Probabilistic Clustering of Interval Data

Paula Brito[1] *, A. Pedro Duarte Silva[2], José G. Dias[3]

December 22, 2013

[1] FEP & LIAAD INESC TEC, Universidade do Porto, PORTUGAL

[2] FEG and CEGE, Universidade Católica Portuguesa (Porto), PORTUGAL

[3] Instituto Universitário de Lisboa (ISCTE-IUL), BRU, PORTUGAL

Abstract

In this paper we address the problem of clustering interval data, adopting a model-based approach. To this purpose, parametric models for interval-valued variables are used which consider configurations for the variance-covariance matrix that take the nature of the interval data directly into account. Results, both on synthetic and empirical data, clearly show the well-founding of the proposed approach. The method succeeds in finding parsimonious heterocedastic models which is a critical feature in many applications. Furthermore, the analysis of the different data sets made clear the need to explicitly consider the intrinsic variability present in interval data.

Keywords: clustering methods; finite mixture models; interval-valued variable; intrinsic variability; symbolic data.

*Corresponding author. Fac. de Economia do Porto, Rua Dr. Roberto Frias, 4200-464 Porto, PORTUGAL, Fax: 00351-5505050; e-mail: mpbrito@fep.up.pt

1 Introduction

Symbolic Data, introduced by E. Diday in the late eighties of the last century (see, for instance, [2], [22] [34]), is concerned with the analysis of data presenting intrinsic variability, which should be explicitly taken into account. This happens, in particular, when huge data bases are aggregated on the basis of some descriptors that define groups of interest - which constitute the statistical units to be analyzed. It is also the case when the entities under analysis are not single elements, but rather classes or concepts, for instance, not a particular car, but a car model, not a particular flight, but the airport traffic, not the particular flower I am picking, but the flower species. In all these cases, we are dealing with statistical units which have inherent variability that should be taken into account. The alternative of representing variable values by central measures such as averages, medians or modes entails a too important loss of information. Symbolic Data Analysis (henceforth, SDA) provides a convenient framework to represent data with such variability. New variable types were introduced that allow for the representation of the intrinsic variability of the data.

As in the classical case, symbolic variables may be either numerical or categorical. Different kinds of numerical and categorical variables may then be considered, and classical numerical and categorical variables are just special cases of symbolic variables. A numerical (or quantitative) variable is single-valued (real or integer) if it takes one single value of an underlying domain for each entity. It is multi-valued if its values are finite subsets of the domain and it is an interval-valued variable if its values are intervals of $\mathbb{R}$. When an empirical distribution over a set of subintervals is given, the variable is called a histogram-valued variable. As in the classical context, data are presented in a matrix, or data-array, now called a "symbolic data

table", where each row corresponds to a group, or concept, i.e., the entity of interest, and each column corresponds to a "symbolic variable". Non-parametric approaches for symbolic data have been presented in e.g., [2], [22], and [34]); parametric modelling has been discussed, for instance, in [5], [6], and [9].

In this paper we propose a model-based approach for clustering interval data, extending the Gaussian models proposed in [9] to the model-based clustering context. For this purpose, we adapt the EM algorithm to the likelihood maximization in our models, for different covariance configurations. The proposed methodology is illustrated with synthetic data and further explored on real data sets with different characteristics. In recent years finite mixture aka latent class modelling has been applied extensively. Apart from other applications (e.g., density estimation, outlier detection, measurement error modelling), finite mixture models have been extensively used as a clustering technique. In this case one assumes that there is discrete population heterogeneity with K subpopulations or clusters that can be unmixed. Because, each cluster or component is characterized by a specific density function, this approach has been called model-based clustering.

The remaining of the paper is organized as follows. In Section 2 interval-valued variables are formally introduced and different representation of interval-data are considered. Parametric modelling of interval data, which will be used in the sequel, is recalled. Section 3 reviews existing proposals for the non-parametric clustering of interval data. Section 4 describes the proposed methodology for clustering interval data. Section 5 illustrates the procedure using two synthetic data sets. Section 6 reports the application of the method to three data sets of different nature and sizes. The article ends by highlighting the main conclusions, advantages of this model-based clustering

method, and desirable further extensions.

2 Interval data

Interval data occur in various contexts. When describing ranges of variable values, as it is the case, for instance, for daily stock prices or temperature ranges, we obtain *native* interval data; in the aggregation of huge data bases into groups of interest, real values describing the individual observations (the *microdata*) lead to intervals describing the groups formed; descriptions of biological species or technical specifications are often presented in the form of intervals for the different variables.

Let $S = \{s_1, \ldots, s_n\}$, be the set of n entities under analysis. Formally, an interval-valued variable is defined by an application

$$Y : S \to T \text{ such that } s_i \to Y(s_i) = [l_i, u_i]$$

where T is the set of intervals of an underlying set $O \subseteq \mathbb{R}$.

Let I be an $n \times p$ matrix representing the values of p interval-valued variables on S. Each $s_i \in S$ is represented by a p-dimensional vector of intervals, $I_i = (I_{i1}, \ldots, I_{ip}), i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}], j = 1, \ldots, p$ (see Table 1).

TABLE 1 ABOUT HERE

The value of an interval-valued variable Y_j for each $s_i \in S$ is naturally defined by the lower and upper bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modelling purposes an alternative parameterization consists in representing $Y_j(s_i)$ by the MidPoint $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$ and Range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij}.

Consider each interval I_{ij} represented by its MidPoint c_{ij} and Range r_{ij}. The Gaussian model (see [9]) assumes a multivariate Normal distribution for MidPoints C and the logs of the Ranges R, $R^* = \ln(R)$, $(C, R^*) \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_C^t, \boldsymbol{\mu}_{R^*}^t\right]^t$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{CC} & \boldsymbol{\Sigma}_{CR^*} \\ \boldsymbol{\Sigma}_{R^*C} & \boldsymbol{\Sigma}_{R^*R^*} \end{pmatrix}$ where $\boldsymbol{\mu}_C$ and $\boldsymbol{\mu}_{R^*}$ are p-dimensional column vectors of the mean values of, respectively, the MidPoints and Log-Ranges, and $\boldsymbol{\Sigma}_{CC}, \boldsymbol{\Sigma}_{CR^*}, \boldsymbol{\Sigma}_{R^*C}$ and $\boldsymbol{\Sigma}_{R^*R^*}$ are $p \times p$ matrices with their variances and covariances.

This model has the advantage of allowing for the application of classical inference methods; nevertheless it is important to keep in mind that the MidPoint c_{ij} and the Range r_{ij} of the value of an interval-valued variable $I_{ij} = Y_j(s_i)$ are two quantities related to one same variable, and must therefore be considered together. As a consequence, the global covariance matrix should take into account the link that may exist between MidPoints and Ranges of the same or different variables. Intermediate parameterizations between the non-restricted and the non-correlation setup considered for real-valued data are relevant for the specific case of interval data. In this paper, we shall consider the following cases:

1. Non-restricted case: allowing for non-zero correlations among all MidPoints and Log-Ranges;

2. Interval-valued variables Y_j are independent, but for each variable, the MidPoint may be correlated with its Log-Range: $\boldsymbol{\Sigma}_{CC}, \boldsymbol{\Sigma}_{CR^*} = \boldsymbol{\Sigma}_{R^*C}, \boldsymbol{\Sigma}_{R^*R^*}$ all diagonal;

3. MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed: $\boldsymbol{\Sigma}_{CR^*} = \boldsymbol{\Sigma}_{R^*C} = \mathbf{0}$;

4. All MidPoints and Log-Ranges are uncorrelated, both among themselves and between each other: $\mathbf{\Sigma}$ diagonal.

From the Normality assumption it obviously follows that imposing non-correlations with Log-Ranges is equivalent to imposing non-correlations with Ranges. It should be remarked that in Cases 2, 3 and 4, $\mathbf{\Sigma}$ can be written as a diagonal by blocks matrix, after a possible rearrangement of rows and columns. This is particularly important for maximum likelihood estimation. In a full complete setup another case could still be considered, namely, allowing for non-null correlation between the MidPoint of each variable and its Log-Range, but not between MidPoints and Log-Ranges of different variables. This case appears to be less natural, and leads to considerably computational complexity, and will therefore not be considered in the present investigation.

3 Non-parametric clustering

Clustering of interval data has been addressed by several authors, under non-parametric exploratory approaches.

Methods based on dissimilarities, generally adaptations of K-means, have been developed, for instance in [36], [15] and [12]. These approaches propose suitable dissimilarity measures for interval data, and then use the K-means algorithm to obtain a partition that locally optimizes a criterion measuring the fit between the cluster composition and their prototypes. In [36], a City-Block L_1 distance between intervals is used, $d_1(I_i, I_j) = |l_i - l_j| + |u_i - u_j|$, whereas in [15] a L_2 distance is considered: $d_2(I_i, I_j) = \sqrt{(l_i - l_j)^2 + (u_i - u_j)^2}$. In [12] different measures are used and results discussed. *SCLUST* (see [17]) is a module of the *SODAS* package that performs non-hierarchical clustering on symbolic data, using a K-means-

like method; for interval-data the Hausdorff distance between intervals, $d_H(I_i, I_j) = \max\{\{|l_i - l_j|, |u_i - u_j|\}$ is used by default.

