

User Manual for Weather Script in R

Juan Camilo Rivera Palacio
Centro Internacional de Agricultura Tropical (CIAT)
Agricultura Especifica por sitio
j.c.rivera@cgiar.org

April 2018

1 Description

Weather data script have been developed by big data's CIAT team. The objective of this script is treatment of precipitation, temperature, relative humidity and solar irradiance data from weather stations. The treatment consists in identification of outliers, filling missing values and give format to input file.

For using code, the data has to meet some requirements, as shown in the section [1.1]. The inputs are limits for variables, that are explained in the section [1.2].

It is important to note that clean data is used as an input in the predictive models.

1.1 Conditions Data Format

The data format has to meet the following conditions:

- The data is text file, i.e. the extension is .txt.
- The name of file is composed by name station, variable and units and separated for underscore (_). i.e `namestation_variable_units`. For example, `BLOCK123_P_MM.txt`. The symbols for variables are table [1] and the units table are in [2]. Note that each file has only one variable.
- The data can be hourly or daily. If data is hourly, the file is composed by two columns, Date and Value, figure [1]. By default, the date format is `YYYYMMDD` and separator columns is empty. But the user can change these parameters.

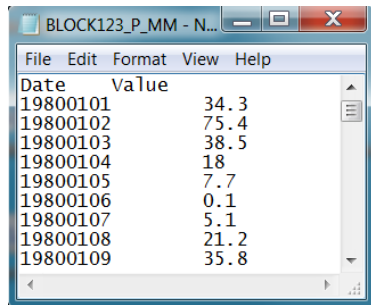
If data is hourly, the file is composed by three columns, Date, Hour and Value, figure [2]. By default, the date format is `YYYYMMDD` and hour format is 24 format, `HH:MM`.

Symbol	Description
P	Precipitation
TX	Maximum Temperature
TM	Minimum Temperature
SR	Solar Radiation
RH	Relative humidity

Table 1: Description of variables

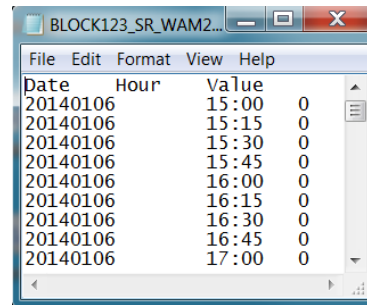
Symbol	Description
MJM2	Megajoules per Square Meter
KWHM2	Kilowatts per Square Meter
FD	Fahrenheit Degrees
MM	Milimeters
NE	A number between 0 to 100
CALCM2	Calories per Squared Centimeter
RH	Relative humidity

Table 2: Units description



Date	Value
19800101	34.3
19800102	75.4
19800103	38.5
19800104	18
19800105	7.7
19800106	0.1
19800107	5.1
19800108	21.2
19800109	35.8

Figure 1: Text file format daily



Date	Hour	Value
20140106	15:00	0
20140106	15:15	0
20140106	15:30	0
20140106	15:45	0
20140106	16:00	0
20140106	16:15	0
20140106	16:30	0
20140106	16:45	0
20140106	17:00	0

Figure 2: Text file format hourly

1.2 Restriction Inputs

The restrictions for inputs or variables are in the figure [4]. Note that restrictions depending on type of data, hourly or daily. Each weather variable has a restriction, i.e. limits for its maximum and minimum value, as shown in figure [3]. Additionally, the units for these variables are in table [2].

Different from weather variables, there are variables for functionality of code, figure [3]. These are:

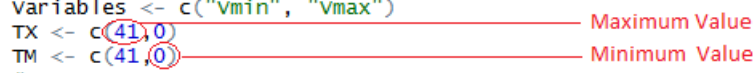
- **Hourly_Daily**. If data is collected daily is 1 and 2 when is hourly.
- **Star_date** and **End_date**. When research starts and ends. The format is "YYYY-MM-DD".
- **Percentage**. It is a percentage of data required to process station. For example, if the percentage is 0.7 that means the station must have 70% of total data.
- **separt**. It is the type of separation between columns of data file and is the same argument sep of function **read.table**.
- **LONG** and **LAT**. When data is hourly, the user must enter spatial location of data zone , longitude (LONG) and latitude (LAT).
- **dist_Station** It is distance in centimeters for clustering stations.

```
#Variables
#Choose Time Data
#If the time is in terms of hours so Hourly_Daily = 1
#If the time is in terms of days so Hourly_Daily = 2

Hourly_Daily <- 2
Start_date <- c("2005-1-1")
End_date <- c("2012-12-31")
Percentage <- 0.7
separt <- ""
date_format <- "%Y%m%d"
dist_Station <- 20000
```

Figure 3: Restrictions for code

```
#Hourly Restrictions as data frame
variables <- c("vmin", "vmax")
TX <- c(41, 0)
TM <- c(41, 0)
..
```



Maximum Value

Minimum Value

Figure 4: Hourly and daily limits

These restrictions are stored as a data frame.

2 Step by step

2.1 Load libraries, sources and folders

Once the libraries and source are loaded, code will create folders, as shown figure [5]. In the Original_Data folder is where input data must be stored.

In the SpatialInformation_InputVariables folder, there is a file called Information_Sptial_Stations.xls. In this file must be registered longitude, latitude and elevation per station, as shown in figure [6].

