Masters in Global Change

School of Biology and Environmental Science

University College Dublin

&

School of Biology

Justus Liebig University

Title:

The Creation of a database and Error Flagging system for Climate data of Angola (Period of 1961 to 1974)

Author: Nídia Loureiro

Student Number: 14205883

Supervisor: Dr. Jon Yearsley

Date of Submission: 14th December 2015

Masters in Global Change

School of Biology and Environmental Science
University College Dublin
&
School of Biology
Justus Liebig University

Title:

The Creation of a Database and Error Flagging System For Climate Data of Angola (Period Of 1961 To 1974)

The thesis is submitted to University College Dublin in part fulfilment of the requirements for the degree of Master of Science (M.Sc.)

Author: Nídia Loureiro
Student Number: 14205883

Supervisor: Dr. Jon Yearsley

Date of Submission: 14th December 2015

Acknowledgements

At first I would like to thanks Dr.Florence (The Director of the Master of Science in Global Change: Ecosystem Science and Policy - School of Biology and Environmental Science Agriculture and Food Science Centre University College Dublin) for all the help, willingness and availability that made me go through this Master course. For the great opportunity to be supervised by Dr. Jon Yearsley (Lectures in evolution ecology in the School of Biology and Environmental Science in UCD.) Thanks for being patient in teaching and helping me thought R's codes and data understanding and for leading me to the right way towards the completions of my thesis, it was more than a pleasure to be working with you. Many thanks for Dr. Conor (Lecturer, in the Computational Mathematics field in the School of Mathematical Sciences in UCD and in collaboration with Meteorology and Climate Science.) due to his thorough knowledge in meteorology helped me to understand more about my data and also referred me to Seamus Walsh who works in the Met. Eireann. With his experience in Meteorological data and Error Checking lead me to the right way of approaching the error checking of the data, many thanks. Thanks to Paul Burns, (Postdoctoral researcher in School of Biology and Environmental Science, Earth Institute, in UCD, under Dr. Jon Yearsley supervision) Thanks for going through my difficulties with coding and for being patient with me as I was hard to understand some times. To my Teacher, Carlos Ribeiro (The coordinator of the Environmental Engineering Course – ISPT) my Scholarship supervisor, goes my huge thanks for the opportunity of a better education and a place in the SASSCAL project which resulted in this internship opportunity. And finally I would like to thank my family for the support, especially my mother and father for the great support in staying with my children. My husband and children for being so understanding, and to my brother and sister in law for the support. My Grandmother for having me every time I had to travel to Ireland and Back home, my uncle, aunt and cousins. Also my friends for the support.

Thanks all for your support and willingness to help me through this my Master's completion.

Table of Contents

List of Figures:

Abstract

Climate data of Angola from the period of 1961 to 1974 was rescued from the Coimbra University Archives and digitized into excel spread sheets. This data resumes summary months of climate variable recordings such as Air Temperature, Humidity, Nebulosity, Precipitations and phenomenon's associated with rain such as lightning and thunderstorms. The data gathers district and stations names and the three geographical coordinates and as well the dates of the occurrence of extremes. In this project the aim involves the compilation of the monthly spread sheets in to a database. Along with this aim emerge the need for the organization of the data in order to be used in the future. The aim of the projects covers as well the association of IDs to each stations and district and the transformation of the geographical coordinates into standardized ones. Besides this cover as well a quality control procedure which flags erroneous data. With the creation of an error checking system using the tolerance and consistency tests to spot errors. The data was imported from Excel into R. and R was chosen for being an open source software, very adaptable and actively supported by a large number of users and also ideal for statistical analyses. In R was possible to fulfil the objective stabled for this project. The database in compiled and ready to a next stage (Correction) which is out of the scope of this project. However it is necessary to notice that the process involving this database is continuous and in this project objectives are met although it is only a first step to the complete and complex task of turning available a database ready to be used.

1. Introduction

Climate data has been recorded for centuries. From rudimentary scriptures, in ancient times, to more precise and accurate readings along the years with the creation of instrumentation and continuous technology improvements, as the liquid thermometer and the more accurate analogical and digital thermometers of today. The World Meteorological Organization (WMO) is the most important organization which has been responsible for the standardization of recording methods. Do it by producing guides, procedures, recommendation and reports about the meteorological practices and climate data rescue, recording and managing besides make it worldwide available.

Angola at the time of the observations was still a colony of Portugal, considered a province of the Portuguese State. Therefore, since Portugal was in accordance with international parameters for collecting meteorological data, it was also Angola. However, from 1975 until 2002 Angola suffered a civil war. This war prevented the collection of meteorological data and eroded the country's meteorological network. These decades of war occurred at the same time of many technological innovations on the meteorological instruments and rules. Therefore, Angola was deficient in this and other branches of development.

In previous years (from 2009) within the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL), Angola step into a rebuilding process towards the re-establishing of the Angola's Meteorological network. Along with new data collection came as well the need for data rescue. With this, meteorological historical data from 1961 to 1974, when Angola was still a Portuguese Colony, was digitized and the records scanned to ensure its safety. Data rescue will continue along with the re-establishment of the meteorological network within the SASSCAL project.

This paper aims is to compile all 14 years of data for the south Angola into a database. I have sub-divided this aim into the following objectives:

(1) Import data;

(2) Create a quality control procedure to flag errors of digitalization (keying mistakes, temperature, precipitation and humidity values outside logical ranges, and missing values),

(3) Create a definitive list of all the station names

(4) Assign identification codes to district and station, and

(5) Create a metadata which describes the data in the database.

It is important to notice that this is a continuous work as many of the procedures will require long time of manual checking which is beyond the end of this project. All the procedures executed in this database will be applied to the data of the north of the country as well, in the future. The long-term goal of this work is to compile a meteorological database for the whole country that follows international best-practice.
These database has the purpose to harbour old data before computer era as in the future will be built a platform to join with the present, future and near past data to it. Therefore many relatively new requirements will not be applied due to not be compliant with the data.

1.1 Background of the study

1.1.1 Brief history and development of meteorology observation and instrumentation

It is important to briefly distinguish the two main important terms when talking about the recording of weather elements such as Meteorology and Climatology. Meteorology is the science which studies the weather (Linacre, 1992) it concerns about weather recording and forecast (Allaby & Allaby, 2009). The science of Climatology studies the weather observation (recordings) of long periods of time, as the term climate describes the variability of the weather in a long period of time (normally 30 years, the minimum standard limit of years for climatological summary studies). Focuses on statistical studies (averages, means, etc.) to describe or make assumption of the weather (Carrega, 2010).

The idea of studying, predicting and recording the weather exists since ancient times, Babylonians did it 3000 years ago (Linacre, 1992; Nebeker, 1995). Wind direction record were taken by the Greeks in (ca. 430 B.C). Precipitation records are known to have started around 1440's (Barry, & Chorley, 2010). The theory of meteorology exists since Aristotle times (ca.340 B.C), before the rise of science in the 16th and 17th century (Linacre, 1992; Nebeker,1995). In the early 1600s, Galileo, started the first temperature observations with the invention of the thermometer, however liquid-in-glass thermometers with calibrated scales were still unavailable, (Fahrenheit) appeared in the early 1700s and/or (Celsius) the 1740s (Barry, & Chorley, 2010). The relative humidity sensor, the hair hydrometer, was invented in 1780 by de Saussure (Barry, & Chorley, 2010). With the creation of instruments, such as the thermometer and barometer in the 17th century it was possible to measure the elements of the weather (Nebeker, 1995). Descartes, Edmond Halley and others, were making almanacs with weather prognostications and making available (Nebeker, 1995).

In the 19th century the cultivation of the science "climatology" joined with the increase of people doing the empirical, theoretical, and practical meteorology activities. Descriptive science (climatology) was the result of studies of empirical meteorology as their focus in average weather, while the focus on the theory based on laws of physics made the branch of dynamical meteorology. Theorists were relying on the relatively small amount of observation to do forecasting and develop practical observation treatises. From the 1870s and after, weather forecasting was established as a profession (Nebeker, 1995). From this time national meteorological services started producing daily forecast.

The establishment of the network of observing stations and the standardisation of observation procedures was essential for meteorology of the 1850's for both Europe and North America (Nebeker, 1995). The telegraph was highly important as a mean of rapid data exchange.

The inter world wars period was crucial for the meteorological development before the 1950s, such as the use of frequencies of different weather types, by Federov in 1921, the concepts of variability of temperature and rainfall, by Gorczynski 1942 and 1945, and microclimatology, the study of the fine climate structure close to the surface by Geiger 1927 (Barry, & Chorley, 2010). The unification of the meteorology had meaning with the computer assistance since 1950s and 1960s (Nebeker, 1995). Later in the 1970s started the recognition of human activity on the environment with as well the realization of the global climate system and the importance of the balanced and dependent relationship between the subsystems such as atmosphere and biosphere. (Barry, & Chorley, 2010).

Although the International Meteorological Organization (IMO) was founded in 1873, only in 1929 it was created the Commission for climatology (CCI). Under IMO umbrella only in 1950, after the Second World War, the WMO was incorporated in the United Nations, as a specialized Agency and IMO successor. The Commission for climatology (CCI) main objectives goes from the collection and managing of data to data transformation, climate forecasts and other climate information (such as projections) into high quality available information (World Meteorological Organization, 2011). WMO and the International Council on Science created the Global Atmospheric Research Programme (GRAP) and the World Climate Research Programme, in the

1980s, leading the climate investigation through coordinated intensive programs of observations, for example the World Ocean Circulation Experiment (WOCE) with the purpose of bringing information on the global currents and global thermohaline circulation (Barry, & Chorley, 2010; World Meteorological Organization, 2011). Since then, WMO has been ahead of the meteorological observation practices. Concerns about data recording, data rescue and management, working as a high authority into the recommendations for the climate data recording (World Meteorological Organization, 2011).

1.1.2 Angola and the meteorology observation

Angola at the time of the observations was still a colony of Portugal, considered a province of the Portuguese State. Therefore, since Portugal was in accordance with international parameters for collecting meteorological data, it was also Angola. The meteorology observation in Portugal started in the 17th century, in a peninsular context the first observations recorded in Lisbon dated from 1724, which were published in the official publication of the Royal Society, The Philosophical Transactions. (Nunes, Alcoforado, & Cravosa, 2014) from 1770 to 1784 are the first observations published in the country (Portugal), however the first meteorological observation test with statistical and climatological purposes dated 1792 (Monteiro, 2001). In Porto where temperature, relative humidity and wind was recorded twice a day. Only in 1854 started the interest on meteorological services as it was established the first international Meteorological Observatory Infante D.Luis (Observatório Meteorológico Infante D.Luis) The Observatory of The Coimbra University started operating ten years later 1863 and the Observatory of Porto University in 1888. Yet in 1864 climatological information was being published and near the end of the century meteorological station were spreading all over the country (Monteiro, 2001). Within this plan it is broaden to a colonial plan with the establishment of an international meteorological system for weather forecast an idea created by Brito Capêlo (1831-1901), in which the whole territory space of the Portuguese State were implementing the network. In the second half of the 19th Century the plan was put in practice in the main capitals of the Portuguese Colonies. In 1857 Angola was incorporated, being done the first bridge between Lisbon and Luanda in 1857 (Nunes, Alcoforado, & Cravosa, 2014).

Unfortunately much of the information on meteorology in Portugal and the Colonies are still in paper in Libraries in Portugal, which limits my access. However, it is mentioned in the Serviços Meteorologicos de Angola (S.M.A), of (1940) that the recording in the Angola Colony was being done since the early 1900, informs as well about instruments and the times which the observations were being taken. According to the source the observations were following international requirements of the (IMO) (S.M.A, 1940).

1.1.3 *The Southern African Science Service Centre for Climate Change and Adaptive Land Management*

The Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) is a Regional Science Service Centre in Southern Africa. It is an initiative that puts together Angola, Botswana, Namibia, South Africa, Zambia, and Germany, responding to the challenges of global change and adding value the whole region (SASSCAL, n.d.). Problem-orientated research on the area to adaptation to climate and change and sustainable management. Provide evidence based advice for all stakeholders and decision makers to improve the livelihoods of people in the region and as a contribution to the creation of a knowledge based society is the overall project mission (SASSCAL, n.d.). Thereby, it vital for SASSCAL to gain extensive knowledge on climate conditions in order fulfil in its response towards its main objects. Beside the implementations of new meteorological network system remain the importance of the past historical weather recordings which are a base for climatological studies (Brunet & Jones, 2011).

1.1.4 *The Instituto Superior Politécnico Tundavala (ISPT) and the SASSCAL*

The Instituto Superior Politécnico Tundavala (ISPT), is a Superior Institute which is the head for the 141 Task of the SASSCAL project in the South region (SASSCAL, 2013). The ISPT is responsible for the maintenance and management of the meteorological stations. The ISPT also has students that are working together on the SASSCAL project, with data digitalization and data analysis. The rescue, processing and management of old data coming along side with the re-establishment of the new AWS is as well an objective of the Task 141. In Particular to Angola and to ISPT, the 141 Task has the aim of development of meteorological observation network system in the Southwest of the Country, covering Namibe province, the western slopes of Serra da Chela and Huíla province (SASSCAL, 2013). New meteorological stations have started recording data in the south region of Angola. Taking in consideration only the 141 Task which covers the Southwest of the Country, covering Namibe province, the western slopes of Serra da Chela and Huíla province. The full scope of SASSCAL It will extend for the whole country. This reliable climate data will be highly useful for applications in Civil Engineering, Agriculture and Aviation, The data will also enable further long and short term weather forecast at regional and national levels. The data will be will be immediately helpful in the building of new infrastructures (i.e. bridges and hydraulic aqueducts) having in account the water regime, helping in farmers towards planning having in account weather information and also preventing human lost from flooding and droughts as well as supportive programmes. (SASSCAL, 2013). It is a late but highly important task since Angola does not possess any available data from the past during colonization (from 1900 to until 1974), as the data exists in Portugal and international library. less developing countries (LDCs) are late in the historical data rescuing due to many factors being among the lack of economic resources the major reason. Besides economic issues Angola hold on war for many years. The overall project is sponsored by the SASSCAL and Germany Government. More information on the SASSCAL project is available in the website ("SASSCAL," n.d. Southern African Science Service Centre for Climate Change and Adaptive Land Management.: http://www.sasscal.org/)

1.2 Data Rescue and database structure recommendation

1.2.1 *The importance and purpose of rescuing and managing data*

The main objective of rescuing and managing climate data is to preserve, organise and provide access to climate data. The National Meteorological Services around the world are engaged in data rescue. Data rescue can be put in words as the process of finding meteorological archives from the past and making them available in computer compatible forms. Data rescue also involves saving both processed and original archives from deterioration (Tan, Burton, & World Meteorological Organization, 2004). In the United States, NOAA's National Climate Data Centre is (and has been) digitizing billions of observations (Tan, Burton, & World Meteorological Organization, 2004). The WMO has many projects for data rescue and management, such as DARE ("CDM_2 WCDMP | WMO," n.d.) (See in http://www.wmo.int/pages/prog/wcp/wcdmp/CDM_2.php) DARE are sponsoring (LDCs) for the rescue of climate data from countries in Africa, Asia, in the Caribbean and South pacific (World Meteorological Organization & Meeting of the CCl Expert Team on the Rescue, Preservation and Digitization of Climate Records, 2008). It is important to create and increase awareness of the essential need to undertake integrated DARE projects, especially among policy-makers, stakeholders and climate data end-users. To consider the long term benefits and not only the costs associated with the improvement of climate data availability (Brunet & Jones, 2011).

This process of bringing historic data provides to the present the foundations for the understanding and assessment of climate variability. It provides the possibility to not just predict extreme climate events but as well to plan strategies for adaptation and mitigations. Rescued data are important to inform climate adaptation policy which results in significant impact on the livelihood of local community, especially in developing countries (Munang, Nkem, & Han, 2013). For preventing loss of lives, goods and properties as well as allowing better planning for crop production combatting hunger in some areas by knowing the extremes of past events (SASSCAL, 2013). Indispensable for the design of buildings for the future, such as roads, bridges and drainages systems (SASSCAL, n.d.; SASSCAL, 2013).

The information on climate of the past is vital for the future predictions supporting the convey of policy responses (Hawkins et al., 2013). It is an absolutely vital source of information to planners, decision makers and researcher (World Meteorological Organization, 2009). It is important for climate studies, researchers and studies such as climate change (World Meteorological Organization, 2008). Understanding the past and the present it essential to better understand, predict and plan responses to global climate change (Brunet & Jones, 2011).

In World Meteorological Organization, (1996) remarks for the Policy which identifies the importance of free and unrestricted exchange of meteorological data and products as a fundamental principle of the World Meteorological Organization. It as well includes the United Nations Framework Convention on Climate Change (IPCC) to promote and cooperate, in full, with the open exchange of information related to the climate system and climate change. The policy implies that it is an obligations of WMO members to facilitate worldwide cooperation in the establishment of observing networks and as well to continuing promotion in the exchange of meteorological and related information.

1.2.2 *Meteorology instruments of the time (1960s and 1970s)*

Any rescued data should provide all the information regarding the instruments used for the recording of climate data. Since this data records do not provide any information of the instruments and observations rules in which the observations were done, it is possible to assume that the same conditions as stated back from 1940s and 1950s did not change much in the 1960s and 1970s, however changes must be done in this documents if literatures that states such practices are available.

The Serviços Meteorológicos de Angola (SMA) ("Meteorological services of Angola") were using the following instruments:

Thermograph (Richard Thermograph); Mercury and alcohol thermometers; Piché Evaporimeter, Rain Gauge (Negretti and Zambra,), relative humidity sensor (Hygrothermograph by Negretti and Zambra). Although our data do not have recordings of solar radiation, at the time it was used the Campbell-Stockes and Jordan Heliographer and air pressure with Richards registering Barometer as well as balloons for upper atmosphere measurements (SMA, 1951).

1.2.4 *Data rescue procedures in Angola (within ISPT - task 141)*

Data Rescue: The Procedure in Angola

This data gathers climate data of Angola of the years of 1961 to 1974. This data was recorded while Angola was a Portuguese Colony. The data was rescued by the Instituto Superior Politécnico Tundavala (ISPT) from the Coimbra University, place where the original archives were saved. The rescue project is linked and sponsored by the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) project within the 141 task. One of the tasks of 141 main Task is the rescue of old climate data. The head of the project in the South, Huíla-Lubango , with Carlos Ribeiro and his team of two people (Nídia Loureiro and Sílvio Filipe) responsible for digitalization of the data and the creation of a database. The data consists of monthly summaries over 14 years. The data are divided into three tables: climatological stations (station names, 3 geographical coordinates, temperature, relative humidity, precipitation and cloudiness), precipitation stations (station names, 3 geographical coordinates and precipitation), evaporation and evapotranspiration of the main river basins (Basins Names, 3 geographical coordinates which Recording, started in 1970).

The digitalization was done manually by typing into Excel spread sheets as a verbatim copy of the original. Optical character recognition (OCR) software, a type of software (program) that can automatically recognise printed text and turn it into a machine readable text format. The software works by analysing a document and comparing it with fonts stored in its database (World Meteorological Organization, 2008) and abandoned as produced no positive results although, this method is widely used as mentioned in Meteorological Organization, (2008). A scan of the documents was also done and there are spread copies of both documents saved. The original paper copies are stored in the ISPT archives

1.2.5 *Procedure on rescued data - First phase*

The procedures to rescue the old data were done according to the guide Guidelines on Climate Data Rescue ((Plummer, Lipa, Palmer, & World Meteorological Organization, 2007; Tan et al., 2004), where the data was digitized into a standard format (Excel spread sheets) and scanned images (jpg images of 300 kb) were taken (figure 1 and 2) and both saved in digital format. The hard copies (Original documents), were stored in a controlled area awaiting for a NMHS to be saved in the future as recommended in Tan et al., (2004).

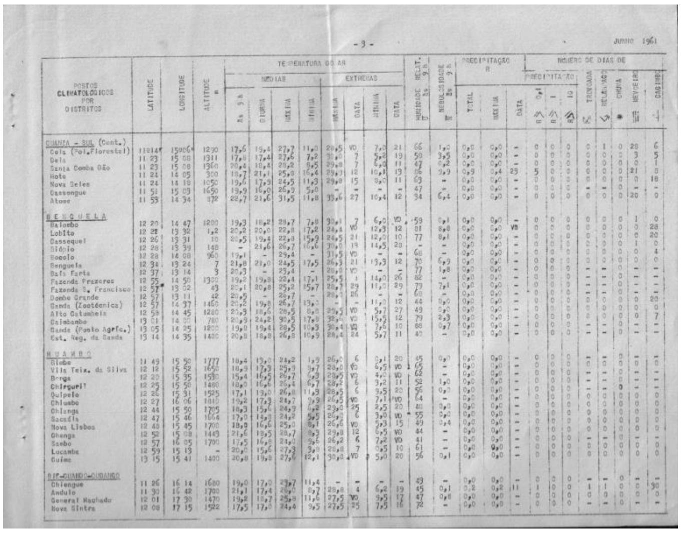
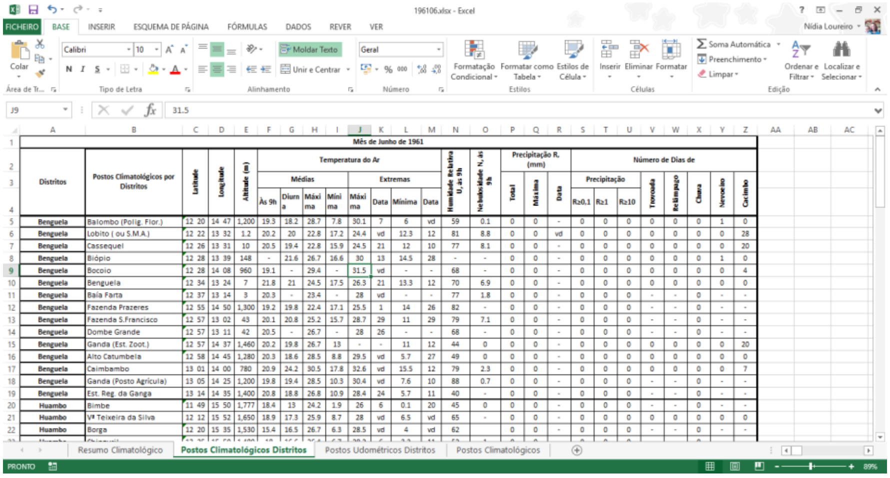


Figure 1- Original Verbatim of the Climatological tables of Angola



Mês de Junho de 1961																								
Distritos	Postos Climatológicos por Distritos	Latitude	Longitude	Altitude (m)	Temperatura do Ar								Humidade Relativa U, às 9h	Nebulosidade N, às 9h	Precipitação R. (mm)			Número de Dias de							
					Médias				Extremas									Precipitação							
					Às 9h	Diurna	Máxima	Mínima	Máxima	Data	Mínima	Data			Total	Máxima	Data	R≥0,1	R≥1	R≥10	Trovoada	Relâmpago	Chuva	Nevoeiro	Cacimbo
Benguela	Balombo (Polig. Flor.)	12 20	14 47	1,200	19.3	18.2	28.7	7.8	30.1	7	6	vd	59	0.1	0	0	-	0	0	0	0	0	0	1	0
Benguela	Lobito (ou S.M.A.)	12 22	13 32	1.2	20.2	20	22.8	17.2	24.4	vd	12.3	12	81	8.8	0	0	vd	0	0	0	0	0	0	0	28
Benguela	Cassequel	12 26	13 31	10	20.5	19.4	22.8	15.9	24.5	21	12	10	77	8.1	0	0	-	0	0	0	0	0	0	0	20
Benguela	Biópio	12 28	13 39	148	-	21.6	26.7	16.6	30	13	14.5	28	-	-	0	0	-	0	0	0	0	0	0	1	0
Benguela	Bocoio	12 28	14 08	960	19.1	-	29.4	-	31.5	vd	-	-	68	-	0	0	-	0	0	0	0	0	0	0	4
Benguela	Benguela	12 34	13 24	7	21.8	21	24.5	17.5	26.3	21	13.3	12	70	6.9	0	0	-	0	0	0	0	0	0	0	0
Benguela	Baía Farta	12 37	13 14	3	20.3	-	23.4	-	28	vd	-	-	77	1.8	0	0	-	0	0	0	-	-	0	-	-
Benguela	Fazenda Prazeres	12 55	14 50	1,300	19.2	19.8	22.4	17.1	25.5	1	14	26	82	-	0	0	-	0	0	0	-	-	0	-	-
Benguela	Fazenda S.Francisco	12 57	13 02	43	20.1	20.8	25.2	15.7	28.7	29	11	29	79	7.1	0	0	-	0	0	0	-	-	0	-	-
Benguela	Dombe Grande	12 57	13 11	42	20.5	-	26.7	-	28	26	-	-	68	-	0	0	-	0	0	0	-	-	0	-	-
Benguela	Ganda (Est. Zoot.)	12 57	14 37	1,460	20.2	19.8	26.7	13	-	-	11	12	44	0	0	0	-	0	0	0	0	0	0	0	20
Benguela	Alto Catumbela	12 58	14 45	1,280	20.3	18.6	28.5	8.8	29.5	vd	5.7	27	49	0	0	0	-	0	0	0	0	0	0	0	0
Benguela	Caimbambo	13 01	14 00	780	20.9	24.2	30.5	17.8	32.6	vd	15.5	12	79	2.3	0	0	-	0	0	0	0	0	0	0	7
Benguela	Ganda (Posto Agrícula)	13 05	14 25	1,200	19.8	19.4	28.5	10.3	30.4	vd	7.6	10	88	0.7	0	0	-	0	0	0	-	-	0	-	-
Benguela	Est. Reg. da Ganga	13 14	14 35	1,400	20.8	18.8	26.8	10.9	28.4	24	5.7	11	40	-	0	0	-	0	0	0	-	-	0	-	-
Huambo	Bimbe	11 49	15 50	1,777	18.4	13	24.2	1.9	26	6	0.1	20	45	0	0	0	-	0	0	0	-	-	0	-	-
Huambo	Vª Teixeira da Silva	12 12	15 52	1,650	18.9	17.3	25.9	8.7	28	vd	6.5	vd	65	-	0	0	-	0	0	0	0	0	0	0	0
Huambo	Borga	12 20	15 35	1,530	15.4	16.5	26.7	6.3	28.5	vd	4	vd	62		0	0	-	0	0	0	-	-	0	-	-

Figure 2 - Excel spread sheets with the digitalized climatological data form the original verbatim

1.2.6 *Database (WMO recommendations) – Second faze*

There are some requirements for the creation of a database. Many WMO guides and reports bring this information (*Climate data management system specifications*, 2014), however, many requirements are not applied to the type of rescuing data as it was developed for automatic and computerized data. The data does not provide information on stations codes and instruments types. The main requirements are met: station and district name and the three geographical coordinates. The information emphasises the important quality checking (QC) which should take place both before entering the data into the database and after the data have been entered (Aguilar, Llansó, & World Meteorological Organization, 2003). WMO advices for proper handling of the database is: periodic checking and a robust and secure system to avoid losing data (for example, from disk failure, problems with power supply, software and network security; Aguilar et al., 2003).

