A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic zoning

Antonio Morales-Esteban
Department of Building Structures and Geotechnical Engineering, University of Seville, Spain,
e-mail: `ame@us.es`

Francisco Martínez-Álvarez
Department of Computer Science, Pablo de Olavide University of Seville, Spain
e-mail: `fmaralv@upo.es`

Sanja Scitovski
Faculty of Civil Engineering, University of Osijek, Croatia
e-mail: `sscitov@unios.hr`

Rudolf Scitovski[1]
Department of Mathematics, University of Osijek, Croatia
Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
e-mail: `scitowsk@mathos.hr`

Abstract. In this paper we construct an efficient adaptive Mahalanobis k-means algorithm. In addition, we propose a new efficient algorithm to search for a globally optimal partition obtained by using the adoptive Mahalanobis distance-like function. The algorithm is a generalization of the previously proposed incremental algorithm [36]. It successively finds optimal partitions with $k = 2, 3, \ldots$ clusters. Therefore, it can also be used for the estimation of the most appropriate number of clusters in a partition by using various validity indexes. The algorithm has been applied to the seismic catalogues of Croatia and the Iberian Peninsula. Both regions are characterized by a moderate seismic activity. One of the main advantages of the algorithm is its ability to discover not only circular but also elliptical shapes, whose geometry fits the faults better. Three seismogenic zonings are proposed for Croatia and two for the Iberian Peninsula and adjacent areas, according to the clusters discovered by the algorithm.

Key words: Adaptive Mahalanobis clustering; Globally optimal partition; DIRECT; Incremental algorithm; Seismogenic zoning;

1 Introduction

Unsupervised learning consists in discovering hidden structures in unlabeled data. The goal of this technique is to find groups of data exhibiting similar behavior. Unlike other methods, no assumption about the data distribution has to be made. Algorithms based on

[1] Corresponding author: Rudolf Scitovski, e-mail: `scitowsk@mathos.hr`, telephone number: +385-31-224-800, fax number: +385-31-224-801

these techniques have been extensively applied in bioinformatics [21], energy management [26, 34], and telecommunications [43].

In this paper a novel incremental clustering algorithm is proposed. The main feature of the algorithm is its ability to discover globally optimal partitions instead of the local ones, similarly to many algorithms and heuristics. The new proposal is based on the algorithm described in [36] and inspired by the k-means algorithm with an adaptive Mahalanobis distance [7, 37]. The choice of using the Mahalanobis distance in k-means is basically a choice to either use full-covariance in the clusters or to ignore them. When the Euclidean distance is used, the clusters are assumed to have the same covariances, i.e. to have circular shapes. If data covariances cannot be represented by identity matrices, the use of the Mahalanobis distance is advised because elliptical shapes are present. In contrast, if two distributions have identity covariance matrices, the Mahalanobis distance reduces to the Euclidean distance. Such is the case of the data analyzed in this paper.

A finite data set $\mathcal{A} \subset \mathbb{R}^n$, $|\mathcal{A}| = m$, is given. A partition of the set $\mathcal{A}$ into $1 \leq k \leq m$ disjoint subsets $\pi_1, \ldots, \pi_k$, such that

$$\bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \qquad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \qquad |\pi_j| \geq 1, \quad j = 1, \ldots, k, \tag{1}$$

will be denoted by $\Pi(\mathcal{A}) = \{\pi_1, \ldots, \pi_k\}$ and the set of all such partitions by $\mathcal{P}(\mathcal{A}, k)$. The elements $\pi_1, \ldots, \pi_k$ of the partition Π are called *clusters in* $\mathbb{R}^n$. In this paper, first an efficient adaptive Mahalanobis k-means algorithm is constructed, which is then used to construct a new efficient algorithm for searching for a globally optimal partition obtained by using the adaptive Mahalanobis distance-like function. An important advantage of the proposed algorithm is that it successively gives optimal partitions with $k = 2, 3, \ldots$ clusters. Therefore, for each $k \geq 2$, it is immediately possible to calculate the value of various validity indexes and in this way to estimate the most appropriate number of clusters in a partition. Additionally, a novel validity index is proposed. This index, combined with some others, will be used to determine the most appropriate number of clusters in a partition.

We apply the algorithm to construct seismogenic zonings for Croatia and the Iberian Peninsula, and by this new method, compact and comprehensive zones have been found. The zones also exhibit a good correlation with the underlying geology.

The remainder of the paper is structured as follows. Section 2 reviews relevant works in this area of knowledge. Section 3 discusses the search for a globally optimal partition and introduces the new algorithm. Its application to create seismogenic zones is described in Section 4. Finally, the conclusions drawn are summarized in Section 5.

2 Related works concerning seismogenic zoning

The seismicity of Croatia is characterized by earthquakes of medium-large magnitude spread all over the country. The first consistent seismogenic zoning can be found in [25], which proposed seventeen zones. However, nowadays most of the works reported in the literature propose less zones. Such is the case of [24], where the authors defined two main zones. The 2010 report by Akkar et al. [1] defined five regional structural units: Adriatic microplate, Adriatic, Dinaric, Supradinaric and Pannonian basin, refining the two aforementioned zones.

The seismicity of the north-west Croatia was analyzed in detail for the first time in [16]. The authors optimized and declustered the catalogue to accurately define eight zones. The same authors improved their approach in 2008 [40] but, this time, the number of proposed zones was five.

Similarly to mainland Croatia, the Iberian Peninsula exhibits a moderate seismicity focused in several zones. One of the most widely used zonings so far was described by Martín [27]. The author defined 27 seismogenic zones. Mezcua et al. [28] have recently proposed a probabilistic seismic hazard analysis for mainland Spain combining some of the zones previously proposed.

Partial zonings of the Iberian Peninsula have also been described. Such is the case of [11], which studied the Granada Basin (south-east of Spain). López-Casado et al. [22] defined four zones for the Betic Cordillera, Rif and nearby regions. The works in [3] and [13] zoned the eastern and southern Spain, respectively. López-Fernández et al. [23] worked on the area between the Pyrenees and the Galicia region.

