ISSN ONLINE 1808-8546/ISSN CD 1808-3765

# REGIONALIZATION OF MONTHLY PRECIPITATION VALUES IN THE STATE OF PARANÁ (BRAZIL) BY USING MULTIVARIATE CLUSTERING ALGORITHMS

**WAGNER ALESSANDRO PANSERA[1]; BENEDITO MARTINS GOMES[2]; MARCIO ANTONIO VILAS BOAS[2]; SILVIO CESAR SAMPAIO[2]; ELOY LEMOS DE MELLO[2] E MANOEL MOISES FERREIRA DE QUEIROZ[3]**

1 Departamento de Construção Civil, Universidade Tecnológica federal do Paraná (UTFPR), Pato Branco-PR, e-mail:pansera@utfpr.edu.br
2 Centro de Ciências Exatas e Tecnológicas, Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel-PR. e-mail:benedito.gomes@unioeste.br, marcio.vilasboas@unioeste.br, silvio.sampaio@unioeste.br, eloy.mello@unioeste.br
3 Centro de Ciências e Tecnológica Alimentar, Universidade Federal de Campina Grande (UFCG), Pombal-PB. e-mail: moises@ccta.ufcg.edu.br.

## 1 ABSTRACT

Rainfall is one of the most important factors in agricultural water management, especially to land farming of Paraná. Regionalization of rainfall data is useful for optimum design of projects and management of water related activities. Therefore, the objectives of this study were to evaluate the potential of multivariate clustering methodologies to identify homogenous regions using time series of monthly precipitation as the classification variable; the quality of the formed regions and estimate precipitation quantiles. k-means, hierarchical and hybrid clustering techniques were used. The rainfall stations in each group were subjected to metrics of homogeneity. The Paraná state could be subdivided into six homogeneous regions of monthly precipitation using the hybrid methodology between k-means and Ward.

**Keywords:** regional frequency analysis, L-moments, cluster analysis

**PANSERA, W. A.; GOMES, B. M.; VILAS BOAS, M. A.; SAMPAIO, S. C; MELLO, E. L. DE; QUEIROZ, M. M. F. DE.**
**REGIONALIZAÇÃO DE VALORES DE PRECIPITAÇÃO MENSAL NO ESTADO DO PARANÁ (BRASIL) USANDO ALGORITMOS DE AGRUPAMENTOS MULTIVARIADOS**

## 2 RESUMO

A chuva é um dos fatores mais importantes na gestão da água na agricultura, especialmente para o estado do Paraná-BR. A regionalização de dados de chuva é útil para otimização de projetos e gestão das atividades relacionadas com a água. Portanto, este estudo teve por objetivo avaliar as metodologias multivariadas de agrupamentos para identificar regiões homogêneas usando as séries históricas de precipitação mensal como variável classificatória, avaliar a qualidade das regiões formadas e estimar quantis de precipitação. As técnicas de

Recebido em 14/05/2013 e aprovado para publicação em 15/09/2015

agrupamento utilizadas foram: k-médias, hierárquico e híbrido. As estações pluviométricas presentes em cada grupo foram submetidas às métricas de homogeneidade e discordância. Foi possível subdividir o estado Paraná em seis regiões homogêneas de precipitação mensal utilizando a metodologia hibrida entre k-médias e Ward.

**Palavras-chave**: Análise de frequência regional, momentos-L, análise de agrupamentos.

## 3 INTRODUCTION

Rainfall is one of the most important factors in agricultural water management, especially to the land farming of Paraná-BR. Regionalization of precipitation is useful for the optimum design and management of water related activities.

The identification of a rainfall spatial pattern is usually an essential need for water resources planning and management. However, the rainfall fluctuation is usually difficult to be fully recognized from year to year and from place to place. Therefore, many present-day hydrologic and climatic studies are trying to find out and develop methods for the regionalization of hydrologic and climatic variables. Regional classification of these variables helps scientists to simplify the hydro-climatic convolution and therefore reduces the massive body of information, observation and variables (MODARRES and SARHADI, 2011).

Regional frequency analysis (RFA) involves the use of multidisciplinary tools to estimate projected precipitation values in locations with no rainfall stations or to improve estimates based on observations from a station of interest. The necessary data for indirect estimates of rainfall are normally transferred to unmonitored locations from a set of similar basins through the process of regionalization. Regionalization identifies spatial units with common attributes and separates them from units that do not possess these attributes. If the identified units cover a continuous geographic area, the area is referred to as a region; the process of creating regions is called regionalization (CUNDERLIK and BURN, 2006).

An important requisite for RFA is the identification of regions that can be used to transfer hydrological information. In this context, a region denotes a set of hydrographic basins that are similar in terms of their hydrological response. The objective of the regionalization process is to identify clusters of regions that are sufficiently similar to justify the combination and transfer of hydrological information from locations within the region.

However, the simple determination of homogenous regions using classification methodologies does not guarantee homogeneity. To verify the homogeneity, Hosking and Wallis (1997) developed the discordancy measure and the test of heterogeneity.

The application of clustering techniques to delineate homogenous regions is not automatic. Users must select (1) the most relevant variables for calculations of the distances between stations, (2) the link function, which strongly affects group formation, and (3) the segment distance in the hierarchical tree, which should reflect a user's objectives by identifying the optimal number of groups (OUARDA et al., 2008).

Therefore, the objectives of this study were (1) to evaluate the potential of multivariate clustering methodologies in identifying homogenous regions using monthly historical series as the classification variable in the state of Paraná, (2) to evaluate the quality of the formed regions, and (3) to estimate precipitation quantiles.

# 4 MATERIALS AND METHODS

## 4.1 Study area

The area chosen for this study is Paraná State in southern Brazil. Its area is about 200.00 km$^2$ and it lies approximately between 22º31' and 26º43' latitude S and between 48º06' and 54º37' longitude W Gr.

The Paraná's climate is subtropical (with mild temperatures), but a small part of its territory has a tropical climate. Annual thermal amplitude is 13ºC, except for the coast, which is 9ºC. According to Köppen classification, it predominates type C climate (mesothermal) and secondly the type A climate (tropical rainy). Annual rainfall lies between 1200 and 1600mm, approximately, except on the coast that can reach 2500 mm. It does not show a well-defined dry season.