1.2.6.1 *Metadata*

Metadata is the information which accompanies the data of each station. It is considered very important once it give all details about the data, types of instruments and methods of observation, times and dates as well all the information regarding the station such as name, location and WMO code number or regional code number, as well as all the changes which might have been done to the stations or its instruments. Metadata informs the users of the conditions in which the data was recorded, compiled and transmitted, as well as the quality control checking applied, which allows users to be more precise about the accuracy of the conclusions of their analysis (Aguilar et al., 2003, Plummer et al., 2007). Lately with the development of Automatic weather stations (AWS) more information is required and gathered to the metadata accompanying any database, however for this rescued data much of the information regarding the procedures of recording, types of instruments and dates of installations is not available and will need further investigation and rescue for the completeness of the metadata. All the information available concerning the database and error checking procedure is present in the metadata of these database which is found in Appendix 2.

1.3 *Quality control*

Quality control is very important to ensure the quality of the data and to eliminate contamination of unrelated factors. It is a process in which mechanisms are used to eliminate many types of errors which can be done by a computer programme however, human checking are indispensable into the quality control procedure and as well to the correction process. All the procedures, changes and observations should be flagged appropriately and explained in the metadata (World Meteorological Organization, 2009).

The main source of errors are: instrumental, observer, data transmission, key entry, data validation process, changing data formats, and data summarization. The most common format errors include miskeying and mistake by operator.

There are a few types of test in which application will depend on the type of data. The World Meteorological Organization, (2011) gives all the information on the tests which should be applied considering the type of data. The WMO Guide insists in the documentation of the procedure and decisions formulated.

Completeness test applied to monthly extreme when there is missing daily data); Consistency test (Internal -physical relationships among climatological elements. Temporal - tests the variation of an element in time. Spatial – comparison of each observation with those taken at the same time at other stations in the area. Summarization (errors can be detected by comparing different summaries of data). Procedures, formulas, and decision criteria should all be documented); Tolerance tests (setting up of upper or lower limits on the possible values of a climatological elements usually compare a value against some standard value with the use of a statistical threshold (World Meteorological Organization, 2009; (World Meteorological Organization, 2011):

2. Methodology

2.1 The data

This database gathers climate data of Angola (South region) of the years of 1961 to 1974. The data was recorded while Angola was a Portuguese Colony. The data was rescued by the Instituto Superior Politécnico Tundavala (ISPT) from the Coimbra University, place where the original archives were saved. The rescue project is linked and sponsored by the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) project within the 141 task.

The data was separated into three different tables as climatological and precipitation stations which provide districts and stations names, the 3 geographical coordinates (Lat, Long and Altitude/elevation) however no identification number is associated to each one. A third table with evaporation and evapotranspiration recordings from the main basins of Angola. In this paper, database creation and error flagging system will only be performed on the climatological table. In the future, same procedure will be applied to the other tables in order compile all the recordings to the database.

The climatological data gathers monthly summaries of recorded climate variables such as Air Temperature, Humidity, Nebulosity, Precipitations and the Number of days which rain thunderstorm and others, this is show in table 2. Each stations is linked to its own District and hold three geographical coordinates, latitude, Longitude and altitude/elevation, this is shown in table 1. The database the variables have Portuguese names as this belongs to Portuguese speaking country. For this paper, only the data base will only focus on the South of the Angola, as the North was still being digitalized. For this paper, focus will fall only on the first table, Climatological Data.

According to meteorological reports of Angola from the 1940 the observation methods were compliant to the IMO (international meteorological organization) where some are described. Since the data in use does not provide such information it is assumed that the procedures would be the same as reported in 1940 (SMA, 1940 and SMA, 1950 and others from 1937 to 1952 available at the NOAA site in http://docs.lib.noaa.gov/rescue/data_rescue_angola.html).

According to SMA, (1940) the time of Observations: were always relative to the legal time of Angola relative to the meridian 15º East of the Greenwich, taken at 9 AM.; (VD - Vários Dias) means Many Days; the sign "dash"(-) means No observation. The mean/average observation derive from the [(Max + Min)/2].

In SMA, (1952) it is announced the separation of the recording and publishing for two publications one which will record all the information from all the station from the whole country and one for the main observatory in the Capital (Luanda) (Observatório João Capelo).

The processing of the data seems to have followed the processing procedures as stated in the point 4.1.2 in WMO, (1983) where in the process stage the data was collected, recorded, checked and later passed to an arithmetic processes and compiled in summaries (monthly in this case). Before recording the data some error checking were applied by meteorologists in order to correct most doubtful errors. It is as well mentioned SMA, (1939) about the collaboration of singular/private people (farmers and exploitations companies located in inhospitable places) which in data appear stations names with name farm, in Portuguese "Fazenda".

Data Attributes	Example	Units	Description
Distrito	Benguela	-	District name
Estacão	Balombo (Polig. Flor.)	-	Station name
Latitude	12 20	(° ′)degrees and minutes	Latitude degree
Longitude	14 47	(° ′)degrees and minutes	Longitude degree
Altitude	1200	meters (m)	Elevation
year	1961	-	year of recording
month	6	-	month of recording (number of the month)

Table 1 Description of data attributes for each station in the database.

Weather Parameter	Data Types	Units	Description
Temperatura do Ar	TempMed.9h	(°C)	Mean Of All The Temperatures Recorded At 9 Hour In A Month Period
	TempMed.7h (*)	(°C)	Mean Of All The Temperatures Recorded At 7 Hour In A Month Period
	TempMed.Max	(°C)	Mean Of The Mean Maximum And Minimum Values [(Mean Minimum + Mean Maximum)/2]
	TempMed.Min	(°C)	Mean Of The Minimum Temperatures Recorded In A Month Period
	TempMed.Diurna	(°C)	Mean Of The Minimum Temperatures Recorded In A Month Period
	TempExt.Max	(°C)	Extreme Maximum Temperature Recorded In The Month
	TempExt.Min	(°C)	Extreme Minimum Temperature Recorded In The Month

Precipitação	Prec.Total.Mm	(mm)	Sum Of All The Precipitation Recorded In A Month
	Prec.Max.Mm	(mm)	The Maximum Amount Of Precipitation Recorded In The Month
Humidade	Humidade.9h	%	Mean Of The Humidity Recorded At 9 Hour Over A Month
	Humidade.7h (*)	%	Mean Of The Humidity Recorded At 7 Hour Over A Month
Nebulosidade	Nebulosidade.9h	0 to 10	Mean Of The Nebulosity Recorded At 7 Hour Over A Month
	Nebulosidade.7h (*)	0 to 10	Mean Of The Nebulosity Recorded At 7 Hour Over A Month
Datas	TempExt.Max.Data	days	Date Of The Extreme Maximum Temperature
	TempExt.Min.Data	days	Date Of The Extreme Minimum Temperature
	Prec.Max.Data	days	Date Of The Extreme Maximum Precipitation
Número de Dias	Prec.Dias.0.1	days	Precipitation R (mm) (R ≥ 0.1)
	Prec.Dias.1	days	Precipitation R (mm) (R ≥1;)
	Prec.Dias.10	days	Precipitation R (mm) (R ≥10)
	Trovoada.Dias	days	Number Of Days Which Was Recorded The Occurrence Of The Phenomenon
	Relâmpago.Dias	days	
	Chuva.Dias	days	
	Nevoeiro.Dias	days	
	Cacimbo.Dias	days	

Table 1 Description of weather variables and its types recorded. (*) When readings were recorded at 7 hour and not at 9 hour.

Many were the inconsistencies found among the data such as the dash (" – ") and trailing (+/*). Different spacing were separating degrees from minutes in latitudes and longitudes coordinates, besides this coordinates do not own the standard format of geographical coordinates. Stations names with many version of the same name such as Benguela (S.M.A.), Benguela (SMA), Benguela (Cidade) which are in fact the same. Districts are at beginning 6 (Benguela, Huambo, Moxico, Bié-Cuando_Cubango, Moçâmedes, and Huíla), later on Huila was divided to Huila and Cunene and Bié separated from Cuando-Cubango accounting 8 districts. All these issues were solved and the methods applied are explained below in this section.

2.2 *Importing data to R*

The first step was to import the data into R (A language and environment for statistical computing. R Foundation for Statistical Computing), (R Core Team, 2015). It was created a script to read excel file into R which required an installation of the package "xlsx" (Dragulescu A., 2014) for this purpose. Find script in Appendix 3. For this step we developed a code in R to read the excel files and compile in the database which was named "climate". When reading the data, there was the need to consider all the variables as character due to the existence of (-) dashes, and (" ") blanks among the variables, VDs (vários dias, many days in English), and the trailing (+/*). VDs appear in the dates for (TempExt.Max.Data, TempExt.Min.Data and Prec.Max.Data) which means that the occurrence do no only happened in one day. Since it a date (numerical) could have been read as such however, it would remove all the VDs from the data risking loosing information.

Data was copied and saved in another name to avoid overwriting data. All the procedures were done in the copy as the original must remain unchanged as mentioned in World Meteorological Organization, (2011) All the procedures described below are explained in the R scripts in Appendix 3, 4 and 5.

2.3 *Cleaning data*

The second step was to clean all the problems found in the first step as well as others specified here. It is important to notice that the variables of the dataset had to be worked individually due to its own characteristics. In Appendix 4 script 2

The first stage of this step was to clean all the data inconsistencies as following:

All dashes (-) set as NAs (missing values) as they are absent readings

All the (" ") blanks were removed from the data

All the comma were changed to dots and any second commas removed. (Portuguese use commas and not dots, changes were made in order to avoid conflicts). It was used a clean function (see script 2 in Appendix 4) to clear this inconsistencies in the data. For the dates (datas) the clean function was a little modified as it would remove VDs. The "clean function" to change all commas to dots, remove all commas in other place and sets no NAs entries with no digits. All procedures are explained in each section where cleaning was applied in script 2 in Appendix 4.

The second stage was to use another function which removed all the (*/+) which were attached to some values of the 9h Mean Temperatures (TempMed.9h). The values with trailing (*/+) in 9h readings meant 7h readings. The "Plus function" was created to take away the trailing (*/+) and assign a new column for those readings named 7h Mean temperature readings (TempMed.7h). The same procedure was done for humidity and nebulosity which had as well the trailing.

Since latitude and longitude had only degrees and minutes separated by a space or blank, a code was developed to separate each degrees from minutes and sum them together after transforming the minute's values by dividing by 60. Table 3 shows how the raw data was (before) and how it was transformed (after). The code used basically separates the degrees and minutes and places into a new columns. In another new column adds them together again (without spacing) and with minutes dived by 60 (lat=lat.deg+lat.min/60).

	Station Name	Latitude	Longitude
Before	Balombo (Polig. Flor.)	12 20	14 47
After	Balombo (Polig. Flor.)	12.3333	14.78333

Table 2 Description of the stations latitudes and longitudes changes. Before entrance shows how was the formatting of the raw data and after shows the results of the changes performed in R.

2.4 *Stations names, District Names and Id Creation*

There are many inconsistencies with the stations names in the data. Many of the station names were written with different versions see table 4 below. The way to solve this problem was by creating a function which could group the station according to their latitude longitude and altitude. This code at first finds all the unique station names in the original list of station names. From one unique station name finds the set of latitudes, longitudes and altitudes and then then uses these locations to find the stations names variants. The function makes a list of all the stations with the variant names, with the same set of latitude and longitude. From this list it was possible to see which were the wrong groups assigned and manually assign each one to the right group. The code fails due to the mistakes find in the data eg.: Caconda and Caconda (Miss. Católica) have exactly the same latitude, longitude and altitude, however are not the same station (see table 4). The main reasons for the failing of the code is the data inconsistencies. This inconsistencies are not only find in the original files as it was as well done in the digitalization process. The groupings of all the stations names follow more or less the same procedure using the list resulted from the code which had 210 station groups and it was produced a new list in which gathered 234 stations, being now the number of stations in the data as shown in table 13. Using this new list it was created IDs for the stations. Following the same procedure of station variants it was done to districts variants in order to find and assign an ID to the right group of districts names. However for district it did not had to be matching with the stations coordinates. The code only creates groups of all the districts variants in a list. From this list it is created a new list with manual sortation. It was, the same way as with the stations assign a code. It was just my decision to assign a code of numbers from 1000 to 9000.

There were 6 districts at the beginning of the recordings Benguela, Huambo, Bié-Cuando-Cubango, Moxico, Moçâmedes and Huíla. The district of Huíla was separated rising a new district called Cunene. The district of Bié-Cuando_Cubango, was divided rising the Bié and Cuando_Cubango districts see table 5.

STATION ID	STATION NAME	ALTITUDE	LONGITUDE	ALTITUDE
80	Caconda	13 43	15 05	1648
80	Caconda	13 43	15 05	1650
81	Caconda (Administração)	13 42	15 03	1656
81	Caconda (Adm.)	13 42	15 03	1656
82	Caconda (Miss. Católica)	13 43	15 05	1650
82	Caconda (Miss. Cat.)	13 42	15 07	1650

Table 3 Example of one station and its names variants. The latitudes, longitude and altitudes.

DISTRITO	DISTRITO.ID
Benguela	1000
Huambo	2000
Bié-Cuando-Cubango	3000
Bié	4000
Cuando-Cubango	5000
Moxico	6000
Moçâmedes	7000
Huíla	8000
Cunene	9000

Table 4 Data district Names and IDs

2.5 *Error checking*

Having into consideration the literature above (WMO, 2009; WMO, 2011) in section 1.3, in this paper it was only possible to carry the tolerance test and the internal consistency test (simple comparison between same variables) having in account the WMO ranges for each instrument/variable and some information about the data itself to develop a flagging error checking. It will be followed the specification as in Andresen et al., (2002) and Westcott et al., (2011). A manual process to correct all the errors will have to be done in the future as it is a long process. A set of 6 test were developed to flag the erroneous data. The tolerance tests (test 1 and test 2) and the concistancy test (tests 3, 4, 5, 6, 7).

The Tolerance test is based on the limits of the climate variable. It is set an upper and/or lower limits on the possible values of a climatological element. This is not only used for checking errors in the data as it is used to check instruments malfunctions. See Table 6 which is explained all the ranges from each of the climate variables. This test 1 is the standard range used globally. Test 2 was set having in account the condition of the data. The lowest temperature of the data is -8.5 and the maximum 42. According to the data the test 2 will spot the same errors as test 1 and will be more accurate by having in account the data in question. In the case of humidity happens the same as there are no value which fall below 10. So it was set an intrinsic range from 10 to 100. This will spot typing mistakes such as 6 or 9 which in the data are may be 60 or 90.

Tolerance Tests

Variables	TEST 1	TEST 2
	Standard Range	**Intrinsic Range**
Air Temperature	From - 80 to +60 °C	From -10 to + 45
Precipitation	(Daily) from 0 to 500.0 mm	(Monthly) 850.0 ≤ Prec.Total ≥ Prec. Maximum
Humidity	From 0 to 100 %	From 10 to 100 %
Nebulosity	From 0 to 10	_
Dates	1 to 31 (days in a month)	_
Number of days	0 to 31	_

Table 5 Description of the Tolerance tests of each weather variable. There is standard range worldwide (test 1) used and an intrinsic range related to the data (test 2). Dash (-) No test performed.

Test 3 spots errors in values which are greater than a smaller variable, eg.: Mean 9h temp. < Mean Maximum temp. The temperature recorded at 9h must always be a smaller value than the maximum temperature. If it is greater means that there is an error.

Test 4 spots errors in values which are smaller than a greater variable, eg.: Mean 9h temp. > Mean Minimum temp. The temperature recorded at 9h must always be greater than the minimum temperature recorded. If it is smaller means that there is an error.

Test 5 is specific for Mean Temperature daytime (temp.Med.diurna) which is the Average of the Mean Max. Temperature (TempMed.Max) with the Mean Min. Temperature (TempMed.Min). There was the need to set this range as the rounding of this values did not followed any rule, therefore it was set a range in which values are considered true if in between a computer average and a threshold of 0.11 decimal places. If the difference between the computer average (CA) an d the Mean Temperature daytime (temp.Med.diurna) is greater than 0.11 than there is an error.

Test 6 detects values which cannot be interpreted as number such as “20.5?”, “2?.6”, which were values that could not be read from the originals.

Test 7 detects error values which are outside parameters such as fractional numbers “1.3” in dates. Dates can only have whole number which range from 1 to 31 days of a month. Any fractional number in these climate variable is and errors. See table 7 below.

Whenever there is a 1 means pass and 2 means failure. For the cases where test were indeed applied and are not present pass or failure (1 or 2) and instead there is a 0 (test not performed) it means that the test was not performed due to not existence of the comparable variable. In table nº 7 it is possible to see that test 3 was indeed applied and there is a result which shows a 0, e.g.:(1010 - 290 - 1.85%) test 3 was indeed applied but did not pass or failed due to lack of comparable variables. With this complex systems it is not only possible to spot which and how many values have passed and failed but as well in which test they have passed and failed and which test was not performed due to lack of comparable variable or due to not being compliant to the weather variable. This flag system show as well the weaknesses of the consistency tests when there are missing values which alert us for the creation and implementation of another flag for failure in performing the consistency tests

CONSISTENCY TESTS

VARIABLE	Type	Test 3	Test 4	Test 5	Test 6	Test 7
TEMPERATURA DO AR	tempMed.9h	tempMed.9h < tempMed.max	tempMed.9h > tempMed.min	-	-	-
	tempMed.7h	tempMed.7h < tempMed.9h	tempMed.7h > tempMed.min	-	-	-
	tempMed.diurna	tempMed.diurna < tempMed.max	tempMed.diurna > tempMed.min	(tempMed.diurna - CA) < 0.11	-	-
	tempMed.min	tempMed.min < tempMed.9h	tempMed.min > tempExt.min	-	-	-
	tempMed.max	tempMed.max < tempExt.max	tempMed.max > tempMed.9h	-	Not Interpretable	-
	tempExt.min	-	-	-	Not Interpretable	-
	tempExt.max	-	-	-	Not Interpretable	-
PRECIPITAÇÃO	prec.total.mm	-	prec.total.mm l ≥ prec.max.mm	-	Not Interpretable	-
	prec.max.mm	prec.max.mm Maximum < prec.total.mm	-	-	-	-
HUMIDADE	humidade.9h	-	-	-	-	-
	humidade.7h	-	-	-	-	-
NUBULOSIDADE	nebulosidade.9h	-	-	-	Not Interpretable	-
	nebulosidade.9h	-	-	-	-	-

DATAS	tempExt.max.data.	-	-	-	Not Interpretable	Outside Parameter
	tempExt.min.data	-	-	-	Not Interpretable	Outside Parameter
	prec.max.data	-	-	-	Not Interpretable	Outside Parameter
NÚMERO DE DIASS	prec.dias.0.1	-	-	-	Not Interpretable	Outside Parameter
	prec.dias.1	-	-	-	Not Interpretable	Outside Parameter
	prec.dias.10	-	-	-	-	Outside Parameter
	trovoada.dias	-	-	-	-	Outside Parameter
	relampago.dias	-	-	-	-	Outside Parameter
	chuva.dias	-	-	-	Not Interpretable	Outside Parameter
	nevoeiro.dias	-	-	-	Not Interpretable	Outside Parameter
	cacimbo.dias	-	-	-	Not Interpretable	Outside Parameter

Table 6 Description of the Consistency tests applied in each weather variable. CA (Computer Average). The dashes (-) mean no test applied.

2.6 *The flagging system*

It was developed a flagging system which give as many combinations as the data passes, fails and did not performed a test. It is added 10, 100, 1000, 10000 100000, 1000000, respectively from test 1 to test 7, if it passes or 20, 200, 2000, 20000, 200000, 2000000 if failed. A zero is added to all the error variable names at the creation of the flagging. 0 will imply that the test was not performed, as shown in table 8. Results will have combinations of tests performed and not the number shown in the table. In results section flagging system results is presented.

