Severe precipitation evaluation in Belo Horizonte - Brazil: Data mining approach

Heloisa Musetti Ruivo[a1], Haroldo F de Campos Velho[b], Saulo R. Freitas[c]

[a]Pan-American Association of Computational Interdisciplinary Sciences (PACIS)
[b]National Institute for Space Research (INPE), São José dos Campos, SP, Brazil
[c]National Aeronautics and Space Administration (NASA), Washington DC, USA

Abstract

Severe rains punish the state of Minas Gerais (Brazil) leaving several cities in emergency from October 2011 to January 2012. With the aim to point out the relevant climatological variables in the occurrence of this event, DM approaches is applied. Statistical analysis is combined with an artificial intelligence technique to identify the most relevant meteorological variables for the local severe rainfall. The p-value statistical technique is employed to select a much smaller subset of climatic variables, preserving the information associated with extreme meteorological events. A decision tree algorithm is used as a model to identify the precipitation severity. The method is tested with the event that occurred early in January 2012. Our results show a good local analysis for extreme precipitation episodes.

Keywords: Severe precipitation, statistical p-value analysis, decision tree algorithm.

1. Introduction

In early 2012, heavy rains battered the central region of the state of Minas Gerais leaving several cities in emergency situation. There was overflowing rivers in several municipalities. Thus, several cities were flooded decreeing emergency situation. According to the CPTEC climanálise bulletin, in Belo Horizonte, capital of Minas Gerais state, the accumulated rainfall in the first ten days of January reached 270 mm, close to 274 mm expected for the entire month. This was considered the third biggest flood in history. Figure 1 shows the bordering neighborhoods that have been hit by the flooding of the River Doce leaving 490 people affected, totaling 126 families, 77 of them distributed in shelters in the city.

[1]E-mail Corresponding Author: helomusettir@gmail.com

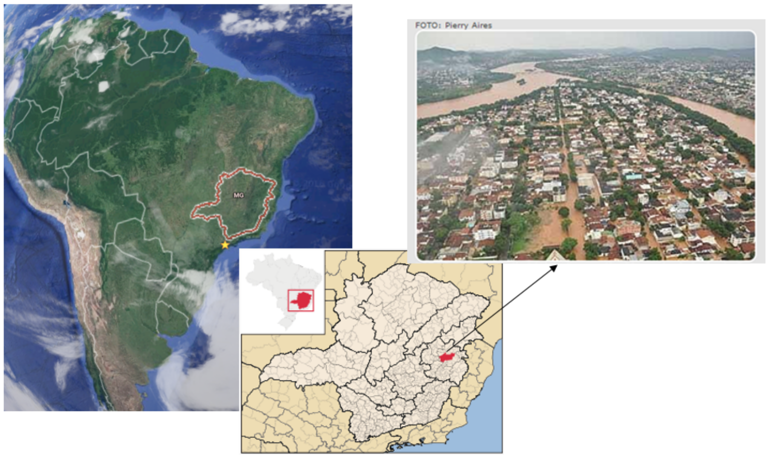

Figure 1 - View of Governador Valadares, and River Doce, located in the state of Minas Gerais, BR.

Heavy rainfall phenomena act over a very short period, and the breadth of heavy rainfall is very short. Thus, the action of the public authority is more critical and necessary in extreme rainfall events, in the sense of mobilizing resources and support actions to the population, requiring almost immediate decision/action (almost "real time") from the decision makers (civil defence). Tragedies caused by lack or excess rain were analyzed in previous studies [1, 2, 3] where Data Mining (DM) methodologies were employed. The current paper employs the same methodology used in previous research, with focus on data science application in extreme weather events. DM methodologies are very suitable to be applied to the large amount of data available to scientists, including climatologists. DM is one step in the more general process of Knowledge Discovery in Databases (KDD), extracting information and transform it into an understandable structure for further use, in order to facilitate a better interpretation of existing data [4].

Our DM approach comprises two steps of knowledge extraction, both of them using statistical tools. The first step uses statistical analysis based on p-value computation with two goals: identify which climatic variables behave differently across pre-defined classes of precipitation intensity, and secondly the complexity reduction of the original dataset [1, 2, 3]. With the analysis of p-values the authors performs a map for each meteorological variable. Therefore, a p-value map associated to different variables can show where a certain meteorological value has a stronger link to the event. The second step consists to design of a Decision Tree (DT).The most influential attributes are used as a predictive DT model. The DT allows for producing a hierarchical structure of importance among the different attributes. Indeed, the DT can be employed as a tool to map the relevance of the attributes associate to a specific event. In the present context, the final result identifies a small subset of climatological variables that may explain or even forecast the extreme event.

