted values are used to compute the refractivity profile according with the equation 5.17 and in general following the procedures of sect.5.3.2. We have chosen to use the atmospheric parameters recorded by a weather station above sea level in order to start the tomography with a better atmospheric model (in a different way respect to the section 5.3.2). The future perspective is to use data from weather balloons, if available, otherwise those of the ground stations. Figure 5.25 shows the wet refractivity

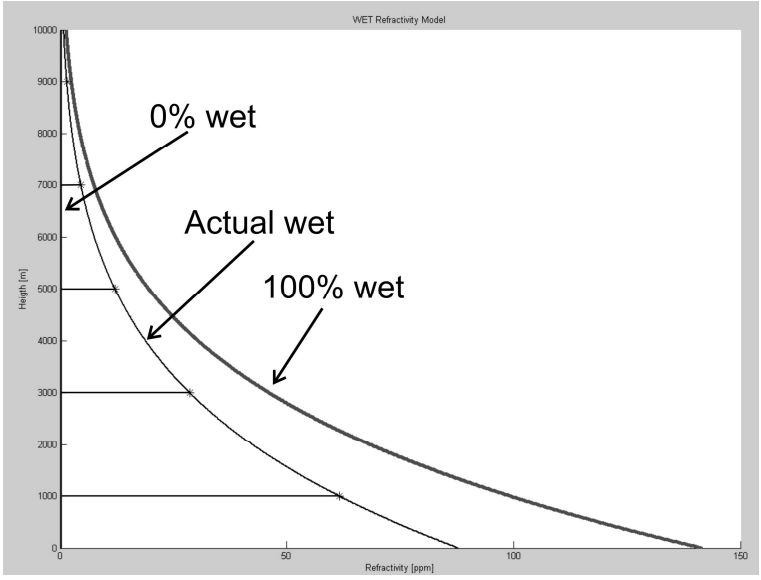


FIGURE 5.25: Three profiles are reported: the first is computed through the actual atmospheric data, the second assuming that the humidity value is always maximum, the third with zero humidity. The black stars imposed over the actual wet profile represents the values of refractivity associated with the layer of the volume builded in the previous section.

profiles. By using the equation 5.17 and the atmospheric data, we compute three different profiles: the first assuming that the humidity value is always maximum, the second is calculated with the actual atmospheric data and the third is zero because from zero humidity assumption. Since the refractivity profiles will be used for the tomography, it is necessary to associate at each voxel a value of wet refractivity. A layered model for the refractivity field is used, for this reason the refractivity values will be calculated only along the z. The black stars imposed over the actual wet profile represents the values of refractivity associated with the layer of the volume build in the previous section. The error was calculated by using the error propagation equation applied to the 5.17.
However, along with the profiles that we have calculated, another refractivity profile called "perturbed profile" is computed. The reason why we need also this "perturbed profile" will be explained later. The anomalies are obtained perturbing the refractivity value of randomly choice voxels by MATLAB. The maximum magnitude of the perturbation is fixed by the user as a percentage of the difference between the limits of the refractivity profiles in the considered layer (derived from figure 5.25). Figure 5.26 shows

an example of the refractivity induced by MATLAB. Dash lines represent the range of the wet refractivity limits, the dash dot line instead is the perturbed refractivity induced by MATLAB.

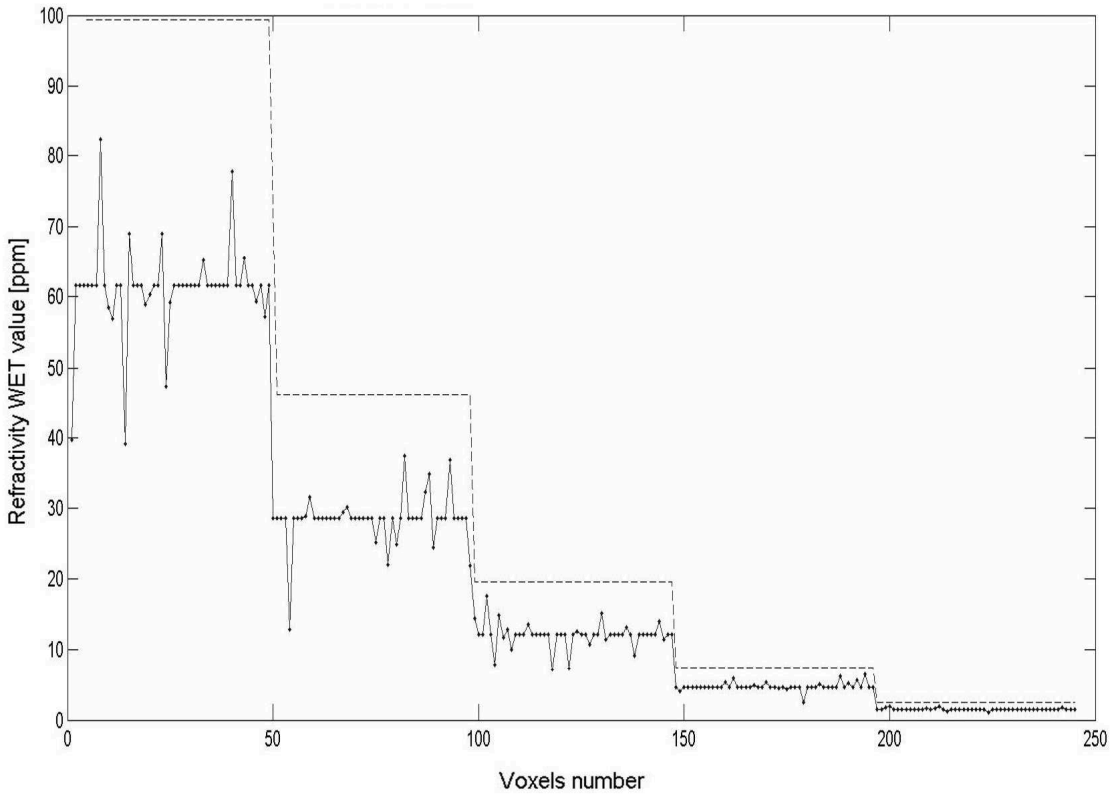


FIGURE 5.26: The wet refractivity anomalies induced by MATLAB. Dash lines represent the range of the wet refractivity, the dot line, instead, are the refractivity perturbed values induced by MATLAB.

Together with the refractivity random model, we introduce three different refractivity structures (as assumed in the previous section):

- the *point model* in which the variation of a single voxel is computed. The maximum magnitude of the perturbation is fixed by the user but, the number of choice voxel per layer is random;
- the *bubble model* in which the variation of eight neighboring voxels is computed. The variation is fixed by the user by choosing their numbers of voxel and the percentage of variation with respect to the admitted range;
- the *random model* in which the code establishes how and how many voxels must be varied, the user chooses only the maximum percentage of variation with respect to the admitted range. This is an evolution (complication) with respect to the "previous checkerboard" structure.

3) The Design Matrix calculation

The discretization of Slanth Wet Delay is defined by the equation 2.136, that we write again:

$$d_i = \sum_{j=1}^{n_{vox}} A_{ij} m_j \tag{5.24}$$

Thus, to determine the refractivity field, distances covered by the signal in each voxel must be determined (A_{ij}). Given the voxel of the current GPS station along with the azimuth and elevation angle of the satellite, we can determine the signal path. For the estimation of distances in each voxel of the grid, the software works on a single couple receiver-satelite and iterates the procedure for the number of satellite and antenna pairs previous selected. This is done through a MATLAB function which is shown in the figure 5.27 as example.

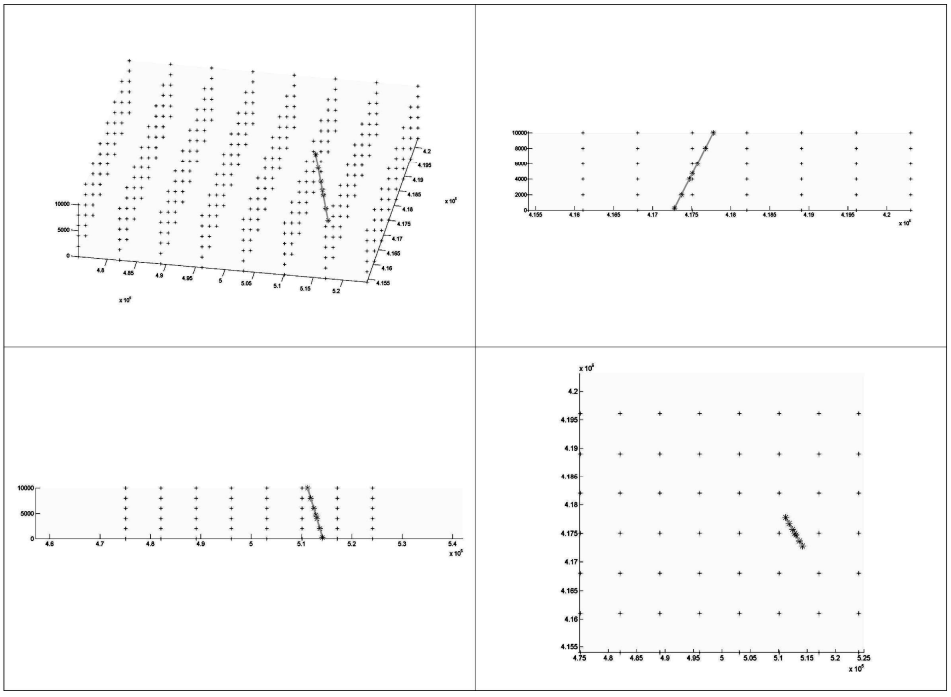


FIGURE 5.27: 3D view and prospects of the intersections between satellite-antenna pair and the voxels.

For each couple of satellite and antenna pair, the software determines the start and the end points in the tomography volume. The coordinates of the intersection between the segment and all the planes inside the tomographic volume are computed and are loaded in a temporary matrix, organized and filtered. Finally, for each crossed voxel, the segment length is computed and loaded in the design matrix (A_{ij}). For each ray path the determination of the segment length (distance) runs until the ray goes out of the grid either on its top or crossing its boundary. The latest observation is important

when we want to write observation equations that can be subdivided in the following two categories:

- rays that leave the grid on the top,
- rays that leave the grid crossing the boundary.

Referring to the first group, we use the equation 2.136. As regards the second, instead, ZWD is influenced only by a thick layer of 10 Km from the Earth surface and, in this volume, we can find all the water vapor of the atmosphere. Since we compute the tomography in a limited space and there is no way to separate the part of the measured GPS ZWD collected inside or outside the investigated volume, rays that leave the grid crossing the boundary, are excluded.

4) Synthetic data for the parameters adjustment
This block of code is very important because, along with the next block, allows to fix a fundamental parameter for tomography: the damping value. During the test phase also, we create synthetic variables useful for the tomography. The variables calculated in this block were obtained using models created in block 2. To facilitate reading, let's start by saying that the synthetic quantities calculated are three: the SW_{clean}, the $SW_{perturbed}$ and the SW_{noise}.
The SW_{clean} has been obtained through the discretization of the slant wet delay, defined by the equation 2.136, which can be written for our purpose as:

$$(SW_{clean})_i = 10^{-6} \sum_{j=1}^{n_{vox}} A_{ij} (N_{actual})_j \tag{5.25}$$

where $(SW_{clean})_i$ represent the Slanth Wet Delay value (in [m]) of the i-th ray inside the tomography volume, A_{ij} is the design matrix computed in the block 3 and N_{actual} represents the refractivity profile, called "actual wet", of the block 2.
The variable $SW_{perturbed}$ has been obtained through the following equation:

$$(SW_{perturbed})_i = 10^{-6} \sum_{j=1}^{n_{vox}} A_{ij} (N_{perturbed})_j \tag{5.26}$$

where $(SW_{perturbed})_i$ represents the perturbed slant wet delay value (in [m]) of the i-th ray inside the tomography volume. Indeed, coming from $N_{perturbed}$, $SW_{perturbed}$ it contains the effect of the perturbation on SWD, by assuming that we choose the shape of perturbation from those itemized in the point 2. This variable has a key role in the

choice of the damping factor.
The variable SW_{noise} has been obtain from $SW_{perturbed}$ as:

$$SW_{noise} = SW_{perturbed} + noise(\%, n\sigma) \tag{5.27}$$

adding a noise to simulate the GAMIT output data: Typically the errors of GAMIT on SWD is about 5 % of value. To generate the noise we start from a series of numbers normally distributed ($\mu = 0, \sigma = 1$) having the same size of $SW_{perturbed}$ array, modified according with the user entry. The user, in fact, can fix the percentage (respect to $SW_{perturbed}$) of noise and the nσ, that is how many σ within the dispersion of the noise has to be limited.

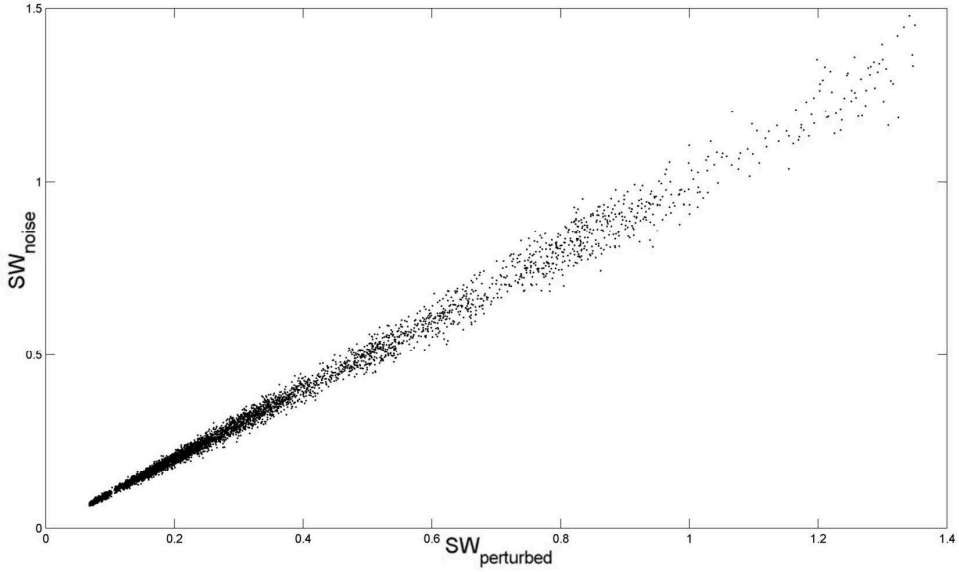


FIGURE 5.28: The comparison between $SW_{perturbed}$ and SW_{noise} is shown. SW_{noise} is obtained from $SW_{perturbed}$, by adding a noise of 10% within 2σ

Figure 5.28 shows an example of the perturbed array by adding a noise of 10% within 2σ. To highlight the noise contribution, we plot $SW_{perturbed}$ versus SW_{noise}. The dispersion of points around the first bisector is the effect of noise on the signal. Finally we calculate $\Delta SW_{synthetic}$ as the difference:

$$\Delta SW_{synthetic} = SW_{noise} - SW_{clean} \tag{5.28}$$

Which isolates the contribution of the slant wet delay due to the perturbation. If there are no anomalies in the troposphere, $\Delta SW_{synthetic}$ will have zero mean. This quantity will be used to perform the tests of tomography and to calculate the damping value in the actual tomographies.