Error Checking And Information for Original Values

LEVEL			EXPLANATION
TEST	PASSED	FAILURE	
0			No test performed
1	1	2	Out of normal range
2	10	20	out of intrinsic range
3	100	200	Out of range, greater than comparable variable
4	1000	2000	Out of range, smaller than comparable variable
5	10000	20000	Out of average range
6	100000	200000	Not interpretable data
7	1000000	2000000	Values Outside Parameter

Table 7 - Description of the values of the flagging system which are add to each test performed

Whenever there is a 1 means pass and 2 means failure. For the cases where test were indeed applied and are not present pass or failure (1 or 2) and instead there is a 0 (test not performed) it means that the test was not performed due to not existence of the comparable variable. In table nº 9 it is possible to see that test 3 was indeed applied and there is a result which shows a 0, e.g.:(1010 - 290 - 1.85%) test 3 was indeed applied but did not pass or failed due to lack of comparable variables. With this flagging systems it is not only possible to spot which and how many values have passed and failed but as well in which test they have passed and failed and which test was not performed due to lack of comparable variable or due to not being compliant to the weather variable. This flag system show as well the weaknesses of the consistency tests when there are missing values which alert us for the creation and implementation of another flag for failure in performing the consistency tests.

2.7 *R Studio programing*

R is a free statistical program language. Run in systems such as Windows, Mac, and Linux (“RStudio | RStudio,” n.d.). Its source code is as well freely available in the internet where many of the questions surrounding R codes can be easily find on the web not only in R websites but as well in many blogs where users around the globe share their knowledge. R was not only chosen to be used in this project for being a free software (this was important and necessary due to Angola’s lack of resources) but mainly for being a great tool for data mining as it is possible to use large databases (Torgo, 2011). Yet the best facet of R lays on the statistical computing which is the main tool in R programming having great graphical outputs and as well endue packages which can be easily downloaded and installed which gives vast possibilities to work in many science and commercial fields (Dalgaard, 2008). In this project the statistical tools were not used however it will be used in the future to spot other errors in the data which could not been done due to time data and time constrains.

In this project it was created a scrip to read excel file into R (A language and environment for statistical computing. R Foundation for Statistical Computing), (R Core Team, 2015) which required an installation of the package “xlsx “for this purpose (Dragulescu A., 2014).It was as needed the installation of the packages to be able to plot the map such as “maps” Becker R., Wilks A., 20015), “sp” (Pebesma et.al., 2013), “scales” (Wickham H., 2015), “reshape“(Wickham, H., 2007).

All of the code done in R will be available in the the Appendix 3 and 4.

3. Results

3.1 *The database*

The database itself cannot be shown in this paper however, for the responsible for the evaluation of this paper, is provided access to a folder where the database can be visualized, as well as all the Excel spread sheets and scanned copies of the original . The scripts developed in R programing are in Appendix 3 and 4.

Table 13 (Appendix 1) can be seen all the stations existent in the database as well as their respective stations name and ID, District name and ID. Table 13 gathers as well the amount of years and months each station gathers in the overall 14 years of recordings. Please see Appendix 1 to view results. In figure 5 it is possible to see the location of the stations in the belonging district in the Angola map.

In the overall the location of the stations are compliant to belonging districts however, in the case of Bié and Cuando_Cubango there are unmatched due to the separation of the district after the existence and recording of the data, this is as well an issue with Huíla and Cunene. I manual sortation will be performed in the future to group the stations into the belonging district to eliminate this issue. In the case of Huila and Benguela there is as well an unmatched as some stations from Huila are inside Benguela district borders. This is an issue to be discussed and solved in the future

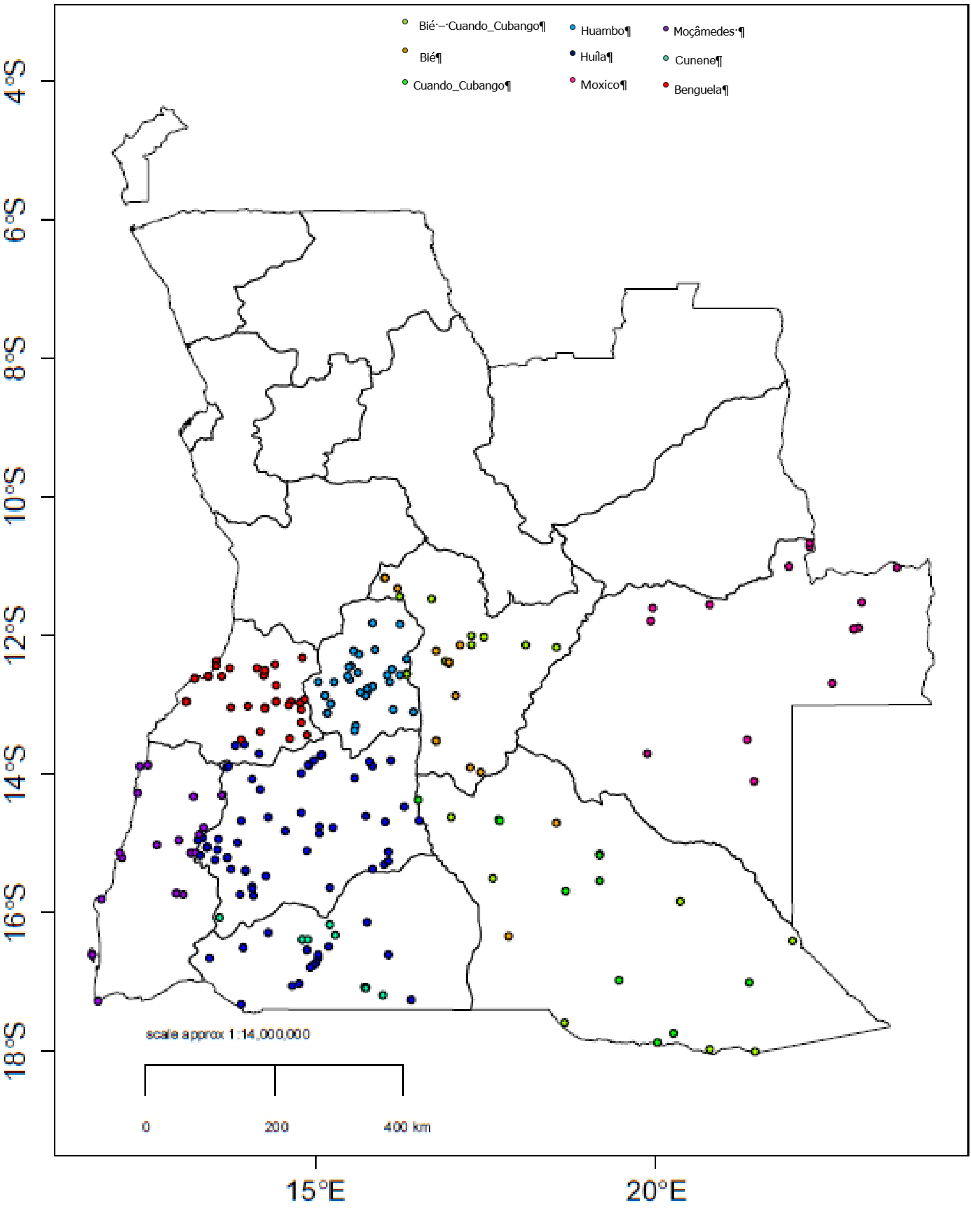


Figure 3 : Map of Angola with the actual administrative division and localization of the stations coloured accordingly to the belonging district.

With the grouping of the stations it is now possible to extract information from one station and visualize results. In the figure below it is displayed the total and maximum precipitation of Cangamba station of Moxico District. There are only four (4) years which gathers the 12 months of the year. It is possible to visualize errors which are visibly obvious. In figure 4 below, in graph a, is possible to note an error in the year 1962 where maximum precipitation is greater than total precipitation. The consistency test 3 and 4 are responsible to spot this types of error where maximum precipitation can never be greater than total precipitation (test 3) and total precipitation can never be smaller than maximum precipitation (test 4). In the overall it possible to see that the data can be used and it seems quite reliable. Graphs do not fallow the precipitation typical bar graph as here the focus fall on the values itself and not on the actual weather information.

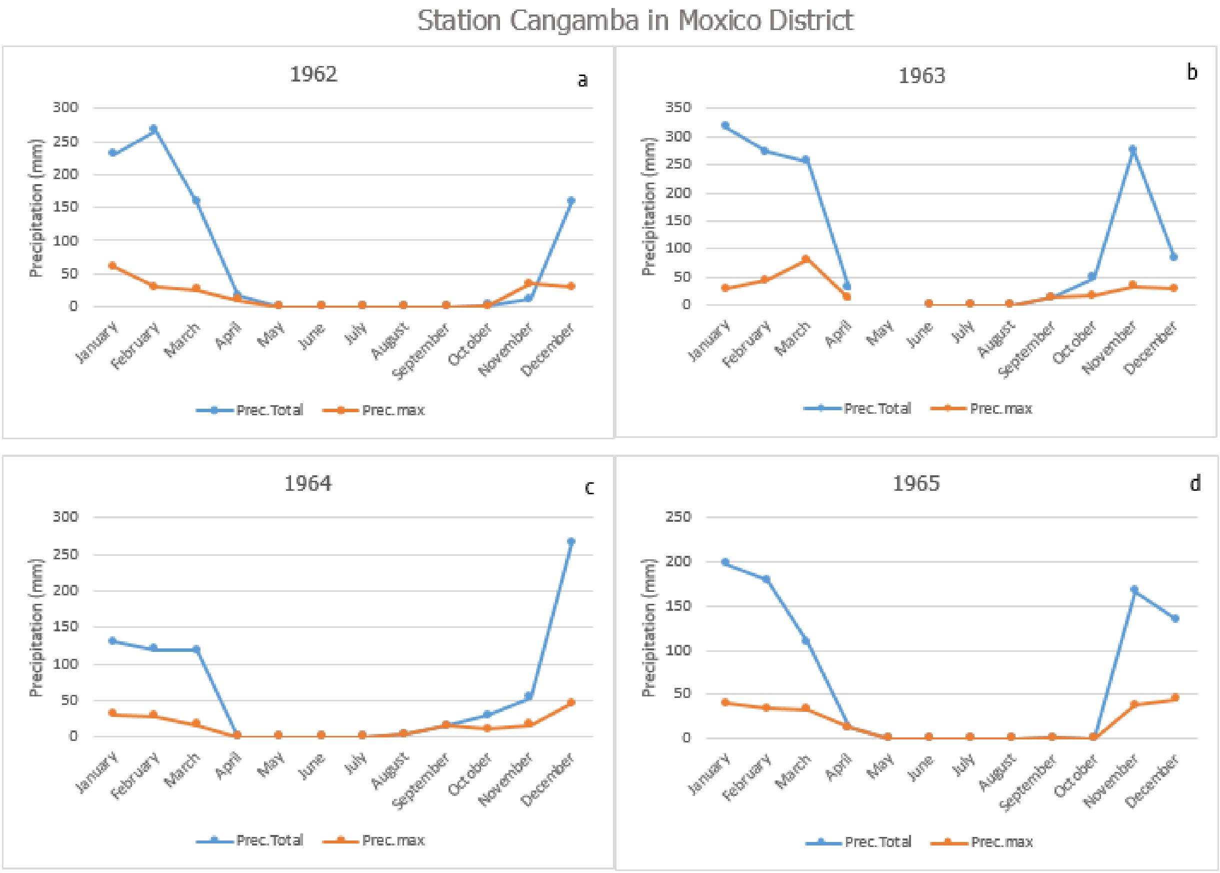


Figure 4: Total and Maximum Precipitation from Cangamba station in Moxico. It is represented the years of 1962 (a), 1963 (b), 1964 (c) and 1965 (d).

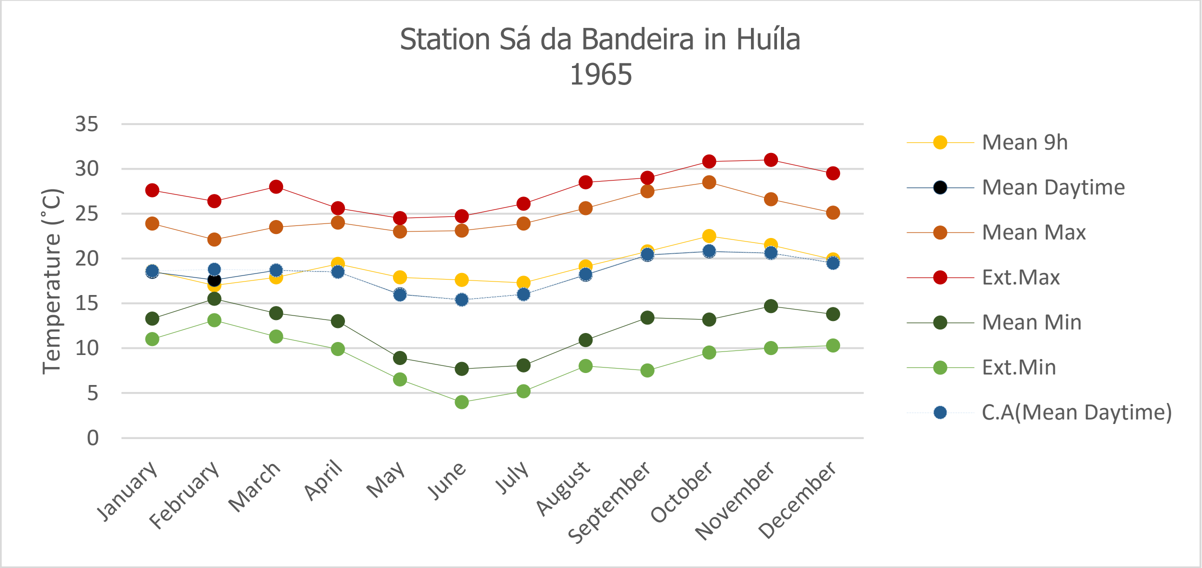


Figure 5 Air Temperature readings of the Sá da Bandeira station in Huíla district. It is plotted the TempMed.9h, TempMed.Max, TempMed.Min, TempMed.Diurna, TempExt.Max, TempExt.Min and the Computer Average (CA) mean daytime.

In figure 5 it is possible to see an error related to the mean daytime (TempMed.Diurna) (black dot) which is not matching the computer mean. This is the type of error which is spotted by test 5. In the overall the figure shows that the data is reasonable, following a normal flow.

3.2 *Error checking Results*

It was created variables with the same name as the climate variables with the prefix Erro (Error in English) to harbour the vales resulted from the flagging system. In the Methods section it is explain which tests were applied and to each of the variable it concerns. It was developed system of error checking which spot different types of errors to the different type of weather variable. For this is important to be understood how the resulted error checking means. In table 9 it possible to see which of the error checking were applied, passed and failed to the weather variable Mean Temperature at 9h. Starting from the right to left of the values on column Tests Results is the sequence of the test applied to climate variable

Testes Results	Erro.Tempmed.9h	Total %
0000	1676	10.69%
0010	216	1.38%
0110	891	5.69%
0210	2	0.01%

1010	290	1.85%
1110	12486	79.68%
1210	24	0.15%
1220	6	0.04%
2010	1	0.01%
2110	59	0.38%
2210	20	0.13%
Total	15671	1

Table 8 Flags resulted from the error checking system applied to the weather variable Mean 9 hour Temperature

In the table above 9 it is possible to see that in the overall data for Mean Temperature at 9h passed all the test. Highlighted in the table there are the values which have passed all of the tests accounting about 80 % of the data. (15671 entries) about 11% of the data for Tempmed.9h did not performed the test for being missing values (meaning that 10.69% of the data of the Mean Temperature at 9h are missing values. Remaining 9.6% which passed test 2, test 2 and 3 and test 2 and 4 and missed test 3 for lack of comparable variable). Under most failed tests 2, 3 and/or 4 accounting 0.7% of the whole data. Note that whenever there is a 2, even if it passed some of the previous or later test it always mean a failure. However, the error may not be in the actual variable but in the variable which was compared with.

Testes Results	Erro.Tempmed.Diurna	Total %
100000	2039	13.01%
100110	3	0.02%
101010	5	0.03%
111110	13147	83.89%
121110	359	2.29%
121210	38	0.24%
121220	3	0.02%
122110	52	0.33%
122210	21	0.13%
200000	4	0.03%
TOTAL	15671	1

Table 9 Flags resulted from the error checking system applied to the weather variable Mean daytime Temperature

In table 10 was applied test 6 which spots values which are not interpretable such as "20.5?" This is the only weather variable which test 5. About 13% of the values have passed test 6 and are missing values, only 0.05% have passed test 2 and 3 or 2 and 4 having missed test 3 due to missing values. About 83 % of the values have passed all tests. About 3 % represent failure, 2.29% failed test 5 which is a considering proportion. Note that in errors spotted from test 3 and 4 does not only mean that the value in question is wrong but instead there might be an error of the values which was compared with. As an Example is Mean Daytime > Mean Minimum if values are respectively (22.9 and 32.5) implies and error in Mean Minimum Temperature and not in Mean Daytime Temperature. So in this case it must be taking in consideration, when correcting the errors, that the values compared must be both analysed. The correction should follow the sequence of the appliance of the tests.

Testes Results	Erro.prec.total.mm	Total %
100000	322	2.05%
100010	134	0.86%
101010	15193	96.95%
101020	2	0.01%
102010	19	0.12%
200000	1	0.01%
TOTAL	15671	1

Table 10 Flags resulted from the error checking system applied to the weather variable Total precipitation

In table 11 above it is seen the same result patter as in the previous testes and descriptions above. In weather variables recorded at 7 hour (tempMed.7h) the missing values account the highest percentage as these observations. There are just a few or singular cases where readings were done at 7 hour instead of 9 hour and whenever there is 7 hour readings there are no 9 hour readings. Results for 7 tempMed.7h can be seen in table 12.

Testes	erro.tempMed.7h	Total %
0000	15431	98.47%
0010	46	0.29%
0210		0.00%
1010	194	1.24%

total	15671	1

Table 11 Flags resulted from the error checking system to the weather variable Mean 7 hour Temperature

4. Discussion

The database is now compiled with the 14 years of monthly summaries which were previously separated in excel spread sheets. The stations and districts names have been organized and are assigned identities to both. The Latitude and longitude are now with coordinates following coordinates standards as before didn't. All the data was submitted to an error checking, assessing tolerance and consistency and are now flagged accordingly. Taking considerations with the results it is possible to assure a success as all the objectives settled at the beginning of the project, validating the approach given. However this project is one of some stages which the data and the database have to pass through. In order to be complete for use it is necessary to be cleared from errors and mentioned inconstancies. WMO requirements are still to be achieved as the data still needs to improve. This is a process which takes time as some of the resolutions requires manual checking. One of the limitation in this project is the station names which are still an issue. Although it was possible to assume some decision and group stations there is still some doubts regarding few differences in the station coordinates which will as well require manual assistance. The IDs created are not compliant to international requirements. This is an issue which must be solved with the Instituto Nacional de Meteorologia e Geofísica de Angola (INAMET Angola) (National Institute of Meteorology and Geophysics of Angola) as they are the legal institution for Meteorological and Climatological issues. Some of the stations already have a WMO code which obeys to international coding mainly the ones in the airport and the oldest and main stations of each city. Once the data base is ready

The data gathers only monthly summaries which restrains the appliance of many tests which spots more errors therefore giving more credibility to the data. This is an important limitation of the data as summaries removes much of the variability during a month. The flagging system fulfil the objective, however for the consistency tests has a weakness. When comparing the data with a missing value the tests are not performed, leaving space for errors in those entries. Once errors are corrected a new error checking and flagging system should be developed for errors where the testes applied were not performed due to missing data.

As mentioned above, this is a continuous process which with time restrains could not be completed in this project. However the improvements to achieve a complete database with standard formats are known and recommend in this paper. Once improvements are concluded many benefits will rise. SASSACAL project, rescuing historical data project of task 141 will be complete (for the data from 1961 to 1974) and the database will be made available in the SASSCAL website. According to WMO policy regarding free exchange of climate data will be fulfilled. This type of information is highly important for policy makers as it contributes to the fundaments of the policy making process. Once such information's has huge implications to policy makers, it should be created policies which sponsors and incentives for this types of projects.

5. Conclusion

The overall task for this project resumes in the compilation of the monthly summaries of climate data recording into a database, the assignment of IDs to the stations and districts, the correction of the latitudes and longitudes into standard versions and a creation of a flagging system which access the quality of the data by performing an error checking system. Besides all previous, a cleaning process to all the missing values and blanks among the data giving the right designation.
The objectives were accomplished as the database is now compiled in which the objective above were successfully completed although many inconsistency with the data requires correction. The correction task it is not an objective of this project as it is a time demanding process. However, in this project it was possible to find the errors which can be considered as a primary process for the continuing work to be done in this database in the future.

Taking into consideration the objective of this project, and the results obtained, let us get to the conclusion that the procedures developed and applied to the data succeeded. The R programing was useful for the task presenting the results need and expected for the completion of this project.

As mentioned in this paper the database was successively compiled and an error checking method applied. However this database is not ready for use as there a few requisites to be completed.

Recommendations:

It is recommended that the errors flagged in this project should be corrected in the future as most are keying errors from the digitalization process. A new plan taking into consideration WMO requirements should be designed in order to clear the database form errors which were not found in this project. It is recommended a continuing research and rescuing to find the instruments and more information on the observation methods and statistical approaches used in the raw data before stating in the record. It is recommended to change the IDs of districts and stations once there will be the merging with the data form the North of the country (after performing the same tasks as stated in this paper) in which full communication with the INAMET is required for the attribution of international IDs taking into consideration WMO requirements .

Future expectations:

With the creation of these database which will mark the beginning of the data rescue and management for the south of Angola. The objective is to integrate the data from the North of the country in which all the procedures done in this stage will be done for the North part of the data. For the near future it is expected to continue with the data rescue from the past (1901 to 1960) and proceed to enlarge as much as possible the old database. Unfortunately data from 1978 until 2002 will never be available due to the civil war that was happening by the time. There will be platform connecting the old and new data being collected since 2014 from the SASSCAL project in Angola as well the integration of earlier data from the National institute of Angola recording since 2012.

References:

Adrian A. Dragulescu (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. http://CRAN.R-project.org/package=xlsx

Aguilar, E., Llansó, P., & World Meteorological Organization, (WMO). (2003). Guidelines on climate metadata and homogenization. WMO. Retrieved from: http://library.wmo.int/opac/index.php?lvl=notice_display&id=11635

Allaby, M. (2007). Encyclopedia of weather and climate (Rev. ed). New York: Facts on File.

Allaby, M., & Allaby, M. (2009). Atmosphere: a scientific history of air, weather, and climate. New York, NY: Facts on File.

Andresen, L., Hellsten, E., Rissanen, P., Palsdóttir, T., & Arason, T. (2002). Quality control of meteorological observations. Retrieved from http://met.no/Forskning/Publikasjoner/MET_report/2002/filestore/08_02.pdf

Barry, R. G., & Chorley, R. J. (2010). *Atmosphere, weather and climate* (9th ed.). London: Routledge.

Burt, P. J. A. (2009). Precipitation: Theory, measurement and distribution, by Ian Strangeways. 2007. Cambridge university press. ISBN-13978-0-521-85117-6, x+290 pp. *Meteorological Applications, 16*(3), 433-433. doi:10.1002/met.105

Carrega, P. (Ed.). (2010). Geographical information and climatology. London: ISTE.

CDM_2 WCDMP | WMO. (n.d.). Retrieved December 12, 2015, from http://www.wmo.int/pages/prog/wcp/wcdmp/CDM_2.php

Dalgaard, P. (2008). Introductory Statistics with R. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-0-387-79054-1

Global Administrative Areas *Boundaries without limits* (2015, December 6). Retrieved from http://gadm.org/home

Hadley Wickham (2015). scales: Scale Functions for Visualization. R package version 0.3.0. http://CRAN.R-project.org/package=scales

Linacre, E. (1992). Climate data and resources a reference and guide. London; New York: Routledge. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=93886

Monteiro, A. (2001). O reconhecimento oficial da importância da climatologia em Portugal (1850-1900).*Revista da Faculdade de Letras. Historia*. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=2322480

Nebeker, F. (1995). *Calculating the weather: Meteorology in the 20th century*. San Diego;London;: Academic Press.