## 4.2 Data

In total, 227 rainfall stations were used. Data from these stations were obtained at the National Water Agency (Agência Nacional de Águas - ANA) through the HIDROWEB information system. The stations were selected according to the following criteria: (1) must have data for the period from 1976 to 2006, (2) must be missing less than 18 values, and (3) must have no more than four consecutive missing values. These criteria helped to identify stations with the same series size and a limited number of missing values. The clustering methodologies were used with data from these stations to obtain homogenous rainfall series that could be described by the same probability distribution. A matrix was created with 227 rows and 372 columns; the rows represented the rainfall stations and the columns represented the monthly rainfall.

## 4.3 Data verification, Identification, Selection, Estimation & Evaluation

The procedure used in this study was based on the methodology proposed by Hosking and Wallis (1997) and included four stages:

1. Verification of the quality of monthly rainfall data,
2. Identification of homogenous regions,
3. Selection of a regional distribution function,
4. Estimation and evaluation of regional quantiles.

Step 1: Verification of quality

A considerable effort was applied to screen and to verify the quality of the rainfall data, with the goal of eliminating false values associated with numerous and variable measurement, reading, and transcription errors.

Step 2: Formation of homogenous groups

*2.1 Cluster analysis*

The data from each station were transformed to fit within the interval [0,1]. The transformation was necessary because of differences in the variance, magnitude, and relative significance of the data (RAO and SRINIVAS, 2006a).

$$x_{ij} = \frac{y_{ij} - y_{j(\min)}}{y_{j(\max)} - y_{j(\min)}} \text{ for } i = 1,2,\ldots,227 \text{ , } j = 1,2,\ldots,372 \text{ ,} \tag{1}$$

where: $x_{ij}$ , denotes the transformed element; $y$ , is the data matrix; $i$, is the row; and $j$ , is the column.

Let $x_{ij}$ and $x_{kj}$ denote transformed observations of stations i and k ($i = k = 1,2,\ldots,227$), referring to j$^{th}$ observation (j = 1, 2,…, 372), the Euclidean distance between stations was evaluated by Eq. (2)

$$d(i,k) = \sqrt{\sum_{j=1}^{372} \left(x_{ij} - x_{kj}\right)^2} \tag{2}$$

When Eq. (2) is used, the matrix of euclidean distances is obtained among $x_i$ stations, with dimension 227x227 and zero diagonal.

Hierarchical clustering is a way to sort and group data by creating a “cluster tree”. The tree is not a single set of clusters, but a multi-level hierarchy, in which clusters at one level are joined as clusters at the next higher level (CORTÉS et al., 2007).

At the beginning of the process, each station is equivalent to a group. According to the following steps, the two nearest groups (or stations) are combined into a new group, so, there is a reduction on the number of groups in a unit at each step. Eventually, all the stations are clustered into a large group. The hierarchical algorithms differ in the way distance between groups of stations is computed.

Single linkage is based on a hierarchy built using the smallest Euclidean distance between one of the stations within one cluster to one of the stations in adjacent clusters. Complete linkage is used to build up the cluster hierarchy and consists of finding the proxy of a cluster, based on finding the station within the cluster which shows the largest Euclidean distance between itself and its nearest surrounding cluster. Average linkage uses the average Euclidean distance between all pairs of stations in clusters a and b (CORTÉS et al., 2007):

$$d(a,b) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} d\left(a_i, b_j\right) \tag{3}$$

In which $d(a,b)$ is the Euclidean distance between cluster a and cluster b; $n_a$, the number of stations in cluster a; $n_b$, the number of stations in cluster b; and, $d\left(a_i, b_j\right)$, the Euclidean distance between stations from cluster a and from cluster b.

In a centroid linkage, the distance between two clusters is the distance between the cluster centroids or means. Firstly, a and b groups (or stations) with the shortest Euclidean distance are grouped, then the matrix of distances is updated by using equation (4).

$$d^2(ab,c) = \frac{n_a}{n_a + n_c} d^2(ac) + \frac{n_b}{n_b + n_c} d^2(bc) - \frac{n_a n_b}{\left(n_a + n_b\right)} d^2(ab) \tag{4}$$

where $d^2(ac)$, $d^2(bc)$ and $d^2(ab)$ are quadratic Euclidean distances between the respective groups and $n_c$ is the number of stations is group c.

The Ward method uses the incremental sum of squares or the increase in the internal sum of squares arising from the junction of two groups. The Ward procedure aims at joining groups without drastically increasing the variation among them, thereby, it produces the most homogenous groups. This technique also separates stations into regions with almost the same size. This ensures that each group has the minimum number of stations to enable the application of the appropriate regional estimate technique (OUARDA et al., 2008). The distance between clusters is given by Eq. (5):

$$d^2(a,b)=\frac{n_a n_b d^2(z_a,z_b)}{n_a+n_b} \tag{5}$$

where: $d^2(z_a,z_b)$ represents the quadratic Euclidean distance between centroids of a and b groups.

Partitioned clustering algorithms divide datasets into groups, often through the minimization of some criterion or error function. The number of groups is typically predefined, but it can be part of an error function. The k-means procedure is a partitioned algorithm based on the squared error criterion. The general objective of this method is to obtain a partition in which the squared errors are minimized for a fixed number of groups (GARCÍA and GONZÁLEZ, 2004).

$$E=\sum_{k=1}^{K}\sum_{j=1}^{372}\sum_{i=1}^{N} d^2\left(x_{ij}^k - x_{\bullet j}^k\right) \tag{6}$$
,

In Eq. (6), $d^2(\ )$ denotes the quadratic Euclidian distance; K denotes the number of clusters; N represents the number of stations in cluster k; $x_{ij}^k$ denotes the rescaled value j in the station i assigned to cluster k; $x_{\bullet j}^k$ is the mean value of month j for cluster k, using Eq. (7):

$$x_{\bullet j}^k=\frac{\sum_{i=1}^{N} x_{ij}^k}{N} \tag{7}$$

In this study, a hybrid of k-means and hierarchical clustering techniques were used. Thus, results from the hierarchical clustering algorithms are used to provide initial cluster centers for the k-means (RAO and SRINIVAS, 2006a). Clustering, in two stages, avoids subjective decisions and identifies a unique solution through iterations and classifications (CHENG and LIAO, 2009).