5) The damping choice

As widely discussed in the section 5.2, the fact that in actual experiments a finite amount of data is available to reconstruct a model with infinitely many degrees of freedom, necessarily means that the inverse problem is not unique in the sense that there are many equivalent models that explain the data. It does not make much sense to make a physical interpretation of a model without acknowledging the errors and the limited resolution of the model. The reason why the estimated model differs from the true model is the non-uniqueness of the inverse problem that causes several (usually infinite) models to fit the data. Technically, this model null-space exits due to inadequate sampling of the model space. Moreover most inverse problems are ill-posed (partly underdetermined and partly overdetermined) and ill-conditioned, which produce large null-spaces and non-unique solutions. Regularization is thus needed, but there is a large ambiguity in its choice, which reflects the fundamental difficulty that one have to faces in solving inverse problems. Starting from the equation 5.14 the solution always requires solving a set of linear algebraic equations in which the parameter γ allows the regularization of the matrix and has to be fixed from user. This is what we will do in this block of code. We will derive the parameter in two different ways: the first using the method of L curves, the second trying to improve the estimation by a method developed during this PhD.

A popular method for determining γ value is plot the curve:

$$\mathcal{L} := (\log||y - Bx_\gamma||, \log||x_\gamma||) : \gamma \geq 0. \tag{5.29}$$

This curve is commonly referred as the L-curve, because for many linear discrete ill-posed problems, the graph of the curve looks like the letter "L". To choose the value of the parameter γ that corresponds to the point at the "vertex" of the L-curve, i.e. the point on the L-curve with curvature of largest magnitude. To determine a suitable value of the regularization a numerical approach is adopted, by computing several points on the L-curve (250 in our study) and evaluating the curvature of the L-curve at these points. An example of the L-curve obtained during the tests is shown in figure 5.29, from with is quite clear that there is no a unique value for the damping factor. This led us to look for another method for choosing the damping value. However, the estimation of the L-curve may suggest the range in which the damping value should be contained. The second method was developed for a better choice of the damping factor. Starting from the non unique values obtained with L-curve, we focus the attention in that limited range. To explain the algorithm for a fine tuning of the damping factor we use figure 5.30a. This figure shows a schematic flow chart of the algorithm. Initially we define two arrays: "percentile"(5×1 elements) and "damping" (n×1 elements), the first contains

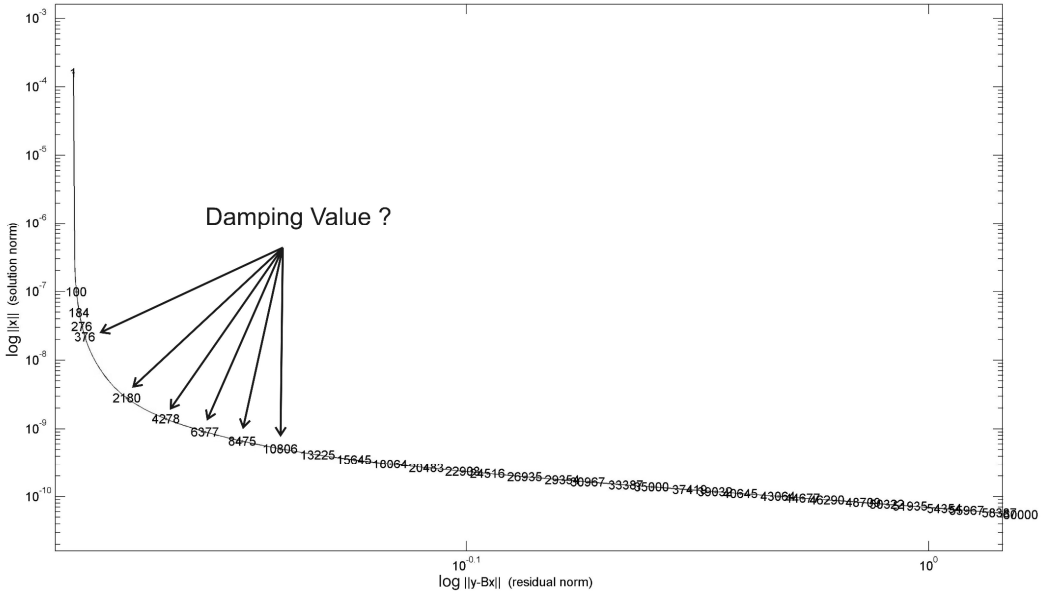


FIGURE 5.29: The L-curve computed to choice the damping factor. The solution norm versus the residual norm is plotted. We cant choice a unique value of damping factor for regularization.

five values of percentile, the second is a set of damping factors. The damping array size is equal to the number of damping values we want to test during processing, in the example 200 (from 376 to 10806), as shown in figure 5.29. Then start two nested loops: the outer directs the array "percentile", the inner the array "damping". Referring to one loop: MATLAB routine computes the resolution matrix (R) and the refractivity field (N_{calc}) by using the value of damping for the regularization of the matrix B. Then computes the spread function for each knot. After having defined the set of best SF values, we take the corresponding R_{jj} values. The spread function values are filtered in the following way: we taking into account the best values of SF, that is the values less than the fixed percentile of the specified loop.

Then, the sum of the differences between the calculated (N_{calc}) and the noised (N_{noise}) model knot values, normalized to the number of samples, is computed. This parameter allow us to perform the fine tuning of the damping factor. The motivation behind this method is that we suppose to know the true refractivity field by simulating a physically acceptable distribution of the field of refractivity, by using the previously explained N_{noise} model. In this way we evaluate, for difference damping values, the discrepancy between the expected and the calculated refractivity field.

The procedure above explained is applied iteratively (5.30a) for five subset of SF population, selected according the 10, 15, 20, 25, 30 percentiles. At the end of this process we obtain the matrix called "choice", of n×5 elements. The matrix "choice" is plotted in figure 5.30b. The minimum of the curves represents the damping value to obtain the best tomography, respect to the simulated refractivity field. Based on the experience,

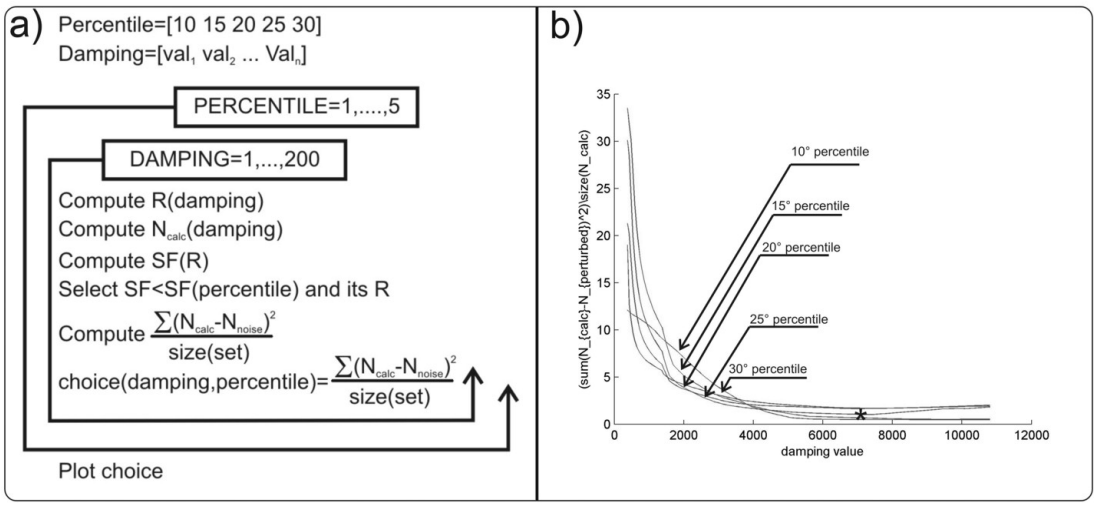


FIGURE 5.30: The method for a fine damping tuning is shown; a) the schematic flow chart of the algorithm used too detect a better damping value; b) the plot of the curves obtained following the idea behind

the percentile value that best represents a real case is between 20 and 25. In the choice of the damping factor, the two curves that have a greater weight are the 20 and 25 percentile.

Finally we try to obtain a damping value without the user choice. This allows us to automate the choice of damping without waiting for input by the user. To do this, we calculate the mean of the curves in figure 5.30b and find the absolute minimum of the new curve. Figure 5.31 shows the result.

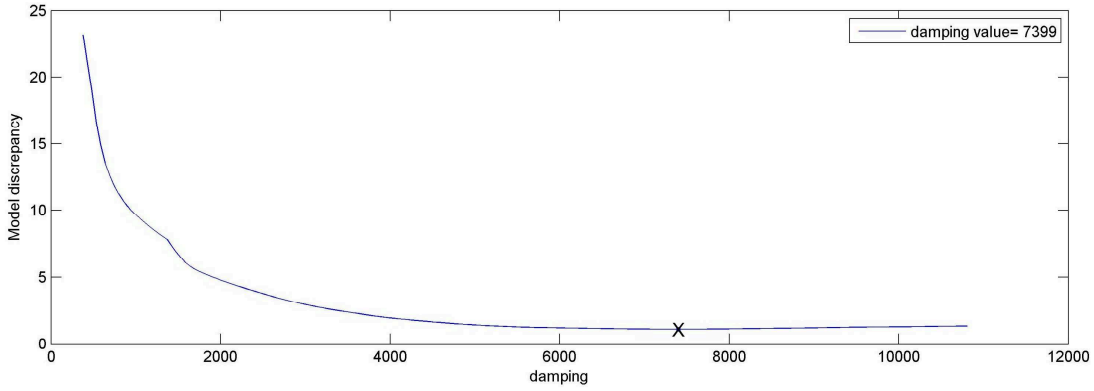


FIGURE 5.31: The mean curve obtained by the figure 5.30b.

6) The tomography

After fixing the geometry, matrices, and set the regularization parameters, tomography is the final step of our study. The tomography, or refractivity field, has been obtained trough the following points:

- ti solve the inverse matrix ($\mathbf{A}^{-1}$) by using Tikhonov regularization

- to compute N_{tomo}
- to compute the resolution matrix $\mathbf{R}$
- to compute the spread function array
- to choice the SF and resolution threshold
- to plot the Tomography

Tikhonov regularization is the most commonly used method of regularization of ill-posed problems, overdetermined or under determined ($\mathbf{A}$ may be ill-conditioned or singular). In statistics, the method is known as ridge regression. An explicit solution is given by:

$$\mathbf{A}^{-1} = (\mathbf{A}^{\mathbf{T}}\mathbf{A} + \mathbf{\Gamma}^{\mathbf{T}}\mathbf{\Gamma})^{-1}\mathbf{A}^{\mathbf{T}} \tag{5.30}$$

where $\mathbf{\Gamma}$ is called "Tikhonov matrix". In many cases, this matrix is chosen as the identity matrix $\mathbf{\Gamma} = \alpha\mathbf{I}$, becoming

$$\mathbf{A}^{-1} = (\mathbf{A}^{\mathbf{T}}\mathbf{A} + \gamma^2\mathbf{I})^{-1}\mathbf{A}^{\mathbf{T}} \tag{5.31}$$

Where γ is the fixed damping factor. This regularization improves the conditioning of the problem, thus enabling a direct numerical solution. As previously said, we used $\Delta SW_{synthetic}$ to perform the inverse tests. Having thus calculated $\mathbf{A}^{-1}$, we compute $\mathbf{N}_{calc}$ trough the equation 5.2:

$$\mathbf{N}_{calc} = \mathbf{A}^{-1}\Delta SW_{synthetic} \tag{5.32}$$

finally we obtain the tomography result with:

$$\mathbf{N}_{tomo} = \mathbf{N}_{calc} + N_{actual} \tag{5.33}$$

Let's introduce the constraint operator matrix or parameter resolution matrix $\mathbf{R}$. This can be regarded as a weighted average filter acting on the true model perturbation vector. It is done by using the equation 5.4. The main utility of a resolution matrix is to provide a measure of the resolution obtainable from the data, and this measure is based on the degree to which the resolution matrix approximates the identity matrix, according to equation 5.6. If $\mathbf{A}$ is not a full rank matrix, $\mathbf{R}$ will not be an identity matrix, which indicates that some parameter in the model cannot be constrained well by

the observation and the inverse problem under such conditions is ill-posed. Figure 5.32 shows the resolution matrix **R** calculated by Tikhonov method for our under determined inverse problem.

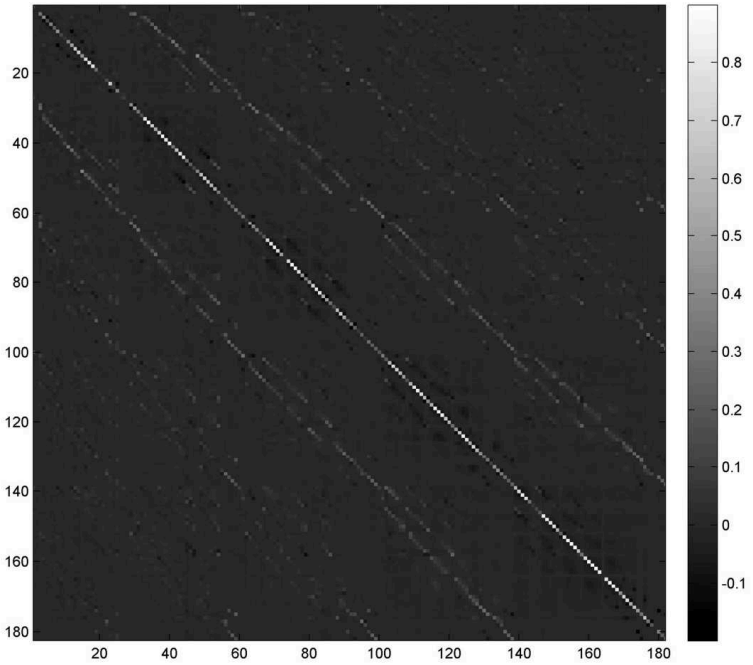


FIGURE 5.32: The resolution matrix **R** calculated by Tikhonov method for the test.

