XXXVIII CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA
2 a 6 de agosto 2009
Juazeiro (BA)/Petrolina (PE)

TÉCNICAS DE MINERAÇÃO DE DADOS PARA DEFINIÇÃO DE ZONAS PLUVIOMETRICAMENTE HOMOGÊNEAS PARA O ESTADO DO RIO GRANDE DO SUL

RAQUEL STUCCHI BOSCHI[1]; STANLEY ROBSON DE MEDEIROS OLIVEIRA[2].

[1]Eng. Agrônoma, mestranda da Faculdade de Engenharia Agrícola, UNICAMP, Campinas - SP – raboschi@yahoo.com.br.
[2]Bacharel em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Campinas - SP – stanley@cnptia.embrapa.br.

Escrito para apresentação no
XXXVIII Congresso Brasileiro de Engenharia Agrícola
2 a 6 de agosto de 2009 - Juazeiro-BA/Petrolina-PE

RESUMO: O objetivo deste trabalho foi identificar zonas pluviometricamente homogêneas no estado do Rio Grande do Sul por meio de técnicas de mineração de dados, para posteriores estudos do comportamento dos veranicos nessa região. Foram utilizadas séries históricas de 16 anos (1988-2003), com dados de precipitação pluviométrica diária de 100 postos pluviométricos adquiridos junto ao site da Agência Nacional de Águas (ANA). Clusterização (agrupamento) foi a tarefa empregada e o programa computacional utilizado para obtenção dos clusters foi o Weka®, sendo o K-Means o algoritmo escolhido. Definiu-se três clusters baseado nas análises de precipitação total, análise da série temporal, cálculo do índice de sazonalidade e consulta ao especialista. O método mostrou-se eficiente para definição de zonas pluviometricamente homogêneas a partir de dados históricos.

PALAVRAS-CHAVE: chuva, clusterização, k-means.

DATA MINING TO IDENTIFY RAINFALL HOMOGENOUS AREAS IN THE STATE OF RIO GRANDE DO SUL

ABSTRACT: The objective of this paper was to identify rainfall homogeneous zones in the state of Rio Grande do Sul by means of techniques of data mining, for further studies of the behavior of the dry spells in this region. Historical series of 16 years were used (1988-2003), with daily rainfall data of 100 pluviometric stations collected from the National Water Agency (ANA). Clustering was the task applied to the data, notably the k-means algorithm. Such an algorithm is available at Weka®, a machine learning environment used in data mining applications. Tree clusters were defined, based on the analyses of total precipitation, analysis of time series, calculation of the sazonality index and the specialist´s advice. The proposed method seemed to be efficient for the definition of rainfall homogeneous zones from historical data.

KEYWORDS: rain, clustering, k-means.

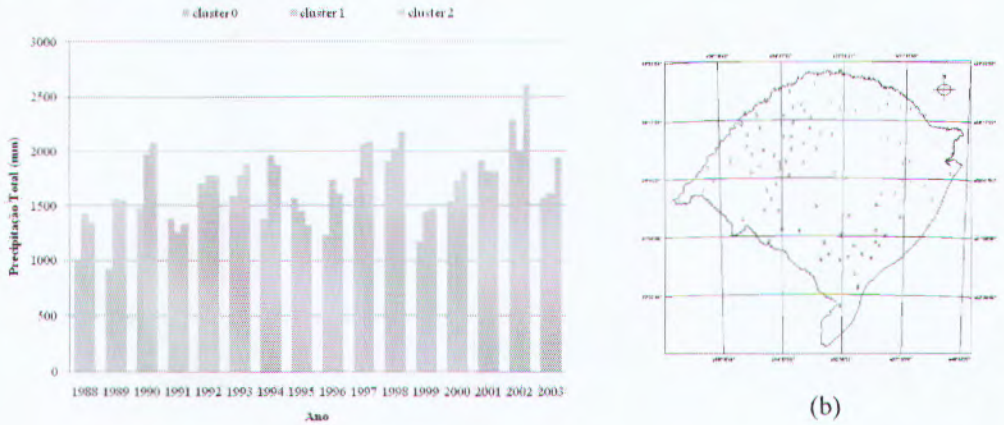
INTRODUÇÃO: A variabilidade temporal da precipitação, observada durante o verão no Rio Grande do Sul é apontada como a principal causa nas variações dos rendimentos das principais culturas do estado. O Rio Grande do Sul é responsável por cerca de 25% da produção de grão no país, sendo as culturas de primavera-verão responsáveis por 90% da produção total de grão do Estado. O cultivo é feito predominantemente em condições não irrigadas, dependendo, portanto do regime de chuvas. Segundo BERLATO (1992), embora as chuvas no estado do Rio Grande do Sul sejam bem distribuídas, apresentam uma alta variabilidade interanual, podendo comprometer o rendimento dos cultivos. A ocorrência de veranicos, que se caracterizam por períodos de estiagem em plena estação chuvosa, surge como um dos principais fatores limitantes, uma vez que dependendo da sua duração e