2. Metodology of Data Mining

Data Mining is the computational process of discovering patterns in large data sets involving methods such as artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Fayyad and co-authors [5] assert that DM is a step in the KDD process. The KDD consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns over the data. Here, a DM approach is employed to comprises two steps of knowledge extraction: class-comparison, and decision trees. These methods are applied successively to reduce the complexity of the original dataset and identify a much smaller subset of climatic variables that may explain the event being studied.

The class-comparison method uses a statistical analysis for comparing two or more pre-defined classes in a time series of climatic grid box values. The objective is to determine which variables in our data set behave differently across pre-defined classes of precipitation. There are several methods for checking whether differences in variable values are statistically significant. Here it is used a t-test that is converted into probability, known as p-value [6]. The climatological attributes with smaller p-value are used to construct a decision tree algorithm.

The DT algorithm consists of a collection of training cases, each having a tuple of values for a fixed set of attributes (independent variables) and a class attribute (dependent variable). The aim is to generate a map that relates an attribute value to a given class. There are several decision tree (DT) algorithms available. Here we used the J4.8, a Java implementation of the C4.5 algorithm, from the WEKA package [7]. The J4.8 algorithm relies on a partition heuristic that maximizes the information gain ratio, the amount of information generated by testing a specific attribute. This approach allows to identify attributes with the greatest discrimination power among classes, and select those that will generate a tree that is both simple and efficient [8, 9].

3. Results

Class-comparisson and DT methods are applied to reduce the complexity of the original dataset and identify a much smaller subset of climatic variables that may explain the event being studied. The results will be presented in the form of p-value maps for each climatological variable. This means that the p-values at a given grid point can be interpreted as the probability that the observed difference between classes for this variable is the product of mere chance. Clearly, coherent patterns of low p-values (darker areas) attract attention. The methodology is employed to the study of the extreme precipitation occurred in Belo Horizonte, Brazil in early 2012.

The entire data set used in this study comprises 32,674 time series. Gridded data cover a region delimited by latitudes 15^oS and 30^oS, and longitudes 50^oW and 35^oW. Pentad-averaged anomalies were used in the analysis. Anomalies were computed relative to the mean values over the period 2000-2013 (14 years). Surface- and pressure-level atmospheric fields have a spatial resolution of $0.5^o \times 0.5^o$ taken to 12 UTC – data extracted from the ECMWF climate reanalysis [10]. ECMWF uses its forecast models and data assimilation systems to reanalyze archived observations, creating global data sets describing the recent history of the atmosphere, land surface, and oceans. The climatological variables of surface used in the analysis were: sea surface temperature, and mean seal level pressure. The other variables at the pressure levels 925, 850, 500, 300 hPa were: Cloud cover, Geopoential, Specific Humidity, Air Temperature, Zonal wind, Meridional wind, and Omega (vertical wind component).

The precipitation series was extracted from INMet site, where the average was computed over 4 stations located over the affected region. For classification purposes, the pentads of the precipitation time series is divided into classes of precipitation intensity: strong, moderate, and light rainfall. The standard t-test was applied, as recommended for applications with two classes: strong (precipitation greater than 8 mm), and moderate (precipitation between 0 and 8 mm). The red dot point in the figures below indicate the location of the INMet stations.

Figure 2 shows a dense dark area of low p-values for air temperature at 925 and 500 hPa coming from the north area. Low p-value area of Specific Humidity at 500 hPa and 300 hPa are displayed in Figure 3 which covers an area involving the Alantic Ocean and the region under study. It can be seen in Figure 4 a dense dark area for cloud cover to medium altitudes over the region. It is also observed a low p-value area for omega at 925 hPa in he affected area and at 500 hPa on the ocean that spreads to the continent (Figure 5). The isolines in Figure 5 represents the value of Omega during the period (pentad) of the drought (earlier 2012). During the extreme rainfall episode, omega values are negative over the most affected region (red dot) – see the isolines. It is well known that upward vertical motion over the continent are linked to precipitation, under certain conditions (moisture, pressure field, etc). Figure 6 shows a dense area of p-values lower in the region above the affected area. The arrows represent the resulting wind in the pentad of the related event date. The arrows indicate the wind coming from the northwest direction.

