Acta Iguazu
ISSN: 2316-4093

Data Mining Techniques for Rainfall Regionalization in Parana State

Jonathan Richetti[1], Elizabeth Giron Cima[1], Jerry A. Johann[2], Miguel Angel Uribe-Opazo[2]

[1]PhD Student, Western Paraná State University, Cascavel – PR - BR, phone: +55 45 3220-7320, email:
[2]PhD Professor, Western Paraná State University, Cascavel – PR - BR, phone: +55 45 3220-7320, email:

Email autor correspondente: j_richetti@hotmail.com
Artigo enviado em 14/09/2017, aceito em 20/01/2018.

Abstract: The prevalence of agro-meteorological data for specific regions serve as parameters for agricultural and related climate studies. This study aims to regionalize the rainfall in the State of Paraná (Southern Brazil) through data mining techniques with ECMWF (European Centre for Medium Range Weather Forecasts) data from 1989 to 2013. The algorithms k-means and Simple EM (Expectation Maximization) for clustering were applied in Weka software, version 3.6. The quality of the clustering was determined with the J48 classification algorithm applied using training set. The decision tree presents similarity indexes and errors measures to determine the best number of cluster for this case. As results 6 regions of homogeneous rainfall in the state of Paraná were presented.

Keywords: cluster, Weka, k-means algorithm, EM algorithm.

Técnicas de Mineração de Dados para Regionalização da Precipitação no Estado do Paraná

Resumo: A prevalência de dados agrometeorológico para regiões específicas servem como parâmetros para estudos agrícolas, do clima e afins. O objetivo deste estudo foi regionalizar a precipitação no estado do Paraná (sul do Brasil) através de técnicas de mineração de dados com dados do ECMWF (Centro Europeu para Previsões Meteorológicas de Médio Alcance) de 1989 para 2013. Os algoritmos de *k-means* e *simple EM* (maximização de expectativa) para *clusters* foram aplicados no software Weka, versão 3.6. A qualidade do agrupamento foi determinada com o algoritmo de classificação J48 aplicado usando o conjunto de treinamento. A árvore de decisão apresenta índices de similaridade e erros de medidas para determinar o melhor número de cluster para este caso. Os resultados apresentam 6 regiões de precipitação homogênea no estado do Paraná.

Palavras chave: algoritmo k-means, cluster, Weka, algoritmo EM.

Introduction

The diverse distributions of spatial-temporal variations in rainfall can directly impact agriculture (ROMANI, 2010). Therefore, understanding of spatial behavior of rainfall in a state that produce 18.5 million tons for the crop-season 2015/2016 is important. Parana

state alone accounts for 18.3% of the Brazilian soybean production (CONAB, 2016), a higher production that China (12.0 million tons), the 4th largest soybean producer in the world, (FAOSTAT, 2016). Besides this impact on crop yield that may be associated with characteristics of rainfall one of the most important parameters for the hydrological regime is rainfall, which needs to be studied in space and time (GOYAL; GUPTA, 2014). Thus, identifying homogeneous rainfall regions is importance for agricultural planning, hydrological studies and watershed managements. This process of identifying homogeneous rainfall regions is called rainfall regionalization.

For a good regionalization, spatially distributed rainfall measures are necessary. Paraná state has a good distribution of pluviometric stations. Nevertheless, some other regions do not present the same distribution and availability. For these cases a solution might be the use of ECMWF (European Centre for Medium Range Weather Forecasts) database (EUROPEAN UNION, 2014). For a rainfall regionalization, data mining techniques are essential and present in most works. Bothale and Katpatal (2014) used an agglomerative hierarchical clustering to regionalize the annual precipitation in the Pranhita basin (India). Golian et al. (2010) used two methods for clustering, a fuzzy k-means algorithm and a supervised classification based on Jenk's optimization method, concluding that the number of groups (clusters) was sensitive to the number of stations used, reducing the number of stations increased the number of groups. Michaelides et al. (2001) used neural network for precipitation variability classification concluding that the method can be used successfully to identify and classify similar precipitation patterns. Muñoz-Diaz and Rodrigo (2004) found spatiotemporal patterns of rainfall in Spain by Principal Component Analysis. For the state of Paraná Pansera et al. (2015) used data from the Brazilian National Water Agency (ANA – *Agência Nacional de Águas*) and subdivided the state in six homogeneous regions of monthly precipitation using a hybrid methodology of k-means and Ward method.