Note that all reviewed works built the zonings using different geological assumptions and some human-made decisions. For this reason, an objective and robust new methodology that aims to discover seismogenic zones is proposed in this paper.

3 Searching for a globally optimal partition

A data set $\mathcal{A} \subset [\alpha, \beta] \subset \mathbb{R}^n$, where $\alpha = (\alpha_1, \ldots, \alpha_n), \beta = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$ is given, and to each data point $a^i \in \mathcal{A}$ a weight $w_i > 0$ is associated.

If components a^i_s, $s = 1, \ldots, n$ of the data point a^i are not of equal range, i.e. if numbers $\beta_1 - \alpha_1, \ldots, \beta_n - \alpha_n$, are mutually significantly different, they should first be normalized. This can be achieved by transforming the set $\mathcal{A}$ into the set $\mathcal{B} = \{T(a^i) \colon a^i \in \mathcal{A}\} \subset [0,1]^n$ using the mapping $T \colon [\alpha, \beta] \to [0,1]^n$, where

$$T(x) = D(x - \alpha), \qquad D = \operatorname{diag}\left(\tfrac{1}{\beta_1 - \alpha_1}, \ldots, \tfrac{1}{\beta_n - \alpha_n}\right). \tag{2}$$

After clustering the set $\mathcal{B}$, the obtained results will be transformed again into $[\alpha, \beta]$ by the mapping $T^{-1} \colon [0,1]^n \to [\alpha, \beta]$, $T^{-1}(x) = D^{-1}x + \alpha$.

If $d \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ = [0, +\infty\rangle$ is some distance-like function (see, e.g., [19, 38]), then to each cluster $\pi_j \in \Pi$ its center c_j can be associated as follows:

$$c_j = \underset{x \in [\alpha, \beta]}{\operatorname{argmin}} \sum_{a^i \in \pi_j} w_i d(x, a^i). \tag{3}$$

After that, by introducing the objective function $\mathcal{F} \colon \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$ the quality of a partition and the search for the *globally optimal k-partition* can be defined by solving the following optimization problem:

$$\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi), \qquad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i\in\pi_j} w_i d(c_j, a^i). \tag{4}$$

Conversely, for a given set of centers $c_1, \ldots, c_k \in \mathbb{R}^n$, by applying the minimal distance principle, the partition $\Pi = \{\pi_1, \ldots, \pi_k\}$ of the set $\mathcal{A}$ can be defined, which consists of clusters:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k.$$

Therefore, the problem of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the following *global optimization problem* for a Lipschitz continuous function (see, e.g., [33, 37, 38])

$$\operatorname*{argmin}_{c_1,\ldots,c_k\in[\alpha,\beta]} F(c_1, \ldots, c_k), \qquad F(c_1, \ldots, c_k) = \sum_{i=1}^{m} w_i \min_{1\leq s\leq k} d(c_s, a^i). \tag{5}$$

The solutions of (4) and (5) coincide [36, 37].

3.1 Adaptive Mahalanobis *k*-means algorithm

The best known algorithm for searching for a locally optimal partition is the k-means algorithm [35]. In this subsection an efficient adaptive Mahalanobis k-means algorithm is constructed. This algorithm is able to recognize spherical clusters and very elongated elliptical clusters which come from some line segments (see Example 1).

Algorithm 1. `(Adaptive Mahalanobis` k`-means)`

`Step 0:` Input $m \geq 1$, $1 \leq k \leq m$, $I = \{1, \ldots, m\}$, $J = \{1, \ldots, k\}$, $\mathcal{A} = \{a^i \in \mathbb{R}^n : i \in I\}$, $w_1, \ldots, w_m > 0$;
Define the vectors $\alpha, \beta \in \mathbb{R}^n$ with components $\alpha_l = \min\limits_{i\in I} a^i_l$, $\beta_l = \max\limits_{i\in I} a^i_l$, $l = 1, \ldots, n$;
Choose mutually different assignment points $z_1 \ldots, z_k \in [\alpha, \beta]$;

`Step 1: (Assignment step)` Define LS-clusters

$$\pi_j = \pi(z_j) = \{a^i \in \mathcal{A} \colon \|z_j - a^i\| \leq \|z_s, a^i\|,\ \forall s \in J\}, \quad j \in J;$$

`Step 2: (Update step)` Let $c = (c_1, \ldots, c_k)$ be the vector of centroids

$$c_j = \tfrac{1}{W_j} \sum_{a^i\in\pi_j} w_i a^i, \quad W_j = \sum_{a^i\in\pi_j} w_i, \qquad j \in J;$$

Determine the objective function value $F_0 = \sum\limits_{j=1}^{k} \sum\limits_{a^i\in\pi_j} w_i \|c_j - a^i\|^2$;
Determine the covariance matrices $S_j = \frac{1}{W_j} \sum\limits_{a^i\in\pi_j} w_i (c_j - a^i)(c_j - a^i)^T$, $\quad j \in J$;
For each cluster π_j define the Mahalanobis distance-like functions

$$d_M^{(j)}(x, y, S_j) := \sqrt[n]{\det S_j}\, (x - y)^T S_j^{-1} (x - y), \quad j \in J; \tag{6}$$

`Step 3: (Assignment step)` For each $j \in J$ define new clusters

$$\hat{\pi}_j = \hat{\pi}(z_j) = \{a^i \in \mathcal{A} \colon d_M^{(j)}(c_j, a^i, S_j) \le d_M^{(s)}(c_s, a^i, S_s),\ \forall s \in J\}, \quad j \in J;$$

`Step 4: (Update step)` Let $\hat{c} = (\hat{c}_1, \dots, \hat{c}_k)$ be the vector of centroids

$$\hat{c}_j = \tfrac{1}{\hat{W}_j} \sum_{a^i \in \hat{\pi}_j} w_i a^i, \quad \hat{W}_j = \sum_{a^i \in \hat{\pi}_j} w_i, \qquad j \in J; \tag{7}$$

