11-1-2023

Dra. Ing. Agrim. Patricia A Rosell

centro de ingeniería mendoza argentina (cima)

Data management for Meteorology

Index

[1 Data structure 3](#_Toc124349740)

[1.1 Hierarchy 3](#_Toc124349741)

[1.2 Country 4](#_Toc124349742)

[1.3 Provider 4](#_Toc124349743)

[1.4 Year 4](#_Toc124349744)

[1.5 Filenames 4](#_Toc124349745)

[2 00\_programas 4](#_Toc124349746)

[3 parametros.txt 5](#_Toc124349747)

[4 Providers data structure 6](#_Toc124349748)

[4.1 RAW 6](#_Toc124349749)

[4.2 NORM 6](#_Toc124349750)

[4.3 GRAPH 7](#_Toc124349751)

[4.4 aux 7](#_Toc124349752)

[4.5 STATS.txt 7](#_Toc124349753)

[4.6 README\_provider 8](#_Toc124349754)

[5 Current providers 8](#_Toc124349755)

[5.1 ARG 8](#_Toc124349756)

[5.1.1 SMN 8](#_Toc124349757)

[5.2 URY 8](#_Toc124349758)

[5.2.1 INUMET 8](#_Toc124349759)

[6 Useful programs for general purposes 8](#_Toc124349760)

[6.1 outliers.R 8](#_Toc124349761)

[6.2 genera\_graficos.R 9](#_Toc124349762)

[6.3 grafica\_variables.R 9](#_Toc124349763)

[7 How to run R scripts from LINUX terminal: 9](#_Toc124349764)

Data management for meteorology

User Manual

Last update: 11th January 2023

¡Welcome!

In this manual you will find information about the structure of data, programs for cleaning and storing data, how to add new information and running the programs. This project was made for the Neutral Atmosphere Analysis Center of SIRGAS of Centro de Ingeniería Mendoza Argentina (CIMA).

If you have any doubts, don't hesitate to ask any questions.

**Contact me!**

**Patricia A. Rosell, Ph.D.**

**​​**[**patricia.rosell@ingenieria.uncuyo.edu.ar**](mailto:patricia.rosell@ingenieria.uncuyo.edu.ar)

**CIMA Team**

**Chair**

María Virginia Mackern Oberti, Ph.D

**Researchers**

María Laura Mateo, Ph.D

María Fernanda Camisay, Ph.D

Patricia Alejandra Rosell, Ph.D

**Interns**

Agostina Ulloa

Pablo Hurtado

# Data structure

In this section you can find the general structure of data. Even if you wish to update current data or add new one, you should follow this structure.

**Main directory:** Set a directory where the data will be stored.

**Master folder:** Datos\_MET

## Hierarchy

Data is stored for each meteorological station in an accumulated file per year and according to the provider. The general structure is as follows:

Datos\_MET

**Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente**|\_**COUNTRY**

|\_**Provider 1**

|\_**RAW**

|\_**NORM**

| |\_YEAR 1

| |\_Station name 1

| |\_Station name 2

| |\_Station name 3

| |\_etc..

| |\_YEAR 2

| |\_Station name 1

| |\_Station name 2

| |\_Station name 3

| |\_etc..

|\_**GRAPH**

|\_ **\_aux**

Figure . Example of the structure

|\_**STATS.txt**

|\_**README\_provider**

In the next sections, you can find information about each folder and file involved.

## Country

Names of countries are expressed using the nomenclature of ISO 3166 alpha-3. If you want to add a new countries should follow this nomenclature. To find it, visit the site: <https://www.iso.org/obp/ui/#search>

## Provider

The institution that gives the data. For every new provider, it is necessary to create a folder with the official acronym in capital letters. An abstract of information about the new provider should be added in this readme (see section 6). If the institution doesn't have an official acronym, a new one should be created (also in capital letters and as short as possible).

## Year

Data is stored by year. The programs will read the information available and will store a file per each station with the accumulated data.