### *2.2 Validation of the clusters*

The cophenetic correlation coefficient (CCC) is used to measure the degree to which a hierarchical structure of a dendrogram represents the multidimensional relationships of the input data in two dimensions (RAO and SRINIVAS, 2006b). The cophenetic correlation measures the degree of fit between the original dissimilarity matrix and the matrix that results from clustering. This correlation is equivalent to the Pearson's correlation between the original dissimilarity matrix and the matrix obtained after the construction of a dendrogram. Thus, values closer to 1 indicate less distortion from the clustering of individuals (BEAVER and PALAZOĞLU, 2006).

In the Davies-Bouldin index, a good partition yields groups with high values for the separation between classes and the density of classes. The index is a function of the ratio between the sum of the inter-class dispersion and the between-class separation (equation (8)).

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left\{\frac{S_i + S_j}{d\left(z_i, z_j\right)}\right\}, \tag{8}$$

Where: K is the number of groups; $S$, the average distance of all stations in cluster to the centroid of group; and $d\left(z_i, z_j\right)$, the Euclidean distance between the centers of groups. The Davies-Bouldin index yields low values for good clusters, thereby denoting compact and well-separated groups.

The Dunn index identifies sets of groups that are compact and well-separated.

$$D = \min_{1\leq i\leq K}\left\{\min_{1\leq i\leq K, j\neq i}\left\{\frac{\delta\left(C_i, C_j\right)}{\max\limits_{1\leq i\leq K}\Delta\left(C_k\right)}\right\}\right\}, \tag{9}$$

Where: $\delta\left(C_i, C_j\right)$ represents the Euclidean distance between clusters $C_i$ and $C_j$ calculated in equation (10), and $\Delta\left(C_k\right)$ represents the diameter of group $C_k$ given in equation (11). The value for which $D$ is maximized is taken as the ideal number of clusters.

$$\delta\left(C_i, C_j\right) = \max_{x_i\in C_i, x_j\in C_j}\left\{d\left(x_i, x_j\right)\right\} \tag{10}$$

$$\Delta\left(C_k\right) = \max_{x_i, x_j\in C_k}\left\{d\left(x_i, x_j\right)\right\}, \tag{11}$$

where $d\left(x_i, x_j\right)$ is the Euclidean distance between stations $x_i$ and $x_j$.

### *2.3 Measurement of heterogeneity*

Measurements of heterogeneity Hn (n=1,2,3) are used to assess the spatial homogeneity of regions with the same underlying distribution and different local scale factors (HOSKING and WALLIS, 1997). The observed dispersions and the simulated L-moments for a given group of stations are used in this process. A region is considered to be "acceptably homogeneous" if Hn≤1, "possibly heterogeneous" if 1<Hn<2, and "definitely heterogeneous" if Hn>2. Large positive values of H1 indicate that the observed L-moments are more dispersed than the values predicted by the homogeneity hypothesis. H2 measures the similarity between local and regional estimates, and large H2 values indicate large deviations between local and regional estimates. H3 indicates the alignment of regional and local estimates. H1 is the main measurement of heterogeneity because H2 and H3 rarely produce values greater than 2, even for roughly heterogeneous regions (HOSKING and WALLIS, 1997; YANG et al., 2010; NGONGONDO et al., 2011). Details for the calculations of Hn measurements are presented in Hosking and Wallis (1997).

### *2.4 Verification of the quality of groups*

The discordancy measure ($D_i$) from Hosking and Wallis (1997) was used as a quality-control tool to identify the stations for which the sample L-moments were significantly different from the pattern observed in the other locations of a region.

Outlier stations were removed from the group, and the measurements of heterogeneity and discordancy were recalculated. This procedure was repeated until no outlier stations were observed.

### Step 3: Selection of the regional frequency distribution

When analyzing a large geographical area that has been divided into various homogeneous regions, specifying the frequency distribution of a region can affect the distributions of other regions. If a particular distribution fits well with the data of the majority of the regions, this distribution can be used for all regions, even though this distribution may not produce the best fit for data from one or more regions. In these cases, instead of using a probabilistic model with three parameters**,** either a Kappa distribution with four parameters or a Wakeby distribution with five parameters can be selected, and both are more robust against incorrect specifications of regional frequency curves (HOSKING and WALLIS, 1997)

Some advantages of using Wakeby distribution are pointed out by Hosking and Wallis (1997): (i) it can mimic the shapes of many commonly used skew distributions (e.g., extreme-value, Log-Normal, Pearson type III); (ii) there is a heavy upper tail and it can give rise to data sets containing occasional high outliers.

### Step 4: Estimates of regional quantiles

At the index-flood method, if $q(F)$ is the dimensionless T-year monthly precipitation value estimated for the homogeneous region with $N$ sites, and $\mu_i$ is the index flood for site i, then the estimate of the T-year event at-site i, $Q_i(F)$, can be described by Saf (2010) (equation (12)):

$$Q_i(F) = \mu_i q(F) \tag{12}$$

In this study, $\mu_i$ is supposed to be the mean of monthly rainfall at-site frequency distribution, and $q(F)$ is the regional monthly quantile of non-exceeding probability F. The sample mean at-site i estimates that a given month is $\hat{\mu}_i = \sum_{j=1}^{31} Q_j \Big/ 31$, and the dimensionless rescaled data are $q_{ij} = Q_{ij} / \hat{\mu}_i$ , $j = 1,2,\ldots,31$, $i = 1,2,\ldots,N$ are the basis for estimating $q(F)$.

Hosking and Wallis (1997) appraised an index-flood method in which the parameters are individually estimated at each site and suggested using a weighted average of the at-site estimates. So, in this study, the stations have the same amount of data:

$$\hat{\theta}_k^R = N^{-1}\sum_{i=1}^{N}\hat{\theta}_k^i \tag{13}$$

Where $N$ is the number of stations, $\hat{\theta}_k^i$ is the L-moment of interest. Substituting these estimates into $q(F)$ produces the estimated regional quantile $q(F) = q(F;\theta_1^R,\ldots,\theta_p^R)$.