Fuzzy clustering has been developed by different authors. The first fuzzy clustering method for interval data has been proposed in [23]. Other approaches followed, see [37], [14], [19], and [30]. Fuzzy K-means methods for interval data generally result from adapting the classical fuzzy c-means algorithm, using appropriate distances, as is done for the crisp algorithms.

Other extensions, using adaptive distances [16] or based on multiple dissimilarity matrices [18] have also been investigated.

A method based on Poisson point processes has been proposed in [28]. The first part of the method consists in a monothetic divisive clustering procedure where the cutting rule uses an extension of the Hypervolumes criterion to interval data. The pruning step uses two likelihood ratio tests based on the homogeneous Poisson point process, the Hypervolumes test and the Gap test, leading to a decision tree. A merging procedure then allows improving the clustering obtained in the first step.

A method for conceptual ascending hierarchical or pyramidal clustering has been proposed in [7] and [8], which may be summarized as follows: for each candidate cluster, a description is built, generalizing the descriptions corresponding to the clusters to be merged, a candidate cluster is eligible only if this new description covers all cluster elements and none other. When two given clusters are merged, it is described, for each variable, by the minimum interval that covers them. Each cluster formed is hence associated with a conjunction of properties on the descriptive variables, which constitutes a necessary and sufficient condition for cluster membership. To choose among the different aggregations meeting the above condition, a "generality degree" evaluates the proportion of the representation space covered by the

considered description; it is computed variable-wise and the values for each variable are then combined to obtain a measure of the variability of the full description. The aggregation leading to the cluster with lower generality is selected.

A monothetic clustering method using a divisive approach is proposed in [11]; each cluster formed is again associated with a conjunction of properties on the descriptive variables, constituing a necessary and sufficient condition for cluster membership. The method uses a criterion that measures intra-class dispersion using distances appropriate to interval-valued variables. The algorithm successively splits one cluster into two sub-clusters, according to a condition expressed as a binary question on the values of one variable; the cluster to be split and the condition to be considered at each step are selected based on the minimization of the intra-cluster dispersion on the next step.

Approaches that use Kohonen maps for clustering interval data have also been developed; in the *SODAS* software Kohonen maps are constructed by the module SYKSOM - see [3] and [4]. Other approaches are investigated in [26], [13], and [39].

Clustering and validation of interval data are discussed in [27].

However, none of above described proposals has taken a model-based approach.

4 Model-based clustering of interval data

Let $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ denote a sample of size n, p represent the number of interval-valued variables, and x_{ij} indicate the observed value for variable j in observation i, with $i = 1, ..., n$, $j = 1, ..., 2p$. The finite mixture model

with K components for $\mathbf{x}_i = (x_{i1}, ..., x_{i,2p})$ is defined by

$$f(\mathbf{x}_i; \boldsymbol{\varphi}) = \sum_{k=1}^{K} \tau_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k), \tag{1}$$

where component proportions τ_k are positive and sum to one; and $\boldsymbol{\theta}_k$ denotes parameters of the conditional distribution of cluster k. Model parameters are $\boldsymbol{\varphi} = (\boldsymbol{\tau}, \boldsymbol{\theta})$, with $\boldsymbol{\tau} = (\tau_1, ..., \tau_{K-1})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$. The number of free parameters in vectors $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ are $d_{\boldsymbol{\tau}} = K - 1$ and $d_{\boldsymbol{\theta}}$, respectively. The number of free parameters is $d_{\boldsymbol{\varphi}} = d_{\boldsymbol{\tau}} + d_{\boldsymbol{\theta}}$.

For continuous metric data, finite mixtures of Gaussian distributions have been extensively applied [32]. For this specification, the conditional distribution is given by $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_s$ are the mean vector and covariance matrix, respectively. For instance, heteroscedastic Case 1 contains $d_{\boldsymbol{\varphi}} = Kp(2p + 3) + K - 1$ free parameters.

Maximum likelihood (ML) parameter estimation involves the maximization of the log-likelihood function: $\ell(\boldsymbol{\varphi}; \mathbf{x}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i; \boldsymbol{\varphi})$, a problem that can be tackled by the Expectation-Maximization (EM) algorithm [20]. E-step computes the joint conditional distribution of the missing data given observed data and provisional estimates of model parameters. In the M-step, standard complete data ML methods are used to update the unknown model parameters using an expanded data matrix with the estimated densities of the missing data (posterior cluster probabilities) as weights.

An important modelling issue is the selection of the number of components (K). We use the Bayesian Information Criterion (BIC) [35] given by

$$BIC = -2\ell(\hat{\boldsymbol{\varphi}}; \mathbf{x}) + d_{\boldsymbol{\varphi}} \ln(n), \tag{2}$$

where $d_{\boldsymbol{\varphi}}$ is the number of free parameters in the model. We notice that BIC is a consistent criterion, whereas the AIC is a biased estimate of the true number of latent classes ([29], and [21]).

In model-based clustering of interval data, $X_i = \left[C_i^t, R_i^{*t}\right]^t$ is defined as the $2p$ dimensional column vector comprising all the MidPoints and Log-Ranges for s_i, and the "complete" data are considered to be $y_i = (x_i, z_i)$, where $z_i = (z_{i1}, \ldots, z_{iK})$ is assumed as the "missing" data, with

$$z_{ik} = \begin{cases} 1 & if\ observation\ x_i\ belongs\ to\ group\ k \\ 0 & otherwise \end{cases}$$

It is well known (see, *e.g*, [32], [24], [10]) that in this case the E-step consists in replacing z_{ik} by the estimated conditional probabilities, $\hat{z}_{ik}$ and the M-step consists in the maximization of

$$F(\boldsymbol{\varphi}|\mathbf{x}_1, \ldots, \mathbf{x}_n, \hat{\mathbf{z}}) \ = \ \sum_{k=1}^{K}\sum_{i=1}^{n} \hat{z}_{ik} \ln(\tau_k\ \phi(\mathbf{x}_i|\boldsymbol{\theta}_k)) \quad = \\ \sum_{k=1}^{K}\sum_{i=1}^{n} \hat{z}_{ik} \ \left(\ln\ \tau_k - p\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^t\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right) \tag{3}$$

In all models and cases, the updating formulas for τ_k and $\boldsymbol{\mu}_k$ are

$$\hat{\tau}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ik}}{n} \ \ ; \ \ \hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ik}\mathbf{x}_i}{\sum_{i=1}^{n}\hat{z}_{ik}} \ \ ; \tag{4}$$

and $\hat{\boldsymbol{\Sigma}}$ (homocedastic models), $\hat{\boldsymbol{\Sigma}}_k$ (heterocedastic models) can be updated, by maximizing respectively

$$\text{constant} \ -\frac{n}{2}\ \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\ tr\mathbf{E}\boldsymbol{\Sigma}^{-1} \tag{5}$$

$$\text{constant} \ -\frac{n_k}{2}\ \ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}\ tr\mathbf{E}_k\boldsymbol{\Sigma}_k^{-1} \tag{6}$$

with $n_k = \sum_{i=1}^{n}\hat{z}_{ik}$, $\mathbf{E}_k = \sum_{i=1}^{n}\hat{z}_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t$ and $\mathbf{E} = \sum_{k=1}^{K}\mathbf{E}_k$.

In the unrestricted case the M-step formulas for $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}}_k$ are obviously the classical ones, $\hat{\boldsymbol{\Sigma}} = \mathbf{E}/n$, $\hat{\boldsymbol{\Sigma}}_k = \mathbf{E}_k/n_k$. In [9] it is shown that when $\boldsymbol{\Sigma}$ is restricted to be a diagonal by blocks matrix, with q blocks:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{(1)} & & & \\ & \boldsymbol{\Sigma}^{(2)} & & \mathbf{0} \\ \mathbf{0} & & \dots & \\ & & & \boldsymbol{\Sigma}^{(q)} \end{pmatrix}$$

(5) is maximized when $\boldsymbol{\Sigma}^{(h)} = \hat{\boldsymbol{\Sigma}}^{(h)} = \mathbf{E}^{(h)}/n$, for $h = 1, \dots, q$ where $\mathbf{E}^{(h)}$ is the block of $\mathbf{E}$ corresponding to $\boldsymbol{\Sigma}^{(h)}$. The same argument can show that (6) is maximized by $\boldsymbol{\Sigma}_k^{(h)} = \hat{\boldsymbol{\Sigma}}_k^{(h)} = \mathbf{E}_k^{(h)}/n_k$, with $\mathbf{E}_k^{(h)}$ defined in the same way.

In our implementation, to try avoiding local optima, each search of the EM algorithm is replicated many times from different starting points. The first starting point is based on the z_{ik} estimates returned by the unrestricted model of the R *MCLUST* package [25], and the remaining ones are obtained from random perturbations applied to the current best solution.

