Statistical Analysis from Time Series Related to Climate Data

Ascensión Hernández Encinas, Araceli Queiruga Dios, Luis Hernández Encinas, and Víctor Gayoso Martínez

***Abstract*—Due to the development that different techniques of data analysis have achieved in recent years, some of them applied to the particular case of climatic variables obtained from the province of Salamanca (Spain), are detailed and analyzed in this study. The analysis was performed by using R software, specifically developed for statistical computing. The aims of this work are to obtain some statistical indicators that allow us to make a climatologic study as complete as possible in the mentioned area.**

***Index Terms*—Climate data, climate variables, R software statistical index.**

I. Introduction

Researches on climate extremes require daily data, but these data must be precise and of good quality, which is not an easy task. The data used in our study have been obtained from the Spanish government meteorology agency (AEMET), in particular from two of its observatories near the town of Béjar, in the south of Salamanca (Spain): La Hoya and Navamuño.

The daily data we have used are related to precipitation, maximum temperature and minimum temperature from 1992 to 2009 (see Fig. 1 and Fig. 2).

In addition, we have studied the data according to four different classifications:

- Monthly survey of extreme temperatures, average temperature and accumulated rainfall.
- Study of the seasons.
- The monthly average temperature and accumulated rainfall.
- Average temperature and total accumulated rainfall annually.

The aims of this work are to obtain some statistical indicators that allow us to make a climatologic study as complete as possible in the mentioned area. We present the consequences of the use of specific statistical software to obtain the results on spatial and temporal variability of warm days, or those exceeding the 90th percentile of maximum temperature, and cold nights, or those falling below the 10th percentile of minimum temperature [1].

The results of this study enable us to assure that the difference between the maximum and the minimum temperature had increased in these places, the precipitation is less intense nowadays, and the number of frost days is increasing.

This work is organized as follows: In Section II the data analysis is detailed. Section III is dedicated to the main part of the paper: The calculation of the climate extremes to make possible the analysis of the data. Finally, conclusions are stayed in Section IV.

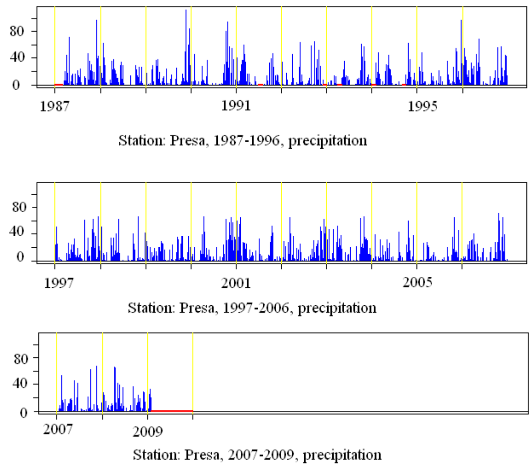

Fig. 1. Precipitation amount data from Navamuño included missing data (in red colour).

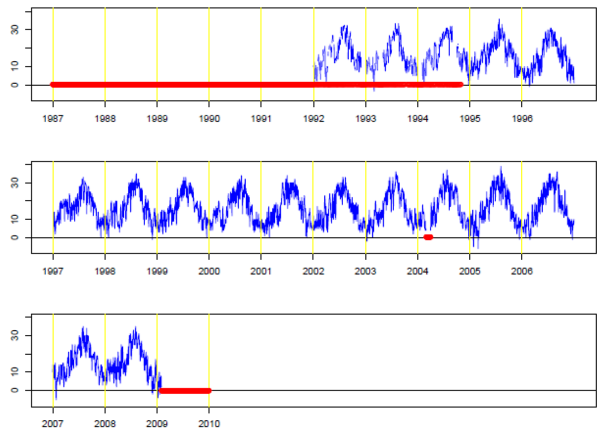

Fig. 2. Daily maximum temperature from Navamuño with missing data (in red colour).