Nunes, M. de F., Alcoforado, M. J., & Cravosa, A. (2014). A. Meteorologia e as observações instrumentais: a emergência da construção de redes internacionais XVIII-XIX. Retrieved from http://www.rdpc.uevora.pt/handle/10174/13356

Becker, R., Wilks, A., R version Brownrigg, R. Enhancements by Minka, T. and Deckmyn, A. (2015). maps: Draw Geographical Maps. R package version 3.0.0-2.http://CRAN.R-project.org/package=maps

Plummer, N., Lipa, W., Palmer, S., & World Meteorological Organization, (WMO). (2007). Guidelines on climate data management. WMO. Retrieved from http://library.wmo.int/opac/index.php?lvl=notice_display&id=16656

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/

Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio,2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

RStudio | RStudio. (n.d.). Retrieved December 13, 2015, from https://www.rstudio.com/products/RStudio/

SASSCAL, (n.d.). Southern African Science Service Centre for Climate Change and Adaptive Land Management. Retrieved December 13, 2015, from: http://www.sasscal.org/

SASSCAL.(2013). Integrated Science Plan - Task Description. Development of Meteorological Observation Conditions in Angolan Southwest – Province of Namibe and slopes of Serra da Chela (Task ID 141) [Online] Retrieved from :http://www.sasscal.org/downloads/Task_Description/Task_141_Description_for_Web_20130826.pdf

Serviço Meteorológico de Angola (1939). (SMA, 1939). *Elementos Meteorológicos e Climáticos de 1951.* Província de Angola. Provided by the NOAA/ESRL Physical Sciences Division, Boulder Colorado from their Web site at http://docs.lib.noaa.gov/rescue/data_rescue_angola.html

Serviço Meteorológico de Angola (1951). (SMA, 1951). *Elementos Meteorológicos e Climáticos de 1951.* Província de Angola. Provided by the NOAA/ESRL Physical Sciences Division, Boulder Colorado from their Web site at http://docs.lib.noaa.gov/rescue/data_rescue_angola.html

Serviço Meteorológico e Magnéticos de Angola (1940). (SMA, 1940). *Elementos Meteorológicos e Climáticos de 1940.* Província de Angola. Provided by the NOAA/ESRL Physical Sciences Division, Boulder Colorado from their Web site at http://docs.lib.noaa.gov/rescue/data_rescue_angola.html

Tan, L. S., Burton, S., & World Meteorological Organization, (WMO). (2004). Guidelines on climate data rescue. WMO. Retrieved from http://library.wmo.int/opac/index.php?lvl=notice_display&id=11637

Torgo, L. (2011). *Data mining with R: learning with case studies*. Boca Raton: Chapman & Hall/CRC.

Westcott, N., Andsager, K., Stoecker, L., Spinar, M., Smith, R., Obrecht, R., O'Connell, D. (2011). Quality Control of 19th Century Weather Data. *Contract Report 2011-02*. Retrieved from http://www.isws.uiuc.edu/pubdoc/CR/ISWSCR2011-02.pdf

Wickham, H. Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 2007.

World Meteorological Organization (Ed.). (1983). Guide to Climatological Practices (2nd ed): World Meteorological Organization.

World Meteorological Organization (Ed.). (2009). Guide to Climatological Practices (3rd ed) : World Meteorological Organization.

World Meteorological Organization (Ed.). (2011). Commission for Climatology: over eighty years of service. Geneva: World Meteorological Organization.

World Meteorological Organization, (WMO). (2008). Proceedings of the International workshop on rescue and digitization of climate records in the Mediterranean basin. WMO. Retrieved from http://library.wmo.int/opac/index.php?lvl=notice_display&id=16665

World Meteorological Organization. (2011). Guide to climatological practices. Geneva, Switzerland: World Meteorological Organization.

Appendix 1

District Name	District ID	Station Name	Station ID	Begin year	End Year	Total months	Total years	% of years in the 14 years of data
Benguela	1000	Balombo (Polig. Flor.)	1	1961	1974	137	11.4	81%
Benguela	1000	Lobito (ou S.M.A.)	2	1961	1974	147	12.2	87%
Benguela	1000	Cassequel	3	1961	1974	145	12.1	86%
Benguela	1000	Biópio	4	1961	1974	143	11.9	85%
Benguela	1000	Bocoio	5	1961	1974	116	9.7	69%
Benguela	1000	Benguela	6	1961	1974	134	11.2	80%
Benguela	1000	Benguela (S.M.A.)	7	1968	1974	60	5	36%
Benguela	1000	Baía Farta	8	1961	1967	59	4.9	35%
Benguela	1000	Fazenda S.Francisco	9	1961	1974	105	8.8	63%
Benguela	1000	Dombe Grande	10	1961	1970	51	4.2	30%
Benguela	1000	Ganda (Est. Zoot.)	11	1961	1973	128	10.7	76%
Benguela	1000	Ganda (Posto Agrícula)	12	1961	1963	26	2.2	16%
Benguela	1000	C.Estudos da Ganda	13	1963	1974	120	10	71%
Benguela	1000	Est. Reg. da Ganga	14	1961	1974	136	11.3	81%
Benguela	1000	Alto Catumbela	15	1961	1974	137	11.4	81%
Benguela	1000	Caimbambo	16	1961	1974	138	11.5	82%
Benguela	1000	V.Mariano Machado (Buça)	41	1963	1970	21	1.8	13%
Benguela	1000	Monte Belo	125	1961	1974	131	10.9	78%
Benguela	1000	Lomaum	141	1962	1974	129	10.8	77%
Benguela	1000	Cubal	143	1962	1974	130	10.8	77%

Benguela	1000	Cubal (C.F.B.)	144	1963	1973	91	7.6	54%
Benguela	1000	Baia Farta	147	1962	1964	2	0.2	1%
Benguela	1000	Tenda Moco	148	1962	1968	58	4.8	34%
Benguela	1000	Catengue (C.F.B)	151	1963	1974	119	9.9	71%
Benguela	1000	Fazenda Fernando Alberto	156	1964	1973	41	3.4	24%
Benguela	1000	Faz.Nelly (Chinene)	160	1964	1967	10	0.8	6%
Benguela	1000	Congoia.Faz.Beira Alta	161	1970	1973	31	2.6	19%
Benguela	1000	Fazenda Santa Isabel	165	1965	1974	52	4.3	31%
Benguela	1000	Faz. Santa Eugenia	167	1965	1966	6	0.5	4%
Benguela	1000	Cavaco (Cent. De Estudos)	194	1968	1974	69	5.8	41%
Benguela	1000	Canjola	203	1971	1974	33	2.8	20%
Benguela	1000	Fazenda Prazeres	220	1961	1972	38	3.2	23%
Benguela	1000	Chicuma	225	1962	1974	132	11	79%
Benguela	1000	Fazenda Santa Ana	229	1972	1972	1	0.1	1%
Bié	4000	Chinguar (Adm. Com.)	24	1963	1974	105	8.8	63%
Bié	4000	Silva Porto Int. Cereais	48	1972	1973	7	0.6	4%
Bié	4000	Catota	50	1961	1973	27	2.2	16%
Bié	4000	Catota-Missão Evangélica	51	1972	1973	14	1.2	9%
Bié	4000	Catabola Mis. Evangélica	52	1974	1974	2	0.2	1%
Bié	4000	Longa	136	1961	1974	87	7.2	51%
Bié	4000	Mucundi	140	1962	1968	48	4	29%
Bié	4000	Mis.Cat.do Vouga	149	1962	1964	6	0.5	4%
Bié	4000	Colónia Penal (Capolo)	170	1966	1973	78	6.5	46%
Bié	4000	Fazenda Etapa?	208	1972	1974	14	1.2	9%

Bié	4000	Calucinda	216	1974	1974	1	0.1	1%
Bié	4000	Calucinga	232	1970	1971	9	0.8	6%
Bié - Cuando - Cubango	3000	Chinguar	22	1961	1963	9	0.8	6%
Bié - Cuando - Cubango	3000	Andulo	39	1961	1974	137	11.4	81%
Bié - Cuando - Cubango	3000	General Machado	40	1961	1970	75	6.2	44%
Bié - Cuando - Cubango	3000	Nova Sintra (ou Catabola)	43	1961	1974	123	10.2	73%
Bié - Cuando - Cubango	3000	Coemba (Miss. Católica)	44	1961	1974	137	11.4	81%
Bié - Cuando - Cubango	3000	Ceilunga (ou C. de Estudos)	45	1961	1974	133	11.1	79%
Bié - Cuando - Cubango	3000	Silva Porto (ou Cidade)	46	1961	1974	122	10.2	73%
Bié - Cuando - Cubango	3000	Silva Porto (S.M.A)	47	1962	1974	130	10.8	77%
Bié - Cuando - Cubango	3000	Chitembo	49	1961	1974	99	8.2	59%
Bié - Cuando - Cubango	3000	Cuchi (Miss. Cat.)	53	1961	1974	87	7.2	51%
Bié - Cuando - Cubango	3000	Serpa Pinto	54	1961	1962	19	1.6	11%
Bié - Cuando - Cubango	3000	Capico (Miss. Cat.)	56	1961	1962	9	0.8	6%

Bié - Cuando - Cubango	3000	Mavinga	57	1961	1974	128	10.7	76%
Bié - Cuando - Cubango	3000	Mis. Stª Cruz do Cuando	58	1961	1966	43	3.6	26%
Bié - Cuando - Cubango	3000	Cuangar	59	1961	1974	90	7.5	54%
Bié - Cuando - Cubango	3000	Dirico	60	1961	1974	118	9.8	70%
Bié - Cuando - Cubango	3000	Mucusso	139	1962	1968	41	3.4	24%
Bié - Cuando - Cubango	3000	Luiana	146	1962	1962	1	0.1	1%
Bié - Cuando - Cubango	3000	Chiengue	222	1961	1964	5	0.4	3%
Bié-Cuando-Cubango	3000	Chinguar (C.F.B)	23	1963	1973	94	7.8	56%
Bié-Cuando-Cubango	3000	Nova Sintra	42	1961	1963	11	0.9	6%
Bié-Cuando-Cubango	3000	Chamavera (Dirico)	61	1963	1970	20	1.7	12%
Bié-Cuando-Cubango	3000	Chingue	128	1961	1961	1	0.1	1%
Bié-Cuando-Cubango	3000	Cuito Cuanavale	129	1961	1974	89	7.4	53%
Bié-Cuando-Cubango	3000	Munhango	152	1963	1973	95	7.9	56%

Cuando-Cubango	5000	Serpa Pinto (S.M.A.)	55	1963	1974	121	10.1	72%
Cuando-Cubango	5000	Cuito Canavale (FAP)	130	1973	1974	4	0.3	2%
Cuando-Cubango	5000	Sambio	163	1965	1973	25	2.1	15%
Cuando-Cubango	5000	Munué	164	1966	1967	12	1	7%
Cuando-Cubango	5000	Calonga	166	1965	1966	4	0.3	2%
Cuando-Cubango	5000	Cutato	188	1968	1974	38	3.2	23%
Cuando-Cubango	5000	Baixo Longa	227	1968	1974	35	2.9	21%
Cuando-Cubango	5000	Ponto de Passagem	228	1967	1967	2	0.2	1%
Cunene	9000	Cáfu (Posto Zoot. do Cunene)	112	1971	1974	25	2.1	15%
Cunene	9000	Pereira d'Eça (S.M.A.)	124	1973	1974	9	0.8	6%
Cunene	9000	Mucope - Loana	198	1973	1974	2	0.2	1%
Cunene	9000	Mucope - S. Adm. Civil	199	1973	1974	6	0.5	4%
Cunene	9000	Taca	209	1972	1974	14	1.2	9%
Cunene	9000	Manquete	218	1974	1974	2	0.2	1%
Cunene	9000	Chiede	219	1974	1974	2	0.2	1%
Huambo	2000	Bimbe	17	1961	1974	113	9.4	67%
Huambo	2000	Chiumbo	18	1961	1974	114	9.5	68%
Huambo	2000	Vª Teixeira da Silva	19	1961	1974	127	10.6	76%
Huambo	2000	Borga	20	1961	1974	138	11.5	82%

Huambo	2000	Chinguril	21	1961	1970	89	7.4	53%
Huambo	2000	Quipeio	25	1961	1974	129	10.8	77%
Huambo	2000	Chianga	26	1961	1970	27	2.2	16%
Huambo	2000	Chianga(C.Estudos)	27	1963	1974	116	9.7	69%
Huambo	2000	Sacaála (Per. Flor.)	29	1961	1962	12	1	7%
Huambo	2000	Sacaala	30	1961	1974	125	10.4	74%
Huambo	2000	Nova Lisboa	31	1961	1961	4	0.3	2%
Huambo	2000	Nova Lisboa (ou S.M.A.)	32	1961	1974	142	11.8	84%
Huambo	2000	Chenga	33	1961	1965	50	4.2	30%
Huambo	2000	Chenga / Fazenda	34	1965	1974	87	7.2	51%
Huambo	2000	Lucamba	35	1961	1961	6	0.5	4%
Huambo	2000	Fazenda Lucamba	36	1961	1970	21	1.8	13%
Huambo	2000	Cuima	37	1961	1973	89	7.4	53%
Huambo	2000	Cuima- Poligno Florestal	38	1974	1974	1	0.1	1%
Huambo	2000	Canjangue	126	1961	1963	25	2.1	15%
Huambo	2000	Canjangue (Fazenda)	127	1963	1974	115	9.6	69%
Huambo	2000	Catanga	142	1962	1974	130	10.8	77%
Huambo	2000	Catabola	145	1962	1974	129	10.8	77%
Huambo	2000	Sanguengue	157	1964	1970	32	2.7	19%
Huambo	2000	Seminário Caála	159	1964	1973	75	6.2	44%
Huambo	2000	Faz. Munana, Cuma	173	1966	1971	51	4.2	30%
Huambo	2000	Vila Nova	179	1966	1974	81	6.8	49%
Huambo	2000	Chinhama	183	1967	1967	5	0.4	3%
Huambo	2000	Mungo	185	1967	1974	39	3.2	23%
Huambo	2000	Catanga Lonjongo	200	1970	1970	1	0.1	1%
Huambo	2000	Alto Hama	207	1972	1972	2	0.2	1%

Huambo	2000	Faculdade de Medicina Veter.	210	1973	1974	7	0.6	4%
Huambo	2000	Bela Vista	213	1973	1974	4	0.3	2%
Huambo	2000	Colonato S.Jorge do Cub.	215	1974	1974	1	0.1	1%
Huambo	2000	Sambo	221	1961	1972	52	4.3	31%
Huambo	2000	Fazenda Sanga	233	1972	1973	12	1	7%
Huambo	2000	Granja Belém	234	1966	1967	9	0.8	6%
Huíla	8000	Chiange	28	1964	1974	103	8.6	61%
Huíla	8000	Cupacaça	79	1961	1974	134	11.2	80%
Huíla	8000	Caconda	80	1961	1966	58	4.8	34%
Huíla	8000	Caconda (Administração)	81	1966	1974	61	5.1	36%
Huíla	8000	Caconda (Miss. Cat.)	82	1966	1974	84	7	50%
Huíla	8000	Uaba 1	83	1961	1974	141	11.8	84%
Huíla	8000	Uaba 2	84	1961	1974	138	11.5	82%
Huíla	8000	Sangueve (Miss. Cat.)	85	1961	1967	72	6	43%
Huíla	8000	Galangue	86	1961	1973	136	11.3	81%
Huíla	8000	Impulo	87	1961	1974	113	9.4	67%
Huíla	8000	Cué	88	1961	1974	142	11.8	84%
Huíla	8000	Quilengues (Adm.)	89	1961	1974	123	10.2	73%
Huíla	8000	Quilengues (Zoot.)	90	1961	1974	143	11.9	85%
Huíla	8000	Vila Artur de Paiva	91	1961	1974	127	10.6	76%
Huíla	8000	Dongo	92	1961	1973	110	9.2	66%
Huíla	8000	Hoque	93	1961	1974	122	10.2	73%
Huíla	8000	V. Paiva Couceiro (ou Quipungo)	94	1961	1974	103	8.6	61%
Huíla	8000	Humpata (E.Z.S)	95	1961	1974	137	11.4	81%
Huíla	8000	Humpata (Agrícola)	96	1961	1964	20	1.7	12%

Huíla	8000	Humpata (Serv. Ag. E Flores.)	97	1973	1974	3	0.2	1%
Huíla	8000	Centro Est. Humpata	98	1963	1974	120	10	71%
Huíla	8000	Sá da Bandeira	99	1961	1974	147	12.2	87%
Huíla	8000	Huíla (Miss. Católica)	100	1961	1974	144	12	86%
Huíla	8000	Cassinga	101	1961	1972	21	1.8	13%
Huíla	8000	Tchivinguiro	102	1961	1974	146	12.2	87%
Huíla	8000	Chibia	103	1961	1967	60	5	36%
Huíla	8000	Chibia (Vila João de Almeida)	104	1968	1974	68	5.7	41%
Huíla	8000	Jau (Missão Católica)	105	1961	1974	124	10.3	74%
Huíla	8000	Quihita (Miss. Católica)	106	1961	1973	128	10.7	76%
Huíla	8000	Quihita - E. H. n.º 5	107	1973	1974	3	0.2	1%
Huíla	8000	Mulondo	108	1961	1974	106	8.8	63%
Huíla	8000	Mupa (Miss. Católica)	109	1961	1974	126	10.5	75%
Huíla	8000	Cahama	110	1961	1974	133	11.1	79%
Huíla	8000	Cáfu	111	1961	1971	112	9.3	66%
Huíla	8000	Cunene - Centro de Est.	113	1969	1974	53	4.4	31%
Huíla	8000	Otchinjau	115	1961	1969	68	5.7	41%
Huíla	8000	Chiulo	116	1963	1964	19	1.6	11%
Huíla	8000	Chiulo (Miss. Católica)	117	1961	1974	108	9	64%
Huíla	8000	Roçadas	118	1961	1965	6	0.5	4%
Huíla	8000	Roçadas (ou S.M.A.)	119	1961	1974	140	11.7	84%
Huíla	8000	Namuculungo	120	1961	1967	58	4.8	34%
Huíla	8000	Namuculungo (ou C.E.I.L.A.)	121	1961	1967	18	1.5	11%
Huíla	8000	Vila Pereira d'Eça	122	1961	1971	86	7.2	51%

Huíla	8000	V. Pereira d'Eça (ou S.G.M.)	123	1961	1974	96	8	57%
Huíla	8000	Chibemba	132	1961	1974	133	11.1	79%
Huíla	8000	Oncócua	133	1961	1974	81	6.8	49%
Huíla	8000	Cafima	134	1961	1970	75	6.2	44%
Huíla	8000	Chingoroi	135	1961	1974	105	8.8	63%
Huíla	8000	Melunga - Chiede	138	1962	1973	27	2.2	16%
Huíla	8000	Chimbolelo	150	1963	1974	37	3.1	22%
Huíla	8000	Nonhe	154	1963	1964	8	0.7	5%
Huíla	8000	Chiveio	155	1963	1964	9	0.8	6%
Huíla	8000	Fazenda Sumbo	162	1964	1965	9	0.8	6%
Huíla	8000	Dongoena	168	1965	1966	12	1	7%
Huíla	8000	Bambi	171	1966	1973	67	5.6	40%
Huíla	8000	Peu-Peu	172	1966	1974	75	6.2	44%
Huíla	8000	Catembulo	175	1966	1966	2	0.2	1%
Huíla	8000	Cuvelai-Matala	176	1966	1974	75	6.2	44%
Huíla	8000	Jamba (Cassinga-Norte)	178	1966	1974	83	6.9	49%
Huíla	8000	Chicuaqueia	181	1966	1974	76	6.3	45%
Huíla	8000	Micosse - Matala	182	1966	1973	35	2.9	21%
Huíla	8000	Vila Folgares	184	1967	1974	72	6	43%
Huíla	8000	Handja	186	1967	1969	16	1.3	9%
Huíla	8000	Vila da Matala	189	1968	1972	49	4.1	29%
Huíla	8000	Rio da Areia	191	1968	1970	10	0.8	6%
Huíla	8000	Senge	193	1969	1973	19	1.6	11%
Huíla	8000	Fazenda Mumba	195	1969	1972	37	3.1	22%
Huíla	8000	Gambos	196	1969	1974	48	4	29%
Huíla	8000	Mucope	197	1970	1973	43	3.6	26%

Huíla	8000	Bimbe (Est. Zoot. Do Sul)	202	1971	1973	32	2.7	19%
Huíla	8000	Chipindo	205	1971	1974	31	2.6	19%
Huíla	8000	Chitado	211	1961	1966	52	4.3	31%
Huíla	8000	Cangolo	212	1973	1974	4	0.3	2%
Huíla	8000	Gando - P. N. Bicuari	217	1974	1974	2	0.2	1%
Huíla	8000	Tchamutete	226	1966	1974	61	5.1	36%
Huíla	8000	Mokete	230	1967	1968	3	0.2	1%
Moçamedes	7000	Virei	75	1963	1965	19	1.6	11%
Moçamedes	7000	Foz do Cunene	114	1962	1971	91	7.6	54%
Moçamedes	7000	Posto Exp. do Lungo	137	1962	1974	103	8.6	61%
Moçâmedes	7000	Lola	67	1961	1972	100	8.3	59%
Moçâmedes	7000	Vila Arriaga	68	1961	1972	121	10.1	72%
Moçâmedes	7000	Caracul	69	1961	1974	143	11.9	85%
Moçâmedes	7000	Bruco	70	1961	1973	80	6.7	48%
Moçâmedes	7000	Bruco (Escola de Reg. Ag.)	71	1973	1974	7	0.6	4%
Moçâmedes	7000	Chão da Chela	72	1961	1968	67	5.6	40%
Moçâmedes	7000	Moçamedes (ou S.M.A.)	73	1961	1974	146	12.2	87%
Moçâmedes	7000	Porto Alexandre	74	1961	1973	98	8.2	59%
Moçâmedes	7000	Curoca Norte	76	1961	1973	92	7.7	55%
Moçâmedes	7000	Muve - Virei	77	1970	1974	45	3.8	27%
Moçâmedes	7000	Baía dos Tigres	78	1961	1968	43	3.6	26%
Moçâmedes	7000	São Nicolau	177	1966	1974	82	6.8	49%
Moçâmedes	7000	Santa Marta	180	1966	1973	75	6.2	44%
Moçâmedes	7000	Giraul	187	1968	1974	28	2.3	16%
Moçâmedes	7000	Baía dos Tigres (S.M.A.)	190	1968	1974	61	5.1	36%

Moçâmedes	7000	Capagombe - Munhias	192	1969	1974	36	3	21%
Moçâmedes	7000	Cacanda - C. de Estudos	204	1971	1974	20	1.7	12%
Moçâmedes	7000	Lucira	231	1961	1974	99	8.2	59%
Moxico	6000	V. Teixeira de Sousa	62	1961	1974	143	11.9	85%
Moxico	6000	V. Teixeira de Sousa(C.F.B.)	63	1966	1974	28	2.3	16%
Moxico	6000	Luso (ou S.M.A.)	64	1961	1974	147	12.2	87%
Moxico	6000	Cangamba	65	1961	1974	102	8.5	61%
Moxico	6000	Vila Gago Coutinho	66	1961	1973	107	8.9	64%
Moxico	6000	Cavungo	131	1961	1974	145	12.1	86%
Moxico	6000	Mucussueje	153	1963	1974	99	8.2	59%
Moxico	6000	Caianda	158	1964	1966	13	1.1	8%
Moxico	6000	Lumbala	169	1966	1971	22	1.8	13%
Moxico	6000	Lutembo	174	1966	1968	15	1.2	9%
Moxico	6000	Cameia	201	1971	1974	19	1.6	11%
Moxico	6000	Cazombo	206	1971	1974	9	0.8	6%
Moxico	6000	Fazenda Piloto	214	1973	1974	3	0.2	1%
Moxico	6000	Cazombo	223	1961	1973	90	7.5	54%
Moxico	6000	Cazombo (F.A.P.)	224	1967	1974	74	6.2	44%

Table 12 Descriptions of the Districts and Stations names and IDs. This tables resumes the year which each station started recording (Begin year) and the last year of the recordings (End year). Total months are the sum of the months from the start until the end of the recordings of each station. The total amount of year derived from the total moths of data collection. There are in the last column the percentages of years in the total years of the data 14 years (from 1961 to 1974).