5 Synthetic data sets

This section applies the model-based clustering procedure to two synthetic interval data sets. As the structure of the problems is known, we can further understand the performance of the proposed method.

5.1 Synthetic data set 1

The first synthetic data contains 2000 observations (n) and two interval-valued variables (p). We assume three components (K) with same size ($\tau_k = 1/3$). Conditional on the cluster (k), the simulated values are given

by $\mathbf{X}_i|k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ \ln 2 \\ 0 \\ \ln 2 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} 6 \\ \ln 4 \\ 1.5 \\ \ln 1 \end{bmatrix}, \ \boldsymbol{\mu}_3 = \begin{bmatrix} 5 \\ \ln 2 \\ 4.5 \\ \ln 2 \end{bmatrix}, \ \textit{and} \ \boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & - & 0.2 & - \\ - & 0.5 & - & 0.2 \\ 0.2 & - & 0.5 & - \\ - & 0.2 & - & 0.5 \end{bmatrix}.$$

The setting assumes homoscedasticy, i.e., clusters share the same covariance matrix. This example specifies a Case 3 configuration, where Midpoints and Log-Ranges are not associated for the same variable, but we allow that Midpoints and Log-Ranges for different variables may be associated. Figure 1 depicts the first 20 observations of this data set. It also adds the centroids of these three components.

FIGURE 1 ABOUT HERE

Table 2 reports model selection results. As competing models to the data generating process described above, we allow different homoscedastic configurations (Cases 1, 2, and 4) under the same number of components (K). BIC identifies the correct configuration, i.e., Case 3.

TABLE 2 ABOUT HERE

Model estimates show that the component sizes are well retrieved, i.e., proportions estimates ($\hat{\tau}_k$) are 0.335, 0.330, and 0.335. Observations are correctly allocated into clusters and proportions of the mixture are virtually

identical. Mean vector and covariance matrix estimates are:

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} 0.98 \\ 0.74 \\ 0.01 \\ 0.72 \end{bmatrix}, \ \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 5.00 \\ 0.69 \\ 4.50 \\ 0.70 \end{bmatrix}, \ \hat{\boldsymbol{\mu}}_3 = \begin{bmatrix} 6.00 \\ 1.42 \\ 1.54 \\ 0.01 \end{bmatrix}, \ \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.52 & - & 0.22 & - \\ - & 0.50 & - & 0.21 \\ 0.22 & - & 0.50 & - \\ - & 0.21 & - & 0.49 \end{bmatrix}.$$

We notice that $\ln 2$ and $\ln 4$ are 0.693 and 1.386, respectively. Overall, this model-based clustering method retrieves the true values.

5.2 Synthetic data set 2

Regarding the second synthetic data set, we consider four variables (p) and 1000 observations. It specifies a two-component mixture model with $\tau_k = 1/2$. This synthetic data set is simulated under Case 2, i.e., Midpoints and Log-Ranges are associated only within each variable. Conditional on the cluster (k), the simulated values are given by $\mathbf{X}_i|k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ 0 \\ -1 \\ 0 \\ -1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 & - & - & - & - & - & - \\ 0.8 & 1 & - & - & - & - & - & - \\ - & - & 1 & 0.8 & - & - & - & - \\ - & - & 0.8 & 1 & - & - & - & - \\ - & - & - & - & 1 & 0.8 & - & - \\ - & - & - & - & 0.8 & 1 & - & - \\ - & - & - & - & - & - & 1 & 0.8 \\ - & - & - & - & - & - & 0.8 & 1 \end{bmatrix}.$$

Analyzing model selection from these four configurations of the covariance structure (Table 3), we conclude that BIC identifies the correct one (Case 2).

TABLE 3 ABOUT HERE

Component-sizes are 0.509 and 0.491 for component 1 and 2, respectively. Conditional mean estimates and the unconditional covariance matrix estimate are

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} -0.95 \\ 0.05 \\ -1.02 \\ 0.01 \\ -1.01 \\ 0.02 \\ -1.02 \\ 0.001 \end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 1.02 \\ 0.01 \\ 1.06 \\ 0.07 \\ 1.08 \\ 0.09 \\ 0.98 \\ -0.01 \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.00 & 0.82 & - & - & - & - & - & - \\ 0.82 & 1.03 & - & - & - & - & - & - \\ - & - & 1.14 & 0.91 & - & - & - & - \\ - & - & 0.91 & 1.11 & - & - & - & - \\ - & - & - & - & 1.04 & 0.81 & - & - \\ - & - & - & - & 0.81 & 0.98 & - & - \\ - & - & - & - & - & - & 1.00 & 0.81 \\ - & - & - & - & - & - & 0.81 & 1.02 \end{bmatrix}.$$

Overall, this method retrieves the data generating process for Case 2. All estimates are close to the true values.

6 Applications

In this section we apply the proposed model to three data sets of different nature and size: the first one concerns meteorological data in the USA, the second one is on income and debt variables, and the last and the last comes from a Portuguese employment survey. In the first application we deal with *native* interval data, since data are directly available in the form of minima and maxima values. In the other two applications, interval data result from the aggregation of micro data. However, the resulting number of observations differs considerably between them. While the Income-debt data set

comprehends 297 observations, in the Unemployment data only 58 observations are available. This latter case illustrates a common situation where the choice of a parsimonious model is particularly important and the use of the BIC value should not be understood as a quest for the "true" model but rather as the selection of a parameter subset comprising the most relevant ones for the clustering problem. The main characteristics of these data sets are summarized in Table 4. To highlight the added value of the interval-data model-based approach, we compared our results with those provided by the well-known *MCLUST* [25] methodology for model-based clustering of real data.

Table 4 ABOUT HERE

In each case, from the interval data matrix, and for each interval-valued observation $Y_j(s_i) = [l_{ij}, u_{ij}]$, MidPoints c_{ij} and Log-Ranges r^*_{ij} were computed. Then, models were estimated for partitions with different numbers of components, in both homocedastic and heterocedastic setups, and considering the different proposed configurations for the variance-covariance matrix (Cases 1 to 4) (see Section 2). To minimize the effect of local optima, in each case 1000 different random starting points were used for the EM algorithm; BIC values were computed for each solution.

6.1 USA meteorological data

Our first application concerns temperatures and pluviosity measured in 282 meteorological stations in the USA. We consider the temperature ranges in January and July, and the annual pluviosity range measured in each station. All these values are based on 30 years averages (1970-2000). Data

were retrieved from the USA National Environmental Satellite, Data and Information Service, at http://www1.ncdc.noaa.gov/pub/data/ccd-data (files nrmmin.txt, nrmmax.txt, nrmpct.txt); Temperatures are represented in the Fahrnheit scale, Pluviosity is measured in Inches. Table 5 shows the observed data for some stations, indicating also the corresponding State.

TABLE 5 ABOUT HERE

The method described in Section 4 was applied to these data set, resulting in the BIC values given in Table 6.

TABLE 6 ABOUT HERE

The lowest BIC value is observed for the unrestricted (Case 1) heterocedastic solution with six clusters and is shown in Figure 2 (Negative Longitude values indicate Longitude West). Component proportions, componentwise mean-vectors and variances are given in Table 7.

FIGURE 2 ABOUT HERE

TABLE 7 ABOUT HERE

In Cluster 1 we find mainly the stations in the desertic inland areas, with high intrinsic temperature variability, and very low pluviosity, also with low intrinsic variability. Cluster 2, consisting of the Alaska stations, shows low temperatures, both in winter and in summer, with low intrinsic variability (low Log-Ranges). The stations in Cluster 3 are mostly in the Southeast; this cluster is the warmest in the summer. Cluster 4, formed by stations in the Northeast and Midwest, is the second coldest in the winter. Cluster 5 groups the Pacific Islands, and San Jose of Puerto Rico, characterised by high temperatures, close to those of Cluster 3 in July but much higher in January, and high pluviosity with high intrinsic variability. The West coast stations, together with Key-West (FL) are grouped in Cluster 6, presenting mild temperatures all year round, and low pluviosity but with relatively high intrinsic variability. It becomes clear that clusters are differentiated not only by the MidPoint but also by the Log-Range variables, putting in evidence the importance of taking intrinsic variability of data into account. Moreover, Cluster 2 presents very high variance values for the January MidPoint variable, while Cluster 1 and Cluster 6 present very high variance values for the July MidPoint variable. Moreover, Cluster 2 has a high variance for the Log-Range of the pluviosity and Cluster 6 for the Log-Range of the temperature in July. This stark difference illustrates well the need of a heterocedastic setup for these data.

For comparison purposes, Ward hierarchical clustering, using the Euclidean distances on standardized data has also been applied to this data set. To decide upon the number of components to retain, we have considered the explained inertia of each partition. These values are presented in Table 8 for partitions between 2 and 10 classes. Based on the values of the explained inertia a partition in 4 clusters appears to be a natural one;

for comparison purposes we consider the partition in six clusters. Figure 3 shows the partition in 4 clusters and Figure 4 the partition in 6 clusters obtained by the Ward method.