## 5 RESULTS AND DISCUSSION

### 5.1 Formation of homogenous groups

5.1.1 Hierarchical algorithms

As shown in Table 1, for the Simple, Complete, and Average connection methodologies, the cophenetic coefficients were greater than 0,7, which indicates that these methods produce less distortion in the distance matrix than the other approaches.

**Table 1.** Cophenetic correlation coefficient of clustering hierarchical methodologies

| Method | Cophenetic correlation coefficient |
|---|---|
| Centroid | 0,65 |
| Simple | 0,71 |
| Complete | 0,75 |
| Average | 0,80 |
| Ward | 0,55 |

The Ward and Centroid methods showed the worst performance, which indicates that these approaches increase the distortion of the distance matrix in each step of the algorithm, especially in the first step.

As shown in Figure 1, the Simple, Centroid, and Ward methodologies display monotonicity in the objective function (Eq. 6), which makes locating the local minima difficult. In contrast, the Complete and Average methodologies yield easily identifiable oscillations that facilitate the localization of local minima. Despite these results, the Simple, Centroid, Average, and Complete linking methodologies do not produce similarly sized and high-quality groups for subsequent uses because they are sensitive to atypical values associated with months with extreme precipitation values. Rao and Srinivas (2006b) found that the simple linking methodology tended to form one large group and various small groups, and they concluded that the cophenetic coefficient inefficiently identifies the number of groups.

The Ward methodology yielded the best results. At each step, the algorithm divided the groups with the greatest number of stations, thereby creating more homogenous groups of a quality suitable for subsequent analyses. The analysis of the relevance of the number of groups was performed using this method alone, Figure 2.

**Figure 1**. Graph of the objective function for the hierarchical clustering methodologies

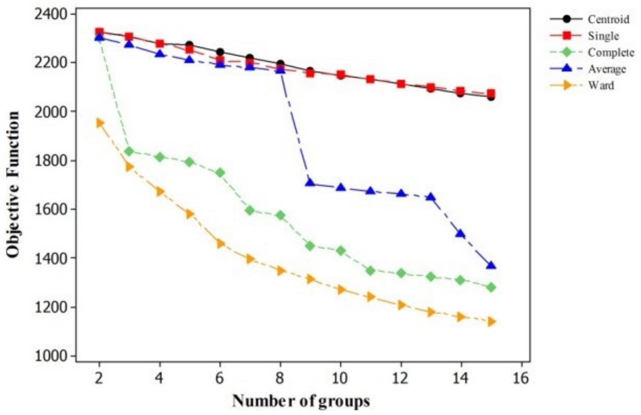
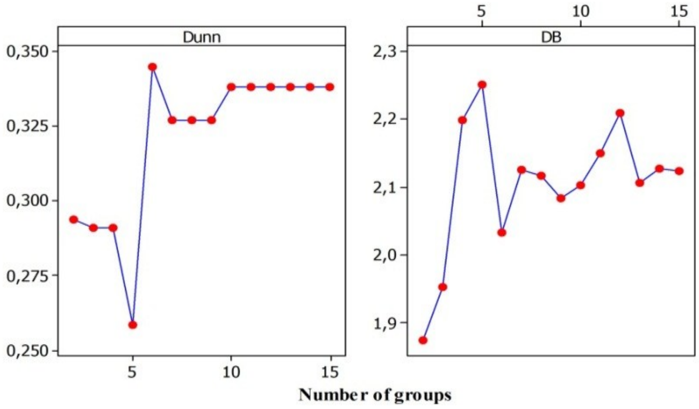


**Figure 2.** Dunn and Davies-Bouldin Indices as function of the number of groups



For the Dunn index, the ideal partition was easily identifiable. In this case, the value was six because the index was at a global maximum at this point. The Davies-Bouldin index yielded a global minimum for two clusters and a local minimum at six clusters; however, two groups did not produce selections with a large number of stations. These selections were therefore not useful for hydrological regionalization.

5.1.2 K-means algorithm

The objective function (Figure 3 and Eq. 6) behaviors of the hybrid Centroid, Simple, Complete, and Average methods were similar. The hybrid Ward and k-means method produced the lowest values for the objective function, independent of the number of groups; in addition, the curves overlapped and visibly indicated monotonicity. A comparison of the hybrid Centroid, Simple, and Average methodologies with the hierarchical algorithms revealed a sensitive reduction in the objective function. For the hybrid Ward and Complete methodologies, no substantial reduction in the objective function was observed compared with the hierarchical algorithms. The groups and objective functions of the k-means, hybrid Ward, and hierarchical Ward methods were similar. In contrast, the other hierarchical methodologies were significantly affected by the k-means method (i.e., the hybrid algorithm). This conclusion is similar to that of Rao and Srinivas (2006b).

Only the k-means and hybrid Ward methodologies were used to evaluate the pertinence of the number of groups, as they did not display a tendency to form groups with less than five stations, which is necessary when calculating the discordancy measure.

The Dunn index varied (Figure 5) the most between 5 and 6 groups and that its global maximum appeared between 10 and 13 groups. The Davies-Bouldin index had a global minimum for two groups and a local minimum for six groups, and the global maximum appeared with 5 groups. Thus, the indices show that six groups produced the best partition for the hybrid Ward method.