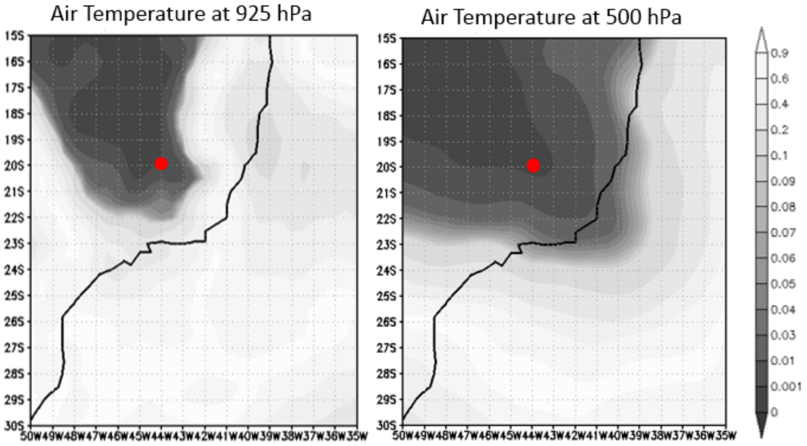
Air Temperature at 925 hPa
Air Temperature at 500 hPa
15S
16S
17S
18S
19S
20S
21S
22S
23S
24S
25S
26S
27S
28S
29S
30S
50W 49W 48W 47W 46W 45W 44W 43W 42W 41W 40W 39W 38W 37W 36W 35W
0.9
0.6
0.4
0.2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.01
0.001
0

Figure 2 - P-VALUES field for air temperature at 925 and 500 hPa in Minas Gerais, BR.

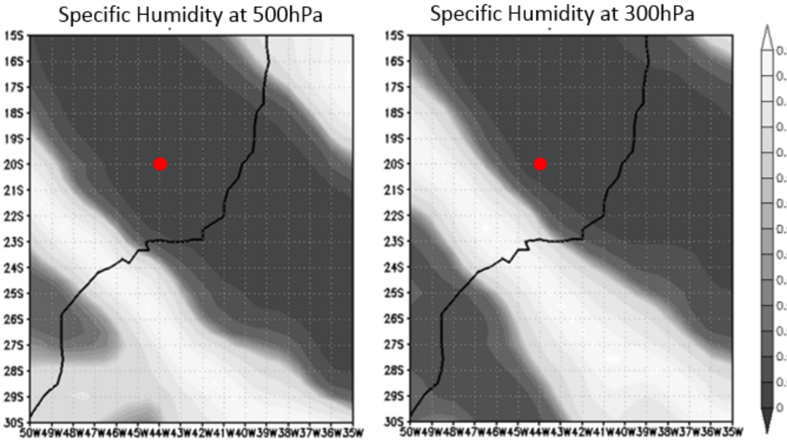
Specific Humidity at 500hPa
Specific Humidity at 300hPa
15S
16S
17S
18S
19S
20S
21S
22S
23S
24S
25S
26S
27S
28S
29S
30S
50W 49W 48W 47W 46W 45W 44W 43W 42W 41W 40W 39W 38W 37W 36W 35W
0.9
0.6
0.4
0.2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.01
0.001
0

Figure 3 - P-VALUES field for Specific humidity at 500 and 300 hPa in Minas Gerais, BR.

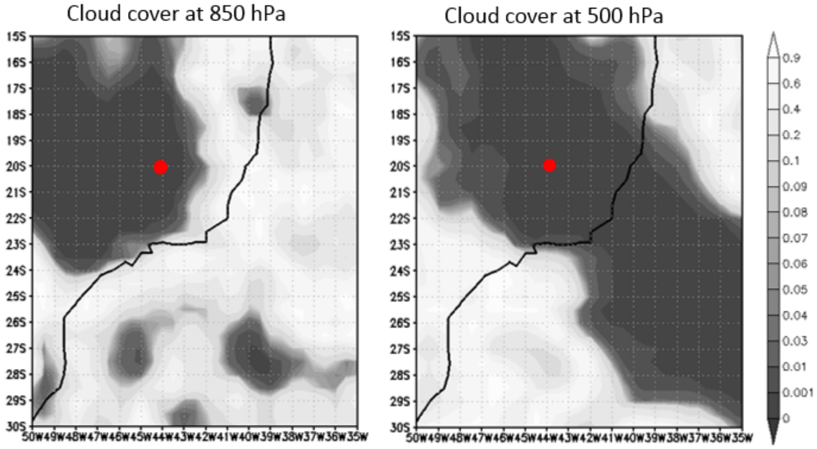

Figure 4 - P-VALUES field for Cloud cover at 850 and 500 hPa in Minas Gerais, BR.