TABLE 8 ABOUT HERE

FIGURE 3 ABOUT HERE

FIGURE 4 ABOUT HERE

These two partitions are clearly different from the one obtained by the model-based method, which appears more natural. In particular, the model-based method separates cold Alaska from warm Pacific-Islands and Puerto-Rico, which is not the case for the Ward method. This may be explained by the fact that the Ward method somehow imposes a similar covariance structure for all clusters, while some natural clusters, the Alaska one being the most obvious example, require larger variances than others. Therefore heterocedastic models seem to be required to properly model these data.

We have also applied the SCLUST algorithm from the *SODAS* package (see [22]), which is based on the K-means methodology, using the Hausdorff distance to compare interval observations. Figure 5 shows the corresponding partition in 6 clusters. As it was the case for the Ward method, this partition

does not seem very natural. The Alaska and the arid regions are well identified, but the remaining clusters are difficult to interpret. In particular, the Pacific Islands and the West coast are scattered by several groups. Probably this is again a consequence of the somehow homocedastic structure imposed.

FIGURE 5 ABOUT HERE

Finally, we applied *MCLUST* to the Midpoints data, and obtained the heterocedastic seven cluster solution shown in Figure 6.

FIGURE 6 ABOUT HERE

As it can be observed, the results are comparable to our solution, with the main exception of the Alaska cluster, which does not appear well defined.

6.2 Income-debt data

In a second application, we used survey data included as sample in the SPSS package (named "customer.dbase"). Among the large set of variables available, we focused on income and debt variables: Household Income (HI), Debt to Income Ratio ($\times$ 100) (DIR), Credit Card Debt (in thousands) (CCD) and Other Debts (OD). The 5000 individual observations have been aggregated on the basis of Gender (F, M), Age Category (18-24, 25-34, 35-49, 50-64, more than 65 years old), Level of education (did not complete high school, high-school degree, some college, college degree, post-undergraduate degree), and Job Category (managerial and professional, sales and office, service, agricultural and natural resources, precision production, craft, repair, operation, fabrication, general labour), leading to 297 groups described by the intervals of observed values on these four variables. The objective

is to investigate how the different sociological groups are related regarding income and debt variables. Table 9 shows the observed data for some groups.

TABLE 9 ABOUT HERE

The method described in Section 4 was applied to these data. BIC values are reported in Table 10.

TABLE 10 ABOUT HERE

The lowest BIC value is observed for the solution in nine components, with a heterocedastic setup and Case 2, i.e., independent interval-valued variables. Component proportions and the component-wise mean-vectors are given in Table 11.

TABLE 11 ABOUT HERE

Clusters 1 and 2 are those where groups (our statistical units) present the lower Household Incomes and lower Credit-Card Debt and Other Debts, all with the lower variability (measured by the Log-Ranges). Furthermore, Cluster 1 presents the lowest Debt-Income Ratio, with low intrinsic variability. Cluster 2, although similar to Cluster 1 in terms of Household Income, has much larger Debts, and higher variability in all variables. Groups in Cluster 6 have the second highest Household Incomes, Credit-Card Debt

and Other Debts and the largest Debt-Income Ratio, all with large variability. Cluster 9 clearly stands out as the cluster of groups where the Household Income is higher; it is also the one where the Credit-Card Debt and Other Debts are higher, all these variables presenting a high variability. The other clusters show intermediate patterns. As it may be seen in the Appendix, the estimated variance-covariance matrices are clearly different across groups, with noteworthy though different correlations between MidPoints and Log-Ranges of the same interval-valued variables. All these correlations are positive, showing that the higher the MidPoint of an interval-valued variable the higher the corresponding intrinsic variability.

MCLUST applied to the Midpoints data produced a heterocedastic five cluster solution; the consideration of Log-Ranges together with Midpoints in this example clearly allows for a finer partitioning of the groups.

6.3 Employment survey data

The third application concerns the Portuguese Employment Survey, from the 1^{st} semester of 2008. The quite large original micro data set comprehends a total of 42226 records. We only considered people who were unemployed at the time of the survey (had no job and were looking for one), i.e., 1540 cases, and focused on the two following variables: activity time (in years)(AT) and unemployment time (in months) (UT). These micro data were gathered on the basis of Gender (M, F), Region (North, Centre, Lisbon and Tagus Valley (TV), South), Age-Group (15-24, 25-44, 45-64, over 65) and Education (Basic or less, Secondary, Higher), leading to 58 sociological groups, which constitute the statistical units of interest to be analysed.

Table 12 shows the observed data for some of these groups.

TABLE 12 ABOUT HERE

The method described in Section 4 was applied to these data set, resulting in the BIC values in Table 13. Empty cells indicate cases where the number of observations does not allow the estimation of all the parameters.

TABLE 13 ABOUT HERE

The lowest BIC value is reached for the five-component solution, with a heterocedastic setup and Case 2, i.e., independent interval-valued variables. Component proportions and component-wise mean-vectors are given in Table 14.

In this application, where the number of observations is relatively low, a restricted though heterocedastic model has been identified as the best fit. This clearly illustrates the point that the method chose the best parameter configuration for clustering, preferring a heterocedastic (and therefore heavier in the number of parameters) model to a "lighter" homocedastic one, but picking up a restricted configuration for the variance-covariance matrix, where interval-valued variables are assumed independent. Choosing Case 2 as opposed to Case 3, also means that correlation between the two parts of the interval-variables is considered more important than correlation between different variables.

TABLE 14 ABOUT HERE

Table 15 shows the composition of each cluster. Cluster 2 is a cluster of young people with secondary or higher education from both genders, which explains that they still have not worked long, and are looking for a new job for a short time. Cluster 1 has similar demographics, with groups of slightly higher age. Activity time and specially unemployment time are higher, and have higher intrinsic variability (measured by the corresponding Log-Ranges). Cluster 3 is mainly formed by groups with low education, with ages between 25 and 64 years old - they have worked for a large time already and are having a very hard time in finding a new job. In Cluster 4, the activity time is high, with large intrinsic variability, but specially, the unemployment time is the highest (MidPoint more than twice as large as the second one, observed in Cluster 3). This cluster gathers groups with basic education or less and age above 45 years old, the only exception being women aged 25-44 from the North. It is known that in this region many textile companies which used to hire young women closed doors. Finally, the groups forming Cluster 5 have a long activity time, are no longer young, but have a secondary or higher education level. Therefore, although the MidPoint of Activity Time is at the same level of Cluster 3, they are looking for a job for a shorter time. It should also be noticed that clusters differ not only in terms of MidPoints, but also in terms of Log-Ranges, i.e., variability inherent to the data differs from cluster to cluster. This stresses the fact that when analysing data with intrinsic variability, the use of just a central measure (average, median) would not capture all the pertinent aspects of the reality. Furthermore, heterocedastic models were preferred to homocedastic ones, showing that different dispersions amongst clusters are also relevant - as can also be confirmed from the variance values in Table 14.

TABLE 15 ABOUT HERE

Applying *MCLUST* to the Midpoints of the two interval-valued variables produced a heterocedastic three cluster partition. Two of those clusters correspond roughly to our clusters 2 and 4; however the remaining groups are not separated, in particular the cluster of higher educated people is not identified. Again, the consideration of Log-Ranges together with Midpoints provides a finer partitioning, with important distinctive characteristics.

7 Conclusion

In this paper we proposed a probabilistic approach to the clustering of interval data. The proposed framework relies on parametrizations that take into account the inherent variability of the relevant data units and the relation that may exist between this variability and the corresponding value levels. To this aim the EM algorithm was suitably adapted. Using both synthetic and empirical data sets the pertinence of the methodology proposed was demonstrated.

In particular, its flexibility to identify heterocedastic models, even in situations with limited information, was put in evidence. Moreover, considering special configurations of the variance-covariance matrix, adapted to specific nature of interval data, proved to be the adequate approach. The presented study also made clear the need to consider both the information about position (conveyed by the MidPoints) and intrinsic variability (conveyed by the Log-Ranges) when analysing interval data.

Further research can compare this proposed framework with a multi-

level setting as MidPoints and Log-Ranges are clustered within variables. In particular, our fixed effects approach can gain additional insights by contrasting it with a random effects approach to between and within variables variability. Other research lines comprise the use of alternative models for interval data. In particular, our approach relies on a Gaussian assumption and may not be advisable when Midpoints / Log-Ranges have highly skewed distributions. In such case, our approach can be easily adapted to models based on the families of skew distributions, e.g., skew-Normal and skew-t (see [1], [31], and [38]). On the other hand to mitigate influence of possible outliers robust parameter estimators (see, e.g., [33]) may replace the traditional maximum likelihood ones. Finally, extensions to other kinds of symbolic data should be investigated.

Acknowledgements

The authors would like to thank the editor and two anonymous reviewers for their constructive comments, which helped improving the manuscript.