**Figure 3.** Objective functions for k-means methodology and its hybrids forms

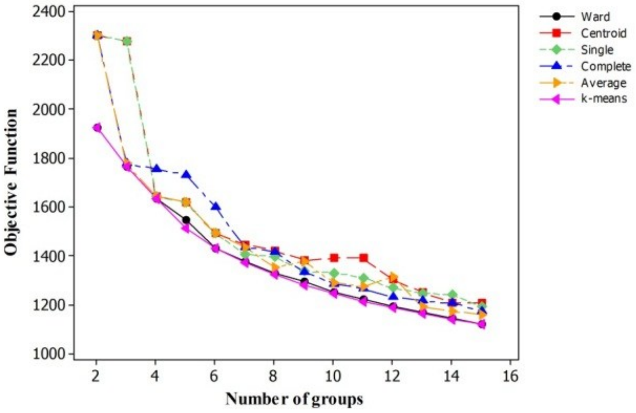
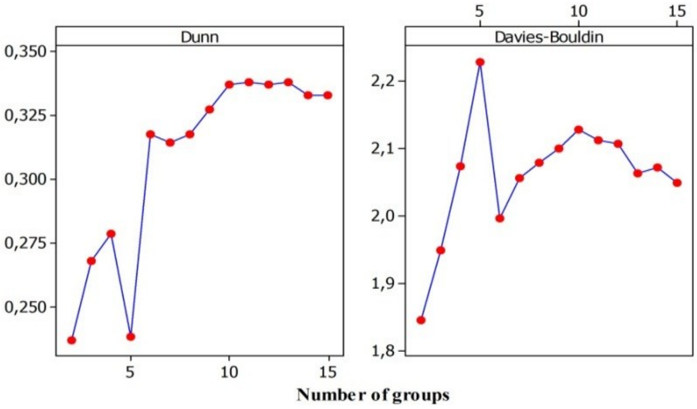
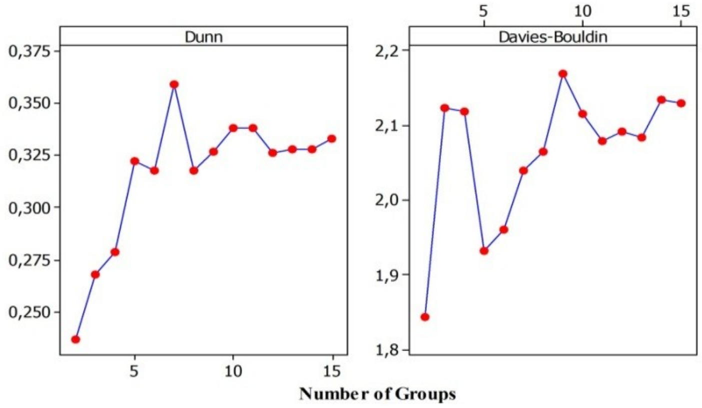


**Figure 4.** Dunn and Davies-Bouldin indices as functions of the number of groups for the hybrid Ward methodology



In Figure 5, the Dunn index has a global maximum with seven groups, whereas the Davies-Bouldin index has a global minimum for two groups and a local minimum for five groups. In addition, it has a global maximum with nine groups. For the k-means methodology, the indices did not corroborate for the same number of groups; thus, the three solutions were checked. This is because the k-means method uses random seeds to clustering.

**Figure 5.** Dunn and Davies-Bouldin indices as functions of the number of groups for k-means methodology



## 5.2 Tests for discordancy and heterogeneity

Five solutions were obtained: one for the hierarchical Ward method, one for its hybrid form, and three for the k-means method. Five, six, seven, and nine groups were obtained as the ideal partitions. Each group was tested for heterogeneity and discordancy. The tests were conducted with 372 records for all series, and no data were preferred for any specific period.

### 5.2.1 Hierarchical algorithms

Table 2 shows the six outlier stations; less than 3% of the stations were outliers. Groups three and six did not contain outlier stations. Following the discordancy test, the groups were tested for heterogeneity, as shown in Table 2, groups 1, 3, and 6 can be considered to be homogenous, whereas group 4 can be considered to be acceptably homogenous according to the heterogeneity measure. Groups 2 and 5 should be considered to be definitely heterogeneous.

**Table 2.** Discordancy and heterogeneity tests for solution with six groups obtained by hierarchical Ward algorithm

| Group | N | Outlier station (ID) | D | H1 | Implication of H1 test |
|---|---|---|---|---|---|
| 1 | 30 | x38<br>x102 | 3,11<br>3,18 | 1,89 | possibly heterogeneous |
| 2 | 82 | x109 | 6,24 | 4,10 | definitely heterogeneous |
| 3 | 36 | - | - | 1,89 | possibly heterogeneous |
| 4 | 25 | x212 | 3,24 | 0,39 | acceptably homogeneous |
| 5 | 49 | x182<br>x196 | 4,03<br>6,71 | 5,39 | definitely heterogeneous |
| 6 | 5 | - | - | 1,21 | possibly heterogeneous |

Saf (2010) indicated that outlier stations are the main source of errors in regionalization. Thus, to evaluate the performance improvement in the heterogeneity measurements, the outlier stations were removed, and new discordancy and heterogeneity tests were performed to yield new outlier stations. The procedure was repeated until outlier stations were no longer obtained, Table 3.

**Table 3.** Outlier stations and heterogeneity measurement for hierarchical Ward methodology with six groups

| Group | Outlier Station (ID) | D | H1 | Implication of H1 test | Implication of outlier station |
|---|---|---|---|---|---|
| 1 | x103 | 3,45 | -0,12 | acceptably homogeneous | remove |
|  | - | - | -0,11 | acceptably homogeneous | do not remove |
| 2 | x77<br>x49 | 3,55<br>3,07 | 4,08 | definitely heterogeneous | remove |
|  | - | - | 4,10 | definitely heterogeneous | do not remove |
| 4 | x106 | 3,29 | -0,65 | acceptably homogeneous | remove |
|  | - | - | -0,75 | acceptably homogeneous | do not remove |
| 5 | x195 | 3,37 | 1,07 | possibly heterogeneous | remove |
|  | - | - | 0,14 | acceptably homogeneous | do not remove |

Note: Group 6 was not evaluated again because it had the minimum number of stations in discordancy measure, as indicated by Hosking and Wallis (1997). N – Number of stations, ID – Identifier, D – Discordancy measure, H1 – heterogeneity measure

A considerable improvement was observed in the heterogeneity measurements, except for group 2, even after the outlier stations were removed. Groups 1, 3, and 6 changed from possibly homogenous (Table 2) to acceptably homogenous (Table 3). Group 5 showed the largest effect upon removal of the outlier stations; with the removal of these stations, the group changed from definitely heterogeneous (Table 2) to acceptably homogenous (Table 3). Approximately 5% of the studied stations were removed.

For group 2, no improvements in the performance of the heterogeneity measurement were observed. This result may be attributed to the group's size. To solve this problem, the stations in this group were classified using the Ward method into two groups. After this division, the stations were again subjected to tests of discordancy and heterogeneity (Table 4).