An identity resolution matrix indicates that the model is recovered exactly and that the inversion was perfect. Instead, if a resolution matrix is a nearly diagonal, each element is the weighted sum of nearby values. If one parameter can be almost perfectly resolved, corresponding row vector of the parameter resolution matrix should have the value of unity for that parameter and zero for others. On the other hand, if a parameter cannot be resolved at all, the row vector consists of random numbers without dominating one. The resolution matrix describes the distribution of information for each voxel, such that each row is the averaging vector for a parameter. The relative size and pattern of the off-diagonal elements show how the information is smeared. For a voxel to be adequately resolved, its resolution should be peaked and should have no significant contribution from voxels that are not adjacent. For a synoptic way of comparing the resolution, we calculate the spread function [Backus and Gilbert, 1967], [Menke, 1984], which compresses each row of the resolution matrix into a single number that describes how strong and peaked the resolution is for that voxel. As described previously, the spread function (SF) is a better way to describe the resolution than by solely examining the diagonal element since the diagonal resolution is very dependent on the grid spacing and the damping. The spread function $SF(j)$ for a voxel $j-th$ is computed from all the elements $R(j,k)$ of the corresponding row of the resolution matrix, weighted by their distance $D(j,k)$ from

the voxel, and $R(j,j)$ is the diagonal element. This spread function is to be calculated for j-th parameter is:

$$SF(j) = ||R(j,j)||^{-1} \sum_{k=1}^{m} D(j,k)R^2(j,k) \tag{5.34}$$

Where m represents the number of parameters. This combines two issues in evaluating resolution: the amount of information and the amount of smearing. The first factor makes the spread function small for voxels that have large diagonal resolution values. The summed terms make the spread function large for voxels that have significant averaging from other voxels and particularly large if the contributing voxels are more distant. Since the spread function combines these factors, it does not have physical units, likewise the resolution matrix which also does not have physical units. The range of values obtained for the same data set will vary depending on the chosen damping and grid spacing.

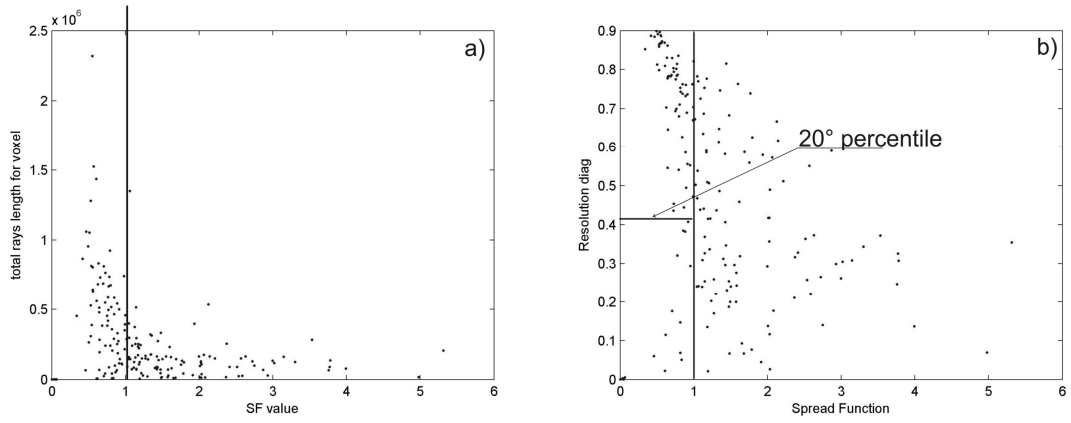


FIGURE 5.33: Threshold fixing for spread function and resolution matrices. a) A plot of the total length versus the spread function of the averaging vector for each of the 245 parameters. The line at SF=1 is the upper limit for values of the spread function considered to be acceptable. b) A plot of the Resolution diagonal value versus the spread function of the averaging vector for each of the 245 parameters. The line at SF=1 is reported from the figure a), then a line of 20° percentile is the lower limit for values of resolution considered to be acceptable.

Approximately, for low values of spread function the model velocity is representative of the volume surrounding a given voxel, for moderate values the velocity may be averaging a larger volume, and small features may not be imaged or may appear as broader features. The highest values of spread function indicate little or no information. Figure 5.33 shows the method used for choosing the threshold of spread function and resolution. Figure 5.33a shows a plot of the total length versus the spread function value for each of the 245 parameters of this test. In the figure 5.33 we choose the SF=1 as the upper limit for values of the spread function considered to be acceptable. In fact, well sampled

voxels (largely total length) generally correspond to smaller values of SF. Then, in figure 5.33b, a plot of the Resolution diagonal value versus the spread function of the averaging vector for each of the 245 parameters is shown. The threshold of SF=1 is reported and calculate the 20° percentile of the resolution voxel values. We fix the 20° percentile value as the lower limit for acceptable resolution value. Figure 5.34 shows a typical synoptic

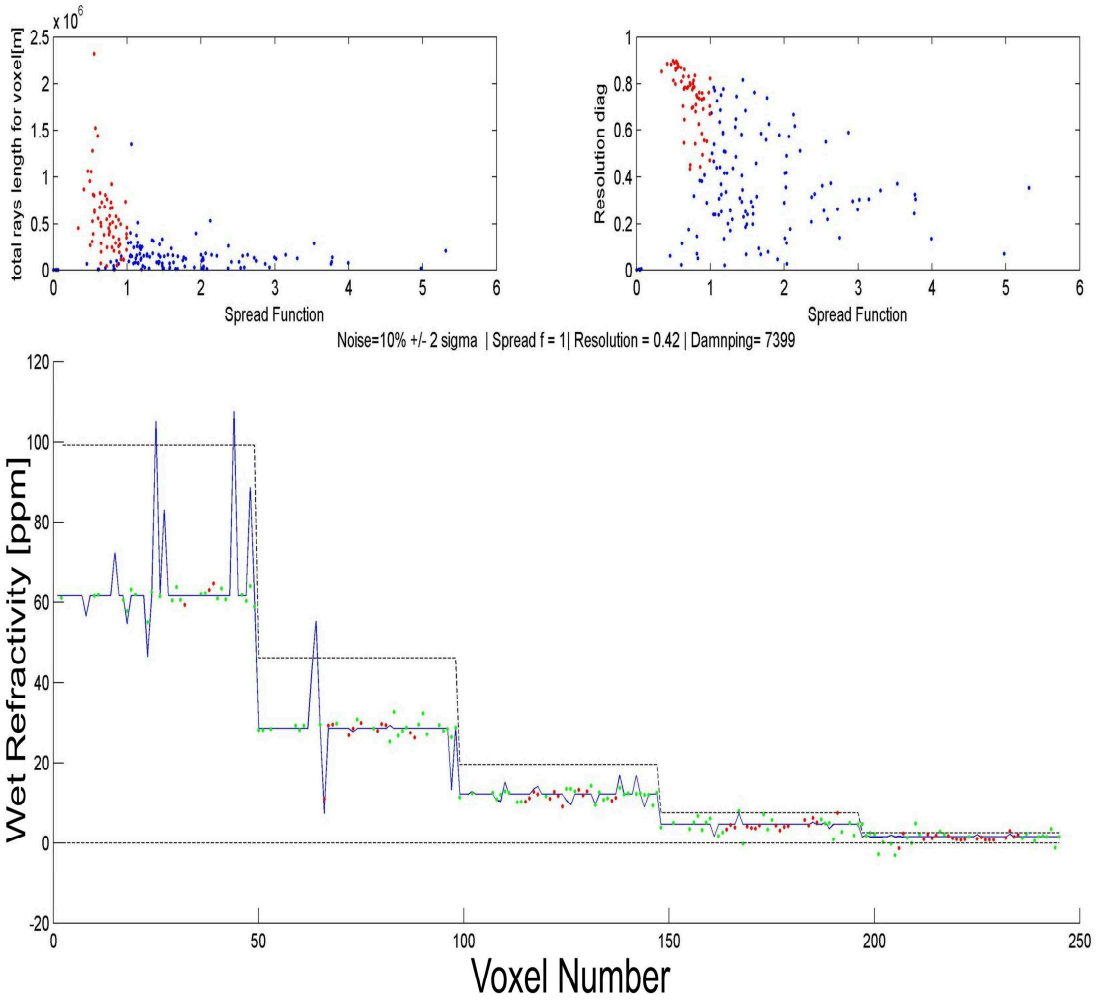


FIGURE 5.34: Numeric results of the test tomography. The red points of the upper left and right figures represent the acceptable parameters. The bottom figure instead shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the random wet refractivity anomaly induced by MATLAB. The green points are all voxels in which it is possible to calculate a refractivity value. The red points, instead, represent the refractivity values of the voxels considered to be acceptable according the fixed threshold, in other words, the tomography's result.

plot of the tomography. The red points of the upper left and right figures represent the acceptable parameters and are a direct consequence of the figures 5.33. The bottom figure instead shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the random wet refractivity anomaly induced by MATLAB.

The green points are all voxels in which it is possible to calculate a refractivity value. The red points, instead, represent the refractivity values of the voxels considered to be acceptable according the fixed threshold, in other words, the result of the tomography.

5.4.2 The tests phase

The procedures developed have to be evaluated by carrying out some tests to establish if the MATLAB code is suitable to obtain the refractivity field with respect to the starting (expected) velocity model, in order to obtain the tomography. We carried out a few tests, by assuming a set of "structures" (i.e. volume in which the refractivity varies from the surrounding space) of well-defined shapes (e.g. points, bubble and random). The tests are aimed at verifying if the software is able to identify the structures by inverting the theoretical noised slant wet delay obtained as explained in the previous sections. All the "structures" were tested by using the same geometrical configuration, as shown in figure 5.35.

The configuration with 34 GPS receivers corresponds to the permanent GPS network

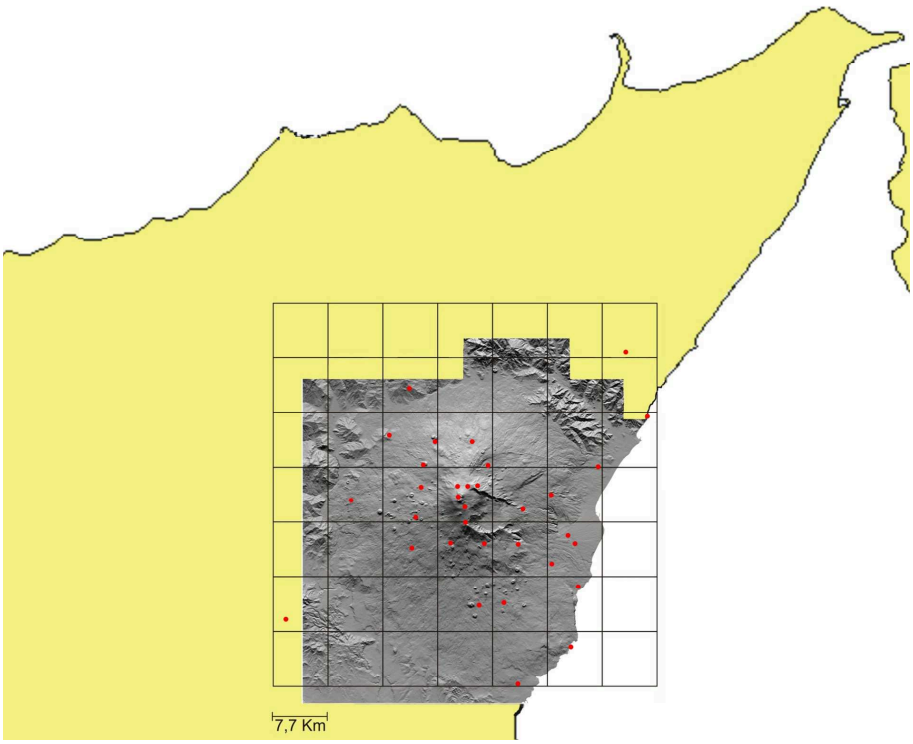


FIGURE 5.35: Plan view of the GPS network configuration used in the test. the black crosses represent the voxels limit. The configuration with 34 stations corresponds to the permanent GPS network until the first quarter of 2013.

until the second half of 2011. We decided do not using different GPS satellite spatial

configurations, as performed in the test with SIMULPS, because the variability of the well-resolved knots with satellite positions was investigated previously.

The "point model" test.

The first test was led with the simplest model we can imagine: a "point model", that is a model in which the variation of a single voxel for each layer is applied. The user chooses the maximum percentage change, compared to the half-distance of the allowable physical difference (the dash lines of figure 5.34). The MATLAB code selects the voxel and, by multiplying it by an integer between -1 and 1, reverses, cancel or leaves unchanged the sign of the perturbation. As explained earlier, we compute $SW_{perturbed}$ and

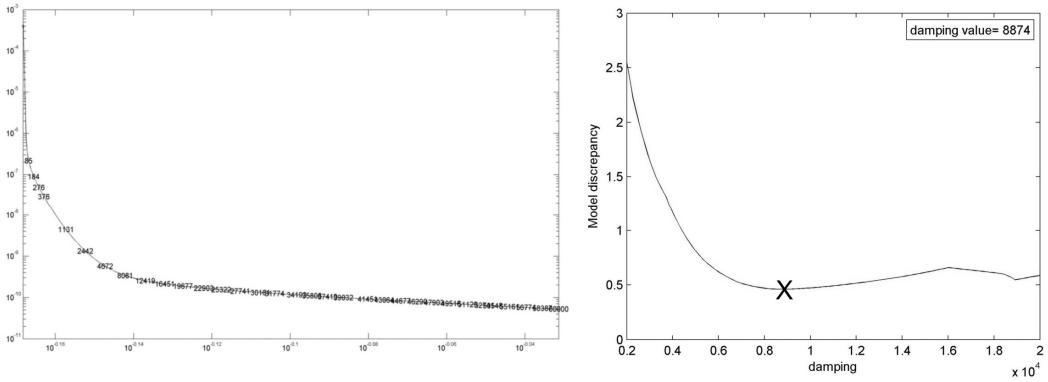


FIGURE 5.36: Damping curve in "point model" test. On left the L-curve computed to choice the damping factor; on right the mean curve of discrepancy from the model

contaminate our data with a normal random noise (equation 5.27) of 10% within 2σ, creating (SW_{noise}). Although the high value of 10%, we decided to test the process with errors higher than those normally returned by GAMIT output. The damping value is choice with the help of figure 5.36, and with the procedure of point 5, section 5.3.1. On the left of figure 5.36, the L-curve method allows us to narrow the search for the correct value of damping. On the right, the earlier explained method, allows us, through the search of the minimum, to determine the correct damping value. Applying the damping value of 8874 we obtain the figure 5.37. Following the method explained at point 6 of section 5.3.1, we set the threshold and colored in red the points of the upper plots of figure 5.37.

