DOI: 10.5902/2179460X11604
Revista do Centro de Ciências Naturais e Exatas - UFSM
Revista Ciência e Natura, Santa Maria
EISSN: 2179-460X, Edição Esp. Dez. 2013, p. 231 - 234

# P-VALUE AND DECISION TREE FOR ANALYSIS OF EXTREME RAINFALL

Heloisa M. Ruivo, Haroldo F. de Campos Velho, Fernando M. Ramos, Gilvan Sampaio

Instituto Nacional de Pesquisas Espaciais (INPE), Brazil.

heloisa.ruivo@lac.inpe.br, haroldo@lac.inpe.br, fernando@lac.inpe.br, gilvan.sampaio@inpe.br

**Abstract:** Severe weather is a challenge issue, with impact on human life and on the economy. Data mining techniques can be employed for analysis of extreme events. The p-value statistical method and decision tree are applied to study of the extreme rainfall event during 2008 in the Santa Catarina state (Brazil).

**Resumo:** Tempo severo é um grande desafio, que tem forte impacto na vida dos cidadãos e na economia. Técnicas de mineração de dados podem ser empregadas para análise de eventos extremos. O método estatístico do valor-p e árvores de decisão são aplicados ao estudo de chuva intensa no estado de Santa Catarina (Brasil) em 2008.

## INTRODUCTION

Two relevant issues are the extreme meteorological events and the amount of data available today. The first issue is due to the impact of such events on the society. The second one is a scientific challenge for developing techniques to deal with giant amount of data – this issue is also called the *Data Science*.

Data mining techniques are employed here to analyze an extreme rainfall event at November/2008 in Santa Catarina state, Brazil (Coelho et al., 2012). The goal is to identify the relevant climatological factors linked of such event. Two methods are applied: the p-value statistical technique, and decision trees. The statistical method is implemented in the computational tool called BRB-ArrayTools, used in cancer research, but adapted to environmental problems. Decision tree algorithms (implemented in the WEKA software package) are used to automatically generate classifier tool. The methodology described here is absolutely general, and it can be easily applied for other extreme events and/or other regions (Ruivo, 2012).

DOI: 10.5902/2179460X11604
Revista do Centro de Ciências Naturais e Exatas - UFSM
Revista Ciência e Natura, Santa Maria
EISSN: 2179-460X, Edição Esp. Dez. 2013, p. 231 - 234

## THE P-VALUE STATISTICAL TECHNIQUE

The p-value method is associated with the hypothesis tests. In general, two basic hypotheses are formulated: null hypothesis (H0), and alternative hypothesis (H1). The null hypothesis is to be tested. If the H0 is not approved, the H1 hypothesis is considered accepted. The method was developed to deal with non-Gaussian statistics.

In the cancer research, the functional genetics searches the connection between a certain gene and a cancer feature. For testing if a given gene is linked with a cancer disease, the researchers have used the p-value testing. The package BRB-ArrayTools, version 3.7.0, was developed by the Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute, USA. This is a free software (see: http://linus.nci.nih.gov/brb/download.html).

## DECISION TREES

Decision trees (DT) is one kind of learning machine algorithm (Witten and Frank, 2005). A fraction of the data set is used to configure the classifier: the training phase. A DT is a method consisting of nodes, or *leafs*, used for testing attributes. Each possible output from a simple test (node) corresponds to follow on a branch of the tree, which leads to another node, meaning another test, and so on. Most of DTs are automatically configured by using the quantity of information. The Shanon entropy is used to evaluate the amount of information. The WEKA package was used, and it available in the internet (http://www.cs.waikato.ac.nz/~ml/weka/).

## ANALYSIS FOR EXTREME RAINFALL USING DATA MINING TOOLS

The meteorological data is from the NCEP/NCAR re-analysis (Kalnay et al., 1996). The data set embraces the period January-1999 up to December-2010. Meteorological data used: sea surface temperature (C), pressure at 1000 hPa, air surface temperature (C), specific moisture (g/kg) at levels 850 and 1000 hPa, Omega (Pa/s) at levels 100, 200, 300, 400, 500, 600, 700, 850, 1000 hPa, geopontential height (m), horizontal wind ($ms^{-1}$) at levels 200, 500, 850, 1000 hPa, cloud covering (%). The region considered is inside of the box 30W- 60W × 20S-50S.