Determine the objective function value $F_1 = \sum\limits_{j=1}^{k} \sum\limits_{a^i \in \hat{\pi}_j} w_i d_M^{(j)}(\hat{c}_j, a^i, S_j)$;

Determine the covariance matrices $\hat{S}_j = \frac{1}{\hat{W}_j} \sum\limits_{a^i \in \hat{\pi}_j} w_i (\hat{c}_j - a^i)(\hat{c}_j - a^i)^T, \quad j \in J$;

For each cluster $\hat{\pi}_j$ define the Mahalanobis distance-like functions

$$d_M^{(j)}(x, y, \hat{S}_j) := \sqrt[n]{\det \hat{S}_j}\,(x-y)^T \hat{S}_j^{-1}(x-y), \quad j \in J; \tag{8}$$

`Step 5:` If $F_1 < F_0$, set $F_0 = F_1$, $c = \hat{c}$ and $S_j = \hat{S}_j$ for each $j \in J$ and go to *Step 3*;
Else set $c_j^\star = \hat{c}_j$, $\forall j \in J$ and STOP.

Remark 1. If $\operatorname{rank}\{\hat{c}_j - a^i \colon a^i \in \hat{\pi}_j\} = n$, then the matrices $\hat{S}_j$ from `Step 4` are positive definite and therefore there exist $\hat{S}_j^{-1}$ and $\det \hat{S}_j > 0$.

Furthermore, the centers $\hat{c}_j$ of clusters $\hat{\pi}_j$ from `Step 3` are defined by

$$\hat{c}_j = \underset{x \in [\alpha, \beta]}{\operatorname{argmin}} \sum_{a^i \in \hat{\pi}_j} w_i d_M^{(j)}(x, a^i, S_j).$$

Since the unique stationary point of the function $x \mapsto \sum\limits_{a^i \in \hat{\pi}_j} w_i d_M^{(j)}(x, a^i, S_j)$ is determined by

$$\sum_{a^i \in \hat{\pi}_j} w_i S_j^{-1}(x - a^i) = 0,$$

it follows that the centers $\hat{c}_j$ coincide with centroids given by (7) in `Step 4`.

Finally, the sequence of objective function values $F_0, F_1, \dots$ obtained in `Step 2` and `Step 4` is monotonically decreasing and attains its minimal value $F^\star$ in finitely many steps. Since $\sum\limits_{a^i \in \hat{\pi}_j} w_i (\hat{c}_j - a^i)^T \hat{S}_j^{-1} (\hat{c}_j - a^i) = n \cdot \hat{W}_j$, the coefficient $\sqrt[n]{\det S_j}$, resp. $\sqrt[n]{\det \hat{S}_j}$, in distance-like functions (6), resp. (8), is essential for the monotonicity property of the sequence of objective function values [37].

Example 1. *A synthetic data set is constructed similarly to [39]. Let us choose two points* $C_1 = (3, 2)$*,* $C_2 = (8, 6) \in \mathbb{R}^2$*, and in the neighborhood of each point* C_j *generate* 100 *random points by using binormal random additive errors with mean vector* $\mathbf{0} \in \mathbb{R}^2$ *and corresponding covariance matrices* $\Sigma_1 = \frac{1}{2}\begin{bmatrix} 2 & -1 \\ -1 & .9 \end{bmatrix}$*,* $\Sigma_2 = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$*. Let us also choose three line segments* $l_1 = [(1, 9), (6, 9)]$*,* $l_2 = [(3, 5), (5, 9)]$*,* $l_3 = [(3, 6), (7, 2)]$ *and in a neighborhood of each generate* 100 *normally distributed random points. In this way, one obtains the partition* Π_0 *and the data set* $\mathcal{A} = \{a^i \in \mathbb{R}^2 \colon i = 1, \dots, m\} \subset [0, 10]^2$*, with* $m = 500$ *random points (see Fig. 1a).*

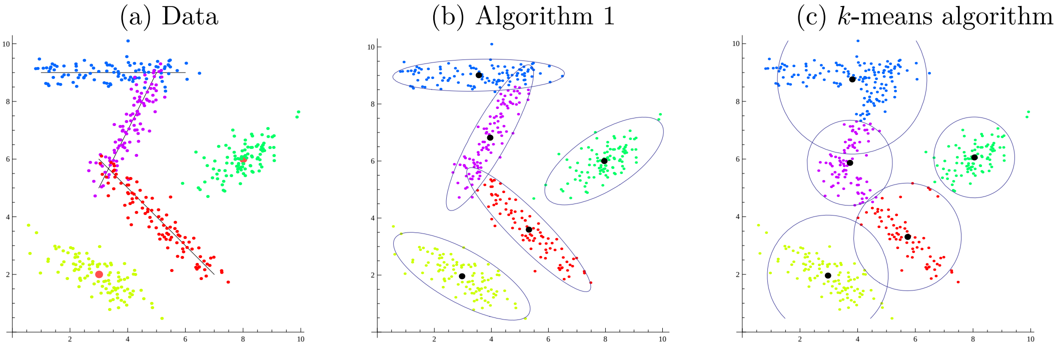


Figure 1: Applications of `Algorithm 1` to the data from Example 1

By applying Adaptive Mahalanobis k-means Algorithm 1 with initial centers $(2,2)$, $(9,5)$, $(3,9)$, $(4,7)$, $(5,4)$, *the optimal partition* $\Pi^\star$ *is obtained (Fig. 1b). Ellipses in the figure include* 95% *of points belonging to the corresponding clusters.*

The Rand and the Jacard index (see Table 1), as well as the confusion matrix $S(\Pi_0,\Pi^\star)$, *show recognition of the original partition very well.*