All the important thresholds are summarized in the title of numerical tomography result (5.37 bottom). The plots shows the numerical results of tomography as wet refractivity value versus the number of voxel. The dash black line represent the physical acceptable limits for solution, blue line is the point refractivity model induced by MATLAB. The green points represent all the voxels in which we compute the refractivity value and the red points are the results of the tomography together with the errors. The abscissa reports the number of voxels (49 per layer in this test), we remember that the voxels are numbered from the lowest to the highest layer, from south to north and from west

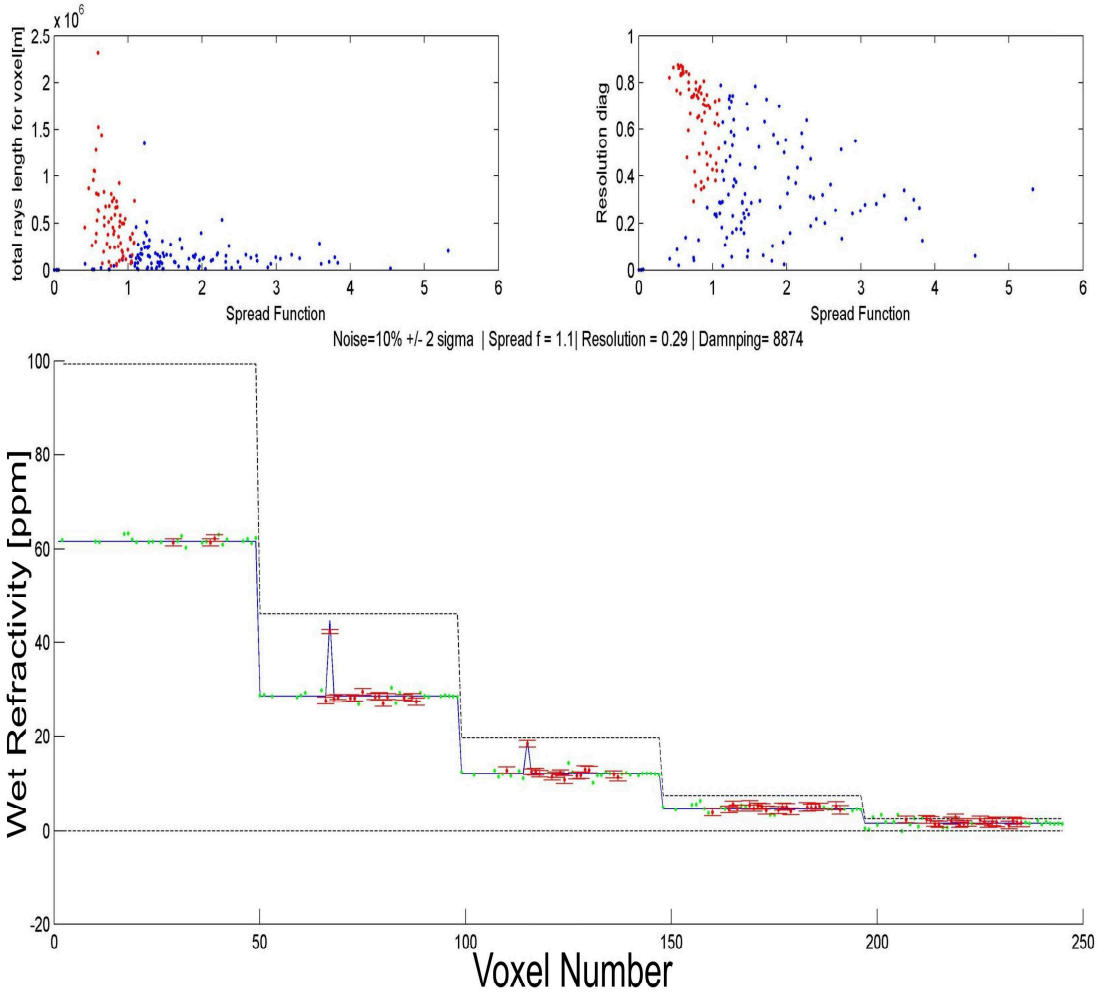


FIGURE 5.37: Numeric results for "point model" test. The red points of the upper left and right figures represent the well resolved points. The bottom figure shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the random wet refractivity anomaly assumed by MATLAB. The green points are all voxels in which it is possible to calculate a refractivity value. The red points, together with their errors, represent the refractivity values of the well resolved voxels.

to east.

The MATLAB code that creates the random anomalies, fixed two anomalies in the layer 2 and 3, respectively at 3000 and 5000 meters above sea level. The solution is very close to the model, this is surely the best achieved goal of this procedure but not the only one. Equally important is that all the solutions can be found within the original model, although initially with a high noise. In addition, there are no reverberations nearby the perturbed voxels, as it was for the SIMULPS.

The "bubble model" test.

The second test performed is a model called "bubble" that differs from the previous one because the bubble is composed by 8 perturbed voxels, 4 adjacent for each layer, and can be considered a more realistic model than the point model. Obviously this test is

more difficult than the previous one. Also in this test, the user chooses the maximum percentage change, compared to the half-distance of the allowable physical difference. We

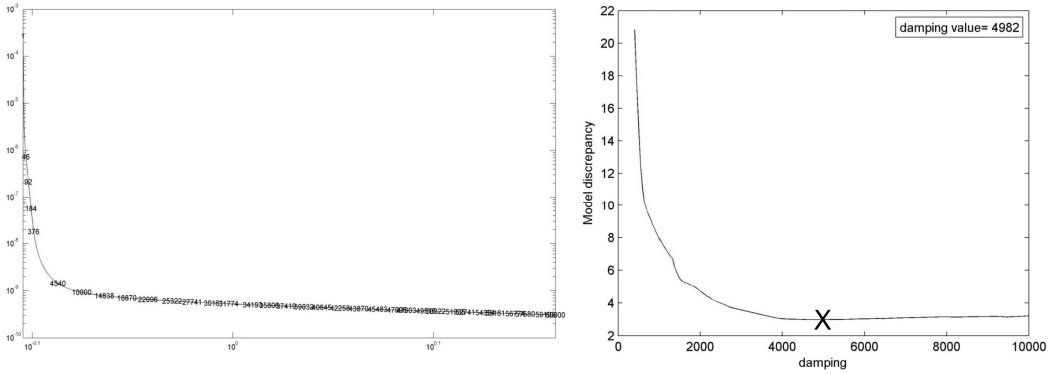


FIGURE 5.38: Damping curve in "bubble model" test. On left the L-curve computed to choice the damping factor; on right the mean curve of discrepancy from the model

compute the SW_{noise} by using the same noise values and plot the damping curve in figure 5.38, finding a minimum at 4982. Figure 5.39 shows the results obtained by inverting the SW_{noise}. The effect of a more complex model, causes a lower damping value, then, this causes different choice for the parameters SF and consequently R (upper plots of 5.39). The lower graph as usual, represents the results of the inversion, demonstrating again an excellent agreement between model and computed values.

The tomography reveals the right perturbations in the right place, the discrepancy of the anomalies between the model and the calculated values are exact in the layer at 5000 [m] and a little less in the layer at 3000 [m]. There are no reverberation in the adjoining layers.

Caused by the error associated with the measures, this test suggest that in the future would be useful to set the top layer (9 Km) at a lower height, since its physical variation is comparable to the error. The result is definitely good.

The "random model" test.

The third test represents the hardest test for the software. It is called "random" and differs a lot from the others. The water vapor moves like bubbles in the atmosphere, so maybe, this test has no physical counterpart, rather it is convenient to evaluate the sensitivity of the software by using such complex anomalies. The random model is thus created: the code establishes how and how many voxels must be perturbed, then the change of the sign and the magnitude of perturbation are applied randomly. The user chooses only the maximum percentage change, compared to the half-distance of the allowable physical difference. We compute the SW_{noise} by using the same previous noise values and plot the damping curve in figure 5.40. A damping value of 4363 allow us to compute the figure 5.41. The random model appears to be very complex (blue line of figure 5.41), there are a lot of variations, alternately, both positive and negative.

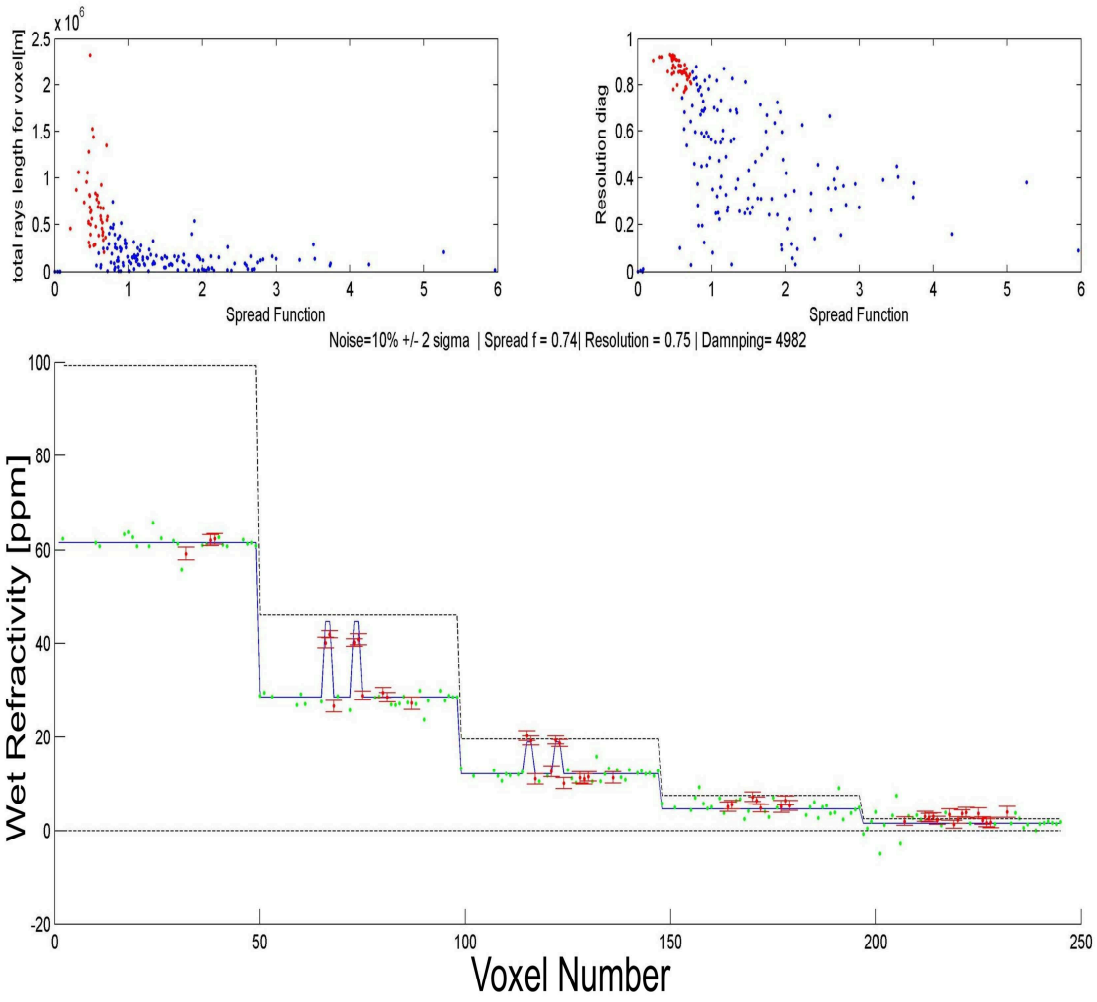


FIGURE 5.39: Numeric results for "bubble model" test. The red points of the upper left and right figures represent the well resolved points. The bottom figure shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the random wet refractivity anomaly induced by MATLAB. The green points are all voxels in which it is possible to calculate a refractivity value. The red points, together with their errors, represent the refractivity values of the well resolved voxels.

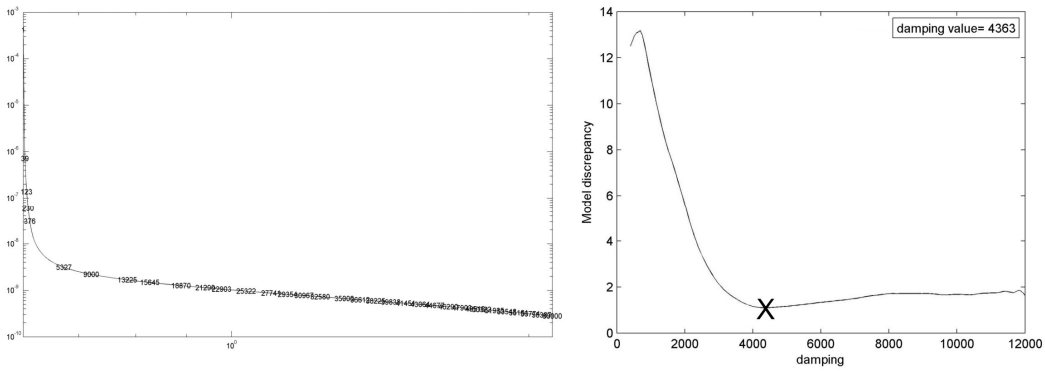


FIGURE 5.40: Damping curve in "random model" test. On left the L-curve computed to choice the damping factor; on right the mean curve of discrepancy from the model

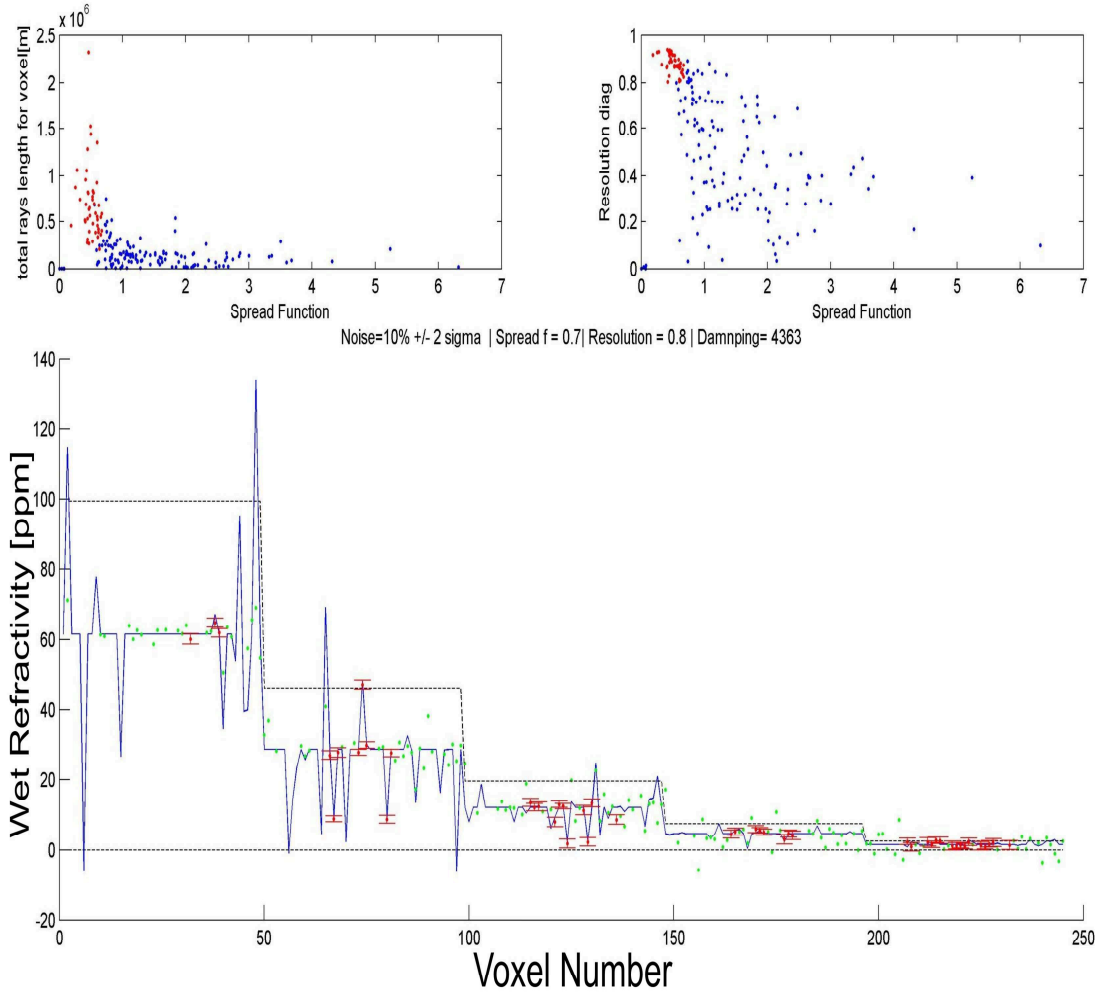


FIGURE 5.41: Numeric results for "random model" test. The red points of the upper left and right figures represent the well resolved points. The bottom figure shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the random wet refractivity anomaly induced by MATLAB. The green points are all voxels in which it is possible to calculate a refractivity value. The red points, together with their errors, represent the refractivity values of the well resolved voxels.