**Table 4.** Outlier stations and heterogeneity measurement for subdivisions of group 2 obtained by hierarchical Ward methodology with six groups

| Group | N | Outlier station (ID) | D | H1 | Implication of H1 test |
|---|---|---|---|---|---|
| 2a | 39 | - | - | 0,22 | acceptably homogeneous |
| 2b | 40 | - | - | 1,44 | possibly heterogeneous |

N – Number of stations, ID – Identifier, D – Discordancy measure, H1 – heterogeneity measure

Outlier stations were not found in the subdivision of group 2. Group 2a was classified as acceptably homogenous, whereas group 2b was classified as possibly homogenous. Thus, the subdivision of group 2 revealed two groups that were suitable for hydrological regionalization. This result is in agreement with Cannarozzo et al. (2009), who showed that even after the elimination of all outlier stations, homogeneity was not obtained, thus necessitating reclassification to obtain homogeneity.

### 5.2.2 K-means algorithm

For the k-means algorithm, the solutions were tested with seven and nine groups. With five groups, homogenous groups could not be formed by the removal of the outlier stations because one of the groups had 90 stations. Subdivision of such a group would have been required to obtain homogeneity, as demonstrated for Ward's hierarchical methodology.

**Table 5.** Outlier stations and heterogeneity measurement for solution with seven groups obtained by the k-means

| Group | N | Outlier Station (ID) | D | H1 | Implication of H1 test | Implication of outlier station |
|---|---|---|---|---|---|---|
| 1 | 47 | x38<br>x46<br>x102<br>x103 | 3,97<br>3,15<br>3,47<br>3,20 | 2,64 | definitely heterogeneous | remove |
|  | 43 | - | - | 0,74 | acceptably homogeneous | do not remove |
| 2 | 18 | x117 | 3,10 | 0,34 | acceptably homogeneous | remove |
|  | 17 | - | - | -1,18 | acceptably homogeneous | do not remove |
| 3 | 18 | x109 | 4,58 | 0,17 | acceptably homogeneous | remove |
|  | 17 | - | - | 0,06 | acceptably homogeneous | do not remove |
| 4 | 23 | x212 | 3,02 | 0,64 | acceptably homogeneous | remove |
|  | 22 | x106 | 3,01 | -0,40 | acceptably homogeneous | remove |
|  | 21 | - | - | -0,53 | acceptably homogeneous | do not remove |
| 5 | 28 | - | - | 0,85 | acceptably homogeneous | do not remove |
| 6 | 46 | x182<br>x196 | 4,11<br>7,03 | 4,74 | definitely heterogeneous | remove |
|  | 44 | - | - | -0,30 | acceptably homogeneous | do not remove |
| 7 | 5 | - | - | 1,21 | possibly heterogeneous | do not remove |

N – Number of stations, ID – Identifier, D – Discordancy measure, H1 – heterogeneity measure

As shown in Table 5, ten stations were removed, representing less than 5% of the total of 227 stations. Initially, groups 2, 3, 4, and 5 were classified as acceptably homogenous and thus remained after the removal of the outlier stations. However, for group 5, the measurement of heterogeneity became too negative. For groups 1 and 6, which were initially classified as definitely heterogeneous, a significant improvement in the heterogeneity measurement was observed following the removal of the outlier stations, and the classification was acceptably homogenous. Group 7 did not contain outlier stations and was classified as possibly homogenous.

**Table 6.** Outlier stations and heterogeneity measurement for solution with nine groups obtained by the k-means

| Group | N | Outlier Station (ID) | D | H1 | Implication of H1 test | Implication of outlier station |
|---|---|---|---|---|---|---|
| 1 | 38 | x38 | 3,76 | 1,83 | possibly heterogeneous | remove |
|  | 37 | - | - | 0,89 | acceptably homogeneous | do not remove |
| 2 | 31 | x76 | 3,08 | -0,33 | acceptably homogeneous | remove |
|  | 30 | - | - | -0,56 | acceptably homogeneous | do not remove |
| 3 | 29 | - | - | -0,43 | acceptably homogeneous | do not remove |
| 4 | 21 | - | - | -0,10 | acceptably homogeneous | do not remove |
| 5 | 23 | x212 | 3,02 | 0,64 | acceptably homogeneous | remove |
|  | 22 | x106 | 3,01 | -0,40 | acceptably homogeneous | remove |
|  | 21 | - | - | -0,53 | acceptably homogeneous | do not remove |
| 6 | 28 | x182<br>x196 | 3,52<br>4,76 | 4,32 | definitely heterogeneous | remove |
|  | 26 | x195 | 3,24 | -0,76 | acceptably homogeneous | remove |
|  | 25 | - | - | -1,95 | acceptably homogeneous | do not remove |
| 7 | 26 | x109 | 4,27 | -0,67 | acceptably homogeneous | remove |
|  | 25 | - | - | -0,99 | acceptably homogeneous | do not remove |
| 8 | 5 | - | - | 1,21 | possibly heterogeneous | do not remove |
| 9 | 26 | - | - | -0,20 | acceptably homogeneous | do not remove |

N – Number of stations, ID – Identifier, D – Discordancy measure, H1 – heterogeneity measure

### 5.2.3 Hybrid algorithm

As shown in Table 7, to obtain homogenous groups, 13 stations were eliminated, representing approximately 6% of the stations. Initially, groups 2, 3, and 4 were classified as acceptably homogenous; group 1 was classified as possibly homogenous; and group 5 was classified as definitely heterogeneous. Following the removal of the outlier stations, groups 1 and 5 became acceptably homogenous, and the heterogeneity measurements of these groups, especially group 5, improved significantly. For groups 2, 3, and 4, no significant improvements in the heterogeneity measurements were observed, as these groups were already classified as acceptably homogenous and thus remained after the removal of the outlier stations. Group 6 was classified as possibly homogenous. Outlier stations were not observed in this group.

The stations x38, x106, x109, x182, x196, and x212 are (Table 8) outliers in all tested situations. Therefore, these stations are removed in subsequent analyses, as their elimination is independent of the clustering algorithm and the number of groups adopted.