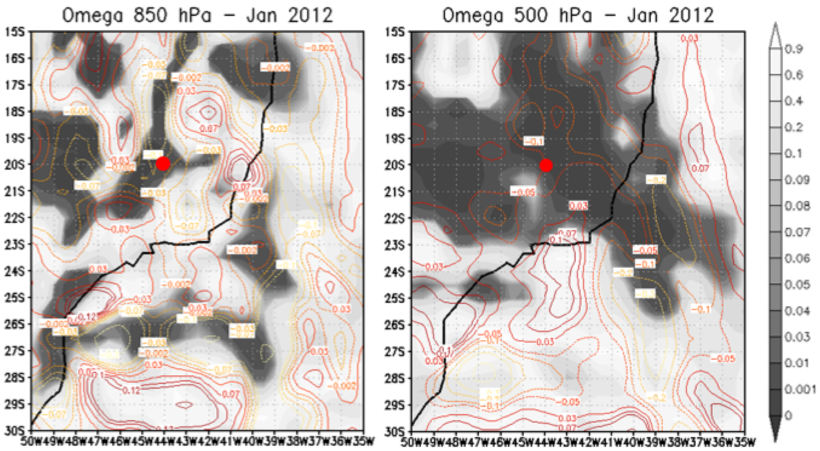

Figure 5 - P-VALUES field for Omega at 850 and 500 hPa in Minas Gerais, BR.

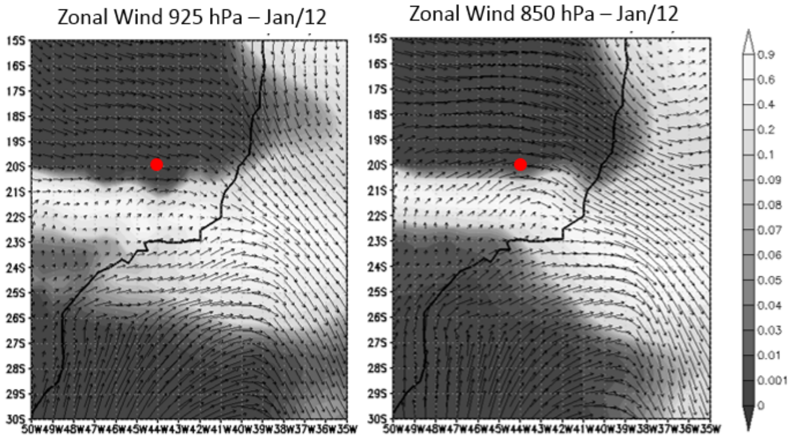
Zonal Wind 925 hPa – Jan/12
Zonal Wind 850 hPa – Jan/12
15S
16S
17S
18S
19S
20S
21S
22S
23S
24S
25S
26S
27S
28S
29S
30S
50W 49W 48W 47W 46W 45W 44W 43W 42W 41W 40W 39W 38W 37W 36W 35W
0.9
0.6
0.4
0.2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.01
0.001
0

Figure 6 - P-VALUES field for Zonal Wind at 925 and 850 hPa in Minas Gerais, BR.

The decision tree with the J4.8 algorithm was created with confidence factor used for pruning (0.25), and number of instances per leaf (2). The DT configuration was computed with the 24 different climatological variables, considering 5 different coordinates for each variable, with smallest p-values, performing 120 climatological attributes. For the DT classifier, the precipitation time series were divided in two classes: light (values below 5), and strong (values above 5). According to our result, the most influential feature is the Omega – the feature appears in the first and third level of the DT.

Omega at 850 hPa in the coordinate 47.5 W and 16 S is reveled as the key feature. Specific humidity at 500 hPa on the coordinates (46 W,16.55 S) and (46 W,18 S) are features in the second level of influence. At third level in the DT, Omega values at 300 hPa, 500 hPa, and 850 hPa already are able to define the precipitation status. If the DT classifier does not indicate the precipitation class, other features are employed to determine the precipitation condition, following the DT procedures.

The training set comprised data from the year 2000 up to 2002 and from 2004 up to 2009 (in earlier 2003 there was an event of rainfall, then this year was extracted from the training set). The years 2003, and from 2010 up to 2013 were used to evaluate the DT performance. As a predictor, the DT (Figure 7) was able to identify the extreme rainfall occurred in earlier 2012 January as well as the 2003 rainfall.