Although so complex, tomography's results detect the anomalies, confirming the right choice of the threshold parameters.

In the lowest layer (5.41 x axis 1-49), unfortunately only 3 points can be well resolved. In this layer the big changes do not intersect with the well resolved points. The only slight variation crossed, however, could be detected. Regarding the layer to 3000 [m] s.l.m.(5.41 x axis 50-98) however, well resolved nodes intersect three perturbations, all perfectly revealed. This is an excellent result. The layer at 5000 [m] (5.41 x axis 99-147) has a close behavior respect to the layer 3000, exception for the rightmost point that report a wrong value. Also in this layer the result is excellent . Layers 7000 (5.41 x axis 148-196) and 9000 [m] (5.41 x axis 197-245) show no particular evidence, a small anomaly is identified in the 7000 layer [m]. For these two layers, The most important

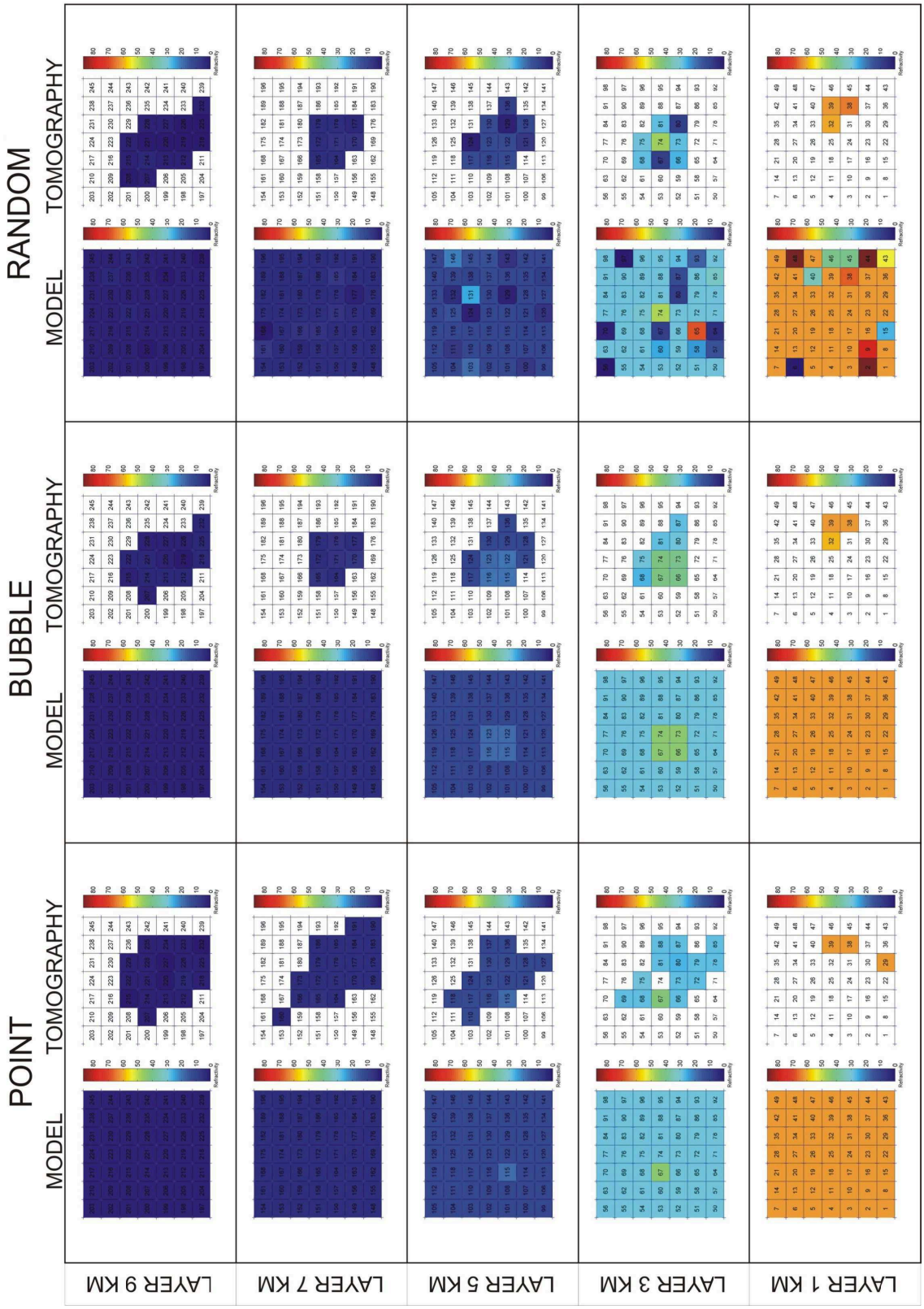


FIGURE 5.42: Tomographic images of the same results reported in figure 5.37, 5.39 and 5.41. Each tomography images show only the voxels over the computed threshold.

thing is that there are no well resolved points outside the physical range.

Above expectations, although very complex, the software is able to recover the anomalies intersected.
Figure 5.42 reports the graphic images of the tomography results plotted in the figures 5.37, 5.39 and 5.41, and already commented. The plots is divided into three parts, which refer to the three tests (point, bubble, random). Each of the three parts is divided into two columns: the left column represents the reference model, the one on the right the result of tomography. The tomography images show only the voxels over the threshold, the others are blank. Although the image doesn't give any numeric information, they allow to compare all tests performed in a synoptic way. According to the scheme in Figure 5.35, we can geographically locate the detected anomalies.

5.4.3 The August 12, 2011 test case

Now we are ready to apply the methods studied in an actual case: August 12, 2011 10:00 am was chosen as a test case. Following the way of the previous tests, the damping factor was calculated, initially by using the L-curves method (fig. 5.43 right), then in a narrow range (fig. 5.43 left). From the figure 5.43 right, it is clear that, after the damping value

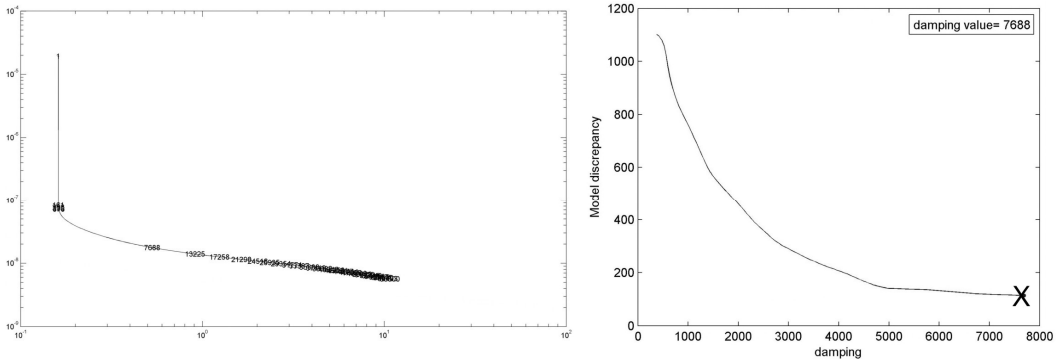


FIGURE 5.43: Damping curve referred to August 12, 2011 test case. On left the L-curve computed to choice the damping factor; on right the mean curve of discrepancy from the model, computed in a narrow range.

of 5000 the model discrepancy remains fairly constant until the minimum value of 7688. The computed tomography is plotted in figure 5.44.

As summarize in the 5.44, the SF value of 0.68 allow to resolve 58 voxels. Although most of them are placed in the upper layers, a clear evidence of anomaly is detected in the four well resolved points of layer 1. The layer 3 shows an indistinct cluster of perturbed points (both positives and negatives), initially we have thought that the scattered points distribution around the starting refractivity model was wrong. The layer 5 does not have relevant characteristics, it's just a little drier than the model.

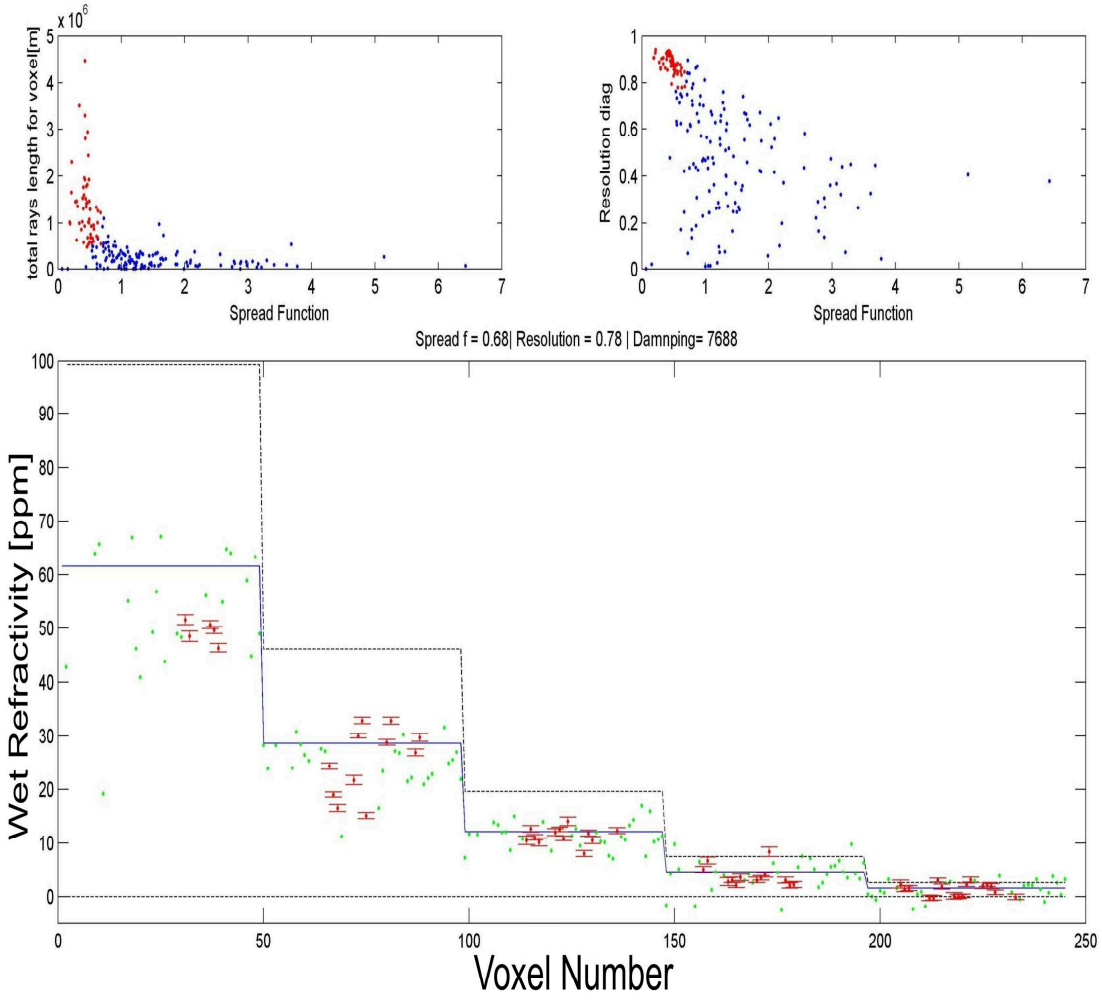


FIGURE 5.44: Numeric tomography results referred to August 12, 2011 test case. The red points of the upper left and right figures represent the well resolved points. The bottom figure shows the numerical result of tomography plotted as wet refractivity value versus number of voxel. The dash black lines represent the physical acceptable limits for solution, blue line is the starting wet refractivity model. The green points are all voxels in which it is possible to calculate a refractivity value. The red points, together with their errors, represent the refractivity values of the well resolved voxels.

The layers 7 and 9, excluding the point of exceeding the limit fixed interval, show the same behavior as those of layer 5. If we integrate the information of 5.44, through their georeferencing (fig.5.45), everything becomes clear. Layer 1 detects a bubble of more dry air than the surroundings in the south-east flank of Etna, between 0 and 2000 [m] a.s.l. (1000 represents the center of voxel). The indistinct cluster of perturbed points in the layer 3 becomes an interesting cluster of perturbed points. It is clear that, points that have the same behavior are side by side, demonstrating that on August 12, 2011 at 10.00 am, the north-west flank of Etna is drier than the East flank. Again the values of side by side voxels like 66 and 72 are similar. The east flank of Etna is a little bit wetter then the starting model. All this conclusions are referred to the volume between 2000-4000 [m] a.s.l. Layer 5 has the same characteristic of layer 3 but much smaller in

magnitude. Layer 7 and 9 do not give relevant information, excluding the voxels 173 and 222 that are geographically overlapping and show a positive trend of water vapor content.