The first author acknowledges the support of Project NORTE-07-0124-FEDER-000059 within the North Portugal Regional Operational Programme (ON.2 - O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). The second author is grateful for the financial support from Fundação para a Ciência e Tecnologia, through project PEst-OE/EGE/UI0731/2011. The third author would like to thank the Fundação para a Ciência e a Tecnologia (Portugal) for its financial support through Grant PTDC/EGE-GES/103223/2008.

References

[1] R.M. Basso, V.H. Lachos, C.R.B. Cabral and P. Ghosh, Robust mixture modeling based on scale mixtures of skew-normal distributions, *Computational Statistics & Data Analysis*, 54 (12) (2010), 2926–2941.

[2] H.-H. Bock and E. Diday (editors), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, Heidelberg, 2000.

[3] H.-H. Bock, Clustering methods and Kohonen maps for symbolic data, *Journal of the Japanese Society of Computational Statistics*, **15 (2)** 2002, 217–229.

[4] H.-H. Bock, Visualizing symbolic data by Kohonen maps, in: *Symbolic Data Analysis and the Sodas Software*, E. Diday and M. Noirhomme-Fraiture (eds.), Wiley, 2008, pp. 205–234.

[5] H.-H. Bock, Probabilistic modeling for symbolic data, in: *COMPSTAT 2008, Proceedings in Computational Statistics*, P. Brito (ed.), Physica-Verlag, Heidelberg, 2008, pp. 55–65.

[6] H.-H. Bock, Analyzing symbolic data: problems, methods, and perspectives, in: *Cooperation in Classification and Data Analysis*, A. Okada, T. Imaizumi, H.-H. Bock, W. Gaul (eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer Verlag, Heidelberg, 2009, pp. 3–12.

[7] P. Brito, Use of pyramids in Symbolic Data Analysis, in: *New Approaches in Classification and Data Analysis*, E. Diday et al. (eds.), Springer-Verlag, Berlin-Heidelberg, 1994, pp. 378–386.

[8] P. Brito, Symbolic objects: order structure and pyramidal clustering, *Annals of Operations Research* **55** (1995), 277–297.

[9] P. Brito and A.P. Duarte Silva, Modelling interval data with Normal and Skew-Normal distributions, *Journal of Applied Statistics* **39 (1)** (2011), 3–20.

[10] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, *Pattern Recognition* **28 (5)** (1995), 781–793.

[11] M. Chavent, A monothetic clustering method, *Pattern Recognition Letters* **19 (11)** (1998), 989–996.

[12] M. Chavent, F.A.T. De Carvalho, Y. Lechevallier and R. Verde, New clustering methods for interval data, *Computational Statistics* **21 (2)** (2006), 211–229.

[13] Anderson B. dos S. Dantas and F.A.T. De Carvalho, Adaptive Batch SOM for Multiple Dissimilarity Data Tables, in: *ICTAI 2011*, 2011, pp. 575–578.

[14] F.A.T. De Carvalho, Fuzzy c-means clustering methods for symbolic interval data, *Pattern Recognition Letters* **28** (2007) 423–437.

[15] F.A.T. De Carvalho, P. Brito and H-H. Bock, Dynamic clustering for interval data based on L_2 distance, *Computational Statistics* **21 (2)** (2006), 231–250.

[16] F.A.T. De Carvalho Y. Lechevallier, Partitional clustering algorithms for symbolic interval data based on single adaptive distances, *Pattern Recognition* **4 (7)** (2009), 1223-1236.

[17] F.A.T. De Carvalho, Lechevallier, Y. and Verde, R.,Clustering methods in symbolic data analysis, in: E. Diday, E. and M. Noirhomme-Fraiture (eds.), *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester, 2008, pp. 182–203.

[18] F.A.T. De Carvalho, Y. Lechevallier and Filipe M. De Melo, Partitioning hard clustering algorithms based on multiple dissimilarity matrices, *Pattern Recognition* **45 (1)** (2012), 447–464.

[19] F.A.T. De Carvalho and C.P. Tenorio, Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances, *Fuzzy Sets and Systems* **161 (23)** (2010), 2978–2999.

[20] A.P.Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B-Methodological* **39 (1)** (1977) 1–38.

[21] J.G. Dias, Latent class analysis and model selection, in: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (eds.), *From Data and Information Analysis to Knowledge Engineering*, Springer-Verlag, Berlin, 2006, pp.95–102.

[22] E. Diday and M. Noirhomme-Fraiture (editors), *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester, 2008.

[23] Y. El-Sonbaty and M.A. Ismail, Fuzzy clustering for symbolic data, *IEEE Transactions on Fuzzy Systems*, **6 (2)** (1998), 195–204.

[24] C. Fraley and A.E. Raftery, How many clusters? Which clustering method? Answers via model-based clustering analysis, *The Computer Journal* **41 (8)** (1998) 578–588.

[25] C. Fraley, A. Raftery, T.B. Murphy and L. Scrucc, *MCLUST Version 4 for R: Normal Mixture Modelling for Model-Based Clustering, Classification and Desnsity Estimation*, Techical Report no. 597, Department of Statistics, University of Washington, 2012.

[26] C. Hajjar and H. Hamdan, Self-organizing map based on Hausdorff distance for interval-valued data, in: *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2011, pp. 1747–1752.

[27] A. Hardy and J. Baune, Clustering and validation of interval data, in P. Brito et al (eds.), *Selected Contributions in Data Analysis and Classification*, Springer, Heidelberg, 2007, pp. 69-82.

[28] A. Hardy and N. Kasaro, A new clustering method for interval data, *Mathématiques et Sciences Humaines*, **187** (2009), 79–91.

[29] C.M. Hurvich and C.L. Tsai, Regression and time series model selection in small samples, *Biometrika* **76 (2)** (1989), 25–43.

[30] J.-T. Jeng, C.-C. Chuan, C.-C. Tseng, and C.-J. Juan, Robust interval competitive agglomeration clustering algorithm with outliers, *International Journal of Fuzzy Systems*, **12 (3)**(2010), 227–236.

[31] T.-I. Lin, Robust mixture modeling using multivariate skew t distributions, *Statistics and Computing*, **20 (3)** (2010), 343–356.

[32] G.J. McLachlan, D. Peel, **Finite Mixture Models**, Wiley, New York, 2000.

[33] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev, Robust fitting of mixtures using the trimmed likelihood estimator, *Computational Statistics and Data Analysis*, **17(3)** (2007), 299–308.

[34] M. Noirhomme-Fraiture and P. Brito, Far Beyond the Classical Data Models: Symbolic Data Analysis, *Statistical Analysis and Data Mining* **4 (2)** (2011), 157–170.

[35] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics* **6** (1978), 461–464.

[36] R.M.C.R. De Souza and F.A.T. De Carvalho, Clustering of interval data based on City-Block distances, *Pattern Recognition Letters* **25 (3)** (2004), 353–365.

[37] P. D'Urso and P. Giordani, A weighted fuzzy c-means clustering model for fuzzy data, *Computational Statistics & Data Analysis*, **50 (6)** 2006, 1496–1523.

[38] I. Vrbik, P.D. McNicholas, Parsimonious skew mixture models for model-based clustering and classification, *Computational Statistics & Data Analysis*, **71** (2014), 196–210.

[39] M.-S. Yang, W.-L. Hung and D.-H. Chen, Self-organizing map for symbolic data, *Fuzzy Sets and Systems* , **203** (2012), 49–73.

APPENDIX

TABLE 15 ABOUT HERE

TABLE 16 ABOUT HERE

TABLE 17 ABOUT HERE

TABLE 18 ABOUT HERE

TABLE 19 ABOUT HERE

TABLE 20 ABOUT HERE

TABLE 21 ABOUT HERE

TABLE 22 ABOUT HERE

TABLE 23 ABOUT HERE

TABLE 1

Table 1: Matrix I of interval data

	Y_1	...	Y_j	...	Y_p
s_1	$[l_{11}, u_{11}]$	...	$[l_{1j}, u_{1j}]$	...	$[l_{1p}, u_{1p}]$
...	...		...		...
s_i	$[l_{i1}, u_{i1}]$	...	$[l_{ij}, u_{ij}]$	...	$[l_{ip}, u_{ip}]$
...	...		...		...
s_n	$[l_{n1}, u_{n1}]$	...	$[l_{nj}, u_{nj}]$	...	$[l_{np}, u_{np}]$

TABLE 2

Table 2: Model selection (Synthetic data set 1)

Cases	Log-likelihood	# parameters	BIC
Case 1	-10389.6	24	20961.5
Case 2	-10772.7	20	21697.4
Case 3	-10390.1	20	**20932.3**
Case 4	-10773.1	18	21682.9

TABLE 3

Table 3: Model selection (Synthetic data set 2)

Cases	Log-likelihood	# parameters	BIC
Case 1	-10084.0	53	20535.0
Case 2	-10096.0	29	**20393.0**
Case 3	-11985.0	37	24225.0
Case 4	-12122.0	25	24417.0