Algorithm	Rand	Jacard
Algorithm 1	0.868	0.809
k-means algorithm	0.713	0.628

Table 1: Rand and Jacard indexes between the original partition Π_0 and the partition $\Pi^\star$ obtained by Adaptive Mahalanobis k-means Algorithm 1, and the partition $\hat{\Pi}$ obtained by k-means algorithm

$$S(\Pi_0,\Pi^\star)=\begin{bmatrix} 100 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 \\ 0 & 0 & 97 & 3 & 0 \\ 0 & 0 & 12 & 87 & 1 \\ 0 & 0 & 0 & 13 & 87 \end{bmatrix}, \qquad S(\Pi_0,\hat{\Pi})=\begin{bmatrix} 100 & 0 & 0 & 0 & 0 \\ 0 & 96 & 0 & 0 & 4 \\ 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 45 & 55 & 0 \\ 0 & 0 & 0 & 36 & 67 \end{bmatrix}.$$

Fig. 1c depicts the partition $\hat{\Pi}$ *obtained by the application of* k*-means algorithm. Circles in the figure include* 95% *of points belonging to the corresponding clusters. Rand and Jacard indexes are significantly smaller and the confusion matrix* $S(\Pi_0,\hat{\Pi})$ *shows weaker recognition of the initial partition.*

Remark 2. It can be shown that the *Mahalanobis k-means algorithm* essentially coincides with the known *Generalized Mixture Decomposition Algorithmic Scheme* (GMDAS) [39] as a special case of the *Expectation Maximization* algorithm (see [44, page 31]). The efficiency of Algorithm 1, measured by the necessary CPU time, is significantly greater. Let us justify this claim by the following numerical test. We apply both algorithms to the data from Example 1 with $m_1 = 100$, 200, and 500 data points in the neighborhood of each point C_j (note that the number of data set $\mathcal{A}$ is $m = 300$, 600, and 1 500). As can be seen from Table 2, the CPU time for Algorithm 1 is about 20 times shorter than the CPU time for GMDAS.

m	GMDAS	Algorithm 1
300	8.50	0.32
600	22.22	0.72
1 500	48.72	2.85

Table 2: CPU times (sec) for Algorithm 1 and for GMDAS

3.2 A new incremental algorithm

The following incremental algorithm proposed in this paper is based on the incremental algorithm proposed in [36] (see also [2, 20]). An application of some efficient global optimization method for a Lipschitz continuous function is very important for a successful implementation of the proposed algorithm. Among many recent works related to the global optimization problem for Lipschitz continuous functions, let us mention only [31, 33]. In this paper we propose to use the well-known global optimization algorithm `DIRECT` (DIviding RECTangles) (see, e.g., [8, 9, 15, 17]).

An important advantage of all incremental algorithms for searching for an optimal partition lies in the fact that one obtains an optimal partition for each $k \leq k_{\max}$, where $k_{\max}$ is given in advance. Alternatively, given $\epsilon > 0$, the algorithm can yield optimal partitions with $k = 2, 3, \ldots$ clusters until the relative difference of the objective function becomes less than ϵ [36]. This allows the estimation of the appropriate number of clusters in a partition by using various well-known indexes (see [10, 42]).

Algorithm 2. (`A new algorithm for searching for an optimal` k`-partition`)

```
Step 0: Input $m \geq 1,\ I = \{1, \ldots, m\},\ \mathcal{A} = \{a^i \in \mathbb{R}^n \colon i \in I\},\ w_1, \ldots, w_m > 0;\ \epsilon > 0$
        Define the vectors $\alpha, \beta \in \mathbb{R}^n$ with components $\alpha_s = \min_{i \in I} a_s^i,\ \beta_s = \max_{i \in I} a_s^i$;

Step 1: Choose integer $k_{\max}$ and $r < k_{\max}$ different assignment points $c_1^{(0)}, \ldots, c_r^{(0)} \in [\alpha, \beta]$;

Step 2: Determine $\hat{c}_1, \ldots, \hat{c}_r \in [\alpha, \beta]$ by using Algorithm 1 and calculate $\hat{F}_r := F(\hat{c}_1, \ldots, \hat{c}_r)$;
        Set $F_r^\star := \hat{F}_r$;

Step 3: By using the DIRECT algorithm for global optimization determine $\hat{c}_{r+1} \in \operatorname{argmin}_{c \in [\alpha, \beta]} \Phi(c)$,
        where $\Phi(c) = \sum_{i=1}^{m} w_i \min\{\delta_r^i, \|c - a^i\|^2\}$, and $\delta_r^i = \min_{1 \leq s \leq r} \|\hat{c}_s - a^i\|^2$;

Step 4: By using Algorithm 1 with initial centers $\hat{c}_1, \ldots, \hat{c}_r, \hat{c}_{r+1}$ determine the new centers
        $c_1^\star, \ldots, c_r^\star, c_{r+1}^\star$ and calculate $F_{r+1}^\star := F(c_1^\star, \ldots, c_r^\star, c_{r+1}^\star)$;

Step 5: If $\frac{F_r^\star - F_{r+1}^\star}{\hat{F}_r} \geq \epsilon$, set $F_r^\star = F_{r+1}^\star$, $r = r + 1$ and go to Step 3;    Else STOP.
```

The sequence of objective function values $(F_r^\star)$ is monotonically decreasing, bounded below and therefore convergent. Unfortunately, it cannot be asserted that the proposed algorithm gives a globally optimal k-partition, but numerous calculations show that the partition obtained by this algorithm is a satisfactory approximation of the globally optimal partition. Thereby it is important to note that the `CPU` time for the implementation of Algorithm 2 is very short. In what follows, the partition obtained by Algorithm 2 will be simply called an optimal partition.