Therefore, the objective of this study was to determine homogeneous regions considering the dekad data from ECMWF in the periods from 1989 to 2013 using EM and k-means algorithms. And, compare the results with ECMWF data and the Pansera et al. (2015) results obtained from ground stations.

Materials and Methods

Study Area and ECMWF data

The study area comprises the state of Paraná, in south of Brazil, located between parallels 22º 29 'S and 26º 43' S and the meridians 48º 02' W 54 and 38' W (Figure 1). Paraná has Aw, Cwa, Cfa and Cfb Köppen climate classification (APARECIDO et al., 2016). Historical series of ten days rainfall were used, acquired in the ECMWF ERA-Interim database (EUROPEAN UNION, 2014). A total of 286 Virtual Stations (VS) spaced in a regular grid of 25 km (Figure 1) were used.

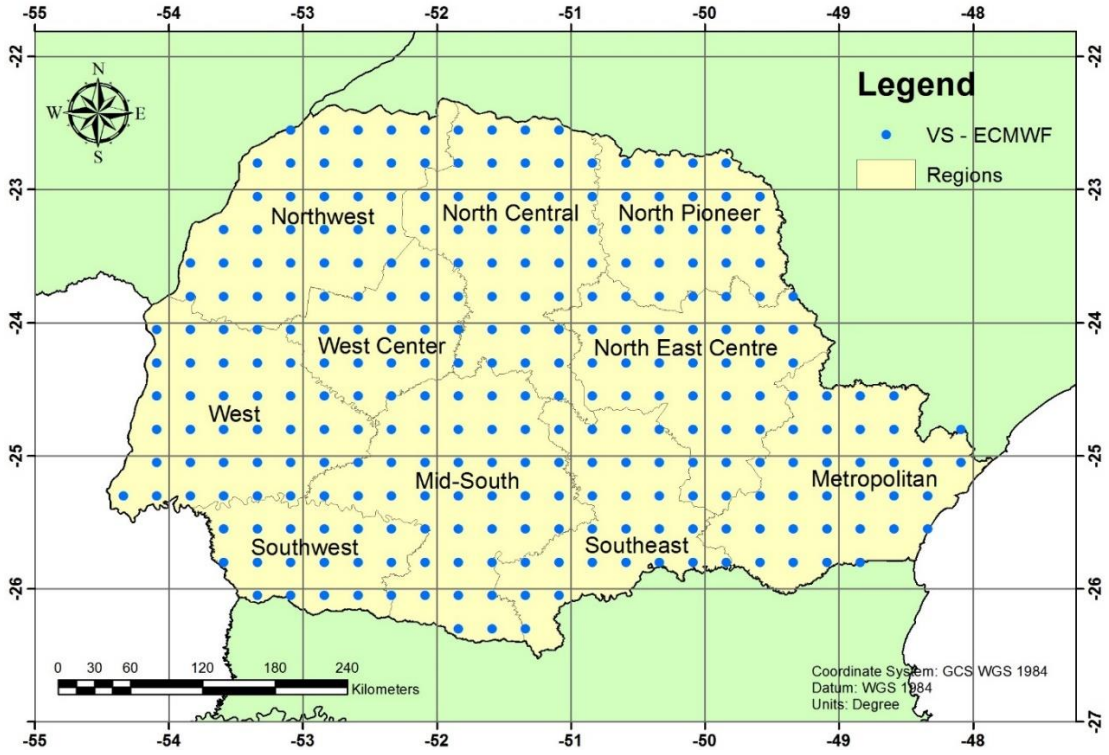

Figure 1. Virtual Stations (VS) at Paraná state from ECMWF.

Algorithms

Simple k-means

K-means is a process for partitioning an N-dimensional population into k sets on the basis of a sample (MACQUEEN, 1967). This process is one of the simplest unsupervised learning algorithms to solve clustering problems. A *k* number of cluster is fixed a priori and the algorithm defines each centroid. After that each information from the data is associated with the nearest centroid when no point is pending, the first clustering of a loop is done. The process is repeated until the there is no more changes in the centroids of the clusters.

(Simple Expectation Maximization)

The expectation maximization (EM) algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. In particular, expectation maximization attempts to find the parameters θ that maximize the log probability $logP(x;\theta)$ of the observed data (DO; BATZOGLOU, 2008).

Decision tree J48

A decision tree is a classification machine learning technique where each object of the data belongs to one of a set of mutually exclusive classes. As an induction task would be to generate all possible decision trees that correctly classify the training set and to select the simplest of them (QUINLAN, 1986). The J48 is the C4.5 algorithm (QUINLAN, 1986) implemented in the Weka software.