## Filenames

The names of each station follow the codes defined by the World Meteorological Organization (WMO - In Spanish: Organización Meteorológica Mundial (OMM)) and the International Civil Aviation Organization (ICAO - In Spanish: Organización de Aviación Civil Internacional (OACI)). For instance: 86580\_SUMU.txt

If a station doesn't have these identifications, it is recommended to create a replacement only for the OACI code. A good way of identifying if the code is the official or a created one is following this suggestion:

**XXYY**

With XX fixed, and YY is two letters representative of the site where the stations are settled.

Ex: NA\_XXVG.txt

In this example, NA stands for "Not available" for the WMO number, XX indicates that the ICAO code is not official, and VG is the representative letter for the site "Vigata" [a fictional town where the Comisario Montalbano lives].

To find out how this information is obtained, please read the section 4.5.

# 00\_programas

This folder is used to store the scripts written for every provider and also for scripts that seek general purposes. The programs for dealing with the data of a specific provider should be saved into a folder named with its acronym (real or not, but always de same). This is because a provider could require several scripts, and all of them should be easily accessible.

Scrips saved outside the provider's folders require normalized data and therefore are for common use. The README of every provider should detail the main use of each script (see section 4.3).

This folder can be used for saving scrips for common. To find out what they do, check the readme inside this folder.

# parametros.txt

This is file and contains a list of paths and parameters for reading, managing, and saving data.

General parameters can be changed by hand before every process, but it is mandatory to check if every path is the correct one. Even though paths can be changed, this action is not recommended.

If you want to save data in a different folder, you can change the path, but remember to change it back, otherwise, you can affect future processing so please, beware!

If a new provider is added and new scripts are made, the paths should be added in this file. Use as a model the existent path name, do not use quotation marks, and use the bar (/) as the separator:

#####################

## General parameters ##

#####################

**overwrite=** Enter "Y" if you want to overwrite data of a specific day, or "N" for not overwriting. Do not forget quotation marks.

**graph\_acum=** Enter "Y" for plotting the accumulated data for each station, or "N" for not plotting (This will save time).

**sd\_outliers=** If graphics are going to be plotted, you should indicate an integer value for the standard deviation (sd) that will be used as a limit for removing outliers. Ex. sd\_outliers=3 - This means that the program will compute the monthly mean and sd of the meteorological data (Only Pressures and Temperature) and will exclude those that exceed the mean+3\*sd and that are below mean-3\*sd. The greater the value, the less restrictive the filter will be.

##############

## Directories ##

##############

**path\_base=**Main directory

**path\_scripts=**Directory for saving the scripts

## ARG - SMN paths ##

# Programms

**smn\_programas=**Path to the location of the SMN scripts

# Folders

**SMN=**Main directory for the SMN data

**smn\_dato\_norm=**Directory where normalized data is stored

**smn\_graph=**Directory where graphics are stored

# Data

**smn\_datohorario\_raw=**Directory where hourly-data is stored

**smn\_datoTP\_raw=**Directory where real-time data is stored

**smn\_historico=**Directory where historic records are stored

# Auxiliary data

**smn\_stats\_norm=**Location of STATS.txt file

**smn\_stats\_raw=**Location of raw stats list file

**smn\_equivalencias=**Location of the names equivalence file

# Providers data structure

Each provider gives data in a different format (RAW). Therefore, a set of scripts should be written to normalize it (NORM).

## RAW

This folder should contain data at its raw level. This mean, without any kind of processing. If there is a different kind of data for each provider, this should be grouped and saved in different folders, with representative (and short) names.

## NORM

This folder stores the normalized data, once processed. Data do not have a header but are sorted as follows:

DATE,PRE\_mar,PRE\_stat,TEMP,HUM,PREC

In which:

DATE: Self-explanatory. Format: YYYY-MM-DD HH:mm:SS

YYYY: Year

MM: Month

DD: Day

HH: Hour (UTC)

mm: minutes

SS: Seconds

PRE\_stat: Pressure at the station level, in heptopascal [hPa]

PRE\_mar: Pressure at sea level, in heptopascal [hPa]

TEMP: Temperature, in centigrades [ºC]

HUM: Humidity, in porcentage [%]

PREC: Precipitations, in millimiters [mm]

Missing data: Is represented with "NA" (Not available).

## GRAPH

Output folder for graphics. The filenames will follow the normalize name.