freqüência, podem ocorrer reduções significativas na produção final das culturas. O uso de técnicas de mineração de dados aparece como uma técnica promissora para definição de zonas homogêneas a partir de séries históricas, uma vez que possibilita a exploração de grandes volumes de dados, visando a descoberta de padrões ou mesmo de novos conhecimentos, fornecendo ainda subsídios para estudo posteriores do comportamento dos veranicos. Com o intuito de contribuir para estudos desse fenômeno, este trabalho tem o objetivo de definir zonas pluviometricamente homogêneas, por meio de técnicas de mineração de dados para o Estado do Rio Grande do Sul.

MATERIAL E MÉTODOS: A área de estudo compreende o Estado do Rio Grande do Sul, situado entre os paralelos 27°03'42" e 33°45'09" de latitude Sul e entre os meridianos 49°42'41" e 57°40'57" de longitude Oeste. Foram utilizadas séries históricas de 16 anos (1988-2003), com dados de precipitação pluviométrica diária de 100 postos pluviométricos adquiridos junto ao site da Agência Nacional de Águas (ANA). A metodologia será baseada no modelo CRISP-DM (Cross Industry Standard Process for Data Mining), proposto por CHAPMAN et al. (2000). O processo compreendeu as fases de entendimento do problema de pesquisa, entendimento dos dados, preparação dos dados, modelagem e avaliação. Clusterização (agrupamento) foi a tarefa empregada visando a obtenção das zonas homogêneas. O programa computacional utilizado foi o Weka® (WITTEN E FRANK, 2005) e o algoritmo utilizado nas análises foi o k-means, proposto por MACQUENN (1967). O k-means é uma técnica na qual os dados são agrupados de acordo com a métrica de distância euclidiana. O "means" refere-se ao centróide do cluster, que é selecionado de forma aleatória, sendo recalculado de forma iterativa até a obtenção do melhor conjunto, visando à minimização da distância entre os componentes do mesmo conjunto e a maximização em relação aos outros grupos formados (REZENDE et al., 2005; AMO, 2004). Utilizou-se valores de k igual a 2, 3 e 4 para obtenção dos clusters. Os dados foram espacializados em ambiente Spring para melhor visualização dos resultados. Na avaliação do número de clusters obtidos, utilizou-se o índice de sazonalidade, a precipitação total e análise das séries temporais. Fez-se a análise das séries temporais por meio da construção de um gráfico e observação do comportamento das chuvas nos clusters definidos.

RESULTADOS E DISCUSSÃO: Optou-se pela distribuição com três clusters, denominados cluster 0, cluster 1 e cluster 2 (fig. 1b). Na tabela 1 pode-se observar a distribuição das estações dentro de cada cluster, assim como a porcentagem de dados faltantes. A precipitação total durante o período de análise (1988-2003) pode ser observada na figura 1a, sendo o cluster 0 o que apresenta o menor volume de chuvas no período analisado com cerca de 24413 mm, seguido pelo cluster 1 com 27620 mm e o cluster 2 com 28693 mm. Observando-se a figura 1a nota-se que o cluster 2 apresenta maior variação durante o período, representando portanto uma área de maior instabilidade. Este fato também pode ser observado no gráfico da evolução das chuvas (gráfico 1). O gráfico 2 representa o índice de sazonalidade do período de 1988 a 2003 para os três clusters obtidos. O índice de sazonalidade da estação mais próxima do centróide está plotado no gráfico no intuito de mostrar sua representatividade, por meio da semelhança de comportamento. Em todos os clusters observa-se uma distribuição uniforme das chuvas durante os meses. Porém, como já constatado nas análises de precipitação total e da série histórica, observa-se uma variação interanual, com picos elevados nos meses de abril e outubro, e grandes quedas nos meses de maio e agosto. As variações entre os clusters obtidos podem ser explicadas de acordo com os fenômenos meteorológicos que atingem a região onde estão localizados. De um modo geral, os maiores volumes de chuva encontrados nos cluster 1 e cluster 2 podem ser explicados pela influência dos sistemas frontais e a atuação dos sistemas tropicais, que são mais intensos nessas regiões. A variabilidade constatada no cluster 2 pode ser atribuída a atuação dos Complexos Convectivos de Mesoescala. Esses sistemas se formam no período noturno sobre o

norte da Argentina e sul do Paraguai e deslocam-se rapidamente atingindo o noroeste do estado com intensa precipitação pluvial.