3.3 Determining the most appropriate number of clusters

Automatic determination of the number of clusters has been one of the most difficult problems in data clustering processes [10, 42]. In simple cases, the number of clusters in a partition is determined by the nature of the problem itself. If the number of clusters in a partition is not given in advance, then it is natural to search for an optimal partition which consists of clusters that are as compact and relatively strongly separated as possible. For determining the most appropriate number of clusters in a partition some of well-known indexes (see [10, 42]) will be adopted for adaptive Mahalanobis clustering and a new index will also be proposed.

Let $\Pi^\star = \{\pi_1^\star, \dots, \pi_k^\star\}$ be an optimal partition of the set $\mathcal{A}$ with weights $w_1, \dots, w_m$ and k clusters $\pi_1^\star, \dots, \pi_k^\star$ with corresponding centers $c_1^\star, \dots, c_k^\star \in \mathbb{R}^n$ and covariance matrices $S_1^\star, \dots, S_k^\star$.

(i) The `Simplified Silhouette Width Criterion (SWC)` will be adopted in the following way. For each $a^i \in \mathcal{A} \cap \pi_r^\star$ the numbers

$$\alpha_{ir} = d_M^{(r)}(c_r^\star, a^i, S_r^\star), \qquad \beta_{ir} = \min_{s \neq r} d_M^{(s)}(c_s^\star, a^i, S_s^\star), \qquad s_i = \frac{\beta_{ir} - \alpha_{ir}}{\max\{\alpha_{ir}, \beta_{ir}\}},$$

are calculated and the `SWC` is defined as the average of s_i:

$$\texttt{SWC}(k) = \tfrac{1}{W} \sum_{a^i \in \mathcal{A}} w_i s_i, \qquad W = \sum_{i=1}^m w_i.$$

More compact and better separated clusters in an optimal partition will result in a greater `SWC` number.

(ii) The `Davies - Bouldin index (VDB)` will be adopted in the following way.

$$\texttt{VDB}(k) = \frac{1}{k} \sum_{j=1}^k \max_{s \neq j} \frac{V(\pi_j^\star) + V(\pi_s^\star)}{d_M^{(j)}(c_j^\star, c_s^\star, S_j^\star)}, \qquad V(\pi_j^\star) = \tfrac{1}{W_j^\star} \sum_{a_s \in \pi_j^\star} w_s d_M^{(j)}(c_j^\star, a^s, S_j^\star), \tag{9}$$

where $W_j^\star = \sum_{a^i \in \pi_j^\star} w_i$. More compact and better separated clusters in an optimal partition will result in a lower `VDB` index.

(iii) The `Calinski - Harabasz index (VCH)` will be adopted in the following way.

$$\mathtt{VCH}(k) = \frac{\mathcal{G}(c_1^\star, \ldots, c_k^\star)/(k-1)}{\mathcal{F}(c_1^\star, \ldots, c_k^\star)/(m-k)}, \tag{10}$$

where

$$\mathcal{F}(c_1^\star, \ldots, c_k^\star) = n \sum_{j=1}^{k} W_j^\star \sqrt[n]{\det S_j^\star}, \qquad \mathcal{G}(c_1^\star, \ldots, c_k^\star) = \sum_{j=1}^{k} W_j^\star \, d_M^{(j)}(c^\star, c_j^\star, S_j^\star),$$

where $c_j^\star = \frac{1}{W_j^\star} \sum\limits_{a^i \in \pi_j} w_i a^i$ are centroids of clusters π_j, and $c^\star = \frac{1}{W} \sum\limits_{a^i \in \mathcal{A}} w_i a^i$ is a centroid of the entire set $\mathcal{A}$. More compact and better separated clusters in an optimal partition will result in a greater `VCH` index.

(iv) Similarly to the *Hypervolume fuzzy index* [12], a new `Area index` will be defined as the sum of weighted areas of the clusters.

$$\mathtt{Area}(k) = \sum_{j=1}^{k} \frac{\det S_j^\star}{W_j^\star} \tag{11}$$

More compact and better separated clusters in an optimal partition will result in a lower `Area index`.

Example 2. *First, Algorithm 2 will be illustrated on the data from* Example 1*. Choose the initial centers* (2, 9) *and* (8, 6)*. Fig. 2 shows the first three of the seven iterations of Algorithm 2. As can be seen in Fig. 3, all indexes very clearly point to the partition with five clusters as the most appropriate partition. Also, the partition with five clusters obtained by using Algorithm 2 coincides with the optimal partition* $\Pi^\star$ *from* Example 1*. Moreover, Rand and Jacard indexes and the confusion matrix coincide with the ones from* Example 1*. All of this allows us to conclude that Algorithm 2 has found the optimal partition.*