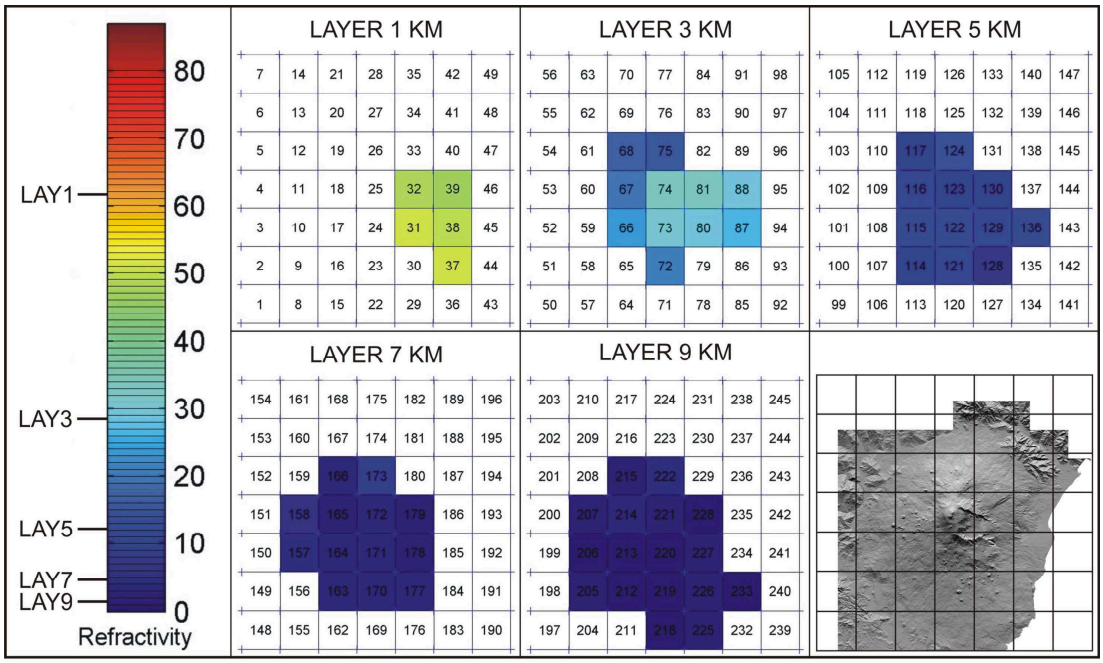


FIGURE 5.45: Damping curve in "random model" test. On left the L-curve computed to choice the damping factor; on right the mean curve of discrepancy from the model

5.4.4 Discussion

The aim of the present study is the measuring the atmospheric anomalies by using GPS data. Our studies start from previous achievements both on seismic tomography and GPS atmospheric sounding for assessing a method to produce a tomography of lower neutral atmosphere over the Mt. Etna area. Initially, infact, a software originally implemented for seismic tomography (SIMULPS) has been modified. The main problem with this approach was to adapt the atmospheric properties to the seismic. We carried out many synthetic tests by considering the geometrical aspect of the problem and optimizing the steps of data processing. Results of the tests show that well resolved area of tomographic images increases with the increasing number of GPS receivers and that actual GPS network is able to reveal the atmospheric anomalies. Unfortunately, we realized that the use of SIMULPS did not give good results for our purpose. About the sensitivity of SIMULPS, this software in fact is not suitable for atmospheric studies. Tests demonstrate that the number of significant digits of the SIMULPS is inadequate. The experiences and knowledge gained by SIMULPS were used for developing a tropospheric tomography software, entirely build during the PhD. The development

environment chosen for this purpose has been MATLAB. We used MATLAB in order to acquire, computing and illustrate the GPS tomography results. All the processing chain was then implemented with particular attention on the refractivity model and the inversion algorithm. The procedures developed has been evaluated by some tests act to establish if the MATLAB code is suitable to obtain the refractivity field with respect to the starting velocity model. We carried out a few tests, by assuming a set of "structures" of well-defined shapes, it was immediately obvious the achieved results in favor of the latter software. The points model has been fully reconstructed without errors, together with the more difficult bubble model. The random model was the last test, extremely difficult to reconstruct, the software was able to solve it with high precision. The results are applied with good results in a real test case. The tomography results indicate clearly important features of the refractivity field of the studied day. Concluding, the created software is able to reveal the tropospheric anomalies to improve, with important information, the results given by SAR technique.

Conclsions and future perspectives

The present work deals with the influence of atmospheric anomalies in the GPS EM wave's propagation useful to extrapolate information about the wet refractivity field to apply to Differential Interferometric SAR (DInSAR) technique for filtering purposes.
Because of its height and the quite variable weather conditions, the estimation of Mount Etna atmospheric anomalies using GPS measurements have noticeable importance to calibrate the SAR interferograms and/or to establish the "effective" ground deformation of the volcanic edifice.
The starting point of this work was to study the effects of different atmospheric layers on radio signals. The great differences between ionized and neutral atmosphere reflect themselves in the refractive index equation and consequently in the EM waves propagation behavior. This led to apply appropriate computing strategies to manage the data sets and solve the problems or reduce the atmospheric effects. On the basis of the existing literature, studies concerning GPS led me to deeper understand its signal structure and potentiality of the GPS system as not simply a satellite positioning system. The fact of being able to be used in single, double or triple differences, depending on the chosen type of analysis, makes it very versatile both for geodetic and environmental purposes.
Advances in technology and new demands on the existing system have now led to efforts to modernize the GPS system and implement the next generation of GPS III satellites and next generation operational control system. Thus there exists a large interest in applying space geodetic techniques, especially GPS and others GNSS, for atmosphere studies. For example to study climate trends, assimilated into numerical weather prediction models to improve weather forecasts, or, as we have done in this study, to estimate the 3D structure of the atmospheric water vapor by applying tomographic methods.
In addition to GPS, other GNSS systems are in use or under development. The Russian GLONASS was developed contemporaneously with GPS, while the European Union Galileo positioning system, the Chinese Compass navigation system and the Indian

IRNSS are currently in different phases of implementation. Along with the growth of the systems, the number of satellites will also increase. The whole family of GNSS is planned to consist of about 120 satellites till 2030. Moreover, the new satellites are capable of transmitting multiple signals in multiple frequency bands. This evolving scenario of the GNSSS systems envisages a continuous and more and more demanding improvement of the applications of the space geodetic techniques to environmental studies (from the monitoring of Earth dynamics to the climate changes).

Thanks to an agreement between the University of Catania and the INGV-OE, the GPS data used in this work were collected by the network of GPS permanent stations named "Etn@net" [Palano et al., 2010]. The processing has been carried out by using dedicated servers and specific software owned by INGV-OE. In particular we analyzed GPS data by using the GAMIT software [Herring et al., 2010], by adopting appropriate processing strategies specifically designed and implemented during this work.

To achieve a high-precision output of the elaborated GPS data, we analyzed a great amount of data by performing a quality check for each GPS station through a specific software developed in MATLAB, which allows to evaluate a great amount of the LC undifferenced phase residual data, by using synoptic plots. This preliminary analysis was very useful to estimate the characteristics of each GPS station of the network and to deduce systematic anomalies that could affect the stations, suggesting a more extensive use in the future to improve the general accuracy of the monitoring activities. The systematic application of this type of analysis on the whole stations of "Etn@net" was carried out for the first time during this study.

The original work of this thesis is divided into two parts: the first concerns the interaction between the GPS EM waves and the volcanic plume emitted from Mount Etna, the second describes how the atmospheric anomalies reflect to the refractivity field, allowing computing a tomography of the atmosphere above Mount Etna.

With regard to the first part, the ability of GPS to detect the volcanic plume presence has been proved only in a few cases worldwide by studying plianian eruptions, with energy ten times greater than the Etna typical events. The Etna volcano is particularly suited for an in-depth investigation on the potentiality of GPS observations to detect volcanic plumes, owing to both the high frequency of explosive episodes over recent years and the cutting-edge GPS monitoring network operating on this volcano. The 4 September 2007 eruption was selected as a good test case, since this event was long-lasting and showed the presence of quasi-stationary eruptive phenomena that remained constant for almost 12 h, producing an quasi-stationary volcanic plume during this period. We proposed two different approaches in order to examine this capability: the first approach, recently proposed in specialized literature [Larson, 2013], is applied on the signal strength of the GPS L2 carrier phase data. The second approach, implemented in this study, is based on statistical analysis of the outcomes of the geodetic processing of GPS data, by analyzing

the single difference post fit phase residual of elaborated signals. The second approach allows to overcome the drawbacks encountered by applying the first one, proving that the GPS is able to detect the volcanic plume generated from a mid-intensity explosive activity. The statistical algorithm implemented in this work indicates which stations and satellites are unambiguously affected by the plume. Knowing that the components mainly affecting the GPS signal are, in order, water vapor, hydrometeors, aerosol and volcanic ash, it is reasonable to conclude that the residuals of the LC observable should present effects more evident respect to SNR signal, which instead is mainly perturbed by obstructions. For this reasons, the second approach is based on the analysis of the LC signal, since the volcanic plume composition is largely made up of volcanic ash, water vapor, volcanic gas and hydrometeors.

As the atmospheric tomography is concerned, we would remind that variations of pressure, temperature and water content cause changes in the effective refractive index along the EM wave path and therefore may corrupt phase change measurement which is at the base of the measurements of the ground deformations, inducing a degradation of the geophysical parameters estimated by DInSAR technique. In this study we presented a method to obtain a 3D electromagnetic waves velocity field, starting from the GPS output data analysis. Since the SAR images have an incidence angle ranging from 20 to 40 degree, the availability of a 3D EM velocity field is of fundamental importance for filtering the DInSAR interferograms, in order to project along any line of sight the effects of the tropospheric anomalies eventually existing on the atmosphere crossed by the actual radar signal.

The implementation of the atmospheric tomography has been performed in two phases. During the first phase, we applied an approach derived from seismic tomography and already tested in early studies for assessing a method to produce a tomography of lower neutral atmosphere over the Mt. Etna area by using GPS data. Thus a software originally implemented for seismic tomography (SIMULPS) has been modified. Despite the use of SIMULPS gave results not completely satisfactory for our purposes, the work carried out and the results obtained were important for many aspects. In particular, they are fundamental to test and assess the number of stations and their optimal configurations as well as the number and location of voxels. Moreover we demonstrated that the analysis of the Spread Function (SF) is an irreplaceable tool for recognizing areas where the results of the topographic inversion can be considered of "good" quality. Finally, by using SIMPLUS we have become familiar with the sensitive parameters for fine tuning a proper tomography and with the definition of the start-up atmospheric model.

Starting from the experience achieved during the first phase, a new code was developed for deriving the tropospheric tomography from the GPS data. A lot of improvements were added, in particular as regarding atmospheric input parameters and the possibility to use data from actual weather measurements, as those extracted from meteorological

balloons. Particular attention was spent to the inversion technique: the Tikhonov regularization was introduced and a new method for the choice of the damping parameter was studied.

The new code was validated by using synthetic tests which assume different structure of atmospheric anomalies and with random noise about twice severe than the typical errors of the GPS. The results of the tests proved that the new code is able to reconstruct the simulated anomalies faithfully. It is very significant the result obtained in reconstructing with high precision a synthetic model, where a randomly distributed set of anomalies were assumed.

The new code was applied to study the structure of the atmosphere in actual case: the period of August 12, 2011 at 10.00 am. The results of the tomography indicate clearly important features of the refractivity field of the studied day. Layer 1 (between 0 and 2000 [m] a.s.l.) detects a bubble of drier air in the south-east flank of Etna, Layer 3 (between 2000 and 4000 [m] a.s.l.) reveals that the north-west flank of Etna is drier than the East flank.

In conclusion, the synthetic tests and the application on actual data sets of the new software demonstrate that it is able to reveal the tropospheric anomalies and thus it is an useful tool to improve the results of the SAR interferometry.

A lot of improving elements can be identified in the processing presented, for further studies. The SNR data must be studied in detail, because a more precise characterization of each station could lead to detect any anomalies. In addition, we think it is opportune to apply the same procedures during a sand storm and heavy raining, in order to evaluate the sensitivity of this analysis to different environmental conditions and, in case, to compare and to understand the differences. As regards the volcanic plume detection, we think that its estimation could be improved through a better representation of the volcanic plume dispersion geometry and a further analysis of GPS data together with deeper studies about the reliability of the statistical tests. A more accurate analysis of these data during volcanic explosive activity could furnish new complementary information and be complementary to the already existing monitoring system of Etna volcanic plumes at INGV-OE, thereby contributing to reduce the risk to air traffic operations. The tomographic inversion method should be optimized, and the limitations of its applicability should be further investigated.

Atmospheric tomography (and the related modeling) has some specific limits: the most relevant are the discretization of the atmosphere in voxels and the projection of the wet zenith delay measurements by using mapping functions which depend on the elevation, azimuth and N-S and E-W gradients values. The first limit might be overcame by improving the computing resources (both hardware and software), in order to consider smaller voxels, and applying new inversion methods like "linear inversion problems with continuous models". The second kind of limitation might be overcame through the

integrated use of residuals and gradients together. This would provide a gradient time-dependent function and therefore a more accurate projection of the wet zenith delay measurements in time and space.

The use of data from meteorological balloons, when they will be available, will give us an accurate atmospheric starting model of the troposphere, much more than that currently used. An important improvement for the tomography results will come also from the comparison with both multispectral satellite images having similar, or better, space resolution and with radiometers, if available. In parallel with all these improvements, the integrated development of GNSS in future years and the increasing number of GPS stations will increase the amount of observables, improving the results of the atmospheric tomography.

Appendix A

Summary of Gamit file name

A - file: ASCII version of the T-file (tabular ephemeris)

B - file: controls the batch mode of data processing

C - file: observed – computed (O-C's), partial derivatives

D - file: driver file of sessions and receivers

E - file: broadcast ephemeris, in RINEX navigation file or FICA Blk 9 format

G - file: orbital initial conditions and non-gravitational parameter values

H - file: adjustments and full variance-covariance matrix for input to GLOBK

I - file: receiver clock polynomial input

J - file: satellite clock polynomial coefficients

K - file: values of receiver clock offset during observation span, from pseudorange

L - file: station coordinates

M - file: controls merging of data (C-) files for solve and editing programs

N - file: data-weight overrides for solve created from autcln.sum.postfit

O - file: record of the analysis (reduced form of Q-file) for post-processing analysis

P - file: record of a model run

Q - file: record of the analysis (solve run)

S - file: simulation controls

T - file: tabular ephemeris

U - file: loading and meterological data for model

V- file: editing output of SCANRMS

W - file: meteorological data in RINEX met-file format

X - file: input observations

Y - file: satellite yaw parameters

Z - file: output meterological data

Bibliography

K. Aki. Overview in seismic tomography. *Ed. London: Chapman & Hall*, pages 1–8, 1993.

M. Aloisi, O. Cocina, G. Neri, B. Orecchio, and E. Privitera. Seismic tomography of the crust underneath the etna volcano, sicily. *Physics of the Earth and Planetary Interiors*, 134:139–155, 2002.

M. Aloisi, M. Mattia, C. Ferlito, M. Palano, V. Bruno, and F. Cannavò. Imaging the multi-level magma reservoir at mt.etna volcano (italy). *Geophys. Res. Lett.*, 38, 2011. URL doi:10.1029/2011GL048488.

D. Andronico, A. Cristaldi, and S. Scollo. The 4–5 september 2007 lava fountain at south-east crater of mt etna, italy. *J Volcanol Geoth Res*, 173:325–328, 2008. URL doi:10.1016/j.jvolgeores.2008.02.004.

M. Aranzulla, F. Cannavò, S. Scollo, G. Puglisi, and G. Immè. Volcanic ash detection by gps signal. *GPS Solutions*, 17(4):485–497, 2013. URL doi:10.1007/s10291-012-0294-4.

G. Backus and J.F. Gilbert. Numerical applications of a formalism for geophysical inverse problems. *Geophys. J.R. Astron. Soc.*, 13:247–276, 1967.

G. Backus and J.F. Gilbert. The resolving power of gross earth data. *Geophys. J.R. Astron. Soc.*, 16:169–205, 1968.

P. Banks and G. Kockarts. Aeronomy. *New York: Academic Press*, 1973.

S. Barbarino. Appunti di campi elettromagnetici. 2013. URL http://www.diit.unict.it/users/campi/.

R.G. Barry and R.J. Chorley. Atmosphere, weather, and climate. 5th ed. Methuen, New York, 1987.