TABLE 4

Table 4: Data sets

Name	Nb. cases	Nb. interval variables	Missing values	Data type
USA metereological data	282	3	NO	Native data
Income-debt data	297	4	NO	Aggregated data
Employment survey data	58	2	NO	Aggregated data

TABLE 5

Table 5: USA meteorological data

Station	State	January Temperature	July Temperature	Annual Pluviosity
HUNTSVILLE	AL	[32.3, 52.8]	[69.7, 90.6]	[3.23, 6.10]
ANCHORAGE	AK	[9.3, 22.2]	[51.5, 65.3]	[0.52, 2.93]
NEW YORK (JFK)	NY	[24.7, 38.8]	[66.7, 82.9]	[2.70, 4.13]
...	...	...	...	...
SAN JUAN	PR	[70.8, 82.4]	[76.9, 87.4]	[2.14, 6.17]

TABLE 6

Table 6: USA meteorological data - BIC values

Nb. components	Homocedastic				Heterocedastic			
K	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
2	5569.8	6032.7	5722.0	6100.0	5078.2	5841.7	5382.4	5952.1
3	5409.3	5893.3	5575.0	5976.8	4808.8	5481.5	5102.9	5658.7
4	5311.3	5745.5	5467.7	5843.1	4678.3	5214.9	4939.7	5379.4
5	5264.3	5680.6	5379.0	5739.2	4658.7	5116.0	4871.2	5275.3
6	5239.8	5611.5	5315.2	5674.9	**4640.4**	5053.2	4863.3	5209.4
7	5184.2	5555.3	5271.8	5596.9	4662.8	5015.6	4860.8	5172.4
8	5151.9	5496.5	5224.8	5547.7	4732.3	5007.2	4842.8	5156.6
9	5125.9	5446.2	5189.4	5478.1	4762.2	4975.7	4856.6	5131.2
10	5085.5	5413.7	5152.8	5446.0	4873.3	5003.2	4866.7	5109.3

TABLE 7

Table 7: USA meteorological data - Component Proportions, Mean-Vectors and Variances

		C1	C2	C3	C4	C5	C6
	Proportions	0.176	0.084	0.175	0.413	0.053	0.098
Mean Values	MIDPT-JAN	30.64	12.51	49.64	25.41	77.65	47.20
	LNRG-JAN	3.08	2.55	3.02	2.84	2.45	2.68
	MIDPT-JUL	74.25	55.33	82.54	73.96	80.57	71.09
	LNRG-JUL	3.42	2.66	2.97	3.04	2.45	2.95
	MIDPT-PREC	1.23	3.88	4.26	3.16	7.76	2.85
	LNRG-PREC	0.38	1.15	1.24	0.77	1.78	1.39
Variances	MIDPT-JAN	94.26	263.52	53.71	76.94	14.33	69.85
	LNRG-JAN	0.07	0.05	0.01	0.02	0.04	0.06
	MIDPT-JUL	50.18	24.13	2.16	15.37	3.01	58.43
	LNRG-JUL	0.02	0.08	0.01	0.02	0.04	0.20
	MIDPT-PREC	0.20	13.03	0.93	0.43	18.14	1.37
	LNRG-PREC	0.23	0.49	0.10	0.13	0.23	0.16

TABLE 8

Table 8: Explained inertia of the partitions obtained by the Ward method

Nb. components	2	3	4	5	6	7	8	9	10
Explained inertia	0.21	0.36	0.48	0.55	0.60	0.63	0.67	0.70	0.72

TABLE 9

Table 9: Income and Debt interval data

Group	HI	DIR	CCD	OD
Male, 8-24 High school degree, Service	$[15, 61]$	$[0.1, 23.4]$	$[0.0, 6.57]$	$[0.02, 7.71]$
Male, 35-49, College degree, Sales and Office	$[19, 190]$	$[1.4, 20.4]$	$[0.04, 16.6]$	$[0.12, 15.39]$
Female, 25-34, Some college Managerial and Professional	$[17, 100]$	$[0.8, 31.7]$	$[0.05, 6.57]$	$[0.09, 7.65]$
...	...	...	...	...

TABLE 10

Table 10: Income-debt data - BIC values

Nb. comp.	Homocedastic				Heterocedastic			
K	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
2	9336.91	10312.07	10304.42	11557.28	8533.95	9222.99	9770.38	10641.34
3	9188.08	9985.10	10117.64	10733.92	7836.8	8057.74	9504.75	9890.29
4	9088.40	9546.50	9959.48	10273.03	7699.80	7772.59	9318.83	9534.24
5	9061.25	9406.52	9885.04	10124.77	7522.50	7281.54	9148.90	9341.71
6	8956.95	9366.33	9829.76	10000.48	7564.48	7055.61	9138.57	9174.65
7	8939.25	9171.05	9713.20	9897.91	—	7011.45	9160.22	9051.77
8	8902.65	9050.61	9654.98	9861.26	—	6859.76	9128.54	8977.98
9	8838.82	8992.64	9590.98	9780.71	—	**6831.01**	9278.78	8925.87
10	8852.52	9054.11	9595.48	9711.42	—	6870.38	9249.67	8924.33

TABLE 11

Table 11: Income-debt data - Component Proportions and Mean-Vectors

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Proportions	0.112	0.193	0.149	0.152	0.083	0.028	0.218	0.053	0.013
Hinc-MidP	35.72	34.25	124.31	78.86	138.88	221.78	86.88	141.69	495.38
Hinc-LogR	3.08	3.50	5.33	4.75	5.40	5.95	4.74	4.90	6.87
DIncR-MidP	7.91	12.51	15.10	13.43	13.06	17.64	11.62	11.39	16.16
DIncR-LogR	2.10	2.99	3.28	3.20	3.04	3.41	2.84	2.01	3.30
CCDbt-MidP	0.75	1.480	6.87	3.34	9.68	16.92	3.26	3.617	40.23
CCDbt-LogR	-0.06	0.93	2.57	1.87	2.90	3.45	1.71	1.46	4.29
ODbt-MidP	1.73	2.65	13.06	6.44	13.09	14.72	5.97	8.73	56.73
ODbt-LogR	0.53	1.49	3.22	2.48	3.08	3.34	2.29	2.23	4.68

TABLE 12

Table 12: Unemployment data

Group	Activity time	Unemployment time
Female, Centre, 15-24, Basic or less	$[0, 4]$	$[3, 49]$
Female, Lisbon and Tagus Valley, 45-64, Higher	$[12, 32]$	$[8, 27]$
Male, North, 15-24, Secondary	$[1, 4]$	$[1, 15]$
Male, South, 25-44, Higher	$[5, 20]$	$[4, 18]$
...	...	...

TABLE 13

Table 13: Unemployement data - BIC values

Nb. components	Homocedastic				Heterocedastic			
K	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
2	1271.3	1288.1	1335.7	1361.8	1203.2	1196.1	1296.4	1305.2
3	1242.9	1265.0	1308.7	1311.5	1137.3	1141.4	1262.6	1268.4
4	1239.0	1237.6	1290.6	1283.2	—	1098.7	1247.3	1239.5
5	1239.4	1238.0	1283.7	1283.8	—	**1080.3**	1234.4	1227.9
6	1243.7	1240.8	1279.2	1279.6	—	—	—	—
7	1241.8	1238.1	1271.8	1278.0	—	—	—	—

TABLE 14

Table 14: Unemployement data - Component Proportions, Mean-Vectors and Variances

		C1	C2	C3	C4	C5
	Proportions	0.271	0.206	0.227	0.103	0.191
Mean Values	AT MidP	8.662	3.785	23.237	33.750	26.627
	AT LogR	2.553	1.600	3.197	3.578	2.748
	UT MidP	31.985	7.495	66.990	150.500	18.869
	UT LogR	4.042	2.110	4.849	5.690	3.060
Variances	AT MidP	17.065	4.963	73.153	57.479	78.060
	AT LogR	0.340	0.544	0.119	0.040	0.182
	UT MidP	101.545	7.841	230.649	468.583	54.774
	UT LogR	0.113	0.685	0.057	0.021	0.225