## aux

Auxiliary data. This is a generic folder, recommended for saving data such as row list of stations, shapefiles, contact info of the providers, among others. Please use this folder to avoid loose files.

Each file should be documented as soon as possible in the readme of each provider.

## STATS.txt

Stations list with a standard format. The header contains the following information:

OACI,OMM,SITE,LAT,LONG,ALT

In which:

* OMM: World Meteorological Organization number (In spanish: Organización Meteorológica Mundial OMM)
* OACI: International Civil Aviation Organization code (In spanish Organización de Aviación Civil Internacional OACI)
* SITE: Name of the site where the station is located
* LAT: Latitude, in decimals degrees
* LONG: Longitude, in decimals degrees
* ALT: Altitude

Missing data is represented with "NA" (Not available)

These lists are not the ones given by the providers. Every provider gives a list in different formats and, therefore, a new one should be created. The raw list of stations can be stored in the auxiliary folder and the new one can be created by hand or with a script.

If the OMM number is not given by the providers, you can find it here:

<https://opendata.dwd.de/climate_environment/CDC/help/stations_list_CLIMAT_data.txt>

Also, if the OACI code is not given, you can create one following the instructions in section 1.5.

## README\_provider

A readme with detailed information about data and programs of one provider. What they do, how to run it, etc.

# Current providers

Up today (see Last update), we have information about the following providers:

## ARG

### SMN

Servicio Meteorológico Nacional (https://www.smn.gob.ar/descarga-de-datos)

## URY

### INUMET

Instituto Uruguayo de Meteorología (https://www.inumet.gub.uy/)

# Useful programs for general purposes

A set of useful programs can be found in the “00\_programas” folder. These scripts are written in R language and can be call from other functions or run independently. To find out how to run an R script, go to section 7. In the current section you can find a description of them.

## outliers.R

If graphics are going to be plotted it is recommended to remove outliers, especially when dealing with real-time data. For running this script, you need three arguments: a data frame with the data organized in columns, the name of the column of interest, and an integer value that represents the threshold that will be used. There is no need to run this program only with meteorological data, you can use it for any kind of data as long as it is in the required format.

## genera\_graficos.R

A function that plots several graphics into one. It requires as an argument, the path to the parameters file and the provider of interest. This program works only with data that follows the format of the normalized data. Even if your data is in the proper format, the only way of using this script is by copying the parameters file and filling it with the paths to your data. This function generates an accumulated data frame for each station and creates a graphic panel for every variable involved.

## grafica\_variables.R

This is not a function itself. Is a compilation of lines of code for different kinds of graphs. Each line should be run individually.

# How to run R scripts from LINUX terminal:

Up today (see Last update), the scrips should be executed in an R terminal by hand. From a Linux terminal, log in to the server of your choice (one with an R package installed), with your user and password. Use the command "ssh" as follows:

ssh paroseell@gps20 <enter>

password \*\*\*\*\*\*\*\*\* <enter>

With:

parosell as the username

gps20 as the server

Then, move to the provider programs folder. Use the "cd" function ("change directory"). Ex. If we want to run the SMN programs:

parosell@gps20: cd /nfs/gps20/Datos\_MET/00\_programas/SMN <enter>

Once in the folder, we open R:

parosell@gps20:/nfs/gps20/Datos\_MET/00\_programas/SMN$ R <enter>

Our terminal will now look like this:

Texto

Descripción generada automáticamente

Figure . Linux terminal with an active session of R

For running a script enter (do not write the > symbol, this is how you will see the terminal):

> source('programita.R') <enter>

"programita" is the name of the program you want to run and will change according to the provider. To find out with program you should run, go to the readme of that provider. Do not forget to add the ".R" extension.

Please note that it is not necessary to change the directory to the programs folder, you can replace that by entering the absolute path to the script. Ex:

> source('/nfs/gps20/Datos\_MET/00\_programas/SMN/programita.R')

If this is the first time you run the script, it will search for the required libraries and install it and/or load it. If an installation is requested, you should choose the mirror server (a pop-up window will appear, and you only have to select the mirror of your choice). That is a one-time action for every user. Once this task ends, the programs will start to process the data and when finishing, the program will inform you with a message.