Clustering Process

Data were normalized and clustered by k-means and EM algorithms. After that, the groups were used as attributes variables and a classifications by J48 algorithm were carried. With that the best number of regions with similar characteristics were determined (Figure 2).

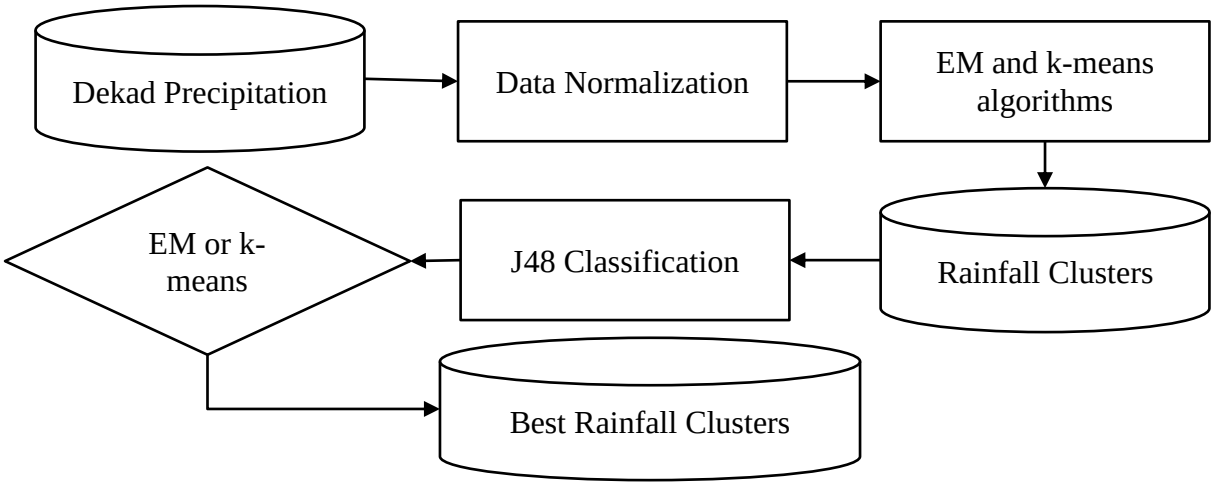

Figure 2. General flowchart of the study.

For the processing and analysis of data the software WEKA, version 3.6, was used. The clusters were generated based on the temporal profile of precipitation from the 286 ECMWF VS. After that, the groups generated were used as target variable for the J48 classification algorithms to check which cluster showed lower errors. This methodology, to determine which was the best group of precipitation region among the groups performed, is presented by Johann et al. (2013). Therefore, the evaluated similarity coefficients were Global Accuracy (GA - %) Instances Misclassified (IM -%), Kappa index, the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Relative Absolute Error (RAE - %) and the Root Relative Squared Error (RRSE - %). The best cluster number presented the highest GA and kappa values and minor errors.

Results and Discussion

A total of 16 rainfall clusters by EM and k-means algorithms (8 clusters each algorithm) were obtained. In order to determine what was the best group of precipitation zones (clusters) the J48 classifier algorithm was used with the use of training set. The group, with best results, i.e., larger adjustments and smaller errors, was the group with six clusters by k-means algorithm (Table 1).

With the six clusters obtained by the k-means algorithm thematic map was build (Figure 3). Where can be observed that the metropolitan area has the same rainfall regime, bounded by Cluster0, with average rainfall during the year of 43.51 mm. The regions: Southwest, the southern part of the Mid-South and Southeast have similar rainfall with an annual average of 48.08 mm grouped by Cluster1. The Northwest region, much of West Center and a part of the North Central have similar arrangements grouped by Cluster2 with an average of 37.31 mm. The Regions: North Pioneer, much of the North Central and North East Centre present annual average of 35.69 mm rainfall in the Cluster3. In addition, the West, southern part of the West-Central and part of South Central have a 43.56 mm rainfall average (Cluster4). Finally, the Cluster5 has an average of 40.22 mm annually.

Table 1. Results from the J48 classifier algorithm with the use of training set.