TABLE 15

Table 15: Unemployment data - Clusters

CLUSTER 1	CLUSTER 2
Female, Centre, 15-24, Basic or less	Female, Centre, 15-24, Secondary
Female, Centre, 25-44, Secondary	Female, Centre, 25-44, Higher
Female, Lisbon and TV, 25-44, Secondary	Female, Lisbon and TV, 15-24, Basic or less
Female, Lisbon and TV, 25-44, Higher	Female, Lisbon and TV, 15-24, Secondary
Female, North, 15-24, Basic or less	Female, North, 15-24, Secondary
Female, North, 25-44, Secondary	Female, South, 15-24, Higher
Female, North, 25-44, Higher	Male, Centre, 15-24, Basic or less
Female, South, 15-24, Basic or less	Male, Centre, 15-24, Secondary
Female, South, 15-24, Secondary	Male, Centre, 25-44, Higher
Female, South, 25-44, Secondary	Male, Lisbon and TV, 15-24, Basic or less
Male, Lisbon and TV, 25-44, Higher	Male, Lisbon and TV, 15-24, Secondary
Male, North, 15-24, Basic or less	Male, North, 15-24, Secondary
Male, North, 25-44, Secondary	
Male, North, 25-44, Higher	
Male, South, 15-24, Basic or less	
Male, South, 15-24, Secondary	
CLUSTER 3	**CLUSTER 4**
Female, Centre, 25-44, Basic or less	Female, Lisbon and TV, 45-64, Basic or less
Female, Centre, 45-64, Basic or less	Female, North, 25-44, Basic or less
Female, Lisboa, and TV, 25-44, Basic or less	Female, North, 45-64, Basic or less
Female, Lisbon and TV, 45-64, Secondary	Female, South, 45-64, Basic or less
Female, North, 45-64, Secondary	Male, Lisbon and TV, 45-64, Basic or less
Female, South, 25-44, Basic or less	Male, South, 45-64, Basic or less
Male, Centre, 25-44, Basic or less	**CLUSTER 5**
Male, Centre, 45-64, Basic or less	Female, Lisbon and TV, 45-64, Higher
Male, Lisbon and TV, 25-44, Basic or less	Female, South, 25-44, Higher
Male, Lisbon and TV, 25-44, Secondary	Female, South, 45-64, Secondary
Male, North, 25-44, Basic or less	Male, Centre, 45-64, Secondary
Male, North, 45-64, Basic or less	
Male, South, 25-44, Basic or less	

TABLE 16

Table 16: Income-Debt data - Component 1 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	294.299	15.764	–	–	–	–	–	–
V2:Hinc-LogR	15.764	1.110	–	–	–	–	–	–
V3:DIncR-MidP	–	–	3.599	0.395	–	–	–	–
V4:DIncR-LogR	–	–	0.395	0.192	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	0.116	0.179	–	–
V6:CCDbt-LogR	–	–	–	–	0.179	0.393	–	–
V7:ODbt-MidP	–	–	–	–	–	–	0.929	0.609
V8:ODbt-LogR	–	–	–	–	–	–	0.609	0.721

TABLE 17

Table 17: Income-Debt data - Component 2 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	80.904	3.795	–	–	–	–	–	–
V2:Hinc-LogR	3.795	0.228	–	–	–	–	–	–
V3:DIncR-MidP	–	–	9.062	0.681	–	–	–	–
V4:DIncR-LogR	–	–	0.681	0.073	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	0.303	0.231	–	–
V6:CCDbt-LogR	–	–	–	–	0.231	0.187	–	–
V7:ODbt-MidP	–	–	–	–	–	–	0.794	0.325
V8:ODbt-LogR	–	–	–	–	–	–	0.325	0.143

TABLE 18

Table 18: Income-Debt data - Component 3 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	1061.778	8.913	–	–	–	–	–	–
V2:Hinc-LogR	8.913	0.080	–	–	–	–	–	–
V3:DIncR-MidP	–	–	8.475	0.664	–	–	–	–
V4:DIncR-LogR	–	–	0.664	0.059	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	3.567	0.524	–	–
V6:CCDbt-LogR	–	–	–	–	0.524	0.078	–	–
V7:ODbt-MidP	–	–	–	–	–	–	4.943	0.360
V8:ODbt-LogR	–	–	–	–	–	–	0.360	0.027

TABLE 19

Table 19: Income-Debt data - Component 4 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	674.648	10.054	–	–	–	–	–	–
V2:Hinc-LogR	10.054	0.168	–	–	–	–	–	–
V3:DIncR-MidP	–	–	5.988	0.440	–	–	–	–
V4:DIncR-LogR	–	–	0.440	0.035	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	0.331	0.097	–	–
V6:CCDbt-LogR	–	–	–	–	0.097	0.029	–	–
V7:ODbt-MidP	–	–	–	–	–	–	3.820	0.603
V8:ODbt-LogR	–	–	–	–	–	–	0.603	0.098

TABLE 20

Table 20: Income-Debt data - Component 5 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	1523.248	13.789	–	–	–	–	–	–
V2:Hinc-LogR	13.789	0.144	–	–	–	–	–	–
V3:DIncR-MidP	–	–	5.460	0.598	–	–	–	–
V4:DIncR-LogR	–	–	0.598	0.106	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	8.444	0.798	–	–
V6:CCDbt-LogR	–	–	–	–	0.798	0.078	–	–
V7:ODbt-MidP	–	–	–	–	–	–	53.712	3.786
V8:ODbt-LogR	–	–	–	–	–	–	3.786	0.278

TABLE 21

Table 21: Income-Debt data - Component 6 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	4982.894	20.273	–	–	–	–	–	–
V2:Hinc-LogR	20.273	0.087	–	–	–	–	–	–
V3:DIncR-MidP	–	–	10.580	0.851	–	–	–	–
V4:DIncR-LogR	–	–	0.851	0.073	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	25.369	1.729	–	–
V6:CCDbt-LogR	–	–	–	–	1.729	0.123	–	–
V7:ODbt-MidP	–	–	–	–	–	–	4.858	0.348
V8:ODbt-LogR	–	–	–	–	–	–	0.348	0.025

TABLE 22

Table 22: Income-Debt data - Component 7 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	1134.334	16.626	–	–	–	–	–	–
V2:Hinc-LogR	16.626	0.308	–	–	–	–	–	–
V3:DIncR-MidP	–	–	10.780	1.082	–	–	–	–
V4:DIncR-LogR	–	–	1.082	0.147	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	1.967	0.609	–	–
V6:CCDbt-LogR	–	–	–	–	0.609	0.196	–	–
V7:ODbt-MidP	–	–	–	–	–	–	5.983	1.026
V8:ODbt-LogR	–	–	–	–	–	–	1.026	0.188

TABLE 23

Table 23: Income-Debt data - Component 8 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	5540.152	45.424	–	–	–	–	–	–
V2:Hinc-LogR	45.424	0.784	–	–	–	–	–	–
V3:DIncR-MidP	–	–	26.234	3.176	–	–	–	–
V4:DIncR-LogR	–	–	3.176	1.173	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	6.603	1.987	–	–
V6:CCDbt-LogR	–	–	–	–	1.987	0.701	–	–
V7:ODbt-MidP	–	–	–	–	–	–	38.454	2.938
V8:ODbt-LogR	–	–	–	–	–	–	2.938	0.429

TABLE 24

Table 24: Income-Debt data - Component 9 Covariance Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1:Hinc-MidP	3509.047	7.942	–	–	–	–	–	–
V2:Hinc-LogR	7.942	0.018	–	–	–	–	–	–
V3:DIncR-MidP	–	–	14.115	1.008	–	–	–	–
V4:DIncR-LogR	–	–	1.008	0.077	–	–	–	–
V5:CCDbt-MidP	–	–	–	–	241.252	7.142	–	–
V6:CCDbt-LogR	–	–	–	–	7.142	0.216	–	–
V7:ODbt-MidP	–	–	–	–	–	–	239.983	4.765
V8:ODbt-LogR	–	–	–	–	–	–	4.765	0.095

CAPTION OF FIGURE 1

First 20 observations and three-cluster means (Synthetic data set 1)

CAPTION OF FIGURE 2

Partition in six clusters obtained by the model-based method for the USA meteorological data

CAPTION OF FIGURE 3

Partition in four clusters obtained by the Ward method for the USA meteorological data

CAPTION OF FIGURE 4

Partition in six clusters obtained by the Ward method for the USA meteorological data

CAPTION OF FIGURE 5

Partition in six clusters obtained by the SCLUST method for the USA meteorological data

CAPTION OF FIGURE 6

Partition in seven clusters obtained by MCLUST for the USA meteorological

MidPoints data

FIGURE 1

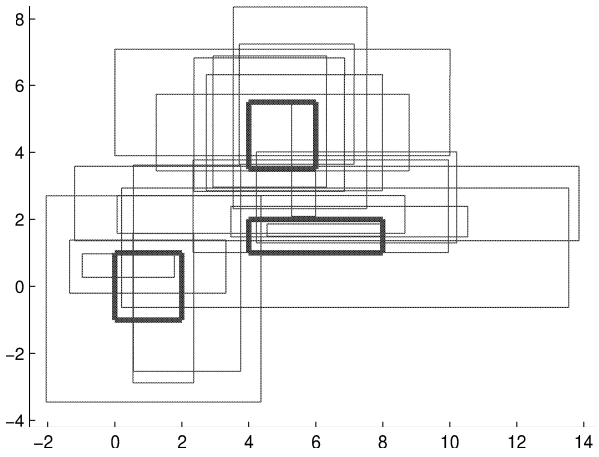

Figure 1: First 20 observations and three-cluster means (Synthetic data set 1)