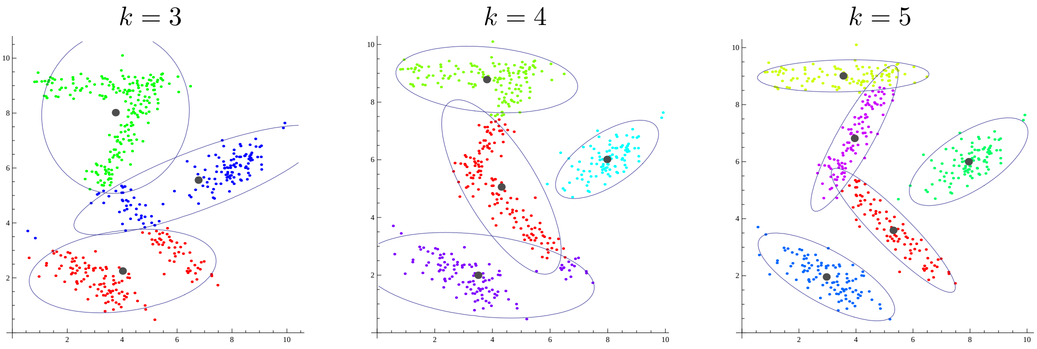


Figure 2: Applications of `Algorithm 2` to the data from Example 1

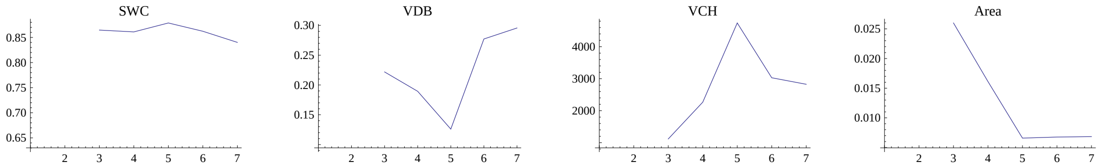


Figure 3: Indexes of optimal partitions for the data from Example 2

Example 3. *The Iris data set*[2] *consists of* 50 *data* (π_1^0) *related to Iris species* Setosa*,* 50 *data* (π_2^0) *to Iris species* Versicolour*, and* 50 *data* (π_3^0) *to Iris species* Virginica*. Each data point is characterized by four attributes:* sepal length in cm, sepal width in cm, petal length in cm, and petal width in cm*. In this way the original partition* $\Pi_0 = \{\pi_1^0, \pi_2^0, \pi_3^0\}$ *and the data set* $\mathcal{A} = \pi_1^0 \cup \pi_2^0 \cup \pi_3^0$ *with weights* $w_i = 1$, $i = 1, \ldots, 150$ *is constructed.*

The implementation of Algorithm 2 with the initial center $(4, 4, 2, 0)$ *is shown in Table 3. The number of elements in the corresponding clusters and the objective function value are shown for the first four iterations of Algorithm 2.*

Iteration	$\lvert\pi_j\rvert$	F
1	$\{150\}$	331.4
2	$\{50, 100\}$	60.6
3	$\{50, 57, 43\}$	42.2
4	$\{50, 48, 40, 12\}$	34.2

Table 3: Implementation of Algorithm 2 on the Iris data set

The most appropriate number of clusters will be determined by indexes mentioned previously (see Fig. 4). Note that only the Area Index is correctly pointing to the partition with three clusters as the most appropriate partition. Confusion matrix $\begin{bmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 7 & 43 \end{bmatrix}$*, Rand Index (0.868) and Jacard Index (0.838) also show very well the coincidence of the obtained optimal partition with three clusters and the original partition. All of this leads to the conclusion that Algorithm 2 has found a globally optimal partition with three clusters which mostly coincides with the original partition* Π_0*.*

Other indexes pointed to the partition with two clusters.

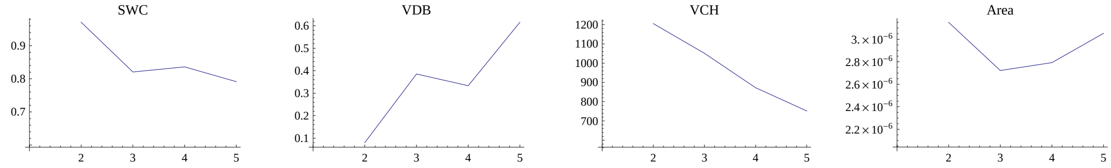


Figure 4: Indexes of optimal partitions for the data from Example 3

[2] UCI Machine Learning Repository, available at `http://archive.ics.uci.edu/ml/datasets/Iris`

4 Application to seismic zoning

Probabilistic Seismic Hazard Analysis (hereinafter referred to as: PSHA) estimates the probabilities of exceeding various ground motion levels at a site, given all possible earthquakes, [5]. The first approach was done by [29]. However, the numerical models initiated by [6] are currently preferred. Before a PSHA can be conducted, in areas of moderate seismic activity, seismogenic zonings have to be created. In this section, zonings for Croatia and the Iberian Peninsula are proposed according to Algorithm 1's output.

4.1 Application to construct the seismic zoning map of Croatia

Algorithm 1 will be applied for earthquake zoning in a wider area of the Republic of Croatia. The information associated with earthquakes around the world since 1971 is public and available at: `http://earthquake.usgs.gov/earthquakes/eqarchives/epic/`. Based on these data, the data set

$$\mathcal{A} = \{a_i = (\lambda_i, \varphi_i) \in \mathbb{R}^2 \colon 13 \leq \lambda_i \leq 20, \quad 42 \leq \varphi_i \leq 47, \quad w_i = M_i \geq 3\}, \tag{12}$$

is determined, consisting of 3184 locations in this area that have been affected by earthquakes of magnitude larger than or equal to 3.0 since 1973. Locations of these earthquakes are depicted in Fig. 5, large magnitude earthquakes are marked by bigger black dots. The abscissae of the points represent the longitudes and the ordinates represent the latitudes.

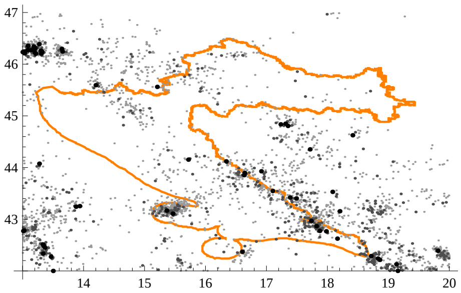


Figure 5: Locations in a wider area of the Republic of Croatia affected by the earthquakes of magnitude larger than or equal to 3.0 since 1973

Since latitudes and longitudes of these data are not of equal range, first they should be normalized according to the transformation (2). After clustering the normalized data, inverse transformation is applied to the obtained results.

In each iteration of Algorithm 1, the corresponding `SWC`, `VDB`, `VCH` and `Area` indexes are calculated. Fig. 7 shows graphs of these indexes. On the basis of these graphs, one can conclude that the partition with 7 clusters is the most appropriate one, but it makes sense to consider also the partitions with 11 or 13 clusters (see Fig. 6).