	EM-6	EM-7	EM-8	EM-9	EM-10	EM-11	EM-12	EM-13
GA	83.59%	90.72%	88.66%	87.63%	91.75%	78.35%	84.53%	88.66%
IM	13.40%	9.28%	11.34%	12.37%	8.25%	21.65%	15.46%	11.34%
Kappa	0.8381	0.8905	0.8694	0.8599	0.9078	0.7610	0.8308	0.8757
MAE	0.0459	0.0265	0.0284	0.0286	0.0165	0.0394	0.0262	0.0174
RMSE	0.2114	0.1628	0.1684	0.1659	0.1284	0.1984	0.1593	0.1321
RAE	16.49%	10.85%	12.99%	14.51%	9.16%	23.73%	17.16%	12.25%
RRSE	56.37%	46.34%	50.63%	52.53%	42.69%	68.60%	57.44%	49.32%
	k-mean-6	**k-mean-7**	**k-mean-8**	**k-mean-9**	**k-mean-10**	**k-mean-11**	**k-mean-12**	**k-mean-13**
GA	***96.91 %***	90.72%	90.72%	90.72%	90.72%	81.44%	95.88%	92.78%
IM	***3.09%***	9.28%	9.28%	9.28%	9.28%	18.56%	4.12%	7.22%
Kappa	***0.9625***	0.8903	0.8903	0.8950	0.8950	0.7943	0.9548	0.9215
MAE	***0.0103***	0.0282	0.0282	0.0206	0.0206	0.0344	0.0069	0.0111
RMSE	***0.1015***	0.1599	0.1599	0.1436	0.1436	0.1820	0.0829	0.1054
RAE	***3.72%***	11.52%	11.52%	10.42%	10.42%	20.78%	4.48%	7.79%
RRSE	***27.18 %***	45.59%	45.59%	45.54%	45.54%	63.11%	29.83%	39.32%

Bold italic: best results

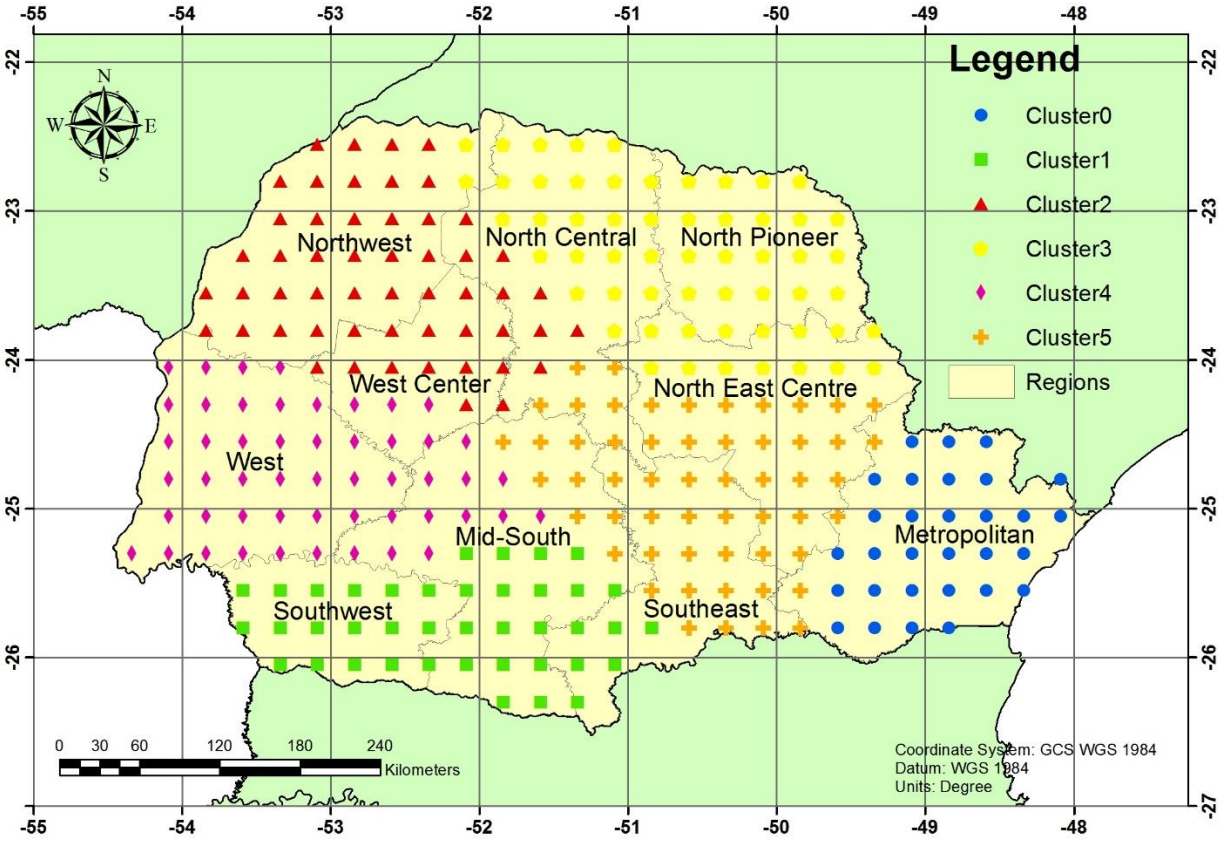

Figure 3. Rainfall clusters for the Paraná state.

