Data Mining Techniques for Rainfall Regionalization in Parana State

Jonathan Richetti¹, Elizabeth Giron Cima¹, Jerry A. Johann², Miguel Angel Uribe-Opazo²

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

Email autor correspondente: j_richetti@hotmail.com

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.

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 than 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-Díaz 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 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.

**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).
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.

<table>
<thead>
<tr>
<th></th>
<th>EM-6</th>
<th>EM-7</th>
<th>EM-8</th>
<th>EM-9</th>
<th>EM-10</th>
<th>EM-11</th>
<th>EM-12</th>
<th>EM-13</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>83.59%</td>
<td>90.72%</td>
<td>88.66%</td>
<td>87.63%</td>
<td>91.75%</td>
<td>78.35%</td>
<td>84.53%</td>
<td>88.66%</td>
</tr>
<tr>
<td>IM</td>
<td>13.40%</td>
<td>9.28%</td>
<td>11.34%</td>
<td>12.37%</td>
<td>8.25%</td>
<td>21.65%</td>
<td>15.46%</td>
<td>11.34%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.8381</td>
<td>0.8905</td>
<td>0.8694</td>
<td>0.8599</td>
<td>0.9078</td>
<td>0.7610</td>
<td>0.8308</td>
<td>0.8757</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0459</td>
<td>0.0265</td>
<td>0.0284</td>
<td>0.0286</td>
<td>0.0165</td>
<td>0.0394</td>
<td>0.0262</td>
<td>0.0174</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.2114</td>
<td>0.1628</td>
<td>0.1684</td>
<td>0.1659</td>
<td>0.1284</td>
<td>0.1984</td>
<td>0.1593</td>
<td>0.1321</td>
</tr>
<tr>
<td>RAE</td>
<td>16.49%</td>
<td>10.85%</td>
<td>12.99%</td>
<td>14.51%</td>
<td>9.16%</td>
<td>23.73%</td>
<td>17.16%</td>
<td>12.25%</td>
</tr>
<tr>
<td>RRSE</td>
<td>56.37%</td>
<td>46.34%</td>
<td>50.63%</td>
<td>52.53%</td>
<td>42.69%</td>
<td>68.60%</td>
<td>57.44%</td>
<td>49.32%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>k-mean-6</th>
<th>k-mean-7</th>
<th>k-mean-8</th>
<th>k-mean-9</th>
<th>k-mean-10</th>
<th>k-mean-11</th>
<th>k-mean-12</th>
<th>k-mean-13</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>96.91%</td>
<td>90.72%</td>
<td>90.72%</td>
<td>90.72%</td>
<td>90.72%</td>
<td>81.44%</td>
<td>95.88%</td>
<td>92.78%</td>
</tr>
<tr>
<td>IM</td>
<td>3.09%</td>
<td>9.28%</td>
<td>9.28%</td>
<td>9.28%</td>
<td>9.28%</td>
<td>18.56%</td>
<td>4.12%</td>
<td>7.22%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.9625</td>
<td>0.8903</td>
<td>0.8903</td>
<td>0.8950</td>
<td>0.8950</td>
<td>0.7943</td>
<td>0.9548</td>
<td>0.9215</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0103</td>
<td>0.0282</td>
<td>0.0282</td>
<td>0.0206</td>
<td>0.0206</td>
<td>0.0344</td>
<td>0.0069</td>
<td>0.0111</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1015</td>
<td>0.1599</td>
<td>0.1599</td>
<td>0.1436</td>
<td>0.1436</td>
<td>0.1820</td>
<td>0.0829</td>
<td>0.1054</td>
</tr>
<tr>
<td>RAE</td>
<td>3.72%</td>
<td>11.52%</td>
<td>11.52%</td>
<td>10.42%</td>
<td>10.42%</td>
<td>20.78%</td>
<td>4.48%</td>
<td>7.79%</td>
</tr>
<tr>
<td>RRSE</td>
<td>27.18%</td>
<td>%</td>
<td>45.59%</td>
<td>45.59%</td>
<td>45.54%</td>
<td>45.54%</td>
<td>63.11%</td>
<td>29.83%</td>
</tr>
</tbody>
</table>

**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.](image)

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