II. Data Analysis and Control

The main starting point of every data processing is to have some good quality data without inhomogeneities [2], [3]. So, the first step in achieving our goals is to get reliable data, by

Manuscript received October 9, 2012; revised January 10, 2013. This work was supported by Fundación Memoria D. Samuel Solórzano Barruso under the project FS/19-2011, and Ministerio de Ciencia e Innovación (Spain) under the grant TIN2011-22668 (Spain).

A. Hernández Encinas and A. Queiruga Dios are with the Department of Applied Mathematics, E.T.S.I.I. of Béjar. University of Salamanca. Fernando Ballesteros, 37700 Béjar, Salamanca, Spain (e-mail: ascen@usal.es, queirugadios@usal.es).

L. Hernández Encinas and V. Gayoso Martínez are with the Department of Information Processing and Coding, Information Security Institute (ISI), CSIC. Serrano 144, 28006 Madrid, Spain (e-mail: luis@iec.csic.es, victor.gayoso@iec.csic.es).

DOI: 10.7763/IJAPM.2013.V3.206

means of quality control.

To apply quality control we use the R software, which is one of the most flexible and powerful statistical package that currently exists to perform statistical tasks of all kinds, from the most elementary to the most advanced. It has been improved from previous versions and it is maintained by some of the most prestigious statesmen. It's free and open source software that can be download and installs easily from the web page http://www.r-project.org/.

In general, speaking about climate data, there are two sources or types of errors that produces inhomogeneous data:

- Those which are derived from the non-stationary observed system, which results because the atmosphere has no cycles in almost all scales.
- The system derived from observer, modified by the environment, the measuring instrument, its location, the person using it, and so on (easier to detect).

At first glance, we notice that the data from Navamuño observatory was the most complete (in the period 1995-2008), but March, 2004 temperature data was lacking. To obtain the missing data we compare the data from both observatories (La Hoya and Navamuño) because they are well related and we used a regression to obtain the lost data. Fig. 3 shows both data: from Navamuño and from La Hoya. It also shows the calculated new data for Navamuño from La Hoya. We used this data obtained from La Hoya, for the lack period of data in Navamuño.

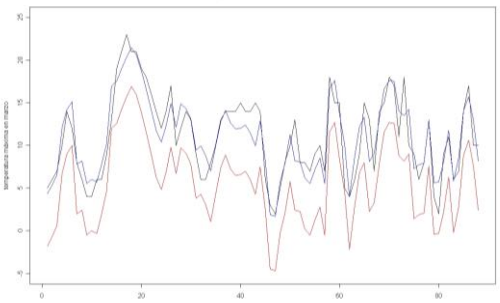

Fig. 3. Temperature data from Navamuño in black colour, La Hoya in red and predicted in blue.

Once we get all data for the time period from 1995 to 2008, we started the process of getting good data. The spatial and time quality of data is a prerequisite in studies of climate variability, particularly when analyzing large data. In studies of climatic extremes daily data must be used. These time series can show errors due to randomness and other discontinuities. For example:

- The precipitation is less than zero.
- The maximum temperature is less than the minimum one for a given day.

We have passed two tests to the data provided by the AEMET. First of all by using ClimDex, in order to verify that any of mentioned errors are presented. Then, we have done the control test of homogeneity (Htest) to determine if there is a change in the tendency of the series. Both software packages (ClimDex and Htest) are based on statistical R package.

A. *ClimDex*

All the information related to Climdex can be found at the ClimDex User's Guide [4]. ClimDex is a program designed to assist researchers in the analysis of climate change and detection. More specifically, ClimDex carries out a four-step analysis process which consists of the following:

1) Quality Control
2) Homogeneity Testing
3) Calculate Indices
4) Region Analysis

Applying the test to the data, we have identified and we have corrected errors and missing values such as the case of maximum temperature lower than the minimum, as is shown in Fig. 4, where the maximum temperature (in red) is minor than the minimum temperature (in black) for some days, negative precipitation, and outliers (if they are not a real value). The test also allows us to calculate climatic indices.