**Table 7.** Outlier stations and heterogeneity measurement for solution with six groups obtained by hybrid Ward algorithm

| Group | N | Outlier Station (ID) | D | H1 | Implication of H1 test | Implication of outlier station |
|---|---|---|---|---|---|---|
| 1 | 40 | x102<br>x103<br>x38 | 3,07<br>3,05<br>3,89 | , | possibly heterogeneous | remove |
| | 37 | - | - | -0,09 | acceptably homogeneous | do not remove |
| 2 | 58 | x77 | 3,12 | 0,50 | acceptably homogeneous | remove |
| | 57 | - | - | 0,50 | acceptably homogeneous | do not remove |
| 3 | 51 | x109 | 6,06 | 0,32 | acceptably homogeneous | remove |
| | 50 | x124 | 3,01 | 0,30 | acceptably homogeneous | remove |
| | 49 | x107 | 3,02 | 0,20 | acceptably homogeneous | remove |
| | 48 | x54 | 3.],27 | -0,31 | acceptably homogeneous | remove |
| | 47 | - | - | -0,33 | acceptably homogeneous | do not remove |
| 4 | 23 | x212 | 3,02 | 0,64 | acceptably homogeneous | remove |
| | 22 | x106 | 3,01 | -0,40 | acceptably homogeneous | remove |
| | 21 | - | - | -0,53 | acceptably homogeneous | do not remove |
| 5 | 50 | x182<br>x196 | 4,16<br>7,02 | 5,35 | definitely heterogeneous | remove |
| | 48 | x195 | 3,96 | 0,57 | acceptably homogeneous | remove |
| | 47 | - | - | -0,47 | acceptably homogeneous | do not remove |
| 6 | 5 | - | - | 1,21 | possibly heterogeneous | do not remove |

N – Number of stations, ID – Identifier, D – Discordancy measure, H1 – heterogeneity measure

**Table 8.** Outlier stations as function of clustering methodology and number of groups

| k-means | | Hybrid | Ward |
|---|---|---|---|
| k=9 | k=7 | k=6 | k=6 |
| x38 | x38 | x38 | x38 |
| | x102 | x102 | x102 |
| | x103 | x103 | x103 |
| x106 | x106 | x106 | x106 |
| x109 | x109 | x109 | x109 |
| x196 | x196 | x196 | x196 |
| x212 | x212 | x212 | x212 |
| x182 | x182 | x182 | X182 |
| x195 | | x195 | x195 |
| | | x77 | x77 |
| | x46 | x54 | x49 |
| | x117 | x107 | |
| x76 | | x124 | |
| 8 | 10 | 13 | 11 |

k= number of groups

### 5.3 Regional estimates of monthly precipitation

In RFA, rainfall estimates for locations with no data are often needed, as seen in Table 9. In this study, values that represented entire homogenous regions were estimated.