Appendix 2

Metadata

Angola Climate Data Period from 1961 To 1974

(South Region)

General Information

This database gathers climate data of Angola (South region) of the years of 1961 to 1974. This data was recorded while Angola was a Portuguese Colony. The data was rescued by the Instituto Superior Politécnico Tundavala (ISPT) from the Coimbra University, place where the original archives were saved. The rescue project is linked and sponsored by the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) project within the 141 task. One of the tasks of 141 main Task is the rescue of old climate data. The head of the project in the South, in Huíla-Lubango , Carlos Ribeiro and the team of two people (Nídia Loureiro and Sílvio Filipe) responsible for digitalization and further creation of a database.

Obs.: This database will in the future gather climate data for the whole country as for now only gathers the South region of Angola.

Name: Angola Climate data

Region: South

Districts: Benguela, Huambo, Moxico, Bié, Cuando–Cubango, Moçâmedes, Huíla and Cunene

Recorded date: 1961 to 1974

Rescued date: 2012

Digitilization date: 2014

Database creation date: 2015

Link to data: It will be linked when fully corrected to the SASSCAL and ISPT websites.

Data types: table 1 and 2 form the main text

Definition of Cacimbo find in Encyclopedia of Weather and Climate by Allaby, (2007)" Cacimbo is a heavy mist or wet fog associated with low stratus cloud and sometimes drizzle (see precipitation) that occurs along the coast of Angola during the dry season. The cacimbo usually forms in the morning and evening and may penetrate inland for some distance. It helps prevent extreme drought. The cacimbo is caused by onshore winds that carry warm air across the cold Benguela Current.". (Allaby, M. (2007).

Purpose:

The database was created in order for climate data information of Angola to become available as there is a new meteorological network is being implanted in the Country. The main purpose it to have documented and make climate data available This is important for the development of many activities such as agriculture, infrastructure buildings and studies such as Climate Change besides other academic and science and as well important to inform policy makers.

How to use the data

Full the description on the content of the data and the methods used I found in the Methods of this paper. Once the database is ready to be used the metadata should stand alone with the database. It should be reformulated as now it would only mean copying all the procedures explained in this paper. Table 1 and 2 explains the climate variables and the type of data it harbours. In methods and discussion sections are explained the limitations and inconsistencies of the data as well as the quality control system (Types of test applied for the error checking and the flagging system use

Use constraints: This data set is not yet in total condition to be used as there are some human and computerized error checking to be performed as well corrections, which requires more time than the project provides.

Metadata

Created by: Nídia Loureiro

Contact email: loureiro.nidia@gmail.com

Contact phone :353 0 899 586 400

Appendix 3

The Script created in R to read the Excel spread sheet from a folder.

Script 1 getclimatedata

```
# A Script to read in XLSX files with climate data from Angola and save them as a data
frame called climate

# Script created by Jon Yearsley and Nídia

rm (list=ls()) # remove all list from the environment
setwd('C:/Users/Nidia/Desktop/Ang_data/data')

require('xlsx')

file.out = 'climate.Rdata'  # Filename used to store data
data.dir = 'C:/Users/Nidia/Desktop/Ang_data/data'  # Directory where data are stored
# define names for the columns to read in
col.names = c('distrito','estacao','latitude','longitude','altitude',
              'tempMed.9h','tempMed.diurna','tempMed.max','tempMed.min',
              'tempExt.max','tempExt.max.data','tempExt.min','tempExt.min.data',
              'humidade.9h','nebulosidade.9h',
              'prec.total.mm','prec.max.mm','prec.max.data',
              'prec.dias.0.1','prec.dias.1','prec.dias.10',
              'trovoada.dias','relampago.dias','chuva.dias','nevoeiro.dias','cacimbo.dias')

# Define data class for each column it was all set to character due to the many
inconsistencies of the data.
col.class = c('character','character','character','character','character',
              'character','character','character','character',
              'character','character','character','character',
              'character','character',
              'character','character','character',
              'character','character','character',
              'character','character','character','character','character')


# List all xlsx files in the data directory
file.list = list.files(path=data.dir, pattern='.xlsx', ignore.case=T,recursive=T,
full.names=T)
n.files = length(file.list)

# Read in an xlsx file
#xlsx.wb = read.xlsx(file.list[[1]], sheetIndex=2, rowIndex=row.index,
colIndex=col.index, header = F)

for (f in 1:n.files) {
  wb = loadWorkbook(file.list[[f]])  # Load in an excel file
  sheets = getSheets(wb)             # Load in sheets from the workbook
  # Start reading from row 5. Read first 26 columns and as many rows as necesary
  tmp.dat =
readColumns(sheets[[2]],startColumn=1,endColumn=26,startRow=5,endRow=NULL,h
eader=F, colClasses=col.class)  # Read in the data
  names(tmp.dat) <- col.names           # Assign a name to each data column

  # Work out year and date of the file
  tmp = strsplit(file.list[f],'/')
  tmp.dat$year =  as.numeric(substring(tmp[[1]][length(tmp[[1]])],1,4))
  tmp.dat$month =  as.numeric(substring(tmp[[1]][length(tmp[[1]])],5,6))

  if (f==1) {
    climate = tmp.dat
  } else {
    climate = rbind(climate, tmp.dat)
  }
}

save(climate,file=file.out)

############ END ###################
```

Appendix 4

Script 2 gathers all the codes developed in R to clean the data, stations and districts variants functions and Id assignment.