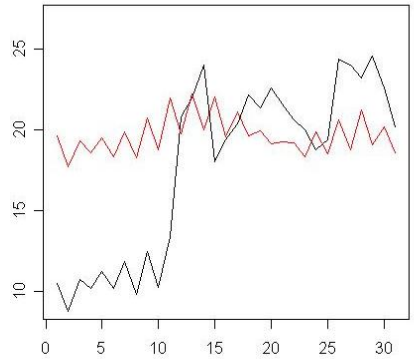

Fig. 4. Magnetization as a function of applied field.

B. *HTest*

All the information related to RHTest can be found at the RHtestsV3 User Manual [5]. RHtest software package detects multiple step change points that might exist in data series, and it can be used as a reference set of homogenous time series well correlated with a base series. However RHtestV3 version detects these changes in tendency even when there is no reference series, as is our case. As a result, the shape of the distribution is often adjusted to the adjustment to the mean. It is possible that winter and summer temperatures are adjusted differently because they belong to the lower and upper quartiles of the distribution, respectively.

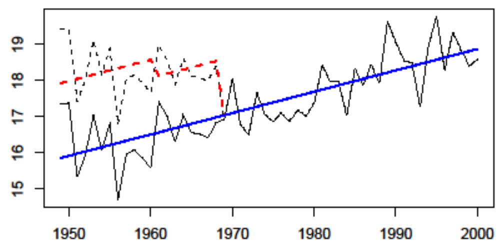

Fig. 5. Changes in tendency detected by RHTest.

When there are this changes in tendency like we can see in the Fig. 5 (red dashed line), it is detected by the script which change this tendency, if it is not identified like one real change and seems to be an error. RHtest considers that the real ones are the latest and adjust the tendency to these ones (blue line).

When we run the package RHtestV3 with our data we obtained the Fig. 6 for the minimum temperature, and something similar for the case of maximum temperature and temperature difference. It can be appreciate that in recent years there is no change in tendency, so that we can consider the data set homogeneous.

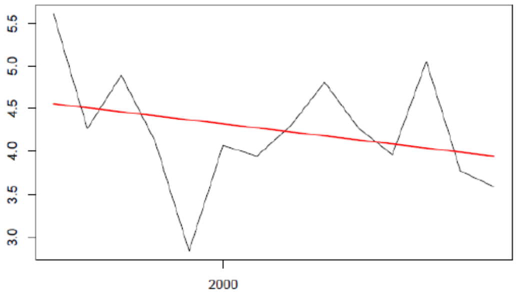

Fig. 6. Annual average minimum temperature tendency.

III. Climate Indices

An index is a set of parameters or indicators that describe a situation. The indices described here are climate indices, i.e., they tell us how the climate evolves in the studied area. A table of indices according to their type is shown in Table I.

Monthly indices are calculated if there exist no more than 3 days missing per month, while the annual values are calculated in a year if not more than 15 days are missing. No annual value will be calculated if one month of data is missing. For indexes with a threshold value, the index is calculated if at least 70% of data are located. For indicators of seasons just as long hot periods or cold periods, a season can continue into the following year.

TABLE I: Climate Indices

Related to frost	FDO, IDO
Related to Temperature	SU25, GSL, TXx, TNx, TXn, TNn, Tn10p, Tx10p, Tn90p, Tx90p, WSDI, CSDI, DTR.
Related to rainfall and humidity	RX1day, RX5day, SDII, R99p, R95p, CDD, PRCPTOT, R10, R20.

In the indices graphs (Fig. 7, Fig. 8, and Fig. 9), there are two lines: a continuous one that is closest to the indices values (linear regression) and also called tendency line and a dashed line which is the polynomial closest to the values (polynomial regression).

A. Indices Related to Frost

Frost is a climatic phenomenon that occurs when the temperature of air below the freezing point of water (0 °C) and makes the water inside the air freezes into ice deposited on surfaces.