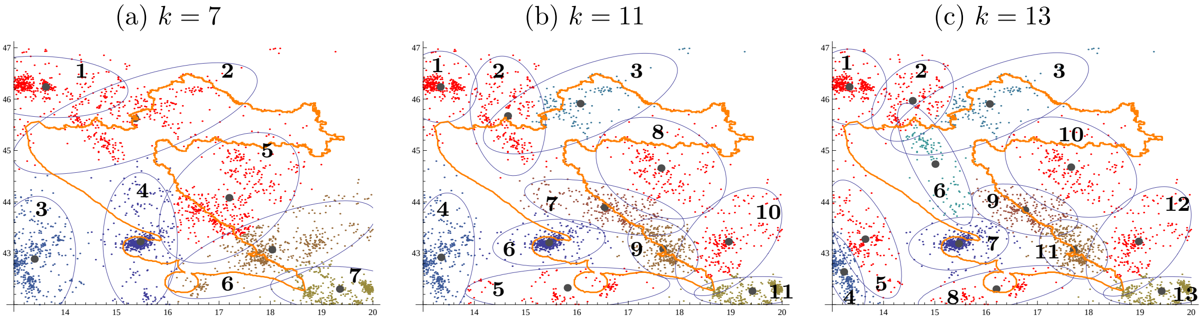

(a) $k = 7$ (b) $k = 11$ (c) $k = 13$

Figure 6: Optimal partitions with the most appropriate number of clusters

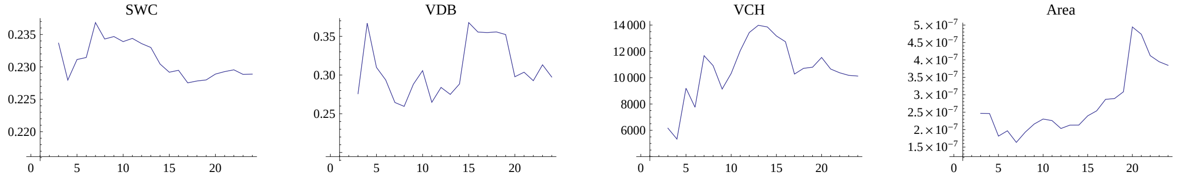


Figure 7: Indexes of optimal partitions for Croatian earthquake data

The following paragraphs provide a geophysical interpretation of the results obtained by the proposed algorithm. Croatia comprises several geotectonic units. The most important are the Pannonian Basin, the eastern Alps, the Dinarides, the transition zone between the Dinarides and the Adriatic Platform, and the Adriatic Platform itself [25]. The seismicity is not uniformly distributed. The most active zone is the coastal part, i.e. the Dinarides, due to the activity generated by the contact between the Adriatic Platform and the Dinarides [32]. North-west Croatia is the most seismically vulnerable zone due to the concentration of population and its economical importance [16].

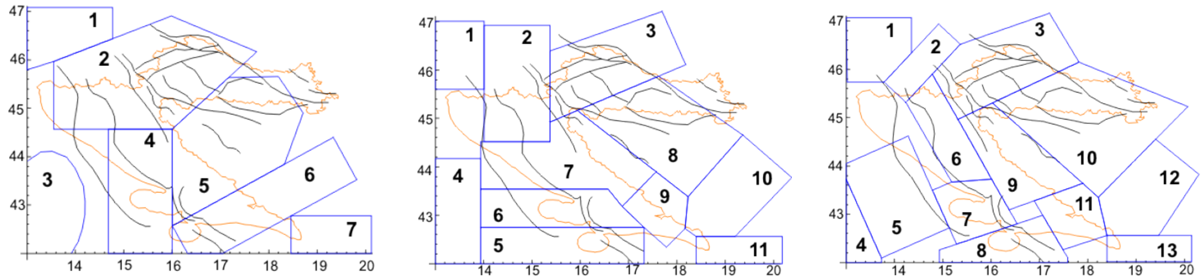
In this paper only the partition with seven zones will be considered. Henceforth, the zones have been labelled as $Z_j^{(7)}$, $j = 1, \dots, 7$ and described in Table 4.

Zone	Description
$Z_1^{(7)}$	Slovenia and northwestern Italy
$Z_2^{(7)}$	Northern Croatia
$Z_3^{(7)}$	Eastern Italy
$Z_4^{(7)}$	Adriatic and Dalmatia
$Z_5^{(7)}$	Dinara and Bosnia and Herzegovina
$Z_6^{(7)}$	Ston–Metković, southern Adriatic and southern Bosnia and Herzegovina
$Z_7^{(7)}$	Dubrovnik and Montenegro

Table 4: Earthquake zoning of Croatia (seven clusters)

$Z_1^{(7)}$ corresponds to Slovenia and northwestern Italy. It is a zone of high seismic activity. Strong historical earthquakes are known to have happened therein. $Z_2^{(7)}$ is a broad zone that matches with the north of Croatia. The seismicity activity is moderate and large earthquakes are unlikely. This vast zone comprises the northern part of the Rijeka–Mt. Velebit fault and the Trieste–Dugi Otok Island fault. It also contains the Fella–Sava–Črnomelj–Bihać fault, the southern marginal fault of the Pannonian basin, the Periadriatic–Drava fault and the Mt. Medvednica fault zone. $Z_3^{(7)}$, eastern Italy, shows a distributed seismic activity. $Z_4^{(7)}$ is a zone of moderate seismic activity. In the Adriatic and Dalmatia zone, earthquakes of magnitude greater than 5 are rare. This zone contains the southern part of the Rijeka–Mt. Velebit fault and the Sinj–Imotski fault. $Z_5^{(7)}$ corresponds to Dinara and Bosnia and Herzegovina. Earthquakes are frequent although of moderate magnitude. It comprises the Banka Luka fault and the Sinj–Imotski fault. $Z_6^{(7)}$ runs from west to east from southern Adriatic to Bosnia and Herzegovina, centered at Ston–Metković. This zone encloses the central part of the Mt. Mosor–Mt. Biokovo fault and the northern part of the Dubovnik fault. Moderate-large magnitude events happen after long periods of time. $Z_7^{(7)}$, Dubrovnik and Montenegro, is a zone characterized by high seismicity with earthquakes of magnitude up to 6.8. The southern part of the Mt. Mosor–Mt. Biokovo fault, the Dubovnik fault and the Adriatic fault can be found therein.