Script 2 dataprocessing

```
### R code for Project ###
# Script modified by_nidia ( created by JonYearley)

#Clear memory to avoid conflictions
rm(list=ls())

#Load in all binary data to R data frame ( path/directory of the data)
load("C:/Users/Nidia/Desktop/Ang_data/data/climate.Rdata")
#load('..clean.Rdata')

#head(climate) # head of the table, first 10 elements
#str(climate) # basic structure is displayed
#summary(climate) #brief summary

#Rename data frame to facilitate and avoid overwrite
d=climate
# remove all the empty or blank read into the data as district entrances.
d = subset(d,!(distrito=="" & estacao==""))
#d = as.character(d)

#Process data...

#### Function definitions
##################################################
# Function to clean up, change all commas to dots,remove all commas in other places
#besides the right place and sets no NAs entries with no digits.
clean <- function(x) { # JY function
  # Convert to character
  x = as.character(x)
  # Replace first apperance of a comma with a decimal point (JY code)
  x = sub(",",".", x)
  # Remove any other commas altogether
  x = gsub(',',", x)
  # Entries with no digit in them set as NA
  x[!grepl("\\d", x)] = NA

  return(x)
}


# Find NA by coercion
# this function helps to find NAs which were added by coercion when settting values to
numeric in error checking
find.coercion <- function(before, after) {
  return(which(!is.na(before) & is.na(after)))
}

# Function which removes all */+ from the data and separate de values putting the
ones
# without the trailing in the X column and the values with the trailing in Y column
plus.fn <- function(x) {
  # A function to split off data with a trailing +
  y = array(NA, dim=c(length(x),1))

  # Find entries containing a + sign
  ind=grep("\\+", x)
  # Substitute values into y and remove + sign
  y[ind] = sub("\\+","",x[ind])
  # Set x values to NA
  x[ind] = NA

  # Find entries containing a * sign
  ind=grep("\\*", x)
  # Substitute values into y and remove * sign
  y[ind] = sub("\\*","",x[ind])
  # Set x values to NA
  x[ind] = NA


  return(cbind(x,y))
}

##########     end of function cleaning and plus       ##############
##### Process tempMed.9h data ############

# Extract data with a trailing */+ and put into column 7h
# only applied to tempMed.9h;nebulosidade.9h and humidade.9h
#clean the data first
d$tempMed.9h = clean(d$tempMed.9h)
d$humidade.9h = clean(d$humidade.9h)
d$nebulosidade.9h = clean(d$nebulosidade.9h)

# use the plus function to create the columns to 7h readings

tempMed = plus.fn(d$tempMed.9h)
d$tempMed.9h = tempMed[,1]
d$tempMed.7h = tempMed[,2]

#Humidity
humid = plus.fn(d$humidade.9h)
d$humidade.9h = humid[,1]
d$humidade.7h = humid[,2]
```

```
#cloudiness
nebul = plus.fn(d$nebulosidade.9h)
d$nebulosidade.9h= nebul[,1]
d$nebulosidade.7h = nebul[,2]

## End of plus function ##############################

###  stations variants function ##########################

# this line makes the search restraine in the districts chosen
#station.variants <- function(station, long, lat, district=NULL, district.list=NULL) { #
for district restriction

d$estacao=as.character(d$estacao)

d$estacao=as.character(d$estacao) # all variables must be characters
station.variants <- function(station, long, lat, alt) {
  # Find all variants and long-lat positions for a station

  station = as.character(station)
  if (is.factor(long)) {long = as.character(long)}
  if (is.factor(lat)) {lat = as.character(lat)}
  if (is.factor(alt)) {alt = as.character(alt)}


  station.list.in = unique(station)  # A list of all station names

  # Loop around all stations and group them roughly into variants that have same set
of long-lats

  #### Loop 1 Start
  i=1    # Set loop counter for loop 1
  while(length(station.list.in)>0) {
    station.variants = station.list.in[1] # Pick first name from the list of stations
    n.variants = length(station.variants)

    #### Loop 2 start
    loop.count = 0
    while (loop.count<100) {
      # Find all stations with this name
      ind = (station %in% station.variants)
      # Find unique long and lat of these
      long.variants = unique(long[ind])
      lat.variants = unique(lat[ind])
      alt.variants = unique(alt[ind])
      # Find station names with these long-lat combinations
      station.variants = unique(station[(long%in%long.variants & lat%in%lat.variants &
alt%in%alt.variants)])
      ###!!!!!!!!!!!!
      # Remove variants that are just blank
      #station.variants = grep("\\w", station.variants, value=T)  # Find variants that
have a letter or a number somewhere

      if (n.variants==length(station.variants)) {
        loop.count=101   # Stop searching if list of variants has not grown
      } else {
        loop.count = loop.count+1
        n.variants = length(station.variants) # Update the number of name variants
      }
    }
    #### Loop 2 end

    if (i==1) {
      station.list.out = list(station.variants)  # Create a list of variants for each station
    } else {
      station.list.out = c(station.list.out, list(station.variants)) # Add to the list if it
already exists
    }

    # Remove the variant names from station.list.in
    ind = which(station.list.in %in% station.variants)   # Find indices for discovered
variants
    station.list.in = station.list.in[-ind]              # Remove them from the station.list.in

    i = i+1 # Update loop counter for loop 1
  }
  #### Loop 1 end

  return(station.list.out)
}


##### End of function definitions ############################

# Clean numerical data
d$longitude = clean(d$longitude)
d$latitude = clean(d$latitude)
d$altitude = clean(d$altitude)

#### Group station names ###########
# Classify stations into groups (each group could be the same station)

station.list.out = station.variants(d$estacao, d$longitude, d$latitude, d$altitude)
n.stations = length(station.list.out)

long.list = vector('list', length=n.stations)
lat.list = vector('list', length=n.stations)
alt.list = vector('list', length=n.stations)
for (i in 1:n.stations) {
  ind = (d$estacao %in% station.list.out[[i]])  # Find indices matching all station name
variants
```

```
  long.list[[i]] = as.character(unique(d$longitude[ind]))     # Find longitude variants
  lat.list[[i]] = as.character(unique(d$latitude[ind]))   # Find latitude variants
  alt.list[[i]] = as.character(unique(d$altitude[ind])) # Find altitude variants
}

################## end of 2nd function station variants-altitude ###

## Developement of a new list using the station.list.out resulted from function of
# station variants

new.station.list=vector("list", length =234)

new.station.list[[1]]=c(station.list.out [[1]])#Balombo (Polig. Flor.)/Balombo
new.station.list[[1]][5]=c(station.list.out [[99]])

new.station.list[[2]]=c(station.list.out [[2]])#Lobito (S.M.A.)/Lobito

new.station.list[[3]]=c(station.list.out [[3]])#Cassequel

new.station.list[[4]]=c(station.list.out [[4]])#BiÃ³pio

new.station.list[[5]]=c(station.list.out [[5]])#Bocoio

new.station.list[[6]]=c(station.list.out [[6]][-2])#Benguela/Benguela-Dis. De Puericult.
new.station.list[[7]]=c(station.list.out [[175]])#Benguela(S.M.A)

new.station.list[[8]]=c(station.list.out [[7]])#BaÃa Farta

new.station.list[[9]]=c(station.list.out [[9]])#Fazenda S.Francisco

new.station.list[[10]]=c(station.list.out [[10]])#Dombe Grande

new.station.list[[11]]=c(station.list.out [[11]])#Ganda (Est. Zoot.)
new.station.list[[12]]=c(station.list.out [[14]][1])#Ganda (Posto AgrÃcola)"
new.station.list[[12]][2]=c(station.list.out [[86]])
new.station.list[[13]]=c(station.list.out [[14]][-1])# Ganda (C.de Estudo da Ganda)
new.station.list[[14]]=c(station.list.out [[15]])#Est. Reg. Da Ganda

new.station.list[[15]]=c(station.list.out [[12]])#Alto Catumbela
new.station.list[[16]]=c(station.list.out [[13]])#Caimbambo

new.station.list[[17]]=c(station.list.out [[16]][-2])#Bimbe

new.station.list[[18]]=c(station.list.out [[16]][2])#Chiumbo
new.station.list[[19]]=c(station.list.out [[17]])#VÂª Teixeira da Silva
new.station.list[[20]]=c(station.list.out [[18]])#Borga

new.station.list[[21]]=c(station.list.out [[19]][1])#Chinguril/Chinguri
new.station.list[[21]][2]=c(station.list.out [[19]][3])

new.station.list[[22]]=c(station.list.out [[19]][2])#Chinguar
new.station.list[[23]]=c(station.list.out [[19]][4])#Chinguar (C.F.B.)
new.station.list[[23]][2]=c(station.list.out [[19]][5])
new.station.list[[23]][3]=c(station.list.out [[19]][6])
new.station.list[[23]][4]=c(station.list.out [[193]])

new.station.list[[24]]=c(station.list.out [[19]][7])#Chinguar (AdministraÃ§Ã£o)"
new.station.list[[24]][2]=c(station.list.out [[19]][8])
new.station.list[[24]][3]=c(station.list.out [[19]][9])
new.station.list[[24]][4]=c(station.list.out [[19]][10])
new.station.list[[24]][5]=c(station.list.out [[123]])

new.station.list[[25]]=c(station.list.out [[20]])#Quipeio

new.station.list[[26]]=c(station.list.out [[21]][1])#Chianga
new.station.list[[27]]=c(station.list.out [[21]][2])#Chianga( C.Estudos)"
new.station.list[[27]][2]=c(station.list.out [[21]][3])
new.station.list[[27]][3]=c(station.list.out [[21]][4])
new.station.list[[27]][4]=c(station.list.out [[21]][5])
new.station.list[[27]][5]=c(station.list.out [[21]][6])

new.station.list[[28]]=c(station.list.out [[21]][7])#Chiange

new.station.list[[29]]=c(station.list.out [[22]][1])#SacaÃila (Per. Flor.)"
new.station.list[[29]][2]=c(station.list.out [[22]][2])

new.station.list[[30]]=c(station.list.out [[22]][3])#C. Est. De SacaÃila
new.station.list[[30]][2]=c(station.list.out [[22]][4])
new.station.list[[30]][3]=c(station.list.out [[88]][1])
new.station.list[[30]][4]=c(station.list.out [[88]][2])
new.station.list[[30]][5]=c(station.list.out [[88]][3])
new.station.list[[30]][6]=c(station.list.out [[88]][4])

new.station.list[[31]]=c(station.list.out [[89]])#Nova Lisboa
new.station.list[[32]]=c(station.list.out [[23]])#Nova Lisboa ( ou S.M.A.)

new.station.list[[33]]=c(station.list.out [[24]][1])#Chenga
new.station.list[[34]]=c(station.list.out [[24]][2])#Chenga ( Fazenda)
new.station.list[[34]][2]=c(station.list.out [[24]][3])

new.station.list[[35]]=c(station.list.out [[25]])#Lucamba
new.station.list[[36]]=c(station.list.out [[105]])#Fazenda Lucamba

new.station.list[[37]]=c(station.list.out [[26]][-5])#Cuima ( Escola TeÃ³filo Duarte
new.station.list[[38]]=c(station.list.out [[26]][5])#Cuima- Poligno Florestal

new.station.list[[39]]=c(station.list.out [[27]])#Andulo

new.station.list[[40]]=c(station.list.out [[28]][1])#Vila General Machado
new.station.list[[40]][2]=c(station.list.out [[28]][2])
new.station.list[[40]][3]=c(station.list.out [[28]][3])
new.station.list[[41]]=c(station.list.out [[28]][4])#Vila Mariano Machado
new.station.list[[41]][2]=c(station.list.out [[28]][5])
new.station.list[[41]][3]=c(station.list.out [[28]][6])

new.station.list[[42]]=c(station.list.out [[29]][2])#Nova Sintra
new.station.list[[42]][2]=c(station.list.out [[29]][3])
new.station.list[[42]][3]=c(station.list.out [[29]][5])

new.station.list[[43]]=c(station.list.out [[29]][1])#Nova Sintra (Catabola)
new.station.list[[43]][2]=c(station.list.out [[29]][4])
new.station.list[[43]][3]=c(station.list.out [[29]][6])

new.station.list[[44]]=c(station.list.out [[30]])#Coemba (Mis. Cat.)

new.station.list[[45]]=c(station.list.out [[31]])#Ceilunga( C.Est.)"
new.station.list[[45]][10]=c(station.list.out [[170]][1])
new.station.list[[45]][11]=c(station.list.out [[170]][2])
new.station.list[[45]][12]=c(station.list.out [[186]])

new.station.list[[46]]=c(station.list.out [[32]][1])#"Silva Porto (Cidade)/Silva Porto
new.station.list[[46]][2]=c(station.list.out [[32]][2])
new.station.list[[46]][3]=c(station.list.out [[32]][3])
new.station.list[[46]][4]=c(station.list.out [[32]][4])
new.station.list[[46]][5]=c(station.list.out [[32]][6])

new.station.list[[47]]=c(station.list.out [[32]][5])#Silva Porto (S.M.A)
new.station.list[[47]][2]=c(station.list.out [[32]][9])
new.station.list[[47]][3]=c(station.list.out [[122]])

new.station.list[[48]]=c(station.list.out [[32]][7])
new.station.list[[48]][2]=c(station.list.out [[32]][8])

new.station.list[[49]]=c(station.list.out [[33]][1])#Chitembo

new.station.list[[50]]=c(station.list.out [[33]][2])#Catota

new.station.list[[51]]=c(station.list.out [[33]][3])#Catota-MissÃ£o EvangÃ©lica

new.station.list[[52]]=c(station.list.out [[33]][4])#Catabola Mis. EvangÃ©lica

new.station.list[[53]]=c(station.list.out [[34]])#Cuchi ( Miss. Cat.)"
new.station.list[[53]][2]=c(station.list.out [[124]])
new.station.list[[53]][3]=c(station.list.out [[132]][1])
new.station.list[[53]][4]=c(station.list.out [[132]][2])

new.station.list[[54]]=c(station.list.out [[35]])#Serpa Pinto
new.station.list[[55]]=c(station.list.out [[125]])#Serpa Pinto(S.M.A.)
new.station.list[[55]][2]=c(station.list.out [[138]])#Serpa Pinto(S.M.A.)

new.station.list[[56]]=c(station.list.out [[36]])#Capico ( Miss. Cat
new.station.list[[56]][2]=c(station.list.out [[100]])

new.station.list[[57]]=c(station.list.out [[37]])#Mavinga/Mavinga (S.M.A)

new.station.list[[58]]=c(station.list.out [[38]])#Mis. Sta. Cruz Cuando"

new.station.list[[59]]=c(station.list.out [[39]])#Cuangar

new.station.list[[60]]=c(station.list.out [[40]])#Dirico

new.station.list[[61]]=c(station.list.out [[171]])# chamavera Dirico
new.station.list[[61]][4]=c(station.list.out [[8]][18])

new.station.list[[62]]=c(station.list.out [[41]][1])#V. Teixeira de Sousa
new.station.list[[63]]=c(station.list.out [[41]][-1])#V. Teixeira de Sousa (C.F.B)

new.station.list[[64]]=c(station.list.out [[42]])#Luso/Luso(S.M.A)

new.station.list[[65]]=c(station.list.out [[43]])#Cangamba
new.station.list[[66]]=c(station.list.out [[44]])#Vila Gago Coutinho
new.station.list[[67]]=c(station.list.out [[45]])#Lola
new.station.list[[68]]=c(station.list.out [[46]])#Vila Arriaga
new.station.list[[69]]=c(station.list.out [[47]])#Caracul

new.station.list[[70]]=c(station.list.out [[48]][1])#Bruco
new.station.list[[71]]=c(station.list.out [[48]][-1])#Bruco - Esc. De Reg. AgrÃcola

new.station.list[[72]]=c(station.list.out [[49]])#ChÃ£o da Chela
new.station.list[[73]]=c(station.list.out [[50]])#MoÃ§amedes(S.M.A.)/MoÃ§amedes
new.station.list[[74]]=c(station.list.out [[51]])#Porto Alexandre

new.station.list[[75]]=c(station.list.out [[52]][3])#Virei
new.station.list[[76]]=c(station.list.out [[52]][-3])#Virei-Curoca Norte

new.station.list[[77]]=c(station.list.out [[184]])#Muve-Virei

new.station.list[[78]]=c(station.list.out [[53]])#BaÃa dos Tigres

new.station.list[[79]]=c(station.list.out [[54]])#CupacaÃ§a/Capere
new.station.list[[79]][3]=c(station.list.out [[8]][30])#Capere

new.station.list[[80]]=c(station.list.out [[55]][1])#Caconda
new.station.list[[81]]=c(station.list.out [[55]][2])#Caconda (AdministraÃ§Ã£o
new.station.list[[81]][2]=c(station.list.out [[55]][4])
new.station.list[[81]][3]=c(station.list.out [[55]][5])

new.station.list[[82]]=c(station.list.out [[55]][3])#Caconda (Miss. Cat.)
new.station.list[[82]][2]=c(station.list.out [[55]][6])
new.station.list[[82]][3]=c(station.list.out [[55]][7])

new.station.list[[83]]=c(station.list.out [[56]][1])#Uaba 1/Uaba Baixo
new.station.list[[83]][2]=c(station.list.out [[56]][3])
new.station.list[[83]][3]=c(station.list.out [[56]][4])
new.station.list[[83]][4]=c(station.list.out [[56]][7])
new.station.list[[83]][5]=c(station.list.out [[56]][9])

new.station.list[[84]]=c(station.list.out [[56]][2])#Uaba 2 /Uaba Alto
new.station.list[[84]][2]=c(station.list.out [[56]][5])
new.station.list[[84]][3]=c(station.list.out [[56]][6])
new.station.list[[84]][4]=c(station.list.out [[56]][8])

new.station.list[[85]]=c(station.list.out [[57]])#Sangueve (Miss. Cat
new.station.list[[86]]=c(station.list.out [[58]])#Galangue

new.station.list[[87]]=c(station.list.out [[59]])#impulo
new.station.list[[88]]=c(station.list.out [[60]])#CuÃ©

new.station.list[[89]]=c(station.list.out [[61]][1])#Quilengues (AdministraÃ§Ã£o
new.station.list[[89]][2]=c(station.list.out [[61]][3])
new.station.list[[89]][3]=c(station.list.out [[61]][4])
new.station.list[[89]][4]=c(station.list.out [[61]][5])

new.station.list[[90]]=c(station.list.out [[61]][2])#Quilengues (Zoot.)
new.station.list[[90]][2]=c(station.list.out [[61]][6])

new.station.list[[91]]=c(station.list.out [[62]])#Vila Artur de Paiva
new.station.list[[92]]=c(station.list.out [[63]])#Dongo
new.station.list[[93]]=c(station.list.out [[64]])#Hoque
new.station.list[[94]]=c(station.list.out [[65]])#V. Paiva Couceiro (Quipungo)

new.station.list[[95]]=c(station.list.out [[66]])#Humpata (E.Z.S)
new.station.list[[95]][5]=c(station.list.out [[111]][1])
new.station.list[[95]][6]=c(station.list.out [[111]][2])
new.station.list[[95]][7]=c(station.list.out [[111]][3])

new.station.list[[96]]=c(station.list.out [[68]][-4])#Humpata (AgrÃcola)
new.station.list[[96]]=c(station.list.out [[68]][-5])
new.station.list[[96]][4]=c(station.list.out [[112]][1])
new.station.list[[96]][5]=c(station.list.out [[112]][2])
new.station.list[[96]][6]=c(station.list.out [[112]][3])
new.station.list[[96]][7]=c(station.list.out [[112]][4])
new.station.list[[96]][8]=c(station.list.out [[106]][1])

new.station.list[[97]]=c(station.list.out [[68]][4])#Humpata (Serv. Ag. E Flores.)"
new.station.list[[97]][2]=c(station.list.out [[68]][5])

new.station.list[[98]]=c(station.list.out [[126]])#Humpata (Centro de Estudos)

new.station.list[[99]]=c(station.list.out [[67]])#SÃi da Bandeira/SÃi da
Bandeira(S.M.A.)

new.station.list[[100]]=c(station.list.out [[69]])#HuÃla (Miss. CatÃ³lica)"
new.station.list[[100]][2]=c(station.list.out [[93]][1])
new.station.list[[100]][3]=c(station.list.out [[93]][2])
new.station.list[[100]][4]=c(station.list.out [[93]][3])
new.station.list[[100]][5]=c(station.list.out [[93]][4])

new.station.list[[101]]=c(station.list.out [[70]])#Cassinga

new.station.list[[102]]=c(station.list.out [[71]])#Tchivinguiro
new.station.list[[103]]=c(station.list.out [[72]][1])#Chibia
new.station.list[[104]]=c(station.list.out [[72]][-1])#Chibia (Vila JoÃ£o de Almeida)"

new.station.list[[105]]=c(station.list.out [[73]])#)Jau (Miss. Cat.)

new.station.list[[106]]=c(station.list.out [[74]])#Quihita (Miss. CatÃ³lica
new.station.list[[106]][2]=c(station.list.out [[94]])
new.station.list[[106]][3]=c(station.list.out [[127]][1])
new.station.list[[106]][4]=c(station.list.out [[127]][2])
new.station.list[[106]][5]=c(station.list.out [[127]][3])
new.station.list[[106]][6]=c(station.list.out [[127]][4])
new.station.list[[106]][7]=c(station.list.out [[127]][5])

new.station.list[[107]]=c(station.list.out [[127]][6])#Quihita - Est. Hidrogeol. NÂº5
new.station.list[[107]][2]=c(station.list.out [[127]][7])
new.station.list[[107]][3]=c(station.list.out [[127]][8])

new.station.list[[108]]=c(station.list.out [[75]])#Mulondo

new.station.list[[109]]=c(station.list.out [[76]])#Mupa (Miss. CatÃ³lica)
new.station.list[[109]][6]=c(station.list.out [[96]][1])
new.station.list[[109]][7]=c(station.list.out [[96]][2])

new.station.list[[110]]=c(station.list.out [[77]])#Cahama

new.station.list[[111]]=c(station.list.out [[78]][-3])#CÃifu

new.station.list[[112]]=c(station.list.out [[78]][3])#CÃifu (Posto Zoot. do Cunene)"
new.station.list[[112]][2]=c(station.list.out [[194]][1])
new.station.list[[112]][3]=c(station.list.out [[194]][2])
new.station.list[[112]][4]=c(station.list.out [[194]][3])
new.station.list[[112]][5]=c(station.list.out [[194]][4])
new.station.list[[112]][6]=c(station.list.out [[194]][5])

new.station.list[[113]]=c(station.list.out [[183]])#Cunene (C. de Estudos)
new.station.list[[113]][5]=c(station.list.out [[8]][29])

new.station.list[[114]]=c(station.list.out [[8]][10])#Foz do Cunene
new.station.list[[115]]=c(station.list.out [[79]])#Otchinjau

new.station.list[[116]]=c(station.list.out [[80]][3])#Chiulo
new.station.list[[117]]=c(station.list.out [[80]][-3])#Chiulo (Miss. CatÃ³lica)
new.station.list[[117]][9]=c(station.list.out [[97]])
new.station.list[[118]]=c(station.list.out [[81]][2])#RoÃ§adas
new.station.list[[119]]=c(station.list.out [[81]][-2])#RoÃ§adas (S.M.A.)
new.station.list[[119]][4]=c(station.list.out [[136]])

new.station.list[[120]]=c(station.list.out [[82]][2])#Namuculungo
new.station.list[[120]][2]=c(station.list.out [[137]])
new.station.list[[121]]=c(station.list.out [[82]][-2])#Namuculungo (C.E.I.L.A)

new.station.list[[122]]=c(station.list.out [[83]][2])#V. Pereira d'EÃ§a
new.station.list[[122]][2]=c(station.list.out [[83]][3])

new.station.list[[123]]=c(station.list.out [[83]][1])#Vila Pereira d'EÃ§a (S.G.M.)"
new.station.list[[123]][2]=c(station.list.out [[83]][4])
new.station.list[[123]][3]=c(station.list.out [[83]][5])

new.station.list[[124]]=c(station.list.out [[200]])#Pereira d'EÃ§a (S.M.A

new.station.list[[125]]=c(station.list.out [[85]])#Monte Belo"

new.station.list[[126]]=c(station.list.out [[87]][1])#Canjangue

new.station.list[[127]]=c(station.list.out [[87]][2])#Canjangue( Fazenda)"
new.station.list[[127]][2]=c(station.list.out [[87]][3])
new.station.list[[127]][3]=c(station.list.out [[135]])

new.station.list[[128]]=c(station.list.out [[90]])#Chingue

new.station.list[[129]]=c(station.list.out [[91]])#CuÃto Cuanavale
new.station.list[[129]][6]=c(station.list.out [[104]])

new.station.list[[130]]=c(station.list.out [[202]])#Cuito Canavale (FAP)"

new.station.list[[131]]=c(station.list.out [[92]])#Cavungo

new.station.list[[132]]=c(station.list.out [[95]])#Chibemba

new.station.list[[133]]=c(station.list.out [[98]])#OncÃ³cua

new.station.list[[134]]=c(station.list.out [[101]])#Cafima
new.station.list[[135]]=c(station.list.out [[102]])#Chingoroi
new.station.list[[135]][2]=c(station.list.out [[8]][4])
new.station.list[[135]][3]=c(station.list.out [[8]][7])
new.station.list[[135]][4]=c(station.list.out [[8]][8])
new.station.list[[135]][5]=c(station.list.out [[8]][12])
new.station.list[[135]][6]=c(station.list.out [[8]][13])
new.station.list[[135]][7]=c(station.list.out [[8]][15])
new.station.list[[135]][8]=c(station.list.out [[8]][19])

new.station.list[[136]]=c(station.list.out [[103]])#Longa
new.station.list[[137]]=c(station.list.out [[107]])#Posto Exp. do Lungo

new.station.list[[138]]=c(station.list.out [[108]])#Melunga Chiede

new.station.list[[139]]=c(station.list.out [[109]])#Mucusso

new.station.list[[140]]=c(station.list.out [[110]])#Mucundi

new.station.list[[141]]=c(station.list.out [[113]])#Lomaum
new.station.list[[141]][2]=c(station.list.out [[6]][2])#Lomuam

new.station.list[[142]]=c(station.list.out [[114]])#Catanga/Catanda

new.station.list[[143]]=c(station.list.out [[115]])#Cubal
new.station.list[[143]][2]=c(station.list.out [[8]][14])

new.station.list[[144]]=c(station.list.out [[8]][17])#Cubal (C.F.B

new.station.list[[145]]=c(station.list.out [[116]])#Catabola do Longonjo/Catabola

new.station.list[[146]]=c(station.list.out [[117]])#Luiana
new.station.list[[147]]=c(station.list.out [[118]])#Baia Farta

new.station.list[[148]]=c(station.list.out [[119]])#Tenda Moco

new.station.list[[149]]=c(station.list.out [[120]])#Mis.Cat.do Vouga
new.station.list[[149]][3]=c(station.list.out [[121]][1])

new.station.list[[150]]=c(station.list.out [[128]])#Chimbolelo

new.station.list[[151]]=c(station.list.out [[129]])#Catengue (C.F.B)"

new.station.list[[152]]=c(station.list.out [[130]])#Munhango

new.station.list[[153]]=c(station.list.out [[131]])#Mucussueje

new.station.list[[154]]=c(station.list.out [[133]])#Nonhe

new.station.list[[155]]=c(station.list.out [[134]])#Chiveio/Chiveio(Cuvelai)
new.station.list[[155]][5]=c(station.list.out [[145]])

new.station.list[[156]]=c(station.list.out [[139]])#Fazenda Fernando Alberto

new.station.list[[157]]=c(station.list.out [[140]])#Sanguengue/Sangueve(Per.Flor.)

new.station.list[[158]]=c(station.list.out [[141]])#Caianda

new.station.list[[159]]=c(station.list.out [[142]])#SeminÃirio CaÃila (!=SacaÃila(29-30))

new.station.list[[160]]=c(station.list.out [[143]][1])#Faz. Nelly
new.station.list[[160]][2]=c(station.list.out [[143]][2])
new.station.list[[160]][3]=c(station.list.out [[143]][3])
new.station.list[[160]][4]=c(station.list.out [[143]][4])

new.station.list[[161]]=c(station.list.out [[143]][5])#Congoia.Faz.Beira Alta
new.station.list[[161]][2]=c(station.list.out [[143]][6])

new.station.list[[162]]=c(station.list.out [[144]])#Fazenda Sumbo

new.station.list[[163]]=c(station.list.out [[146]][1])#Sambio

new.station.list[[164]]=c(station.list.out [[146]][-1])#MunuÃ©

new.station.list[[165]]=c(station.list.out [[147]])#Fazenda Santa Isabel

new.station.list[[166]]=c(station.list.out [[148]])#Calonga
new.station.list[[167]]=c(station.list.out [[149]])#Faz. Santa Eugenia
new.station.list[[168]]=c(station.list.out [[150]])#Dongoena

new.station.list[[169]]=c(station.list.out [[151]])#Lumbala

new.station.list[[170]]=c(station.list.out [[152]])#ColÃ³nia Penal (Capolo)

new.station.list[[171]]=c(station.list.out [[153]])#Bambi

new.station.list[[172]]=c(station.list.out [[154]])#Peu-Peu

new.station.list[[173]]=c(station.list.out [[155]])#Faz. Munana

new.station.list[[174]]=c(station.list.out [[156]])#Lutembo

new.station.list[[175]]=c(station.list.out [[157]])#Catembulo

new.station.list[[176]]=c(station.list.out [[158]])#Cuvelai-Matala

new.station.list[[177]]=c(station.list.out [[160]])#SÃ£o Nicolau

new.station.list[[178]]=c(station.list.out [[161]])#Jamba (Cassinga-Norte

new.station.list[[179]]=c(station.list.out [[162]])#Vila Nova

new.station.list[[180]]=c(station.list.out [[163]])#Santa Marta

new.station.list[[181]]=c(station.list.out [[164]])#Chicuaqueia

new.station.list[[182]]=c(station.list.out [[165]])#Micosse - Matala

new.station.list[[183]]=c(station.list.out [[166]])#Chinhama

new.station.list[[184]]=c(station.list.out [[167]])#Vila Folgares
new.station.list[[185]]=c(station.list.out [[168]])#Mungo

new.station.list[[186]]=c(station.list.out [[169]])#Handja

new.station.list[[187]]=c(station.list.out [[172]])#Giraul

new.station.list[[188]]=c(station.list.out [[173]])#Cutato

new.station.list[[189]]=c(station.list.out [[174]])#Vila da Matala

new.station.list[[190]]=c(station.list.out [[176]])#BaÃa dos Tigres(S.M.A.)"

new.station.list[[191]]=c(station.list.out [[177]])#Rio da Areia
new.station.list[[192]]=c(station.list.out [[178]])#Capagombe - Munhias
new.station.list[[193]]=c(station.list.out [[179]])#Senge/Sangue/Sengue
new.station.list[[193]][4]=c(station.list.out [[8]][33])#Senge/Sangue/Sengue

new.station.list[[194]]=c(station.list.out [[180]])#Cavaco  (Cent. De Estudos
new.station.list[[194]][2]=c(station.list.out [[8]][27])
new.station.list[[194]][3]=c(station.list.out [[8]][25])#Cavaco

new.station.list[[195]]=c(station.list.out [[181]])#Fazenda Mumba

new.station.list[[196]]=c(station.list.out [[182]][1])#Gambos

new.station.list[[197]]=c(station.list.out [[185]][1])#Mucope
new.station.list[[198]]=c(station.list.out [[185]][2]) #Mucope - Loana.

new.station.list[[199]]=c(station.list.out [[201]])#Mucope - S. Adm. Civil"

new.station.list[[200]]=c(station.list.out [[187]])#Catanga Lonjongo

new.station.list[[201]]=c(station.list.out [[188]])#Cameia

new.station.list[[202]]=c(station.list.out [[189]])#Bimbe (Est. Zoot. Do Sul)(Bimbe 17)
new.station.list[[203]]=c(station.list.out [[190]])#Canjola/Canjola - P.F.
new.station.list[[204]]=c(station.list.out [[191]])#Cacanda C. de Estudos
new.station.list[[205]]=c(station.list.out [[192]])#Chipindo
new.station.list[[206]]=c(station.list.out [[195]])#Cazombo
new.station.list[[207]]=c(station.list.out [[196]])#Alto Hama
new.station.list[[208]]=c(station.list.out [[197]])#Fazenda Etape
new.station.list[[209]]=c(station.list.out [[198]])#Taca
new.station.list[[210]]=c(station.list.out [[199]])#Faculdade de Medicina Veter

new.station.list[[211]]=c(station.list.out [[84]])#Chitado

new.station.list[[212]]=c(station.list.out [[203]])#Cangolo /Est. Hidrogeolog.

new.station.list[[213]]=c(station.list.out [[204]])#Bela Vista

new.station.list[[214]]=c(station.list.out [[205]])#Fazenda Piloto
new.station.list[[215]]=c(station.list.out [[206]])#Colonato S.Jorge do Cub."

new.station.list[[216]]=c(station.list.out [[207]])#Calucinda

new.station.list[[217]]=c(station.list.out [[208]])#Gando - P. N. Bicuari

new.station.list[[218]]=c(station.list.out [[209]])#Manquete

new.station.list[[219]]=c(station.list.out [[210]])#Chiede

new.station.list[[220]]=c(station.list.out [[8]][1])#Fazenda Prazeres
new.station.list[[220]][2]=c(station.list.out [[8]][16])
new.station.list[[220]][3]=c(station.list.out [[8]][35])

new.station.list[[221]]=c(station.list.out [[8]][2])#Sambo

new.station.list[[222]]=c(station.list.out [[8]][3])#Chiengue

new.station.list[[223]]=c(station.list.out [[8]][6])#Cazombo
new.station.list[[224]]=c(station.list.out [[8]][23])#Cazombo( ForÃ§a AÃ©rea)
new.station.list[[224]][2]=c(station.list.out [[8]][24])
new.station.list[[224]][3]=c(station.list.out [[8]][32])
new.station.list[[224]][4]=c(station.list.out [[8]][37])

new.station.list[[225]]=c(station.list.out [[8]][11]) #Chicuma

new.station.list[[226]]=c(station.list.out [[8]][20]) #Tchamutete

new.station.list[[227]][1]=c(station.list.out [[8]][26]) #Baixo Longa
new.station.list[[227]][2]=c(station.list.out [[8]][28])

new.station.list[[228]]=c(station.list.out [[8]][22]) #Ponto de Passagem

new.station.list[[229]]=c(station.list.out [[8]][34]) #Fazenda Santa Ana

new.station.list[[230]]=c(station.list.out [[8]][21]) #Mokete

new.station.list[[231]]=c(station.list.out [[8]][5]) #Lucira
new.station.list[[231]][2]=c(station.list.out [[8]][9])

new.station.list[[232]]=c(station.list.out [[8]][31]) #Calucinga

new.station.list[[233]]=c(station.list.out [[8]][36]) #Fazenda Sanga

new.station.list[[234]]=c(station.list.out [[159]][1]) #Granja BelÃ©m

save(new.station.list, file= "stations.Rdata")

## function to check erros in the new.statons.list compared with the station.list.out
```

```
# erro check station names

old.list = unlist(station.list.out)
new.list = unlist(new.station.list)


# Check for duplicate entries in new.station.list
if (length(new.list)>length(unique(new.list))) {new.list[duplicated(new.list)]}   # Show
names that are duplicated

# Check for missing stations
if (any(!(old.list%in%new.list))) {
  omitted = old.list[!(old.list%in%new.list)] # Show the names of these missing
stations

  print('Missing station names')
  print(omitted)
  print(' ')

  # Show original group contaiing each variant
  for (s in 1:length(omitted)) {
    ind = which(sapply(station.list.out,function(x) {any(x%in%omitted[s])}))
    ind2 = which(station.list.out[[ind]]==omitted[s])
    print(paste('Station name variant = ',omitted[s],'. This missing variant is from
station.list.out[[',ind,']][',ind2,']',sep=''))
  }
}

################ End of funtion #####################


####### Creation of a code number ID to each  group of stations ####

d["estacao.id"] = NA

loop.count = 1
for (i in 1:length(new.station.list)) {
  name.group = new.station.list[i]

  for (j in 1:length(name.group[[1]])) {
    ind = (d$estacao %in% name.group[[1]][j])
    d$estacao.id[ind] = loop.count
  }

  loop.count = loop.count + 1
}

#######  end of station variants and Id #####################


## Function to create district variants, same as stations variants
```

```
d$distrito=as.character(d$distrito)

district.variants <- function(district) {
  # Find all variants of district

  district = as.character(district)
  if (is.factor(district)) {district = as.character(district)}


  district.list.in = unique(district)  # A list of all district names

  # Loop around all districts and group them roughly into variants that have same
name

  #### Loop 1 Start
  i=1    # Set loop counter for loop 1
  while(length(district.list.in)>0) {
    district.variants = district.list.in[1] # Pick first name from the list of districts
    n.variants = length(district.variants)

    #### Loop 2 start
    loop.count = 0
    while (loop.count<100) {
      # Find all districts with this name
      ind = (district %in% district.variants)
      # Find unique district and lat of these
      district.variants = unique(district[ind])

      # Find district names with these district combinations
      district.variants = unique(district[(district%in%district.variants)])

      if (n.variants==length(district.variants)) {
        loop.count=101   # Stop searching if list of variants has not grown
      } else {
        loop.count = loop.count+1
        n.variants = length(district.variants) # Update the number of name variants
      }
    }
    #### Loop 2 end

    if (i==1) {
      district.list.out = list(district.variants)  # Create a list of variants for each district
    } else {
      district.list.out = c(district.list.out, list(district.variants)) # Add to the list if it
already exists
    }

    # Remove the variant names from district.list.in
    ind = which(district.list.in %in% district.variants)   # Find indices for discovered
variants
```

```
    district.list.in = district.list.in[-ind]           # Remove them from the district.list.in

    i = i+1 # Update loop counter for loop 1
  }
  #### Loop 1 end

  return(district.list.out)
}


##### End of function definitions ###############################


#### Group district names ###########
# Classify districts into groups (each group could be the same district)


district.list.out = district.variants(d$distrito)
n.districts = length(district.list.out)

district.list = vector('list', length=n.districts)

for (i in 1:n.districts) {
  ind = (d$distrito %in% district.list.out[[i]])  # Find indices matching all district name
variants
  district.list[[i]] = as.character(unique(d$distrito[ind]))     # Find stationitude variants

}

######################## END ######################

## Create a new list of district names and group them into the right groups

new.district.list=vector("list", length =9)

new.district.list[[1]]=c(district.list.out [[1]]) #Benguela
new.district.list[[1]][2]=c(district.list.out [[10]])

new.district.list[[2]]=c(district.list.out [[2]]) #Huambo

new.district.list[[3]]=c(district.list.out [[3]]) #BiÃ© - Cuando_Cubango
new.district.list[[3]][2]=c(district.list.out [[7]])
new.district.list[[3]][3]=c(district.list.out [[12]])

new.district.list[[4]]=c(district.list.out [[8]]) #BiÃ©
new.district.list[[4]][2]=c(district.list.out [[13]])

new.district.list[[5]]=c(district.list.out [[14]]) #Cuando_Cubango
new.district.list[[5]][2]=c(district.list.out [[15]])
new.district.list[[5]][3]=c(district.list.out [[16]])
new.district.list[[6]]=c(district.list.out [[4]]) #Moxico

new.district.list[[7]]=c(district.list.out [[5]]) #MoÃ§Ã¢medes
new.district.list[[7]][2]=c(district.list.out [[9]][1])

new.district.list[[8]]=c(district.list.out [[6]]) #HuÃla

new.district.list[[9]]=c(district.list.out [[17]]) #Cunene
new.district.list[[9]][2]=c(district.list.out [[11]])

save(new.district.list, file= "district.Rdata")


## Function to find which districts are missing, omitted or duplicated

old.list = unlist(district.list.out)
new.list = unlist(new.district.list)

# Check for duplicate entries in new.station.list
if (length(new.list)>length(unique(new.list))) {new.list[duplicated(new.list)]}   # Show
names that are duplicated

# Check for missing districts
if (any(!(old.list%in%new.list))) {
  omitted = old.list[!(old.list%in%new.list)] # Show the names of these missing
stations

  print('Missing district names')
  print(omitted)
  print(' ')

  # Show original group containig each variant
  for (s in 1:length(omitted)) {
    ind = which(sapply(district.list.out,function(x) {any(x%in%omitted[s])}))
    ind2 = which(district.list.out[[ind]]==omitted[s])
    print(paste('district name variant = ',omitted[s],'. This missing variant is from
district.list.out[[',ind,']][',ind2,']',sep=''))
  }
}

################ End of funtion ################

# District ID

# guive a code number ID to each of the District names

d["distrito.id"] = NA

loop.count = 1000
for (i in 1:length(new.district.list)) {
  name.group = new.district.list[i]
  for (j in 1:length(name.group[[1]])) {
    ind = (d$distrito %in% name.group[[1]][j])
    d$distrito.id[ind] = loop.count
  }

  loop.count = loop.count + 1000
}

########### End ##################

#### Rearenge lats and longs must be done at the end as stations list groups ae
done before

#Split latutide degree minutes into two variables and create decimal degree variable

# Start by replace degrees symbol or minutes symbol by blanks

ind = grep("'",d$latitude) # Find strings with a ' in them
d$latitude[ind] = gsub("\\D"," ",d$latitude[ind]) # Replace anything that isn't a digit
with a space