Frost occurs when the relative humidity exceeds 60%, otherwise there will not be enough water in the atmosphere to be deposited on surfaces.

It is also necessary that the wind is not strong; otherwise the water may not be deposited.

There are two indices related to frost: FDO and ID0. FDO is the number of days when the daily minimum temperature is less than 0 ℃. Fig. 7 shows the FDO index of the data acquire from AEMET. In this case the number of days when minimum temperature is less than 0 ℃ has been increasing, with an oscillation of almost biennial. The year that there were fewer days with this temperature was 1995 with 50 days and the year in which the index is higher was 1999, with a total of 115 days.

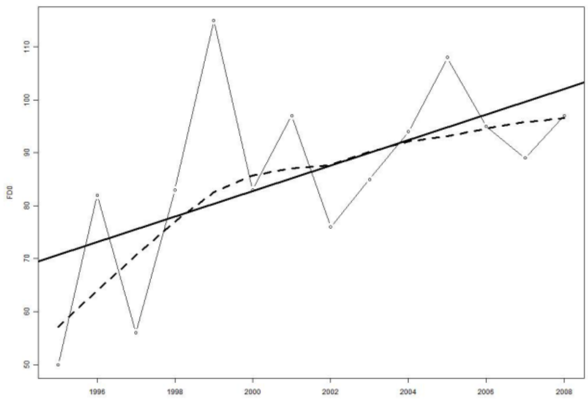
Fig. 7. FDO index representation.

The ID0 index calculates the number of icing days or frost days, when the daily maximum temperature is less than 0 ℃. The ID0 values from the data of this study range between 0 ℃ and 5 ℃. Maximum occurs in the years 2005 and 2008 with 5 days. In 1995, 1997, 1999, 2001 and 2002 there were no days with maximum temperature below 0 ℃.

B. Indices Related to Temperature

Some of the temperature indices studied are the following:

1) SU25 (summer days): This index gives the number of days for which the maximum temperature is higher than 25 ℃. In this study, during the linear regression or tendency line, the number of days is increasing over the years, being 1997 the year with fewer days (37), and the year with the highest number of days with maximum temperature greater than 25 ℃, was 2006 with 86. Before 2006 there was a clear tendency to increase, and from 2007 to 2008 there is another increase.
2) GSL (Grow Season Length): This index related to the length of the growing season, calculates the number of times in which during a period of lower temperature (winter area) there exists at least six days with temperatures above 5 ℃, and during the days of higher temperatures (summer time) there exists at least six days with temperature below 5 ℃. Although the tendency line has negative slope, in recent years, the polynomial regression line rises so the number of days could rise.
3) Tn10p (Cool nights): Percentage of days when TN is less than 10th percentile. The unit used to measure this index is %. The tendency line is rising, so we conclude that the minimum temperatures over the years are lower, thereby increasing the cold nights. The highest value occurs in 2008 with a value above the rest, just over 12 days of cold nights and the lowest occurs in 1996 with only 2 cold nights. The remaining years get values range between 2.03 and 6.95.
4) WSDI: The Warm Spell Duration Index is defined as annual number of days with at least 6 consecutive days when the average of maximum temperatures is greater the 90th percentile. We notice that half of the data related to Béjar area during the reference period, especially in the first part, there was no warm periods, although in the last 4 years there is a tendency to increase. In general the tendency line is clearly upward, thus, we conclude that there are more days with temperatures above the 90th

percentile, so that the temperature increases.

5) DTR: Diurnal Temperature Range is defined as monthly mean difference between maximum temperature TX and minimum temperature TN. The index value for January can be written as

$$DTR_{January} = \frac{\sum_{i=1}^{31}\left(T_{xi} - T_{ni}\right)}{31},$$

where T_{xi} is the maximum temperature the ith day of January and T_{ni} is the minimum temperature the ith day of the same month. A comparison of observed model changes in DTR over the last 50 years show much less reduction in the index in the model simulations due to greater warming of maximum temperatures in the models observed [6]. The growing tendency line indicates that the maximum temperatures are higher, and the minimum are lower (see Fig. 8). The year with a greatest difference of temperatures was 1999, with 13 °C, and the year with the smaller difference was 1996, with 10 °C.