**Figure 7.** Delineation of the six homogenous regions

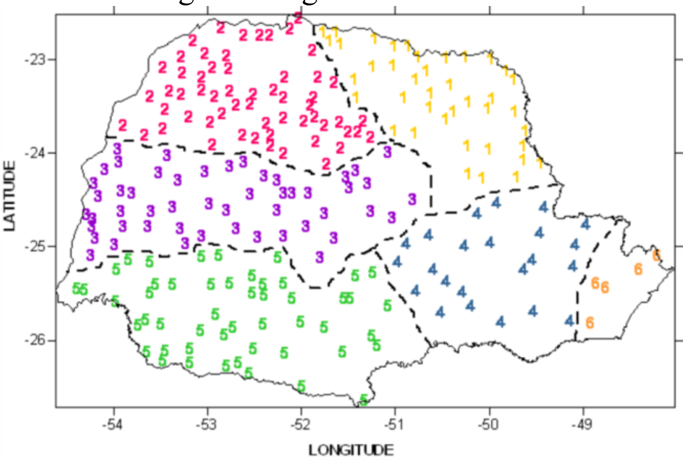


**Table 9.** Regional estimates of monthly rainfall (mm)

| | Group 1 | | | | | | Group 2 | | | | | | Group 3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Return period (years) | | | | | | Return period (years) | | | | | | Return period (years) | | | | | |
| Month | 2 | 5 | 10 | 20 | 50 | 100 | 2 | 5 | 10 | 20 | 50 | 100 | 2 | 5 | 10 | 20 | 50 | 100 |
| Jan | 182 | 296 | 367 | 427 | 492 | 533 | 168 | 266 | 326 | 374 | 426 | 457 | 180 | 268 | 321 | 367 | 421 | 456 |
| Feb | 152 | 235 | 283 | 320 | 359 | 382 | 147 | 220 | 269 | 312 | 363 | 397 | 151 | 238 | 288 | 329 | 371 | 397 |
| Mar | 117 | 178 | 221 | 264 | 322 | 366 | 116 | 173 | 211 | 247 | 291 | 322 | 115 | 184 | 230 | 272 | 319 | 351 |
| Apr | 90 | 139 | 168 | 192 | 216 | 231 | 101 | 158 | 201 | 244 | 305 | 352 | 123 | 196 | 249 | 305 | 382 | 442 |
| May | 92 | 180 | 233 | 277 | 323 | 352 | 100 | 212 | 289 | 358 | 440 | 496 | 143 | 285 | 368 | 435 | 504 | 545 |
| Jun | 63 | 129 | 173 | 218 | 286 | 348 | 76 | 152 | 193 | 226 | 269 | 309 | 110 | 178 | 228 | 280 | 350 | 405 |
| Jul | 43 | 91 | 124 | 156 | 195 | 222 | 51 | 87 | 112 | 140 | 183 | 222 | 82 | 145 | 187 | 227 | 278 | 313 |
| Aug | 30 | 89 | 130 | 166 | 206 | 231 | 39 | 99 | 145 | 191 | 253 | 299 | 68 | 142 | 184 | 217 | 250 | 269 |
| Sep | 112 | 168 | 198 | 227 | 269 | 306 | 119 | 197 | 243 | 283 | 330 | 360 | 147 | 228 | 279 | 327 | 389 | 435 |
| Oct | 114 | 179 | 218 | 250 | 285 | 306 | 133 | 222 | 269 | 302 | 333 | 348 | 183 | 261 | 312 | 366 | 443 | 506 |
| Nov | 120 | 184 | 229 | 272 | 326 | 364 | 123 | 195 | 242 | 284 | 331 | 362 | 156 | 227 | 280 | 335 | 408 | 466 |
| Dec | 170 | 249 | 299 | 344 | 397 | 432 | 168 | 245 | 296 | 342 | 396 | 432 | 168 | 255 | 315 | 371 | 439 | 487 |
| | Group 4 | | | | | | Group 5 | | | | | | Group 6 | | | | | |
| | Return period (years) | | | | | | Return period (years) | | | | | | Return period (years) | | | | | |
| Month | 2 | 5 | 10 | 20 | 50 | 100 | 2 | 5 | 10 | 20 | 50 | 100 | 2 | 5 | 10 | 20 | 50 | 100 |
| Jan | 183 | 257 | 311 | 365 | 440 | 497 | 182 | 269 | 322 | 365 | 412 | 440 | 340 | 495 | 594 | 680 | 776 | 838 |
| Feb | 151 | 213 | 252 | 287 | 329 | 359 | 172 | 265 | 315 | 357 | 401 | 429 | 310 | 444 | 508 | 552 | 589 | 607 |
| Mar | 114 | 177 | 222 | 266 | 324 | 367 | 121 | 195 | 245 | 289 | 342 | 377 | 262 | 365 | 430 | 485 | 544 | 581 |
| Apr | 83 | 124 | 155 | 189 | 236 | 275 | 135 | 208 | 267 | 332 | 425 | 503 | 158 | 244 | 295 | 336 | 379 | 405 |
| May | 92 | 200 | 279 | 354 | 448 | 516 | 149 | 304 | 396 | 472 | 551 | 598 | 111 | 218 | 293 | 364 | 452 | 514 |
| Jun | 94 | 148 | 185 | 226 | 285 | 335 | 145 | 225 | 273 | 315 | 360 | 389 | 95 | 159 | 203 | 244 | 294 | 329 |
| Jul | 95 | 152 | 185 | 218 | 268 | 312 | 116 | 172 | 222 | 290 | 419 | 558 | 117 | 197 | 240 | 279 | 333 | 378 |
| Aug | 65 | 121 | 153 | 181 | 215 | 239 | 93 | 172 | 211 | 243 | 287 | 329 | 78 | 133 | 173 | 212 | 263 | 300 |
| Sep | 131 | 215 | 264 | 303 | 343 | 367 | 154 | 251 | 301 | 338 | 371 | 389 | 166 | 241 | 304 | 371 | 470 | 551 |
| Oct | 142 | 217 | 256 | 286 | 313 | 328 | 210 | 315 | 383 | 442 | 508 | 550 | 183 | 251 | 301 | 352 | 417 | 466 |
| Nov | 117 | 172 | 214 | 258 | 320 | 371 | 167 | 235 | 282 | 341 | 449 | 561 | 178 | 260 | 323 | 388 | 478 | 548 |
| Dec | 148 | 214 | 252 | 283 | 316 | 335 | 168 | 247 | 299 | 347 | 405 | 444 | 219 | 341 | 432 | 521 | 635 | 720 |

## 6 CONCLUSIONS

The obtained answers in this study can state that the hybrid form between Ward algorithm and k-means showed the best results, since the removal of discordant stations allowed the generation of groups of homogenous stations regarding L-moments.

Another important issue recorded in this trial was the effect of discordant stations on L-moments homogeneity according to the studied region as well as the number of stations that make part of the group.

Finally, this study showed problems that can occur when there is a direct choice of the clustering method in RFA.

## 7 REFERENCES

BEAVER, S.; PALAZOĞLU, A. A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. **Atmospheric Environment**, v.40, p. 713-725, 2006.

CANNAROZZO, M.; NOTO, L. V.; VIOLA, F.; LA LOGGIA, G. Annual runoff regional frequency analysis in Sicily. **Physics and Chemistry of the Earth**, v.34, p. 679-687, 2009.

CHENG, C. L.; LIAO, M. C. Regional rainfall level zoning for rainwater harvesting systems in northern Taiwan. **Resources Conservation and Recycling**, v. 53, p. 421-428, 2009.

CORTÉS, J. A.; PALMA, J. L.; Wilson, M. Deciphering magma mixing: The application of cluster analysis to the mineral chemistry of crystal populations. **Journal of Volcanology and Geothermal Research**, v. 165, p. 163-188, 2007.

CUNDERLIK, J. M.; BURN, D. H. Site-focused nonparametric test of regional homogeneity based on flood regime. **Journal of Hydrology**, Amsterdam, v. 318, p. 301-315, 2006.

GARCIA, H. L.; GONZALEZ, I. M.; Self-organizing map and clustering for wastewater treatment monitoring. **Engineering Applications of Artificial Intelligence**, v.17, p.215-225, 2004.

HOSKING, J. R. M., WALLIS, J. R. **Regional frequency analysis – an approach based on L-moments**. Cambridge University Press, New York, p. 224, 1997.

MODARRES, R.; SARHADI, A. Statistically-based regionalization of rainfall climates of Iran. **Global and Planetary Change**, v. 75, p. 67-75, 2011.

NGONGONDO, C. S.; XU, C-Y.; TALLAKSEN, L. M.; Alemaw, B.; Chirwa, T. Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches. **Stochastic Environmental and Research Risk Assessment**, v. 25, p. 939-955, 2011.

OUARDA, T. B. M. J.; BÂ, K.M.; DIAZ-DELGADO, C.; CARSTEANU, A.; CHOKMANI, K.; GINGRAS, H.; QUENTIN, E.; TRUJILLO, E.; BOBEE, B. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. **Journal of Hydrology**, Amsterdam, v. 348, p. 40-58, 2008.

RAO, A. R.; SRINIVAS, V. V. Regionalization of watersheds by fuzzy cluster analysis. **Journal of Hydrology**, Amsterdam, v. 318, p. 57-79, 2006(a).

RAO, A. R.; SRINIVAS, V. V. Regionalization of watersheds by hybrid-cluster analysis. **Journal of Hydrology**, Amsterdam, v. 318, p. 37-56, 2006(b).

SAF, B. Assessment of the effects of discordant sites on regional flood frequency analysis. **Journal of Hydrology**, Amsterdam , v. 380, p. 362-375, 2010.

YANG, T.; SHAO, Q.; HAO, Z. C.; CHEN, X.; ZHANG, Z.; XU, C. Y.; SUN, L. Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. **Journal of Hydrology**, Amsterdam, v. 380, p. 386-405, 2010.