Although a few clusters provide similar average (e.g. Cluster0 and cluster4) is observed that there are differences between the groups over the dekad periods (Figure 5). That is, the rainfall is well distributed and similar during the summer, being January the rainiest period of the state. However, the main difference is between the 10 and 32 dekads (winter and spring) with precipitation volume differences of more than 30 mm in just one dekad (e.g. 17th, 26th and 27th dekads - Figure 5). Which means that the Paraná state has a uniformity of rainfall during the summer, but for the rest of the year the rainfall regimes are different.

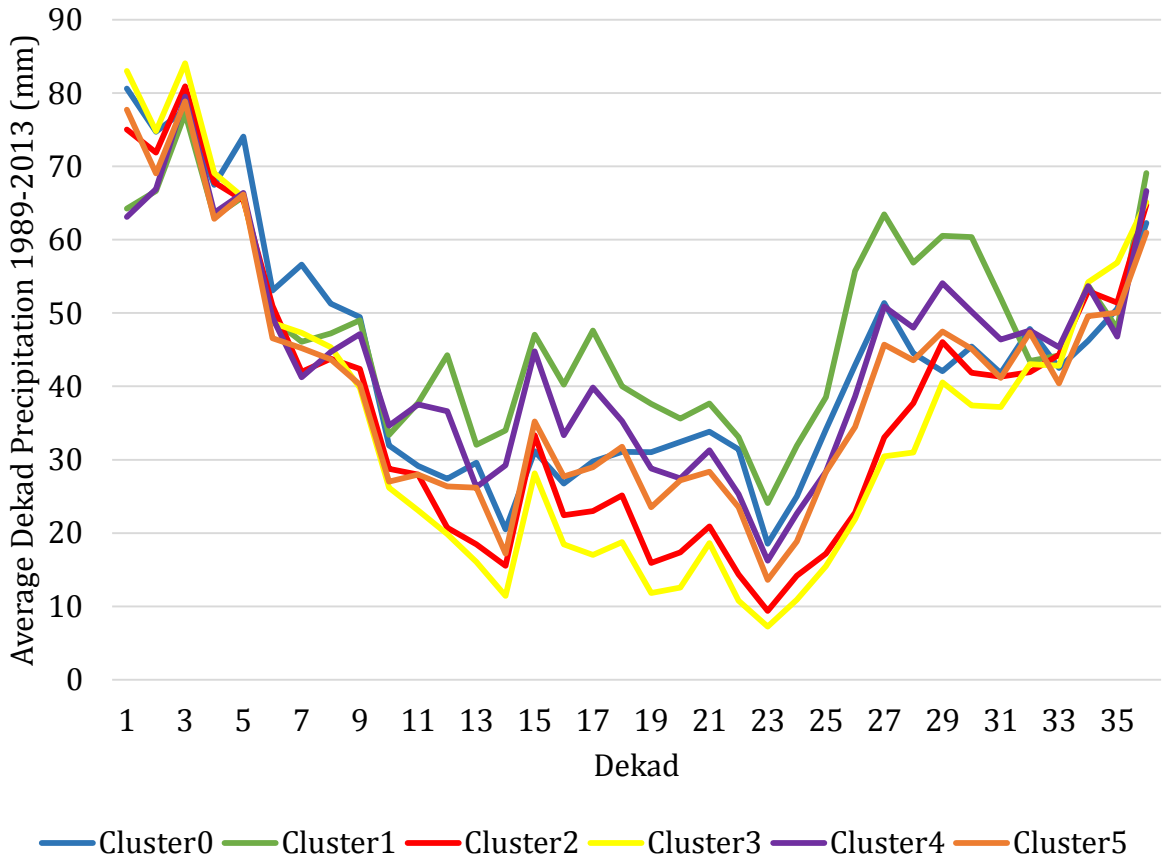

Figure 5. Historical rainfall average 1989 a 2013) by cluster.

This results corroborate with those presented by Pansera et al. (2015), that also presented six clusters in the state with slightly difference at cluster borders. This shows that the use of dekadly data from the ECMWF showed very similar results to those from monthly pluviometric stations.