lat_split=strsplit(d$latitude, split="[[:blank:]]+")

d$lat.deg=NA
d$lat.min=NA

#for (i in 1:10) {d$lat.deg[i]=as.numeric(lat_split[[i]][1])}
#for (i in 1:10) {d$lat.min[i]=as.numeric(lat_split[[i]][2])}
for (i in 1:length(lat_split)) {d$lat.deg[i]=as.numeric(lat_split[[i]][1])}
for (i in 1:length(lat_split)) {d$lat.min[i]=as.numeric(lat_split[[i]][2])}

d$lat=NA
d$lat=d$lat.deg+d$lat.min/60

d$lat=-abs(d$lat)# All latitudes are negative
d$lat=as.character(d$lat)
#Split longitude degree minutes into two variables and create decimal degree variable


ind = grep("'",d$longitude)   # Find strings with a ' in them
d$longitude[ind] = gsub("\\D"," ",d$longitude[ind]) # Replace anything that isn't a
digit with a space

lon_split=strsplit(d$longitude, split="[[:blank:]]+")

d$lon.deg=NA
d$lon.min=NA

for (i in 1:length(lon_split)) {d$lon.deg[i]=as.numeric(lon_split[[i]][1])}
for (i in 1:length(lon_split)) {d$lon.min[i]=as.numeric(lon_split[[i]][2])}
```

```
d$lon=NA
d$lon=d$lon.deg+d$lon.min/60

d$lon=as.character(d$lon)

############################ End ####################


###### Error Flagging #####################################

#  Quality checking and information for original values
#-------------------------------------------------------
#            level            explanation
#-----------------------------------------------------------------------------
#       passed     failed
#-----------------------------------------------------------------------------
#        0                      No test performed
#Test 1   1         2              out of normal range
#Test 2   10        20             out of intrinsique range
#Test 3   100       200            out of range , greater than compareble variables
#Test 4   1000      2000           out of range, smaller than compareble variables
#Test 5   10000     20000          out of average range
#Test 6   100000    200000         not interpretable data
#Test 7   1000000   2000000        Values Outside Parameter
#------------------------------------------------------------------------------

##### precipitation #####

# Precipitation (Amount (daily) 0 - 500 mm)only for maximum temperature which is a daily value
# precipitation Total (Amount (monthly) > precipitation maximum < 850.0(maximum value of precipitation total)
# precipitaion total < 850.0


#Clean precipitation data
#Clean prec.max
d$prec.max.mm = clean(d$prec.max.mm)
old.prec.max.mm = d$prec.max.mm
d$prec.max.mm = as.numeric(d$prec.max.mm)
#clean prec.total
d$prec.total.mm = clean(d$prec.total.mm)

# Tests of prec.total.mm

old.prec.total.mm = d$prec.total.mm
d$prec.total.mm = as.numeric(d$prec.total.mm)#Warning message:NAs introduced by coercion

d$erro.prec.total.mm = 0       # Test not performed at start
```

```
ind.na = is.na(d$prec.total.mm)

# Test 1
# First flag not applied to this variable (prec.total.mm)

# Test 2
ind = d$prec.total.mm>=0 & d$prec.total.mm<=850.0 # TRUE if in the range
d$erro.prec.total.mm[!ind.na & ind] = d$erro.prec.total.mm[!ind.na & ind] + 1*10^1
# Passed test 2
d$erro.prec.total.mm[!ind.na & !ind] = d$erro.prec.total.mm[!ind.na & !ind] + 2*10^1
# Failed test 2

# Test 3
# No test performed

# Test 4
ind = d$prec.total.mm>=d$prec.max.mm    # TRUE if in the range
ind.na = is.na(d$prec.total.mm) | is.na(d$prec.max.mm)
d$erro.prec.total.mm[!ind.na & ind] = d$erro.prec.total.mm[!ind.na & ind] + 10^3    #
Passed test 4
d$erro.prec.total.mm[!ind.na & !ind] = d$erro.prec.total.mm[!ind.na & !ind] + 2*10^3
# Failed test 4

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.prec.total.mm) &  is.na(d$prec.total.mm)
d$erro.prec.total.mm[!is.na(ind.coerce) & !ind.coerce] =
d$erro.prec.total.mm[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.prec.total.mm[!is.na(ind.coerce) & ind.coerce] =
d$erro.prec.total.mm[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

################ End ################################

# Tests of prec.max.mm

d$erro.prec.max.mm = 0       # Test not performed at start
ind.na = is.na(d$prec.max.mm)

# Test 1

ind = d$prec.max.mm>=0 & d$prec.max.mm<=500 # TRUE if in the range
d$erro.prec.max.mm[!ind.na & ind] = d$erro.prec.max.mm[!ind.na & ind] + 1    #
Passed test 1
d$erro.prec.max.mm[!ind.na & !ind] = d$erro.prec.max.mm[!ind.na & !ind] + 2  #
Failed test 1

# Test 2
# Second flag not applied to this variable (prec.total.mm)
# Test 3

ind = d$prec.max.mm<=d$prec.total.mm # TRUE if in the range
ind.na = is.na(d$prec.total.mm) | is.na(d$prec.max.mm)
d$erro.prec.max.mm[!ind.na & ind] = d$erro.prec.max.mm[!ind.na & ind] + 10^2    #
Passed test 3
d$erro.prec.max.mm[!ind.na & !ind] = d$erro.prec.max.mm[!ind.na & !ind] + 2*10^2
# Failed test 3

# Test 4
# fourth flag not applied to this variable (prec.total.mm)

# Test 5
# No test performed

# Test 6
# 6th flag not applied to this variable

############### End ###########################

# erro flaging to humidity and Nebulosity

# Tests of Humidity

# clean was performed erlier
old.humidade.9h = d$humidade.9h
old.humidade.7h = d$humidade.7h

d$humidade.9h = as.numeric(d$humidade.9h)
d$humidade.7h = as.numeric(d$humidade.7h)

#Humidity.9h

d$erro.humidade.9h = 0       # Test not performed at start
ind.na = is.na(d$humidade.9h)

# Test 1

ind = d$humidade.9h>=0 & d$humidade.9h<=100 # TRUE if in the range
d$erro.humidade.9h[!ind.na & ind] = d$erro.humidade.9h[!ind.na & ind] + 1    #
Passed test 1
d$erro.humidade.9h[!ind.na & !ind] = d$erro.humidade.9h[!ind.na & !ind] + 2  #
Failed test 1

## Test 2

ind = d$humidade.9h>=10 & d$humidade.9h<=100 # TRUE if in the range
d$erro.humidade.9h[!ind.na & ind] = d$erro.humidade.9h[!ind.na & ind] + 1*10^1
# Passed test 2
d$erro.humidade.9h[!ind.na & !ind] = d$erro.humidade.9h[!ind.na & !ind] + 2*10^1
# Failed test 2
# Test 3
# third flag not applied to this variable

# Test 4
# 4th flag not applied to this variable

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

################ End ###################################

#Humidity.7h

d$erro.humidade.7h = 0       # Test not performed at start
ind.na = is.na(d$humidade.7h)

# Test 1

ind = d$humidade.7h>=0 & d$humidade.7h<=100 # TRUE if in the range
d$erro.humidade.7h[!ind.na & ind] = d$erro.humidade.7h[!ind.na & ind] + 1    #
Passed test 1
d$erro.humidade.7h[!ind.na & !ind] = d$erro.humidade.7h[!ind.na & !ind] + 2  #
Failed test 1

## Test 2

ind = d$humidade.7h>=10 & d$humidade.7h<=100 # TRUE if in the range
d$erro.humidade.7h[!ind.na & ind] = d$erro.humidade.7h[!ind.na & ind] + 1*10^1
# Passed test 2
d$erro.humidade.7h[!ind.na & !ind] = d$erro.humidade.7h[!ind.na & !ind] + 2*10^1
# Failed test 2

# Test 3
#  3th flag not applied to this variable

# Test 4
# 4th flag not applied to this variable

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

############  End #####################

#Nebulosidade.9h
```

```
old.nebulosidade.9h = d$nebulosidade.9h
d$nebulosidade.9h = as.numeric(d$nebulosidade.9h)#Warning message:NAs
introduced by coercion

d$erro.nebulosidade.9h = 0       # Test not performed at start
ind.na = is.na(d$nebulosidade.9h)

# Test 1

ind = d$nebulosidade.9h>=0 & d$nebulosidade.9h<10 # TRUE if in the range
d$erro.nebulosidade.9h[!ind.na & ind] = d$erro.nebulosidade.9h[!ind.na & ind] + 1
# Passed test 1
d$erro.nebulosidade.9h[!ind.na & !ind] = d$erro.nebulosidade.9h[!ind.na & !ind] + 2
# Failed test 1

## Test 2
#Second flag not applied to this variable

# Test 3
#third flag not applied to this variable

# Test 4
# third flag not applied to this variable

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.nebulosidade.9h) &  is.na(d$nebulosidade.9h)
d$erro.nebulosidade.9h[!is.na(ind.coerce) & !ind.coerce] =
d$erro.nebulosidade.9h[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.nebulosidade.9h[!is.na(ind.coerce) & ind.coerce] =
d$erro.nebulosidade.9h[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

########## End ###################################

#Nebulosidade.7h
old.nebulosidade.7h = d$nebulosidade.7h

d$nebulosidade.7h = as.numeric(d$nebulosidade.7h)

d$erro.nebulosidade.7h = 0       # Test not performed at start
ind.na = is.na(d$nebulosidade.7h)

# Test 1

ind = d$nebulosidade.7h>=0 & d$nebulosidade.7h<10 # TRUE if in the range
d$erro.nebulosidade.7h[!ind.na & ind] = d$erro.nebulosidade.7h[!ind.na & ind] + 1
# Passed test 1
```

```
d$erro.nebulosidade.7h[!ind.na & !ind] = d$erro.nebulosidade.7h[!ind.na & !ind] + 2
# Failed test 1

## Test 2
#Second flag not applied to this variable

# Test 3
#third flag not applied to this variable

# Test 4
# 4th flag not applied to this variable

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

########## End ###################################

# Temperature
# Air temperature -80 - +60?C. Obs.: temperature never goes higher than 45 and lower than -10.
# range (-10 - +45)
#Temperatures 9h < MaxMed temperatures;
#MaxMedtemp. > Temp 9h > MinMed temp;
#Temp.Diurna =  [((Max+Min)/2) - CA] < 0.11
#TempMed.max temperatures > TempMed.9h temperatures;
#TempMed.min temperatures < TempMed.9h temperatures;
#TempExt.max temperatures > TempMed.max temperatures;
#TempEx.min temperatures < TempMed.min temperatures;

#  Temperature range ( -10 to 45) error

#  TempMed.9h

#TempMed.9h was cleaned before
d$tempMed.min = clean(d$tempMed.min)
d$tempMed.max = clean(d$tempMed.max)
d$tempExt.min = clean(d$tempExt.min)
d$tempExt.max = clean(d$tempExt.max)

old.tempMed.9h = d$tempMed.9h
old.tempMed.max = d$tempMed.max
old.tempMed.min = d$tempMed.min
old.tempExt.min = d$tempExt.min
old.tempExt.max = d$tempExt.max

d$tempMed.min=as.numeric(d$tempMed.min)
d$tempMed.max=as.numeric(d$tempMed.max)
d$tempExt.min=as.numeric(d$tempExt.min)
```

```
d$tempExt.max=as.numeric(d$tempExt.max)#NAs by coercion
d$tempMed.9h = as.numeric(d$tempMed.9h)

d$erro.tempMed.9h = 0       # Test not performed at start
ind.na = is.na(d$tempMed.9h)


# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

## Test 2

ind = d$tempMed.9h>(-10) & d$tempMed.9h< 45 # TRUE if in the range
d$erro.tempMed.9h[!ind.na & ind] = d$erro.tempMed.9h[!ind.na & ind] + 1*10    #
Passed test 1
d$erro.tempMed.9h[!ind.na & !ind] = d$erro.tempMed.9h[!ind.na & !ind] + 2*10  #
Failed test 1

# Test 3

ind = d$tempMed.9h < d$tempMed.max # TRUE if in the range
ind.na = is.na(d$tempMed.9h) | is.na(d$tempMed.max)
d$erro.tempMed.9h[!ind.na & ind] = d$erro.tempMed.9h[!ind.na & ind] + 10^2    #
Passed test 2
d$erro.tempMed.9h[!ind.na & !ind] = d$erro.tempMed.9h[!ind.na & !ind] + 2*10^2  #
Failed test 2

# Test 4

ind = d$tempMed.9h > d$tempMed.min # TRUE if in the range
ind.na = is.na(d$tempMed.9h) | is.na(d$tempMed.min)
d$erro.tempMed.9h[!ind.na & ind] = d$erro.tempMed.9h[!ind.na & ind] + 10^3    #
Passed test 2
d$erro.tempMed.9h[!ind.na & !ind] = d$erro.tempMed.9h[!ind.na & !ind] + 2*10^3  #
Failed test 2

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th Flag not applied in this variable

############  End #####################

#tempMed.7h

old.tempMed.7h = d$tempMed.7h
d$tempMed.7h = as.numeric(d$tempMed.7h)

d$erro.tempMed.7h = 0       # Test not performed at start
```

```
ind.na = is.na(d$tempMed.7h)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2

ind = d$tempMed.7h>(-10) & d$tempMed.7h< 45 # TRUE if in the range
d$erro.tempMed.7h[!ind.na & ind] = d$erro.tempMed.7h[!ind.na & ind] + 1*10    #
Passed test 1
d$erro.tempMed.7h[!ind.na & !ind] = d$erro.tempMed.7h[!ind.na & !ind] + 2*10  #
Failed test 1

# Test 3

ind = d$tempMed.7h < d$tempMed.max # TRUE if in the range
ind.na = is.na(d$tempMed.7h) | is.na(d$tempMed.max)
d$erro.tempMed.7h[!ind.na & ind] = d$erro.tempMed.7h[!ind.na & ind] + 10^2    #
Passed test 2
d$erro.tempMed.7h[!ind.na & !ind] = d$erro.tempMed.7h[!ind.na & !ind] + 2*10^2  #
Failed test 2

# Test 4

ind = d$tempMed.7h > d$tempMed.min # TRUE if in the range
ind.na = is.na(d$tempMed.7h) | is.na(d$tempMed.min)
d$erro.tempMed.7h[!ind.na & ind] = d$erro.tempMed.7h[!ind.na & ind] + 10^3    #
Passed test 2
d$erro.tempMed.7h[!ind.na & !ind] = d$erro.tempMed.7h[!ind.na & !ind] + 2*10^3  #
Failed test 2

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

############  End #####################

#tempMed.min

d$erro.tempMed.min = 0       # Test not performed at start
ind.na = is.na(d$tempMed.min)

#applied
#old.tempMed.min = d$tempMed.min
#d$tempMed.min = as.numeric(d$tempMed.min)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2

ind = d$tempMed.min>(-10) & d$tempMed.min< 45 # TRUE if in the range
d$erro.tempMed.min[!ind.na & ind] = d$erro.tempMed.min[!ind.na & ind] + 1*10   #
Passed test 1
d$erro.tempMed.min[!ind.na & !ind] = d$erro.tempMed.min[!ind.na & !ind] + 2*10  #
Failed test 1

# Test 3

ind = d$tempMed.min< d$tempMed.9h # TRUE if in the range
ind.na = is.na(d$tempMed.min) | is.na(d$tempMed.9h)
d$erro.tempMed.min[!ind.na & ind] = d$erro.tempMed.min[!ind.na & ind] + 10^2    #
Passed test 2
d$erro.tempMed.min[!ind.na & !ind] = d$erro.tempMed.min[!ind.na & !ind] + 2*10^2
# Failed test 2

# Test 4

ind = d$tempMed.min > d$tempExt.min # TRUE if in the range
ind.na = is.na(d$tempMed.min) | is.na(d$tempExt.min)
d$erro.tempMed.min[!ind.na & ind] = d$erro.tempMed.min[!ind.na & ind] + 10^3    #
Passed test 2
d$erro.tempMed.min[!ind.na & !ind] = d$erro.tempMed.min[!ind.na & !ind] + 2*10^3
# Failed test 2

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

############  End ##########

#tempMed.max

d$erro.tempMed.max = 0       # Test not performed at start
ind.na = is.na(d$tempMed.max)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2
ind = d$tempMed.max>(-10) & d$tempMed.max< 45 # TRUE if in the range
d$erro.tempMed.max[!ind.na & ind] = d$erro.tempMed.max[!ind.na & ind] + 1*10   #
Passed test 1
```

```
d$erro.tempMed.max[!ind.na & !ind] = d$erro.tempMed.max[!ind.na & !ind] + 2*10 #
Failed test 1

# Test 3

ind = d$tempMed.max< d$tempExt.max # TRUE if in the range
ind.na = is.na(d$tempMed.max) | is.na(d$tempExt.max)
d$erro.tempMed.max[!ind.na & ind] = d$erro.tempMed.max[!ind.na & ind] + 10^2
# Passed test 2
d$erro.tempMed.max[!ind.na & !ind] = d$erro.tempMed.max[!ind.na & !ind] +
2*10^2  # Failed test 2

# Test 4

ind = d$tempMed.max > d$tempMed.9h # TRUE if in the range
ind.na = is.na(d$tempMed.max) | is.na(d$tempMed.9h)
d$erro.tempMed.max[!ind.na & ind] = d$erro.tempMed.max[!ind.na & ind] + 10^3
# Passed test 2
d$erro.tempMed.max[!ind.na & !ind] = d$erro.tempMed.max[!ind.na & !ind] +
2*10^3  # Failed test 2

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.tempMed.max) &  is.na(d$tempMed.max)
d$erro.tempMed.max[!is.na(ind.coerce) & !ind.coerce] =
d$erro.tempMed.max[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.tempMed.max[!is.na(ind.coerce) & ind.coerce] =
d$erro.tempMed.max[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

######### End ################

#tempExt.min

#apllied before
#d$tempExt.min = as.numeric(d$tempExt.min) #NAs introduced by coercion

d$erro.tempExt.min = 0       # Test not performed at start
ind.na = is.na(d$tempExt.min)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2
ind = d$tempExt.min>(-10) & d$tempExt.min< 45 # TRUE if in the range
d$erro.tempExt.min[!ind.na & ind] = d$erro.tempExt.min[!ind.na & ind] + 1*10    #
Passed test 1
d$erro.tempExt.min[!ind.na & !ind] = d$erro.tempExt.min[!ind.na & !ind] + 2*10  #
Failed test 1
```

```
# Test 3

ind = d$tempExt.min< d$tempMed.min # TRUE if in the range
ind.na = is.na(d$tempExt.min) | is.na(d$tempMed.min)
d$erro.tempExt.min[!ind.na & ind] = d$erro.tempExt.min[!ind.na & ind] + 1*10^2    #
Passed test 2
d$erro.tempExt.min[!ind.na & !ind] = d$erro.tempExt.min[!ind.na & !ind] + 2*10^2  #
Failed test 2

# Test 4
# 4th flag not applied to this variable

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.tempExt.min) &  is.na(d$tempExt.min)
d$erro.tempExt.min[!is.na(ind.coerce) & !ind.coerce] =
d$erro.tempExt.min[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.tempExt.min[!is.na(ind.coerce) & ind.coerce] =
d$erro.tempExt.min[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

######### End ################

#tempExt.max

d$erro.tempExt.max = 0       # Test not performed at start
ind.na = is.na(d$tempExt.max)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2

ind = d$tempExt.max>(-10) & d$tempExt.max< 45 # TRUE if in the range
d$erro.tempExt.max[!ind.na & ind] = d$erro.tempExt.max[!ind.na & ind] + 1*10    #
Passed test 1
d$erro.tempExt.max[!ind.na & !ind] = d$erro.tempExt.max[!ind.na & !ind] + 2*10  #
Failed test 1

# Test 3
# 3th plag do not aplly to this variable

# Test 4
ind = d$tempExt.max > d$tempMed.max
ind.na = is.na(d$tempExt.max) | is.na(d$tempMed.max)
d$erro.tempExt.max[!ind.na & ind] = d$erro.tempExt.max[!ind.na & ind] + 1*10^3
# Passed test 1
```

```
d$erro.tempExt.max[!ind.na & !ind] = d$erro.tempExt.max[!ind.na & !ind] + 2*10^3
# Failed test 1

# Test 5
# No test performed

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

########## End ############################

#tempMed.diurna

d$tempMed.diurna = clean(d$tempMed.diurna)
old.tempMed.diurna=d$tempMed.diurna

d$tempMed.diurna=as.numeric(d$tempMed.diurna)

d$erro.tempMed.diurna = 0       # Test not performed at start
ind.na = is.na(d$tempMed.diurna)

# Test 1
# first flag not applied to this variable as an intrinsic range was developed according to
coutry and data limits

# Test 2

ind = d$tempMed.diurna>(-10) & d$tempMed.diurna< 45 # TRUE if in the range
d$erro.tempMed.diurna[!ind.na & ind] = d$erro.tempMed.diurna[!ind.na & ind] + 1*10
# Passed test 1
d$erro.tempMed.diurna[!ind.na & !ind] = d$erro.tempMed.diurna[!ind.na & !ind] +
2*10  # Failed test 1

## Test 2
# 2nd test not applied to the variable

# Test 3
ind = d$tempMed.diurna < d$tempMed.max  # TRUE if in the range
ind.na = is.na(d$tempMed.diurna) | is.na(d$tempMed.max)
d$erro.tempMed.diurna[!ind.na & ind] = d$erro.tempMed.diurna[!ind.na & ind] +
1*10^2    # Passed test 1
d$erro.tempMed.diurna[!ind.na & !ind] = d$erro.tempMed.diurna[!ind.na & !ind] +
2*10^2 # Failed test 1

# Test 4
ind = d$tempMed.diurna > d$tempMed.min  # TRUE if in the range
ind.na = is.na(d$tempMed.diurna) | is.na(d$tempMed.min)
d$erro.tempMed.diurna[!ind.na & ind] = d$erro.tempMed.diurna[!ind.na & ind] +
1*10^3    # Passed test 1
d$erro.tempMed.diurna[!ind.na & !ind] = d$erro.tempMed.diurna[!ind.na & !ind] +
2*10^3 # Failed test 1
```

```
# Test 5
sum=(d$tempMed.max + d$tempMed.min)
ave= sum/2
co=abs(round(ave, digits=1)-d$tempMed.diurna)
# co must be smaller than 0.11, if greater there is an erro, 0.11

ind = co< 0.11  # TRUE if in the range
ind.na = is.na(co)
d$erro.tempMed.diurna[!ind.na & ind] = d$erro.tempMed.diurna[!ind.na & ind] +
1*10^4    # Passed test 1
d$erro.tempMed.diurna[!ind.na & !ind] = d$erro.tempMed.diurna[!ind.na & !ind] +
2*10^4 # Failed test 1


# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.tempMed.diurna) &  is.na(d$tempMed.diurna)
ind.na = is.na(d$tempMed.diurna) | is.na(d$tempExt.max)
d$erro.tempMed.diurna[!is.na(ind.coerce) & !ind.coerce] =
d$erro.tempMed.diurna[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.tempMed.diurna[!is.na(ind.coerce) & ind.coerce] =
d$erro.tempMed.diurna[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

######### End ################

# Number of days erro Flagging between 0 to 31 days

#prec.dias.0.1
d$prec.dias.0.1= clean(d$prec.dias.0.1)

old.prec.dias.0.1 = d$prec.dias.0.1
d$prec.dias.0.1 = as.numeric(d$prec.dias.0.1)#Warning message:NAs introduced by
coercion

d$erro.prec.dias.0.1 = 0       # Test not performed at start
ind.na = is.na(d$prec.dias.0.1)

# Test 1

ind = d$prec.dias.0.1 >=0 & d$prec.dias.0.1 <= 31 # TRUE if in the range
d$erro.prec.dias.0.1[!ind.na & ind] = d$erro.prec.dias.0.1[!ind.na & ind] + 1    #
Passed test 1
d$erro.prec.dias.0.1[!ind.na & !ind] = d$erro.prec.dias.0.1[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable
```

```
## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.prec.dias.0.1) &  is.na(d$prec.dias.0.1)
d$erro.prec.dias.0.1[!is.na(ind.coerce) & !ind.coerce] =
d$erro.prec.dias.0.1[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.prec.dias.0.1[!is.na(ind.coerce) & ind.coerce] =
d$erro.prec.dias.0.1[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$prec.dias.0.1))
d$erro.prec.dias.0.1[!ind] = d$erro.prec.dias.0.1[!ind] + 1*10^6 # Passed test 7
d$erro.prec.dias.0.1[ind] = d$erro.prec.dias.0.1[ind] + 2*10^6   # Failed test 7

######### End ################

#prec.dias.1
d$prec.dias.1= clean(d$prec.dias.1)

old.prec.dias.1 = d$prec.dias.1
d$prec.dias.1 = as.numeric(d$prec.dias.1)#Warning message:NAs introduced by
coercion

d$erro.prec.dias.1 = 0       # Test not performed at start
ind.na = is.na(d$prec.dias.1)

# Test 1

ind = d$prec.dias.1 >= 0 & d$prec.dias.1 <= 31 # TRUE if in the range
d$erro.prec.dias.1[!ind.na & ind] = d$erro.prec.dias.1[!ind.na & ind] + 1    # Passed
test 1
d$erro.prec.dias.1[!ind.na & !ind] = d$erro.prec.dias.1[!ind.na & !ind] + 2  # Failed
test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
```

```
ind.coerce =  !is.na(old.prec.dias.1) &  is.na(d$prec.dias.1)
d$erro.prec.dias.1[!is.na(ind.coerce) & !ind.coerce] =
d$erro.prec.dias.1[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.prec.dias.1[!is.na(ind.coerce) & ind.coerce] =
d$erro.prec.dias.1[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$prec.dias.1))
d$erro.prec.dias.1[!ind] = d$erro.prec.dias.1[!ind] + 1*10^6 # Passed test 7
d$erro.prec.dias.1[ind] = d$erro.prec.dias.1[ind] + 2*10^6   # Failed test 7

######### End ###############

#prec.dias.1
d$prec.dias.10= clean(d$prec.dias.10)

old.prec.dias.10 = d$prec.dias.10
d$prec.dias.10 = as.numeric(d$prec.dias.10)

d$erro.prec.dias.10 = 0       # Test not performed at start
ind.na = is.na(d$prec.dias.10)

# Test 1

ind = d$prec.dias.10 >= 0 & d$prec.dias.10 <= 31 # TRUE if in the range
d$erro.prec.dias.10[!ind.na & ind] = d$erro.prec.dias.10[!ind.na & ind] + 1    # Passed
test 1
d$erro.prec.dias.10[!ind.na & !ind] = d$erro.prec.dias.10[!ind.na & !ind] + 2  # Failed
test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
# 6th flag not applied

# Test 7 ()
ind=grep("\\.", (d$prec.dias.10))
d$erro.prec.dias.10[!ind] = d$erro.prec.dias.10[!ind] + 1*10^6 # Passed test 7
d$erro.prec.dias.10[ind] = d$erro.prec.dias.10[ind] + 2*10^6   # Failed test 7

######### End ##########################
```

```
# trovoada.dias

d$trovoada.dias= clean(d$trovoada.dias)
old.trovoada.dias= d$trovoada.dias
d$trovoada.dias = as.numeric(d$trovoada.dias)

d$erro.trovoada.dias = 0       # Test not performed at start
ind.na = is.na(d$trovoada.dias)

# Test 1

ind = d$trovoada.dias >= 0 & d$trovoada.dias <= 31 # TRUE if in the range
d$erro.trovoada.dias[!ind.na & ind] = d$erro.trovoada.dias[!ind.na & ind] + 1    #
Passed test 1
d$erro.trovoada.dias[!ind.na & !ind] = d$erro.trovoada.dias[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
# 6th flap not applied

# Test 7 ()
ind=grep("\\.", (d$trovoada.dias))
d$erro.trovoada.dias[!ind] = d$erro.trovoada.dias[!ind] + 1*10^6 # Passed test 7
d$erro.trovoada.dias[ind] = d$erro.trovoada.dias[ind] + 2*10^6   # Failed test 7

######### End ###########################

# relampago.dias
d$relampago.dias= clean(d$relampago.dias)
old.relampago.dias= d$relampago.dias
d$relampago.dias = as.numeric(d$relampago.dias)

d$erro.relampago.dias = 0       # Test not performed at start
ind.na = is.na(d$relampago.dias)

# Test 1

ind = d$relampago.dias >= 0 & d$relampago.dias <= 31 # TRUE if in the range
```

```
d$erro.relampago.dias[!ind.na & ind] = d$erro.relampago.dias[!ind.na & ind] + 1    #
Passed test 1
d$erro.relampago.dias[!ind.na & !ind] = d$erro.relampago.dias[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
# 6th flag not applied to this variable

# Test 7 ()
ind=grep("\\.", (d$relampago.dias))
d$erro.relampago.dias[!ind] = d$erro.relampago.dias[!ind] + 1*10^6 # Passed test 7
d$erro.relampago.dias[ind] = d$erro.relampago.dias[ind] + 2*10^6   # Failed test 7

######### End ################

# chuva.dias
d$chuva.dias= clean(d$chuva.dias)

old.chuva.dias = d$chuva.dias
d$chuva.dias = as.numeric(d$chuva.dias)

d$erro.chuva.dias = 0       # Test not performed at start
ind.na = is.na(d$chuva.dias)

# Test 1

ind = d$chuva.dias >= 0 & d$chuva.dias <= 31 # TRUE if in the range
d$erro.chuva.dias[!ind.na & ind] = d$erro.chuva.dias[!ind.na & ind] + 1    # Passed
test 1
d$erro.chuva.dias[!ind.na & !ind] = d$erro.chuva.dias[!ind.na & !ind] + 2  # Failed
test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
```

```
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.chuva.dias) &  is.na(d$chuva.dias)
d$erro.chuva.dias[!is.na(ind.coerce) & !ind.coerce] =
d$erro.chuva.dias[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.chuva.dias[!is.na(ind.coerce) & ind.coerce] =
d$erro.chuva.dias[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$chuva.dias))
d$erro.chuva.dias[!ind] = d$erro.chuva.dias[!ind] + 1*10^6 # Passed test 7
d$erro.chuva.dias[ind] = d$erro.chuva.dias[ind] + 2*10^6   # Failed test 7

######### End ################

# cacimbo.dias
d$cacimbo.dias= clean(d$cacimbo.dias)

old.cacimbo.dias = d$cacimbo.dias
d$cacimbo.dias = as.numeric(d$cacimbo.dias)#Warning message:NAs introduced by
coercion

d$erro.cacimbo.dias = 0       # Test not performed at start
ind.na = is.na(d$cacimbo.dias)

# Test 1

ind = d$cacimbo.dias >= 0 & d$cacimbo.dias <= 31 # TRUE if in the range
d$erro.cacimbo.dias[!ind.na & ind] = d$erro.cacimbo.dias[!ind.na & ind] + 1    #
Passed test 1
d$erro.cacimbo.dias[!ind.na & !ind] = d$erro.cacimbo.dias[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.cacimbo.dias) &  is.na(d$cacimbo.dias)
```

```
d$erro.cacimbo.dias[!is.na(ind.coerce) & !ind.coerce] =
d$erro.cacimbo.dias[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.cacimbo.dias[!is.na(ind.coerce) & ind.coerce] =
d$erro.cacimbo.dias[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$cacimbo.dias))
d$erro.cacimbo.dias[!ind] = d$erro.cacimbo.dias[!ind] + 1*10^6 # Passed test 7
d$erro.cacimbo.dias[ind] = d$erro.cacimbo.dias[ind] + 2*10^6   # Failed test 7

######### End ################

# nevoeiro.dias
d$nevoeiro.dias = clean(d$nevoeiro.dias)

old.nevoeiro.dias = d$nevoeiro.dias
d$nevoeiro.dias = as.numeric(d$nevoeiro.dias)#Warning message:NAs introduced by
coercion

d$erro.nevoeiro.dias = 0       # Test not performed at start
ind.na = is.na(d$nevoeiro.dias)

# Test 1

ind = d$nevoeiro.dias >= 0 & d$nevoeiro.dias <= 31 # TRUE if in the range
d$erro.nevoeiro.dias[!ind.na & ind] = d$erro.nevoeiro.dias[!ind.na & ind] + 1    #
Passed test 1
d$erro.nevoeiro.dias[!ind.na & !ind] = d$erro.nevoeiro.dias[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.nevoeiro.dias) &  is.na(d$nevoeiro.dias)
d$erro.nevoeiro.dias[!is.na(ind.coerce) & !ind.coerce] =
d$erro.nevoeiro.dias[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.nevoeiro.dias[!is.na(ind.coerce) & ind.coerce] =
d$erro.nevoeiro.dias[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$nevoeiro.dias))
```

```
d$erro.nevoeiro.dias[!ind] = d$erro.nevoeiro.dias[!ind] + 1*10^6 # Passed test 7
d$erro.nevoeiro.dias[ind] = d$erro.nevoeiro.dias[ind] + 2*10^6   # Failed test 7

######### End ################

#dates

# tempExt.max.data

d$tempExt.max.data = as.character(d$tempExt.max.data)
# clean function for dates, it does not set VD to NAs
clean.dates <- function(x) { # JY function
  # Convert to character
  x = as.character(x)
  # Replace first apperance of a comma with a decimal point
  x = sub(",",".", x)
  # Remove any other commas altogether
  x = gsub(',','', x)


  return(x)
}

d$tempExt.max.data = clean.dates(d$tempExt.max.data)

null_idxs=d$tempExt.max.data==""
d$tempExt.max.data[null_idxs]=NA
dash_idxs=d$tempExt.max.data=="-"
d$tempExt.max.data[dash_idxs]=NA

old.tempExt.max.data = d$tempExt.max.data
d$tempExt.max.data = as.numeric(d$tempExt.max.data)#Warning message:NAs
introduced by coercion

d$erro.tempExt.max.data = 0       # Test not performed at start
ind.na = is.na(d$tempExt.max.data)
# Test 1

ind = d$tempExt.max.data >= 1 & d$tempExt.max.data <= 31 # TRUE if in the range
d$erro.tempExt.max.data[!ind.na & ind] = d$erro.tempExt.max.data[!ind.na & ind] + 1
# Passed test 1
d$erro.tempExt.max.data[!ind.na & !ind] = d$erro.tempExt.max.data[!ind.na & !ind] +
2  # Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
```

```
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.tempExt.max.data) &  is.na(d$tempExt.max.data)
d$erro.tempExt.max.data[!is.na(ind.coerce) & !ind.coerce] =
d$erro.tempExt.max.data[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.tempExt.max.data[!is.na(ind.coerce) & ind.coerce] =
d$erro.tempExt.max.data[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$tempExt.max.data))
d$erro.tempExt.max.data[!ind] = d$erro.tempExt.max.data[!ind] + 1*10^6 # Passed
test 7
d$erro.tempExt.max.data[ind] = d$erro.tempExt.max.data[ind] + 2*10^6   # Failed
test 7

######### End ################

# tempExt.min.data

d$tempExt.min.data = as.character(d$tempExt.min.data)

# clean function for dates, it does not set VD to NAs
clean.dates <- function(x) { # JY function
  # Convert to character
  x = as.character(x)
  # Replace first apperance of a comma with a decimal point
  x = sub(",",".", x)
  # Remove any other commas altogether
  x = gsub(',',", x)


  return(x)
}

d$tempExt.min.data = clean.dates(d$tempExt.min.data)

null_idxs=d$tempExt.min.data==""
d$tempExt.min.data[null_idxs]=NA
dash_idxs=d$tempExt.min.data=="-"
d$tempExt.min.data[dash_idxs]=NA

old.tempExt.min.data = d$tempExt.min.data
d$tempExt.min.data = as.numeric(d$tempExt.min.data)#Warning message:NAs
introduced by coercion

d$erro.tempExt.min.data = 0       # Test not performed at start
ind.na = is.na(d$tempExt.min.data)
```

```
# Test 1

ind = d$tempExt.min.data >= 1 & d$tempExt.min.data <= 31 # TRUE if in the range
d$erro.tempExt.min.data[!ind.na & ind] = d$erro.tempExt.min.data[!ind.na & ind] + 1
# Passed test 1
d$erro.tempExt.min.data[!ind.na & !ind] = d$erro.tempExt.min.data[!ind.na & !ind] +
2  # Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.tempExt.min.data) &  is.na(d$tempExt.min.data)
d$erro.tempExt.min.data[!is.na(ind.coerce) & !ind.coerce] =
d$erro.tempExt.min.data[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.tempExt.min.data[!is.na(ind.coerce) & ind.coerce] =
d$erro.tempExt.min.data[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$tempExt.min.data))
d$erro.tempExt.min.data[!ind] = d$erro.tempExt.min.data[!ind] + 1*10^6 # Passed
test 7
d$erro.tempExt.min.data[ind] = d$erro.tempExt.min.data[ind] + 2*10^6   # Failed
test 7

###########  End ################################

# prec.max.data

d$prec.max.data = as.character(d$prec.max.data)

# clean function for dates, it does not set VD to NAs
clean.dates <- function(x) { # JY function
  # Convert to character
  x = as.character(x)
  # Replace first apperance of a comma with a decimal point
  x = sub(",",".", x)
  # Remove any other commas altogether
  x = gsub(',','', x)
  return(x)
}

d$prec.max.data = clean.dates(prec.max.data)

null_idxs=d$prec.max.data==""
d$prec.max.data[null_idxs]=NA
dash_idxs=d$prec.max.data=="-"
d$prec.max.data[dash_idxs]=NA

old.prec.max.data= d$prec.max.data
d$prec.max.data = as.numeric(d$prec.max.data)#Warning message:NAs introduced
by coercion

d$erro.prec.max.data = 0       # Test not performed at start
ind.na = is.na(d$prec.max.data)
# Test 1

ind = d$prec.max.data >= 1 & d$prec.max.data <= 31 # TRUE if in the range
d$erro.prec.max.data[!ind.na & ind] = d$erro.prec.max.data[!ind.na & ind] + 1    #
Passed test 1
d$erro.prec.max.data[!ind.na & !ind] = d$erro.prec.max.data[!ind.na & !ind] + 2  #
Failed test 1

## Test 2 #check why this values
# 2nd test not apllied to the variable

## Test 3 #check why this values
# 3rd test not apllied to the variable

## Test 4 #check why this values
# 4th test not apllied to the variable

## Test 5 #check why this values
# 5th test not apllied to the variable

# Test 6 (data exists but not interpretable)
ind.coerce =  !is.na(old.prec.max.data) &  is.na(d$prec.max.data)
d$erro.prec.max.data[!is.na(ind.coerce) & !ind.coerce] =
d$erro.prec.max.data[!is.na(ind.coerce) & !ind.coerce] + 1*10^5 # No coercion
d$erro.prec.max.data[!is.na(ind.coerce) & ind.coerce] =
d$erro.prec.max.data[!is.na(ind.coerce) & ind.coerce] + 2*10^5   # Coerce to NA

# Test 7 ()
ind=grep("\\.", (d$prec.max.data))
d$erro.prec.max.data[!ind] = d$erro.prec.max.data[!ind] + 1*10^6 # Passed test 7
d$erro.prec.max.data[ind] = d$erro.prec.max.data[ind] + 2*10^6   # Failed test 7

d$prec.max.data = old.prec.max.data
##################   End of erro checking
#########################################
```

```
# return to the old data without NAs introduced by coercion
d$prec.total.mm = old.prec.total.mm
# return to the old data without NAs introduced by coercion
d$prec.max.mm = old.prec.max.mm
# return to the old data without NAs introduced by coercion
d$humidade.9h = old.humidade.9h
# return to the old data without NAs introduced by coercion
d$humidade.7h = old.humidade.7h
# return to the old data without NAs introduced by coercion
d$nebulosidade.9h = old.nebulosidade.9h
# return to the old data without NAs introduced by coercion
d$nebulosidade.7h = old.nebulosidade.7h

# return to the old data without NAs introduced by coercion
d$tempMed.9h = old.tempMed.9h
# return to the old data without NAs introduced by coercion
d$tempMed.7h = old.tempMed.7h
# return to the old data without NAs introduced by coercion
d$tempMed.min = old.tempMed.min
# return to the old data without NAs introduced by coercion
d$tempMed.max = old.tempMed.max
# return to the old data without NAs introduced by coercion
d$tempExt.min = old.tempExt.min
# return to the old data without NAs introduced by coercion
d$tempExt.max = old.tempExt.max
# return to the old data without NAs introduced by coercion
d$tempMed.diurna = old.tempMed.diurna

# return to the old data without NAs introduced by coercion
d$prec.dias.0.1 = old.prec.dias.0.1
# return to the old data without NAs introduced by coercion
d$prec.dias.1 = old.prec.dias.1
# return to the old data without NAs introduced by coercion
d$prec.dias.10 = old.prec.dias.10
# return to the old data without NAs introduced by coercion
d$trovoada.dias=old.trovoada.dias
# return to the old data without NAs introduced by coercion
d$chuva.dias = old.chuva.dias
# return to the old data without NAs introduced by coercion
d$relampago.dias=old.relampago.dias
# return to the old data without NAs introduced by coercion
d$nevoeiro.dias = old.nevoeiro.dias
# return to the old data without NAs introduced by coercion
d$cacimbo.dias = old.cacimbo.dias

# return to the old data without NAs introduced by coercion
d$tempExt.max.data=old.tempExt.max.data
# return to the old data without NAs introduced by coercion
d$tempExt.min.data=old.tempExt.min.data
```

```
#Saving cleaned data
save(d,new.station.list,file="clean.Rdata")

# clearing unecessary variables and organizing the database
# rearanging the dataframe

colnames(d) # see all the names to reaorderr the sequence

# remove from database as it was splitted and renamed lat.deg and lat.min
d$latitude=NULL
d$longitude=NULL

colnames(d) # see all the names to reaorderr the sequence after removing previews

# necessary packge to rename variables. Install and require is necessary for R to
performe command
#install.packages("reshape")
require(reshape)
d=rename(d, c(lat="latitude", lon="longitude"))
d=rename(d, c(lat.deg="lat.grau", lon.deg="lon.grau"))

d= d[c("distrito", "distrito.id", "estacao",
"estacao.id","lat.grau","lat.min","latitude","lon.grau","lon.min","longitude","altitude","te
mpMed.9h",
"tempMed.7h","tempMed.max","tempMed.min","tempMed.diurna","tempExt.max","tem
pExt.max.data","tempExt.min","tempExt.min.data","humidade.9h", "humidade.7h",
"nebulosidade.9h","nebulosidade.7h", "prec.total.mm", "prec.max.mm",
"prec.max.data", "prec.dias.0.1", "prec.dias.1", "prec.dias.10", "trovoada.dias",
"relampago.dias", "chuva.dias", "nevoeiro.dias", "cacimbo.dias", "year", "month",
"erro.tempMed.9h", "erro.tempMed.7h", "erro.tempMed.max", "erro.tempMed.min",
"erro.tempMed.diurna", "erro.tempExt.max", "erro.tempExt.max.data",
"erro.tempExt.min", "erro.tempExt.min.data", "erro.humidade.9h", "erro.humidade.7h",
"erro.nebulosidade.9h","erro.nebulosidade.7h", "erro.prec.total.mm",
"erro.prec.max.mm", "erro.prec.max.data", "erro.prec.dias.0.1", "erro.prec.dias.1",
"erro.prec.dias.10", "erro.trovoada.dias", "erro.relampago.dias", "erro.chuva.dias",
"erro.nevoeiro.dias", "erro.cacimbo.dias")]


# Write CSV in R for all access
write.csv(d, file = "Database.csv")


#### END ############################################
```