FIGURE 2

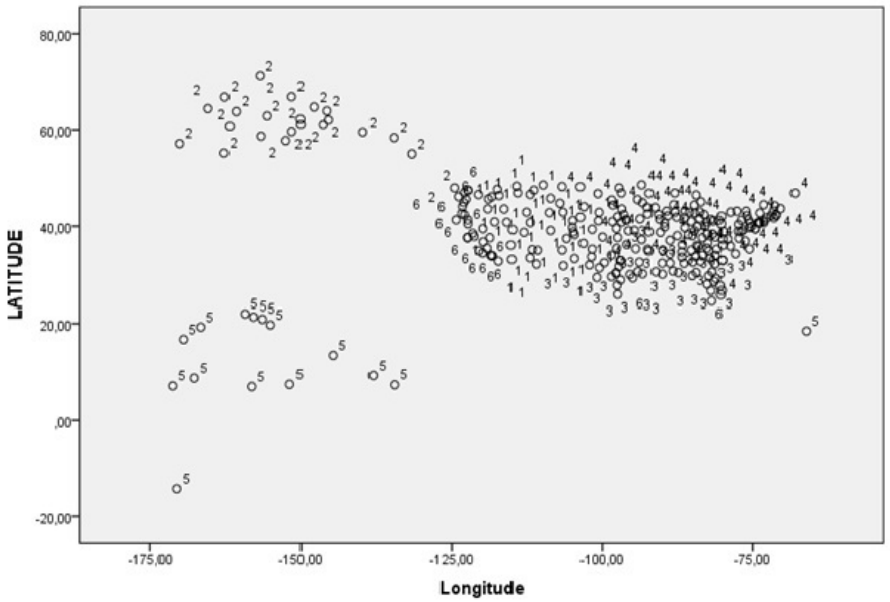

Figure 2: Partition in six clusters obtained by the model-based method for the USA meteorological data

FIGURE 3

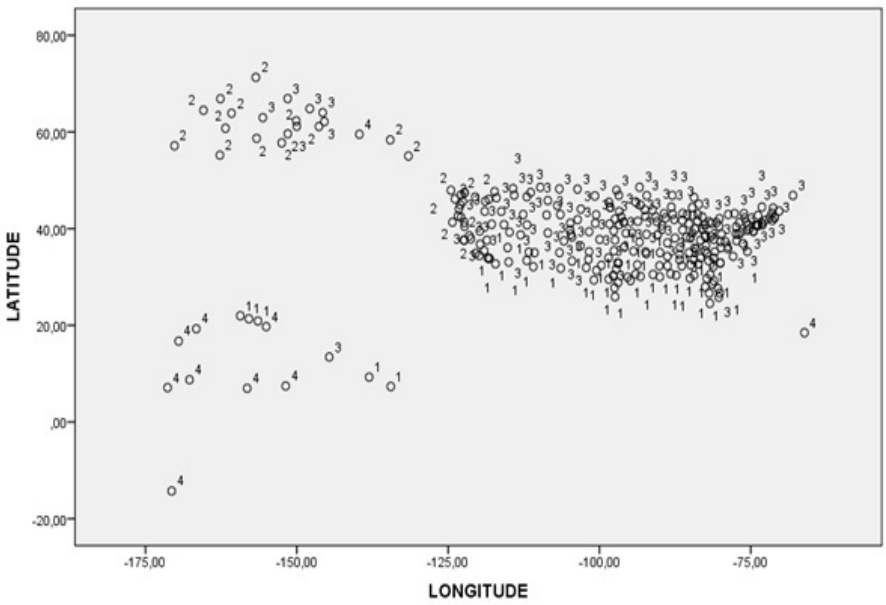

Figure 3: Partition in four clusters obtained by the Ward method for the USA meteorological data

FIGURE 4

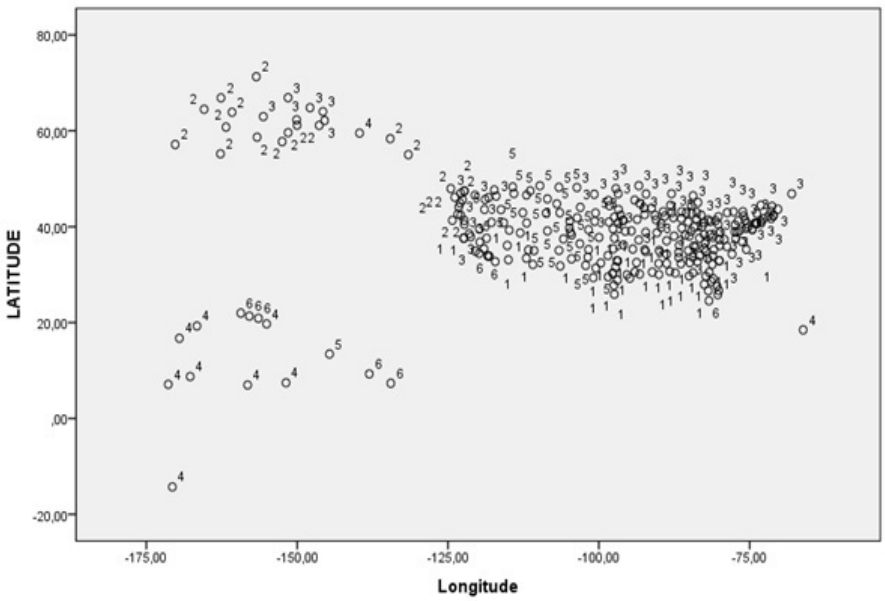

Figure 4: Partition in six clusters obtained by the Ward method for the USA meteorological data

FIGURE 5

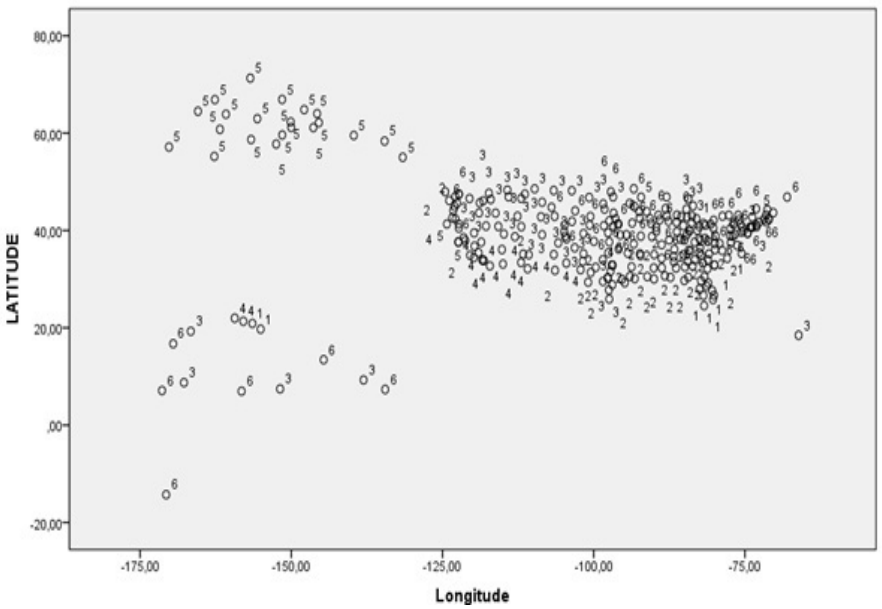

Figure 5: Partition in six clusters obtained by the SCLUST method for the USA meteorological data

FIGURE 6

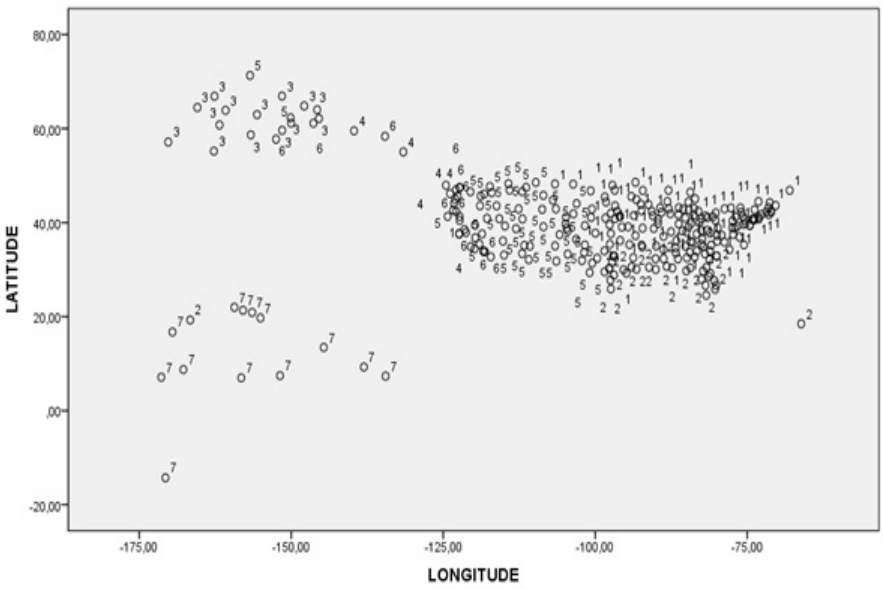

Figure 6: Partition in seven clusters obtained by MCLUST for the USA meteorological data