S. Bassiri and G.A. Haij. Higher-order ionospheric effects on the global positioning system observables and means of modeling them. *Manuscripta geodaetica*, 18:280–289, 1993.

B. R. Bean and G. D. Thayer. Crpl exponential reference atmosphere. *Technical report, U.S. Government Printing Office*, 1959. URL http://digicoll.manoa.hawaii.edu/techreports/PDF/NBS4.

F. Beauducel, P. Briole, and J. Froger. Volcano wide fringes in ers sar interferograms of etna: Deformation or tropospheric effect? *Journal of Geophysical Research*, 105: 16391–16402, 2000.

M. Bevis, S. Businger, T.A. Herring, C. Rocken, R.A. Anthes, and R.H. Ware. Gps meteorology remote sensing of atmospheric water vapor using the global positioning system. *J. Geophys. Res.*, 97(D14):15787–801, 1992. URL doi:10.1029/92JD01517.

M. Bevis, S. Businger, S. Chiswell, T. A. Herring, R. A. Anthes, C. Rocken, and R. H. Ware. Gps meteorology: Mapping zenith wet delays onto precipitable water. *J. Appl. Meteorology*, 33(3):379–386, 1994. URL doi:10.1175/1520-0450(1994)033.

J. Böhm and H. Schuh. Vienna mapping functions. *Leipzig, Germany. Verlag des Bundesamtes für Kartographie und Geodäsie.*, Proc. 16th Working Meeting on European VLBI for Geodesy and Astrometry.:131–143, 2003.

J. Böhm and H. Schuh. Vienna mapping functions in vlbi analyses. *Geophys. Res. Lett.*, 31:L01603, 2004. URL doi:10.1029/2003GL018984.

J. Böhm, A. Niell, P. Tregoning, and H. Schuh. Global mapping function (gmf): a new empirical mapping function based on numerical weather model data. *Geophys. Res. Lett.*, 33L07304, 2006a. URL doi:10.1029/2005GL025546.

J. Böhm, B. Werl, and H. Schuh. Troposphere mapping functions for gps and very long baseline interferometry from european centre for medium-range weather forecasts operational analysis data. *J. Geophys. Res*, 111:B02406, 2006b. URL doi:10.1029/2005JB003629.

J. Böhm, P.J. Mendes Cerveira, H. Schuh, and P. Tregoning. The impact of mapping functions for the neutral atmosphere based on numerical weather models in gps data analysis. *In P. Tregoning and C. Rizos, editors, Dynamic Planet*, 130:837–843, 2007. URL doi:10.1007/978-3-540-49350-1_118.

J. Böhm, R. Heinkelmann, and H. Schuh. Neutral atmosphere delays: Empirical models versus discrete time series from numerical weather models. In H. Drewes, editor, Geodetic Reference Frames - IAG Symposium, volume 134 of IAG Symposia, Munich, Germany,:317–321, 2009a. URL doi:10.1007/978-3-642-00860-3_49.

J. Böhm, J. Kouba, and H. Schuh. Forecast vienna mapping functions 1 for real-time analysis of space geodetic observations. *J. Geodesy*, 83(5), 2009b. URL doi:10.1007/s00190-008-0216-y.

C. Bonadonna, C.B. Connor, B.F. Houghton, L. Connor, M. Byrne, A. Laing, and T.K. Hincks. Probabilistic modeling of tephra dispersal: hazard assessment of a multiphase rhyolitic eruption at tarawera, new zealand. *J Geophys Res*, 110:B03203, 2005. URL doi:10.1029/2003JB002896.

A. Bonforte, G. Puglisi, A. Ferretti, A. Menegaz, C. Prati, and F. Rocca. Modelli atmosferici peril controllo dei movimenti dell'etna. *Alta Frequenza*, 11 (4):34–39, 1999.

A. Bonforte, A. Ferretti, C. Prati, G. Puglisi, and F. Rocca. Calibration of atmospheric effects on sar interferograms by gps and local atmosphere models: first results. *Journ. Atmos. and Sol. Terr. Phys*, 63:1343–1357, 2001.

M. Born and E. Wolf. Principles of optics. *Cambridge Univ. Press, New York, 7thedition*, 1999.

P. Briole, G. Nunnari, G. Puglisi, and J. Murray. The 1989 september-october eruption of mount etna (italy) - some quantitative information obtained by geodesy and tiltmetry. *C. R. Acad. Sci. Série II*, 310(13):1747–1754, 1990.

F.K. Brunner and M. Gu. An improved model for the dual frequency ionospheric correction of gps observations. *Manuscr Geod*, 16:205–214, 1991a.

F.K. Brunner and M. Gu. An improved model for the dual frequency ionospheric correction of gps observations. *Manuscripta geodaetica*, 16:205–214, 1991b.

V. Bruno, M. Aloisi, A. Bonforte, J. Immè, and G. Puglisi. Atmospheric anomalies over mt. etna using gps signal delays and tomography of radio wave velocities. *Annals of Geophysics*, 50(2), 2007.

V. Bruno, Mattia M., M. Aloisi, M. Palano, F. Cannavò, and W.E. Holt. Ground deformations and volcanic processes as imaged by cgps data at mt. etna (italy) between 2003 and 2008. *J. Geophys. Res.*, 117, 2012. URL doi:10.1029/2011JB009114.

K. G. Budden. The propagation of radio waves. *Cambridge University Press, New York, 1 edition*, 1985.

V. Cerveny. Seismic ray theory. *Cambridge University Press, New York*, 2005.

V. Cerveny, L. Klimes, and I. Psencik. Complete seismic-ray tracing in three-dimensional structures. *Seismological algorithms, Doornbos, editor, Academic Press, New York.*, page 89–168, 1988.

C. Champollion, F. Mason, M.N. Bouin, A. Walpersdorf, E. Doerfiinger, O. Bock, and J. van Baelen. Gps water vapour tomography: preliminary results from the escompte field experiment. *Atmospheric Research*, 74:253–274, 2005. URL doi:10.1016/j.atmosres.2004.04.003.

C.C. Chao. The troposphere calibration model for mariner mars 1971. *Technical Report 32–1587 NASA JPL, Pasadena, CA*, 1974.

D.S. Coco, T.L. Guassiran, and C. Coker. Passive detection of sporadic e using gps phase measurements. *Radio Science*, 30(6):1869–1874, 1995.

A.E. Cole, A. Court, and A.J. Kantor. Model atmospheres. *Ed. S.L. Valley. McGraw-Hill Book Company. New York*, Handbook of Geophysics and Space Environments, 1965.

M. Coltelli. Osservazione da satellite e simulazione dell'emissione di cenere. *INGV-OE internal report*, 2007. URL `http://www.ct.ingv.it`.

A. Costa, G. Macedonio, and A. Folch. A three dimensional eulerian model for transport and deposition of volcanic ashes. *Earth Planet Sci Lett*, 241:634–647, 2006. URL `doi:10.1016/j.epsl.2005.11.019`.

J. L. Davis, T. A. Herring, I. I. Shapiro, A. E. E. Rogers, and G. Elgered. Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length. *Radio Sci.*, 20(6):1593–1607, 1985. URL `doi:10.1029/RS020i006p01593`.

J. L. Davis, G. Elgered, A. E. Niell, and C. E. Kuehn. Ground-based measurement of gradients in the "wet" radio refractivity of air. *Radio Sci*, 28(6):1003–1018, 1993a. URL `doi:10.1029/93RS01917`.

J.L. Davis. Atmospheric propagation effects on radio interferometry. Technical Report AFGL-TR-86-0243, Scientific Report No. 1, Air Force Geophysics Laboratory, 1986.

J.L. Davis, G. Elgered, A.B. Niell, and C.E. Kuehn. Ground-based measurement of gradients in the 'wet' radio refractivity of air. *Radio Science*, 28(6):1003–1018, 1993b.

P. Debye. Polar molecules. *Dover, New York*, 1929.

C. Delacourt, P. Briole, and J. Achache. Tropospheric corrections of sar interferograms with strong topography. application to etna. *Geophys. Res. Lett.*, 25:2849–2852, 1998.

R.C. Dixon. Spread spectrum systems 2nd edition. *John Wiley & Sons*, 1984. URL `ISBN0-471-88309-3`.

US DMA. Department of defense world geodetic system 1984 : its definition and relationships with local geodetic systems. *Defense Mapping Agency technical report.*, 1991.

F. Donnadieu. Volcanological applications of doppler radars: A review and examples from a transportable pulse radar in l-band, in doppler radar observations - weather

radar, wind pro filer, ionospheric radar, and other advanced applications. *edited by J. Bech. InTech, Rijeka, Croatia*, page 409 – 446, 2012. URL ISBN:978-953-51-0496-4.

G. Elgered, H.-P. Plag, H. van der Marel, S. Barlag, and J. Nash. Exploitation of ground-based gps for operational numerical weather prediction and climate applications. *COST action 716: Final Report. European Union, Brussels, Belgium*, 2005.

T. R. Emardson and H. J. P. Derks. On the relation between the wet delay and the integrated precipitable water vapour in the european atmosphere. *Meteorol. Appl.*, 7(1):61–68, 2000. URL doi:10.1017/S1350482700001377.

L. Essen and K. D. Froome. The refractive indices and dielectric constants of air and its principal constituents at 24,000 mc/s. *Proc. Phys Soc. B.*, 64(10):862–875, 1951. URL doi:10.1088/0370-1301/64/10/303.

L.H. Estey and C.M. Meertens. Teqc: The multi-purpose toolkit for gps/glonass data. *GPS Solutions*, 3(1):42–49, 1999.

K. Bull et al. The 2009 eruption of redoubt volcano, the report of investigations 2011-5. *Editor Janet R. Schaefer, Published by the State of Alaska*, 2012.

J.R. Evans, D. Eberhart-phillips, and C.H.Thurber. User's manual for simulps12 for imaging vp and vp/vs: a derivative of the "thurber" tomographic inversion simul3 for local earthquakes and explosions. *USGS Open.file Rep.*, pages 94–431, 1994.

R.G. Fleagle and J.A. Businger. An introduction to atmospheric physics. *ed., Academic Press, New York.*, 1980.

U. Fölsche. Tropospheric water vapor imaging by combination of spaceborne and ground-based gnss sounding data. *PhD thesis, Univ. Graz, Graz, Austria*, 1999.

L. P. Gradinarsky, J. M. Johansson, H. R. Bouma, H.-G. Scherneck, and G. Elgered. Climate monitoring using gps. *Physics and Chemistry of the Earth*, 27:335–340, 2002. URL doi:10.1016/S1474-7065(02)00009-8.

W. Gurtner and G. Mader. Receiver independent exchange format version 2. *GPS bulletin*, 3:1–8, 1990.

G.J. Hahn and S.S. Shapiro. Statistical models in engineering. *Wiley, New York*, 1994. URL ISBN-10:0471040657.

T. A. Herring. Modeling atmospheric delays in the analysis of space geodetic data. *Symposium on Refraction of Transatmospheric Signals in Geodesy*, page 157–164, 1992.

T. A. Herring, J. L. Davis, and I. I. Shapiro. Geodesy by radio interferometry: The application of kalman filtering to the analysis of very long baseline interferometry data. *J. Geophys. Res.*, 95(B8):12:561–581, 1990. URL 90JB00683.

T.A. Herring, R.W. King, and S.C. McClusky. Gamit reference manual. *Department of Earth, Atmospheric, and Planetary Sciences Massachusetts Institute of Technology, Cambridge*, 2010. URL http://www-gpsg.mit.edu/.

R. J. Hill, R. S. Lawrence, and J. T. Priestley. Theoretical and calculational aspects of the radio refrac- tive index of water vapo. *Radio Sci*, 17(5):1251–1257, 1982. URL doi:10.1029/RS017i005p01251.

T. Hobiger, R. Ichikawa, Y. Koyama, and T. Kondo. Fast and accurate ray-tracing algorithms for real-time space geodetic applications using numerical weather models. *J. Geophys. Res.*, 113:D20302, 2008. URL doi:10.1029/2008JD010503.

A.G.A. Hoeven, R.F. Hanssen, and B.A.C. Ambrosius. Tropospheric delay estimation and analysis using gps and sar interferometry. *Physics and Chemistry of the Earth*, 27:385 – 390, 2002.

B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. Global positioning system: theory and practice. *Springer, Wien*, 2001.

H.S. Hopfield. Two-quartic tropospheric refractivity profile for correcting satellite data. *J. Geophys. Res*, 74:4487–4499, 1969. URL doi:10.1029/JC074i018p04487.

N. Houlie', P. Briole, A. Nercessian, and M. Murakami. Sounding the plume of the 18 august 2000 eruption of miyakejima volcano (japan) using gps. *Geophys Res Lett*, 32, 2005a. URL doi:10.1029/2004GL021728.

N. Houlie', P. Briole, A. Nercessian, and M. Murakami. Volcanic plume above mount st. helens detected with gps. *Eos Trans AGU*, 86:277, 2005b. URL doi:10.1029/2005EO300001.

G. C. Hulley and E. C. Pavlis. A ray-tracing technique for improving satellite laser ranging atmospheric delay corrections, including the effects of horizontal refractivity gradients. *J. Geophys. Res*, 112(B06417):1–19, 2007. URL doi:10.1029/2006JB004834.

I. Ifadis. The atmospheric delay of radio waves: Modeling the elevation dependence on a global scale. *Technical Report 38L, School Electrical Computer Engineering, Chalmers University of Technology, Göteborg, Sweden*, 1986. URL ISBN:99-0605353-4.

J.V. Iribarne and H.R. Cho. Atmospheric physics. D. Reidel Publishing Company, Dordrecht, Holland., 1980.

A. Karabatic, R. Weber, and T. Haiden. Near real-time estimation of tropospheric water vapour content from ground based gnss data and its potential contribution to weather now-casting in austria. *Adv. Space Res*, 47(10):1691–1703, 2011. URL doi:10.1016/j.asr.2010.10.028.

J.A. Klobuchar. Ionospheric effects on gps. *Global positioning system: theory and applications. Parkinson BW, Spilker JJ (eds)*, 1:485–515, 1996.

A. Kulshrestha. Gps receivers for civil and military applications. *National Conference on Global Positioning System, 21-23 Feb, 1997, IIT Kanpur, India.*, 1997.

K.M. Larson. A new way to detect volcanic plumes. *Geophys. Res. Lett.*, 40:2657–2660, 2013). URL doi:10.1002/grl.50556.

K.M. Larson, E.E. Small, E. Gutmann, A. Bilich, P. Axelrad, and J. Braun. Use of gps receivers as a soil moisture network for water cycle studies. *Geophys. Res. Lett.*, 35, 2008. URL doi:10.1029/2008GL036013.

K.M. Larson, E. Gutmann, V. Zavorotny, J. Braun, M. Williams, and F. Nievinski. Can we measure snow depth with gps receivers? *Geophys. Res. Lett.*, 36, 2009. URL doi:10.1029/2009GL039430.

R.F. Leandro, M.C. Santos, and R.B. Langley. Unb neutral atmosphere models: Development and performance. In National Technical Meeting of The Institute of Navigation, Monterey, California,:564–573, 2006.