(a) (b)

FIGURA 1: (a) Precipitação total por cluster no período de 1988 a 2003; e (b) distribuição espacial dos clusters.

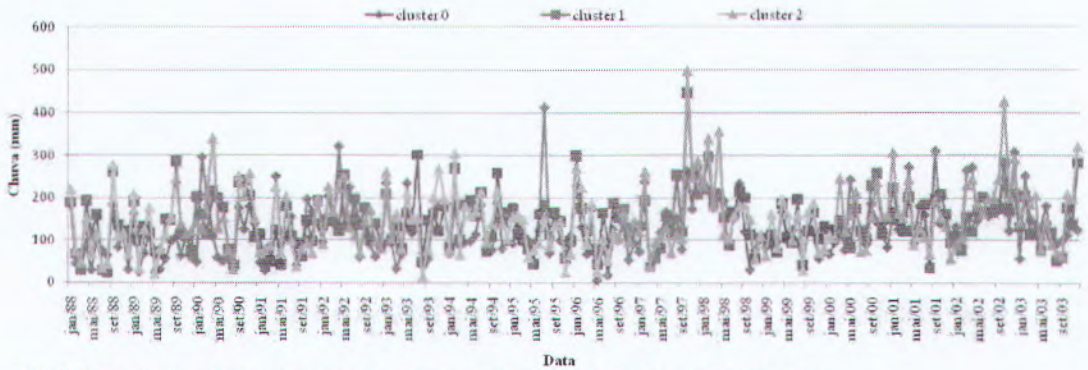


GRÁFICO 1: Evolução da precipitação por cluster no período de 1988 a 2003.

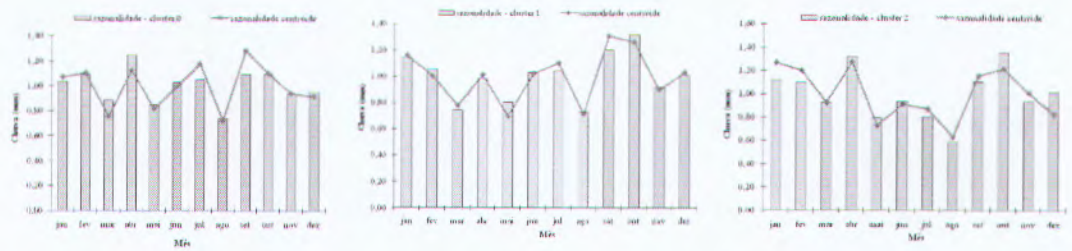


GRÁFICO 2: Índice de sazonalidade do período de 1988 a 2003 para os três clusters obtidos.

TABELA 1: Distribuição das estações nos clusters obtidos.

Cluster	n° de estações	Completas	Incompletas	Dados Faltantes (%)
0	24	16	8	3,40
1	31	17	14	4,23
2	45	24	21	4,3

CONCLUSÕES: Através do presente trabalho pode-se concluir que a mineração de dados mostrou-se eficaz na determinação de zonas pluviometricamente homogêneas a partir de séries históricas. Os clusters apresentam dependência espaço-temporal e a determinação do índice de sazonalidade foi eficiente para a análise dos clusters. A interação direta com especialistas foi fundamental para compreensão e comprovação dos resultados, a fim de validar as técnicas de mineração utilizadas.

REFERÊNCIAS

AMO, S. A. Técnicas de Mineração de Dados. In: **Sociedade Brasileira de Computação**, Universidade Federal da Bahia. (Org.). Jornadas de Atualização em Informática. Salvador: Universidade Federal da Bahia, 2004, v. 2, p.195-236.

BERLATO, M. A.; FONTANA, D. C.; GONÇALVES, H. M. Relação entre rendimento de grãos de soja e variáveis meteorológicas. **Pesquisa Agropecuária Brasileira**, Brasília, v.27, n.5, p.675-702, 1992.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. [Illinois]: SPSS, 2000. 78p.

MCQUEEN, J. B. Some Methods for Classification and Analysis of Multivariate Observations. In: Lecam, L. M.; Neyman, J. editors, **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability**, p. 281-297, University of California Press, Berkeley, CA, 1967.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; DE PAULA, M. F. Mineração de Dados. In: REZENDE, S. O. **Sistemas Inteligentes:** fundamentos e aplicações. 1. ed. São Paulo: Manole, 2005. p. 307-336.

WITTEN, I. H.; FRANK, E. **Data mining**: Practical machine learning tools and techniques. 2nd. ed. San Francisco: Morgan Kaufmann, 2005. 525p.