Conclusions

Generally, it was possible to regionalize the rainfall in the state of Paraná in [I] six regions with homogeneous rainfall distribution. [II] It was possible to trace a rainfall temporal profile of the state for each region. [III] During the summer, rain schemes are similar, i.e., precipitation can be considered equal. [IV] The rest of the year, the regions have different precipitation profiles. [V] Data from the ECMWF proved to be robust and consistent with the reality of the state for precipitation information. [VI] Data mining techniques were efficient and

presented good results for regionalizing rainfall in the state.

References

APARECIDO, L.E.O.; ROLIM, G.S.; RICHETTI, J.; DE SOUZA, P.S.; JOHANN, J.A.: Köppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil. **Ciência e Agrotecnologia**, v. 40, n. 4, p. 405–417, 2016.

BOTHALE, R.; KATPATAL, Y. Spatial and Statistical Clustering Based Regionalization of Precipitation and Trend Identification in Pranhita Catchment , India. **IJIRSET - International Journal of Innovative Research in Science, Engineering and Technology**, v. 3, n. 5, p. 12557–12567, 2014.

CONAB. Companhia Nacional de Abastecimento. **Acompanhamento da safra brasileira - grãos (Safra 2015/16)**. [s.l: s.n.]. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_02_04_11_21_34_boletim_graos_fevereiro_2016_ok.pdf>.

DO, C. B.; BATZOGLOU, S. What is the expectation maximization algorithm? **Nature Biotechnology**, v. 26, n. 8, p. 897–899, 2008.

EUROPEAN UNION. **EC-JRC-MARS data created by MeteoConsult based on ECWMF (European Centre for Medium Range Weather Forecasts) model outputs**. Disponível em: <http://spirits.jrc.ec.europa.eu/?page_id=2869>. Acesso em: 19 ago. 2015.

FAOSTAT. Food and Agricultural Organization - Statistits Department. **World Production**. Disponível em: <http://faostat3.fao.org/home/E>. Acesso em: 1 ago. 2016.

GOLIAN, S.; SAGHAFIAN, B.; SHESHANGOSHT, S.: Ghalkhani, Hossein, S. et al. Comparison of classification and clustering methods in spatial rainfall pattern recognition at Northern Iran. **Theoretical and Applied Climatology**, v. 102, n. 3, p. 319–329, 2010.

GOYAL, M. K.; GUPTA, V. Identification of homogeneous rainfall regimes in northeast region of India using fuzzy cluster analysis. **Water Resources Management**, v. 28, n. 13, p. 4491–4511, 2014.

JOHANN, J. A.; ROCHA, J. V.; OLIVEIRA, S. R. DE M.; RODRIGUES, L. H. A.: Lamparelli, Rubens A. C. Data mining techniques for identification of spectrally homogeneous areas using NDVI temporal profiles of soybean crop. **Engenharia Agrícola**, v. 33, n. 3, p. 511–524, 2013.

MACQUEEN, J. B. **Some Methods for Classification and Analysis of Multivariate Observations**. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. **Anais**...Berkley: 1967Disponível em: <http://www.umiacs.umd.edu/~raghuram/ENEE731/Spectral/kMeans.pdf>. Acesso em: 19 jul. 2017

MICHAELIDES, S. C.; PATTICHIS, C. S.; KLEOVOULOU, G. Classification of rainfall variability by using artificial neural networks. **International Journal of Climatology**, v. 21, n. 11, p. 1401–1414, 2001.

MUÑOZ-DIAZ, D.; RODRIGO, F. S. Spatio-temporal patterns of seasonal rainfall in Spain (1912 – 2000) using cluster and principal component analysis : comparison. **Annales Geophysicae**, v. 22, p. 1435–1448, 2004.

PANSERA, W. A.; GOMES, B. M.; VILAS BOAS, M. A.; SAMPARIO, S. C.; DE MELLO, E. L.: Queiroz, Manoel Moises Ferreira

PANSERA, W. A. et al. Regionalization of Monthly Precipitation Values in the State of Paraná (Brazil) By Using Multivariate Clustering Algorithms. **Irriga**, v. 20, n. 3, p. 473–489, 2015.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, v. 1, p. 81–106, 1986.

ROMANI, L. A. S. **Integrating Time Series Mining and Fractals to Discover Patterns and Extreme Events in Climate and Remote Sensing Databases**. São Paulo - SP: ICMC-USP, 2010.