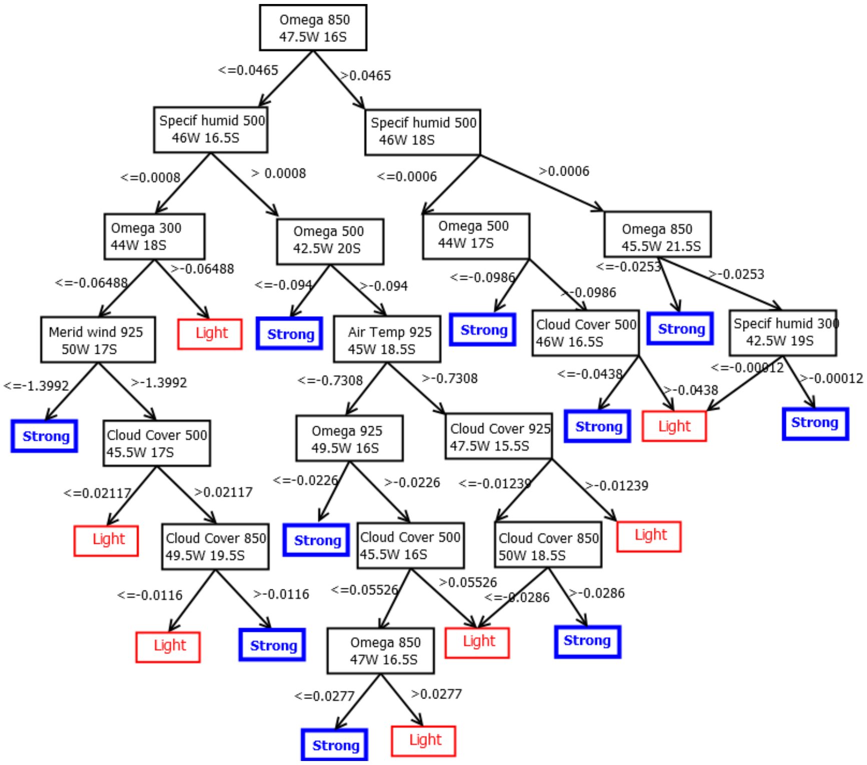

Figure 7 - DT using training set from 2000 up to 2002 and 2004 up to 2009, and test set: 2003 and from 2010 up to 2013.

4. Conclusion

The extreme rainfall event occurred in Minas Gerais States (Brazil) in the year 2012 was analyzed using two techniques of data mining. The aim was to investigate the climatic variables linked to severe precipitation in that region. Among the two techniques, the class-comparison methodology was able to identify the most relevant variables, with significant reduction of the size of the original data set. The p-values maps become an easier interpretation way for specialists, pointing out climatological variable directly related to the extreme event. By complementing the study, the decision tree trained with the reduced dataset was able to correctly classify the case of extreme rainfall in 2012. Overall, the applied data mining procedure has shown to be a promising approach in the investigation of climatic extreme event and the extraction of knowledge from large and complex data sets.

Acknowledgments.Authors thank the Brazilian Agencies for research sup-port (CNPq, FAPESP, CAPES).

References

[1] Ruivo, H. M., Sampaio, G., Ramos, F. M. (2014). Knowledge extraction from large climatological data sets using a genome-wide analysis approach: application to the 2005 and 2010 amazon droughts. Climatic Change, pp. 115.

[2] Ruivo, H. M., Campos Velho, H. F., Sampaio, G., Ramos, F. M., (2015) Analysis of Extreme Precipitation Events Using a Novel Data Mining Approach.AJEE 5(1A) pp.

[3] Ruivo, H. M., Campos Velho, H. F, Ramos, F. M., Sampio G. P-value and decision tree for analysis of extreme rainfall. Ciência e Natura, v. 1, p. 210-213, 2013.

[4] Fayyad U., Piatetsky-Shapiro, G, Smyth, P., Uthurusamy, R., 1996. Advances in Knowledge Discovery and Data Mining. California The MIT Press pp 560.

[5] Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996). ”From Data Mining to Knowledge Discovery in Databases” (PDF). Retrieved 17 August 2016

[6] Simon R. M. et al., 2003. Design and analysis of DNA microarray investigations. Springer Vol 209.

[7] Witten I. H., Frank, E. S., 2000. Data mining: Practical machine learning tools and techniques with java implementation. Morgan Kaufmann Publishers.

[8] Quinlan, J. R. (1993). C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco.

[9] Shannon, C .E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, vol. 27(4), 623666.

[10] Dee, D. P., Uppala, S. M., Simmons, A. J., Ber-risford, P., Poli, P., Kobayashi, S., Andrae, U. et al. The ERA-Interim reanalysis: configuration and perfor-mance of the data assimilation system Version of Record online: 28 APR 2011, DOI: 10.1002/qj.828