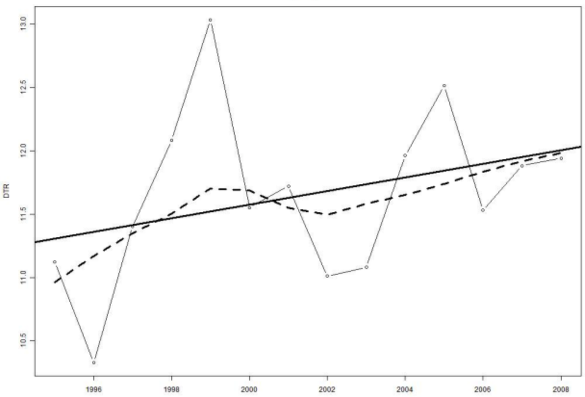

Fig. 8. DTR index representation.

C. Climate Indices of Precipitation and Humidity

We have study the next indices:

1) Rx1day: The Max-1day precipitation amount represents the monthly maximum 1 day precipitation. Although we found 2 quite different years: 1995 (Rx1=96.7mm) and 1999 (Rx1=36,9mm), the maximum precipitation is small and close to 65mm.

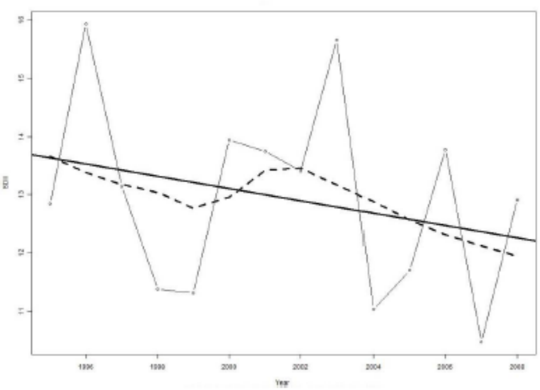

Fig. 9. SDII index representation.

2) SDII: The Simple Daily Intensity Index calculates the total annual precipitation divided by the number of humid days in 1 year. The unit used to measure this index is mm/day. As we can see (Fig. 9), the tendency is negative, and the precipitation is less intense nowadays.

IV. Conclusion

We carried out a study of extreme climatic variables of temperature, mean temperature and accumulated precipitation in Béjar (Salamanca) environment using the statistical package R. In particular, the starting data belong to the meteorological stations at Navamuño and La Hoya from 1987 to 2008.

First of all, we have represented the series and studied if they are complete. As the data were incomplete, we decided to study only the data from the Navamuño observatory in the period 1995-2008 because they were the most complete as the March temperature in 2004 was missing. Since we have that month from the La Hoya observatory and we found that there is a good correlation between data from the two stations, we used linear regression to complete the series of Navamuño, considering as independent variable La Hoya data.

We used two tests implemented in the R software to make the data homogeneous: ClimDex, that determines if there are errors such as the maximum temperature is less than the minimum in a day or accumulative rainfall is negative, and Htest, which checks changes in trends in the series.

We reached the following conclusions from the data:

- The warmest year was 1995 with 11.1 °C and the coldest one 1999 with 9.3 °C. The rainiest year was 1996 with 1735mm. and the least rainy was 2005 with 848.9mm.
- The months of January, June, July and September tend to increase their average maximum temperature in coming years as well as summer and autumn seasons.
- During June, July and September, the average temperatures are falling. Similarly, in the spring and winter seasons.
- The months of February, March, October, November and December the minimum temperature decreases.
- The precipitation tends to increase during spring and summer seasons as well as during February, March, April, June, and October. However it tends to decrease in winter and in January, as well as in July, August, September, November and December.
- The precipitation is less intense nowadays.
- The maximum temperatures are higher, and the minimum are lower

References

[1] C. Rodriguez-Puebla, A. H. Encinas, L. A. García-Casado, and S. Nieto, "Trends in warm days and cold nights over the Iberian Peninsula: relationships to large-scale variables," *Climatic Change*, vol. 100, pp. 667-684, 2010.