The precipitation series is obtained considering averages for each 5 records. The

DOI: 10.5902/2179460X11604
Revista do Centro de Ciências Naturais e Exatas - UFSM
Revista Ciência e Natura, Santa Maria
EISSN: 2179-460X, Edição Esp. Dez. 2013, p. 231 - 234

accumulated precipitation series for the Itajaí river valley is shown in Figure 1.

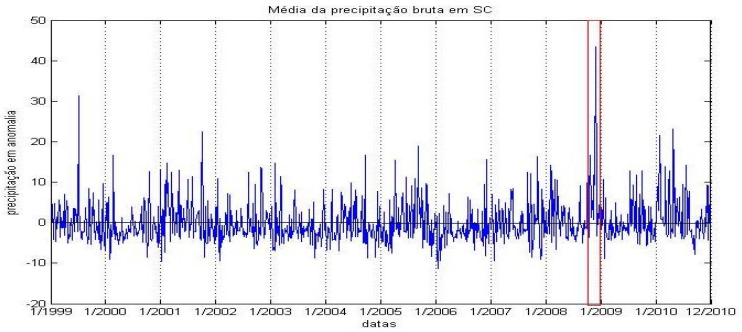
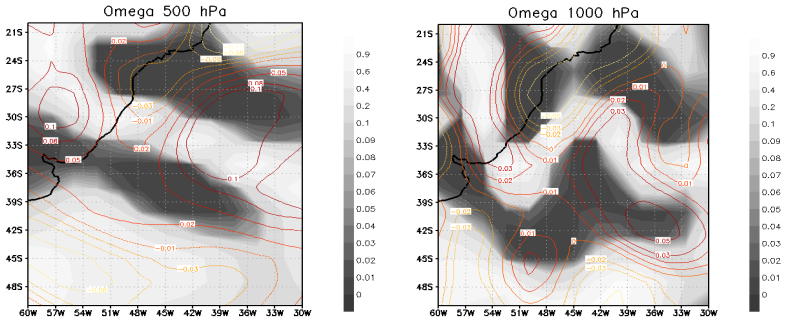



**Figure 1:** Time series for accumulated precipitation: Itajaí river.

Figure 2 displays the computed p-value for the Omega variable over the map. It is clear the cell with positive and negative values for Omega.




(a) (b)

**Figure 2:** p-value field for upward motion anomaly at 500 and 1000 hPa.

Figure 3 shows the DT obtained using J4.8 algorithm (WEKA): 50 meteorological variables with the lowest p-value.

DOI: 10.5902/2179460X11604
Revista do Centro de Ciências Naturais e Exatas - UFSM
Revista Ciência e Natura, Santa Maria
EISSN: 2179-460X, Edição Esp. Dez. 2013, p. 231 - 234

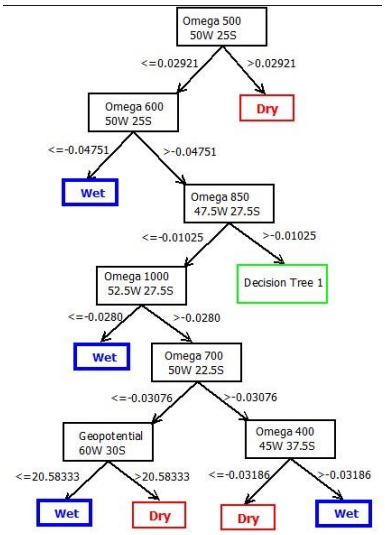



**Figure 3:** DT obtained computing 50 meteorological variables with the lowest p-value.

## REFERENCES

COELHO, C. A. S. et al., Climate diagnostics of three major drought events in the amazon and illustrations of their seasonal precipitation predictions. **Meteorol. Appl.**,vol. 19, 237-255, 2012.

RUIVO, H. M. M**etodologias de Mineração de Dados em Análise Climática**. Doutorado em Computação Aplicada, Instituto Nacional de Pesquisas Espaciais, São José dos Campos (SP), Brasil, 2012.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. Morgan Kaufmann Publishers, 2005.

KALNAY, E.; KANAMITSU, M.; KISTLER, R.; COLLINS, W.; DEAVEN, D. The NCEP/NCAR 40-year reanalysis project. **Bull. Amer. Meteor. Soc.**, vol. 77, 437-470, 1996.