Appendix 5

Script 3 Stations Map

```
#script created by Nídia
# Script modified by JY (7/12/2015) up to ##### mark
# Lines added by JY are flagged by "Added by JY" comments

# Added by JY
rm(list=ls())
load('clean.Rdata')


#  Districts and stations IDs and lats and longs associated.
n.new.stations = length(new.station.list)
stations = data.frame(Station.id=c(1:n.new.stations), Station.Name=NA, lat=NA,
lon=NA, District.Name=NA, District.ID=NA)
for (i in 1:n.new.stations) {
  stations$Station.Name[i]=new.station.list[[i]][1]

  district_id = unique(d$distrito.id[d$estacao %in% stations$Station.Name[i]])
  #print(i)
  #print(district_id)

  # Added by JY
  if (length(district_id)>1) {  # Print a warning if a station name has more than one
district id
    print(paste('Warning: more than 1 district for station ',i, ':
(',stations$Station.Name[i],')'))
  }

  district_id = district_id[1]
  district_names = unique(d$distrito[d$distrito.id %in% district_id])

  # Added by JY
  district_names = district_names[nchar(district_names)>0] # Remove blank names
  #print(district_names)
  stations$District.Name[i] = district_names[1]
  # Added by JY
  stations$District.ID[i] = district_id[1]

  #district_names = unique(d$distrito[ which(d$estacao %in%
stations$Station.Name[i])])
  #stations$District.Name[i] = district_names[1]

  ind = (d$estacao %in% new.station.list[[i]])  # Find indices matching all station
name variants
  long.list = unique(d$lon[ind])     # Find longitude variants
  lat.list = unique(d$lat[ind])   # Find latitude variants
  #alt.list = unique(d$altitude[ind]) # Find altitude variants
```

```
 ## pick alatitude and longitude and altitude
 if (!all(is.na(long.list))) {
  tmp=table(d$lon[ind])
  stations$lon[i] = as.numeric(names(which.max(tmp)))


  # repeat for latitude
  if (!all(is.na(lat.list))) {
   tmp=table(d$lat[ind])
   stations$lat[i] = as.numeric(names(which.max(tmp)))
  }

 }

}

##################
#### No modifications by JY below this line

# Map plotting

library(maps)
library(sp)

# Added by JY
angola <- readRDS("C:/Users/Nidia/Desktop/Ang_data/data/AGO_adm1.rds")
#angola <- readRDS("AGO_adm2.rds")
# map('worldHires','Angola') # Commented by JY
#library(maptools)

# Added by JY
plot(angola, axes=T)

# library(scales)
# map.scale(relwidth=0.2, ratio=FALSE)
map.scale(x=12.5, y=-18.2, ratio=T, metric = T, relwidth=0.15, col ="black",cex =
0.5)
title('Location of stations')

# Added by JY (use rainbow colours)
districts = as.factor(stations$District.Name)
col.map=rainbow(nlevels(districts))
colours = col.map[match(districts, levels(districts))]
points(stations$lon, stations$lat, bg=colours, cex = 0.8, pch=21)
library(scales)

## END ######
```