[2] J. Abaurrea *et. al.*, "Metodología para el control de calidad y homogeneidad de una base de datos de precipitación diaria. El clima entre el mar y la montaña," *Publicaciones de la AEC*, Serie A, no. 4, 2004.

[3] M. Staudt *et. al.*, "Homogenization of long-term monthly Spanish temperature data," *International Journal of Climatology*, vol. 27, pp. 1809-1823, 2007.

[4] *ClimDex User's Guide, v.1.3*. (June, 2001). [Online]. Available: http://www.cccma.seos.uvic.ca/ETCCDI/ClimDex/climdex-v1-3-users-guide.pdf.

[5] X. L. Wang and Y. Feng. (June, 2010). *RHtestsV3 User Manual*. [Online]. Available: http://www.cccma.seos.uvic.ca/ETCCDI/RHtest/RHtestsV3_UserManual.doc.

[6] K. Braganza, D. J. Karoly, and J. M. Arblaster, "Diurnal temperature range as an index of global climate change during the twentieth century," *Geophys. Res. Lett*., vol. 31, L13217, 2004.

A. Hernández Encinas was born in 1959 in Navalmoral (Spain), she was Graduate in Physics at the University of Salamanca (Spain) in 1982, and Ph.D. in Physics at the same university in 1990.

She is a researcher in Climatology, Cryptography (secret-key cryptosystems; public-key cryptography; secret sharing schemes; hash functions and cryptographic protocols), and cellular automata (reversibility; forest fire modelling; epidemic modeling; cryptographic applications). She has supervised 1 doctoral thesis. She has participated in 19 research grants. She is author of 14 papers in international journals, and she has more than 45 contributions to international and national conferences. She is an associate professor from 2002 at the Department of Applied Mathematics at the University of Salamanca.

A. Queiruga Dios was born in Pasajes (north of Spain) in 1969, she was Graduate in Physics at the University of Salamanca (Spain) in 1998, and obtained his Ph.D. in Mathematics at the same university, in 2006. Her major field of study is public key cryptography, together with educational tools and mathematical applications for engineering students.

She is a lecturer in Mathematics at the Department of Applied Mathematics at the College of Industrial Engineering in Béjar (Spain). She has participated in 16 research projects. She is author of 17 papers, and 40 contributions to workshops and conferences. Her current research interests include cryptography (secret, public and protocols), image processing, number theory, climatology and cellular automata, and the applications of mathematics to education.

L. Hernández Encinas was born in 1957 in Navalmoral (Spain), he obtained his Graduate in Mathematics at the University of Salamanca (Spain) in 1980, and his Ph.D. in Mathematics at the same university in 1992.

He is a researcher at the Information Security Institute (ISI), Spanish National Research Council (CSIC) in Madrid (Spain). He has participated in 33 research projects. He is author of 9 books, 6 patents, more than 120 papers, more than 100 contributions to workshops and conferences and he has supervised 3 doctoral thesis. His current research interests include cryptography (secret, public and protocols), algebraic curve cryptosystems, image processing, number theory, climatology and cellular automata.
Dr. Hernández Encinas belongs to the International Association for Cryptologic Research (IACR).

V. Gayoso Martínez was born in 1975 in Madrid (Spain). He obtained his Degree in Telecommunication Engineering from the Polytechnic University of Madrid in 2002, and his Ph.D. in Telecommunication Engineering at the same university in 2010. Since 1998, he has been working in topics related to smart cards, Java technology, and public key cryptography in both private companies and academic institutions.

He is author of 5 papers in international journals and 13 contributions to international and national conferences. He is currently working at the Spanish National Research Council (CSIC).