W.H. Lee and V. Pereyra. Mathematical introduction to seismic tomography. *Ed. London: Chapman & Hall*, Seismic Tomography: Theory and Practice:9–22, 1993.

A. Leick. Gps satellite surveying. 2nd ed., John Wiley & Sons, New York., 1994.

K. Levenberg. A method for the solution of certain nonlinear problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

D.R. Lide. Crc handbook. of chemistry and physics. 78th ed., crc press, boca raton, flor. 1997.

H. J. Liebe. An updated model for millimeter wave propagation in moist air. *Radio Sci.*, 20(5):1069–1089, 1985. URL doi:10.1029/RS020i005p0106.

H. J. Liebe. Mpman atmospheric millimeter-wave propagation model. *Int. J. Infrared Millimeter Waves*, 10(6):631–650, 1989. URL doi:10.1007/BF01009565.

H.J. Liebe, G.A. Hufford, and M.G. Cotton. Propagation modeling of moist air and suspended water/ice particles at frequencies below 1000 ghz. *In Proc. AGARD 52d*

Specialists Meeting of the Electromagnetic Wave Propagation Panel, Palam de Mallorca, Spain, AGARD, page 3.1–3.10, 1993.

J.V. lribarne and W.L. Godson. *Geophysics and Astrophysics Monographs., D. Reidel Publishing Company, Dordrecht, Holland.*

F.K. Lutgens and E.J. Tarbuck. The atmosphere - an introduction to meteorology. Prentice-Hall, Englewood Cliffs, New Jersey, 1979.

J.W. Marini. Correction of satellite tracking data for an arbitrary tropospheric profile. *Radio Science*, 7(2):223–231, 1972. URL doi:10.1029/RS007i002p00223.

J.W. Marini and C.W. Murray. Correction of laser range tracking data for atmospheric refraction at elevation angles above 10 degrees. *Technical Report X-591-73-351, NASA*, 1973.

F.J. Massey. The kolmogorov-smirnov test for goodness of fit. *J Am Stat Assoc*, 46: 68–78, 1951. URL doi:10.2307/2280095.

D. Massonnet, M. Rossi, C. Carmona, F. Adragna, G. Peltzer, K. Feigl, and T. Rabaute. The displacement field of the landers earthquake mapped by radar interferometry. *Nature*, 364:138–142, 1993.

G. McPherson. Statistics in scientific investigation: its basis, application and interpretation. *Springer*, 1990. URL ISBN0-387-97137-8.

C. Meertens, C. Rocken, J. Braun, B. Stephens, C. Alber, R. Ware, M. Exner, P. Kolesnikoff, and A. Niell. Antenna type, mount, height, mixing, and snow effects in high accuracy gps observations, in the global positioning system for the geosciences. *National Academies Press, Washington*, page 211–218, 1997.

V. B. Mendes. Modeling of the neutral-atmosphere propagation delay in radiometric space techniques. *PhD dissertation, Department of Geodesy and Geomatics Engineering Tech. Report No.199, University of New Brunswick, Fredericton, New Brunswick, Canada*, 1999.

V. B. Mendes and R. Langley. Tropospheric zenith delay prediction accuracy for airborne gps high-precision positioning. *In Proc. of ION GPS, Nashville, TN, USA*, 98:337–347, 1998.

W. Menke. Geophysical data analysis: discrete inverse theory. *Academic Press, San Diego.*, 1984.

A. Miller and J.C. Thompson. Elements of meteorology. 3rd ed., Charles E. Merril Publishing Company, Toronto., 1979.

V. Nafisi, M. Madzak, J. Böhm, A. A. Ardalan, and H. Schuh. Ray-traced tropospheric delays in vlbi analysis. *Radio Sci.*, 47:RS2020, 2012a. URL doi:10.1029/2011RS004918.

V. Nafisi, L. Urquhart, M. C. Santos, F. G. Nievinski, J. Böhm, D. D. Wijaya, H. Schuh, A. A. Ardalan, T. Hobiger, R. Ichikawa, F. Zus, J. Wickert, and P. Gegout. Comparison of ray-tracing packages for troposphere delays. *IEEE Trans. Geosci. Remote Sensing*, 50(2):469–481, 2012b. URL doi:10.1109/TGRS.2011.2160952.

S. Nakada, M. Nagai, T. Kaneko, A. Nozawa, and K. Suzuki-Kamata. Chronology and products of the 2000 eruption of miyakejima volcano. *B Volcanol*, 67:205–218, 2005. URL doi:10.1007/s00445-004-0404-4.

A. Niell. Global mapping functions for the atmosphere delay at radio wavelengths. *J. Geophys. Res.*, 101(B2):3227–3246, 1996. URL doi:10.1029/95JB03048.

A. E. Niell. Improved atmospheric mapping functions for vlbi and gps. *Earth Planets Space*, 52:699–702, 2000.

A. E. Niell. Preliminary evaluation of atmospheric mapping functions based on numerical weather model. *Phys. Chem. Earth(A)*, 26:475–480, 2001. URL doi:10.1016/S1464-1895(01)00087-4.

A. E. Niell. Interaction of atmosphere modeling and vlbi analysis strategy. In D. Behrend and K.Baver, editors, International VLBI Service for Geodesy and Astrometry 2006 General Meeting Proceedings, number NASA/CP-2006-214140,, 2006.

F. G. Nievinski. Ray-tracing options to mitigate the neutral atmosphere delay in gps.master's thesis. *University of New Brunswick, Department of Geodesy and Geomatics Engineering*, Technical Report No. 262, 2009. URL http://hdl.handle.net/1882/1050.

T. Nilsson. Path delays in the neutral atmosphere. *Springer Atmospheric Sciences*, Atmospheric Effects in Space Geodesy:73–136, 2013. URL 10.1007/978-3-642-36932-2_3.

T. Nilsson and L. Gradinarsky. Water vapor tomography using gps phase observations: Simulaton results. *IEEE Trans. Geosci. Remote Sensing*, 44(10):2927–2G1, 2006. URL doi:10.1109/TGRS.2006.

T. Nilsson, L. Gradinarsky, and G. Elgered. Water vapour tomography using gps phase observations: Results from the escompte experiment. *Tellus A*, 59:574–682, 2007. URL doi:10.1111/j.1600-0870.2007.00247.x.

T. Ning and G. Elgered. Trends in the atmospheric water vapor content from ground-based gps: The impact of the elevation cutoff angle. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5:744–751, 2012. URL doi:10.1109/JSTARS.2012.2191392.

C. Oppenheimer, D.M. Pyle, and J. Barclay. Volcanic degassing. *Geol Soc Spec Publ, London*, page 213–420, 2003.

J. C. Owens. Optical refractive index of air: Dependence on pressure, temperature and composition. *Appl. Opt.*, 6(1):51–59, 1967. URL doi:10.1364/AO.6.000051.

M. Palano, M. Rossi, F. Cannavo', V. Bruno, M. Aloisi, D. Pellegrino, M. Pulvirenti, G. Siligato, and M. Mattia. Etn@ref: a geodetic reference frame for mt. etna gps networks. *Ann Geophys-Italy*, 53:49–57, 2010. URL doi:10.4401/ag-4879.

T. K. Pany. Development and application of tropospheric gps slant delay models based on numerical weather prediction models and turbulence theory. *PhD thesis, Institute of Engineering Geodesy and Measurements Systems, Graz University of Technology,*, 2002.

R.L. Parker. Geophysical inverse theory. *Princeton University Press, Princeton, New Jersey*, 1994.

J.P. Peixoto and A.H. Oort. Physics of climate. *American Institute of Physics, New York.*, 1992.

D. Perler, A. Geiger, and F. Hurter. 4d gps water vapor tomography: new parameterized approaches. *J. Geodesy*, 85(8):539–550, 2011. URL doi:10.1007/s00190-011-0454-2.

E.J. Petrie, M.A. King, P. Moore, and D.A. Lavallée. Higher order ionospheric effects on the gps reference frame and velocities. *J. Geophys. Res.*, 115(B03417), 2010. URL doi:10.1029/2009jb006677.

P. Poli, P. Moll, F. Rabier, G. Desroziers, B. Chapnik, L. Berre, S. B. Healy, E. Andersson, and F.-Z. El Guelai. Forecast impact studies of zenith total delay data from european near real-time gps stations in météo france 4dvar. *J. Geophys. Res.*, 112:D06114, 2007. URL doi:10.1029/2006JD007430.

J. M. Rüeger. Refractive index formulae for radio waves. *In Proc. XXII FIG International Congress, Washington DC, USA*, 2002a. URL http://www.fig.net/pub/fig_2002/procmain.htm.

J. M. Rüeger. Refractive indices of light, infrared and radio waves in the atmosphere. Technical report, UNISURV S-68, School of Surveying and Spatial Information Systems, The University of New South Wales, Australia, 2002b.

R.R. Rogers and M.K. Yau. A short course in cloud physics. 3rd ed. pergamon. press, new york. 1989.

M. Rothacher, T.A. Springer, S. Schaer, and G. Beutler. Processing strategies for regional gps networks. *Advances in Positioning and Reference Frames, IAG Symposia Series*, 118:93–100, 1998.

J. Saastamoinen. Introduction to practical computation of astronomical refraction. *Bull. Géod.*, 106:383–397, 1972a. URL doi:10.1007/BF02522047.

J. Saastamoinen. Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites agu, in s. w. henriksen et al. washington, d.c. The Use of Artificial Satellites for Geodesy, volume 15:247–251, 1972b.

G. Saito, K. Uto, K. Kazahaya, H. Shinohara, T. Kawanabe, and H. Satoh. Petrological characteristics and volatile content of magma from the 2000 eruption of miyakejima volcano. *B Volcanol*, 67:268–280, 2005. URL doi:10.1007/s00445-004-0409-z.

R. Schmid, M. Rothacher, D. Thaller, and P. Steigenberger. Absolute phase center corrections of satellite and receiver antennas. impact on gps solutions and estimation of azimuthal phase center variations of the satellite antenna. *GPS Solut*, 9:283–293, 2005. URL doi:10.1007/s10291-005-0134-x.

S. Scollo, M. Prestifilippo, G. Spata, M. D'Agostino, and M. Coltelli. Monitoring and forecasting etna volcanic plumes. *Nat Hazard Earth Syst*, 9, 2009. URL doi:10.5194/nhess-9-1573-2009.

S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965. URL doi:10.1093/biomet/52.3-4.591.

R. Snieder and J. Trampert. Wavefield inversion. *Ed. A. Wirgin, Springer Verlag, New York,*, pages 119–190, 2000. URL ISBN978-3-211-83320-9.

F. S. Solheim, J. Vivekanandan, R. H. Ware, , and C. Rocken. Propagation delays induced in gps signals by dry air, water vapor, hydrometeors, and other particulates. *J. Geophys. Res.,*, 104(D8):9663–9670, 1999. URL doi:10.1029/1999JD900095.

G. Strang. Linear algebra and its applications. *Harbourt Brace Jovanovisch Publishers, Fort Worth*, 1988.

A. Tarantola. Inverse problem theory. *SIAM*, 2005. URL ISBN0-89871-572-5.

G. D. Thayer. A rapid and accurate ray tracing algorithm for a horizontally stratified atmosphere. *Radio Sci.*, 1(2):249–252, 1967.

G. D. Thayer. An improved equation for the radio refractive index of air. *Radio Sci*, 9(10):803–807, 1974. URL doi:10.1029/RS009i010p00803.

R. N. Thessin. Atmospheric signal delay affecting gps measurements made by space vehicles during launch, orbit and reentry. *Master's thesis, Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, Cambridge*, 2005. URL http://hdl.handle.net/1721.1/33211.

C.H. Thurber. Earthquake locations and three-dimensional crustal structure in the coyote lake area, central california. *Journal of Geophysical Research*, 88:8226–8236, 1983.

C.H. Thurber and K. Aki. Three dimensional seismic imaging. *Annual Review of Earth and Planetary Sciences*, 15:115–139, 1987.

C.H. Thurber and W.L. Ellsworth. Rapid solution of ray tracing problem in heterogeneous medium. *Bulletin of the Seismological Society of America*, 70:1137–1148, 1980.

C.H. Thurber, H.M. Iyer, and K. Hirahara. Local earthquake tomography: velocities and vp/vs-theory. *Ed. London: Chapman & Hall*, Seismic Tomography: Theory and Practice:563–580, 1993.

D.R. Toomey and G.R. Foulger. Application of tomographic inversion to local earthquake data from the hengill-grensdalur central volcano complex, iceland. *Journal of Geophysical Research*, 94:497–510, 1989.

J. Trampert. Global seismic tomography: the inverse problem and beyond, inverse problems. 14:371–385, 1998.

P. Tregoning and T.A. Herring. Impact of a priori zenith hydrostatic delay errors on gps estimates of station heights and zenith total delays. *Geophys. Res. Lett.*, page 33(L23303), 2006. URL doi:10.1029/2006GL027706.

K. E. Trenberth, A. Dai, R.M. Rasmussen, and D.B. Parsons. The changing character of precipitation. *Bull. Amer. Meteor. Soc.*, 84(9), 2003. URL doi:10.1175/BAMS-84-9-1205.

M. Troller, A. Geiger, E. Brockmann, J.-M. Bettems, B. Bürki, and H.-G. Kahle. Tomographic determination of the spatial distribution of water vapor using gps observations. *Adv. Space Res*, 37(12):2211–2217, 2006. URL doi:10.1016/j.asr.2005.07.002.

F. T. Ulaby, R. K. Moore, and A. K. Fung. Microwave remote sensing active and passive. I, 1981. URL ISBN-10:0890061904.

G. Wadge, P.W. Webley, I.N. James, R. Bingley, A. Dodson, S. Waugh, T. Veneboer, G. Puglisi, M. Mattia, D. Backer, S.J. Edwards S.C. Edwards and, and P.J. Clarke. Atmospheric models, gps and insar measurements of the tropospheric water vapour field over mount etna. *Geophysical Research Letters*, 29, 2002.

J.M. Wallace and P.V. Hobbs. Atmospheric science: An introductory survey. *Academic Press, New York.*, 1977.

J. Wang, L. Zhang, and A. Dai. Global estimates of water-vapor-weighted mean temperature of the atmosphere for gps applications. *J. Geophys. Res*, 110(D21101), 2005. URL doi:10.1029/2005JD006215.

P.W. Webley and G. Wadge. Determing radio wave delay by non-hydrostatic atmospheric modelling of water vapour over mountains. *Physics and Chemistry of the earth*, 29:139–148, 2004.

D. D. Wijaya. Atmospheric correction formula for space geodetic techniques. phd thesis. *University of Technology, Institute of Engineering Geodesy and Measurements Systems, Graz, Austria*, 2010.

S. Williams, Y. Bock, and P. Fang. Integrated satellite interferometry: Tropospheric noise and gps estimates and implications for interferometric synthetic aperture radar products. *J. Geophys. Res.*, 103(B11):27051–27067, 1998. URL di: 10.1029/98JB02794.

H. A. Zebker, P. A. Rosen, and S. Henseley. Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. *Journal Geophys. Res.*, 102:7547–7563, 1997.