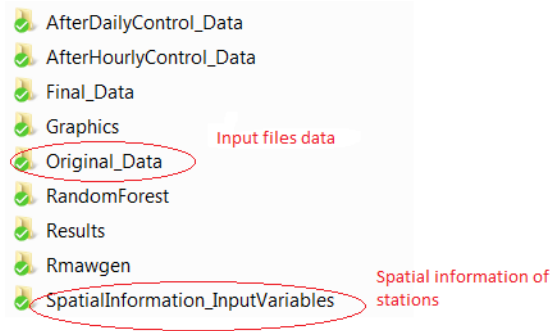


Figure 5: Folders

	A	B	C	D
1	Station_Name	Latitude	Longitude	Altitude
2	11045010	5.690555556	-76.64377778	75
3	11047040	5.697991667	-76.66225833	20.83
4	11120040	7.439444444	-77.11527778	8
5	12010070	7.884444444	-76.64777778	18

Figure 6: Spatial information per station

The AfterDailyControl_Data and AfterHourlyControl_Data are folders that save files with format and without outliers. The process data files will be saved in AfterHourlyControl_Data if data is hourly and AfterDailyControl_Data if data is daily.

2.2 Monitoring Files

The monitoring files are files that save information about results of daily control process, clustering stations and amount data per station.

The results from daily control process is composed by name station, latitude, longitude, altitude, variable, start data and end data. That information will save in Results_DailyControl.xls.

The information of amount data per station is composed by stations that have no minimum amount data and stations with more information, the first is saved in Stations_Delete.xls and the latter in Stations_Few_NA.txt. Note that user must delete stations that don't meet the conditions for amount of data in the Original_Data folder, and then running script from beginning.

The Clustering_Stations.xls is saved all analysis clustering of filled missing values process, this will explain more later.

The files mentioned above are stored in the Results folder, as shown in figure [7].





Name	Date modified	Type	Size
 Clustering_Stations	3/28/2018 11:07 A...	Microsoft Excel Co...	4 KB
 Results_DailyControl	4/18/2018 2:43 PM	Microsoft Excel Co...	5 KB
 Stations_Delete	4/19/2018 7:53 AM	Text Document	1 KB
 Stations_Few_NA	4/19/2018 7:53 AM	Text Document	1 KB

Figure 7: Result folder

2.3 Filled missing values

The code uses package **RMWAGEN** for fill missing values process for precipitation, maximum and minimum temperature. This is based on Vector Auto Regression (VAR). And random forest for solar radiation and relative humidity.

The code has a comment, **# Using RMWAGEN**, for recognize that process will start. It is important to note that the process is manual because there is not a method for recognize automatically which stations can be joined. A good approximation is detected closer stations, the code has an analysis of clustering where stations are joined according their spatial location. This analysis are in files Clustering_Station.xls of Results folder and Clustering_Station.jpeg in Graphics, see figures [8], [9]. On the other hand, random forest is automatic process.

2.4 Results

The final results are restored as matrices in the folder Results. These matrices are files with extension .txt and saved with name station, i.e. stationname.txt.

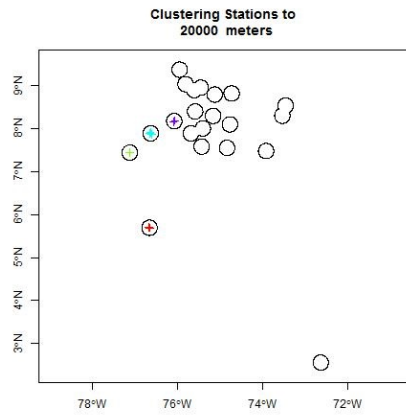


Figure 8: Diagram clustering

	A	B	C	D	E	F
1		Station_Nari	clust	Star_Date	End_Date	Variable_Nari
2	1	11045010	1	1/1/1980	7/31/2016	P
8	7	11120040	2	1/1/1980	4/30/2016	P
9	8	12010070	3	1/1/1980	4/30/2016	P
35	34	25020140	10	1/1/1980	3/31/2016	P

Figure 9: Excel file clustering

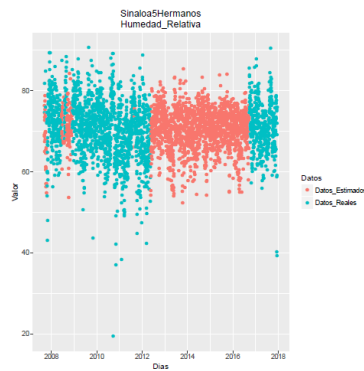


Figure 10: Final Graphic

Date	P	RH	SR	TM	TX
2007-09-10	0.2	74.61	267.01	27	30.9
2007-09-11	0	81.7	504.68	25	33.6
2007-09-12	38.8	76.3901478412698	373.		
2007-09-13	26.6	68.7249635038775	303.		
2007-09-14	26.6	69.1482690471047	471.		
2007-09-15	26.6	67.4128729574869	491.		
2007-09-16	2.4	64.9086557245535	489.7		
2007-09-17	5	73.4596580810261	444.725		
2007-09-18	5	72.6227677074022	454.690		
2007-09-19	2.4	73.4186370916295	432.2		
2007-09-20	0	75.4325875949885	465.253		
2007-09-21	5	78.8176081414142	267.958		
2007-09-22	24.8	77.4313263290278	286.		
2007-09-23	5	74.2138282162021	391.225		

Figure 11: Final Data

Also there are graphs of the final results in the graphics folder.