Finally, the zones have been smoothed according to geology. For that purpose, a new Fig. 8 that includes the faults from Croatia [25] has been depicted.



(a) $k = 7$

(b) $k = 11$

(c) $k = 13$

Figure 8: Smoothed zones for Croatia

4.2 Application to construct the seismic zoning map of the Iberian Peninsula

Algorithm 1 will be applied for earthquake zoning in a wider area of the Iberian Peninsula. Based on the data from the catalogue of the National Geographic Institute of Spain (NGIS: `www.ign.es`), the data set

$$\mathcal{A} = \{a_i = (\lambda_i, \varphi_i) \in \mathbb{R}^2 : \; -12 \le \lambda_i \le 6, \quad 33 \le \varphi_i \le 45, \quad w_i = M_i \ge 3\}, \tag{13}$$

is determined, consisting of 9327 locations in this area that have been affected by earthquakes of magnitude larger than or equal to 3.0 since 1978. Locations of these earthquakes are denoted in Fig. 9. As in Fig. 5, the abscissae of the points represent the longitudes and the ordinates represent the latitudes.

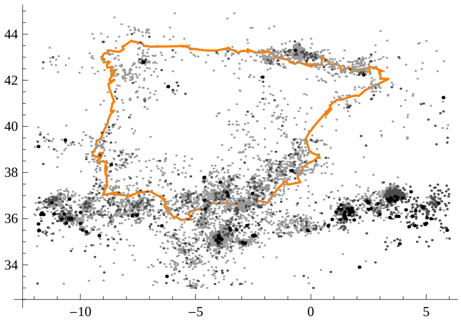


Figure 9: Locations in a wider area of the Iberian Peninsula affected by the earthquakes of magnitude larger than or equal to 3.0 since 1978

Since latitudes and longitudes of these data are not of equal range, first they should be normalized according to the transformation (2). After clustering the normalized data, inverse transformation is applied to the results obtained.

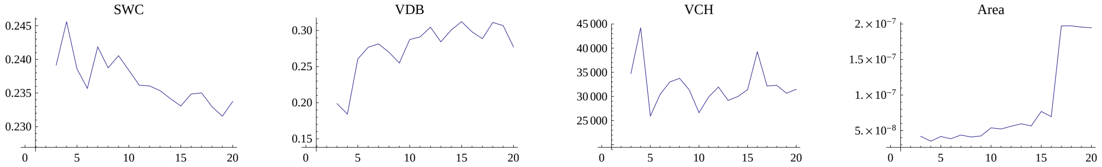
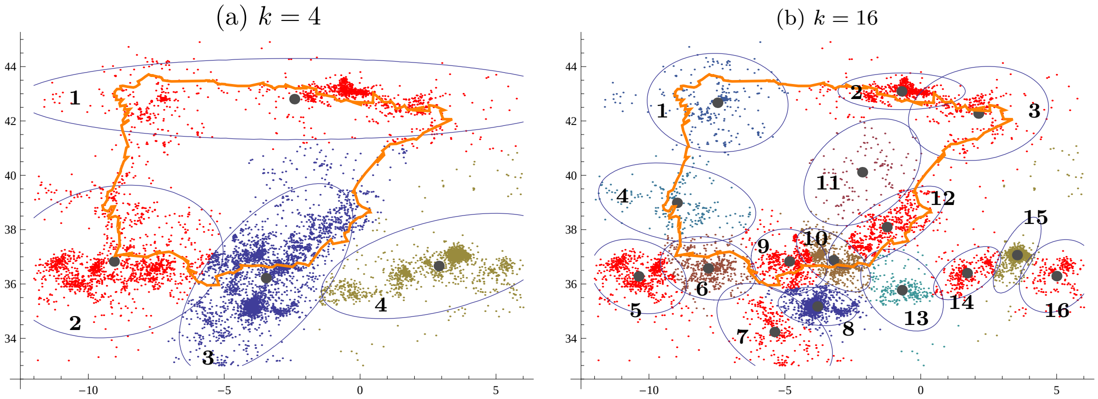


Figure 10: Indexes of optimal partitions for Iberian Peninsula earthquake data

In each iteration of Algorithm 1, the corresponding `SWC`, `VDB`, `VCH` and `Area` indexes are calculated. Fig. 10 shows graphs of these indexes. All indexes point to the partition with four clusters as the partition with the most appropriate number of clusters (see Fig. 11a). The area index and the `VCH` index allow us to conclude that in case of Spanish data the most appropriate number of clusters could be 16, too (see Fig. 11b).


(a) $k = 4$

(b) $k = 16$

Figure 11: Optimal partitions with the most appropriate number of clusters

The following paragraphs provide a geophysical interpretation of the results obtained by the proposed algorithm. The convergence directed NW-SE between the African and the Eurasian plate causes a deformation of the crust of the Iberian Peninsula, the Maghreb, and the adjacent coastal areas of the Mediterranean and the Atlantic [18]. The plate limit is heterogeneous. Continental and oceanic areas are in contact and progressive changes in the stress direction happen. The area corresponding to the Iberian Peninsula and northwest Africa can be considered the most complicated contact area with a moderate seismicity in relation to the magnitude of earthquakes. This area is surrounded on both sides by a frequent seismic activity with very large earthquakes [14, 41]. The seismic activity extends to interplate areas placed far away, such as the northeast of the Iberian Peninsula.

The seismicity of the Iberian Peninsula is characterized by the occurrence of moderate-magnitude earthquakes with a magnitude generally less than 5 [30]. Large earthquakes are separated by long periods of time [4]. Many of the earthquakes are located in the east of the Gibraltar Arch and spread over a diffuse area of approximately 500 km wide, centered in the Alboran Sea, containing parts of the southeast of Spain, the north of Morocco and Algeria.

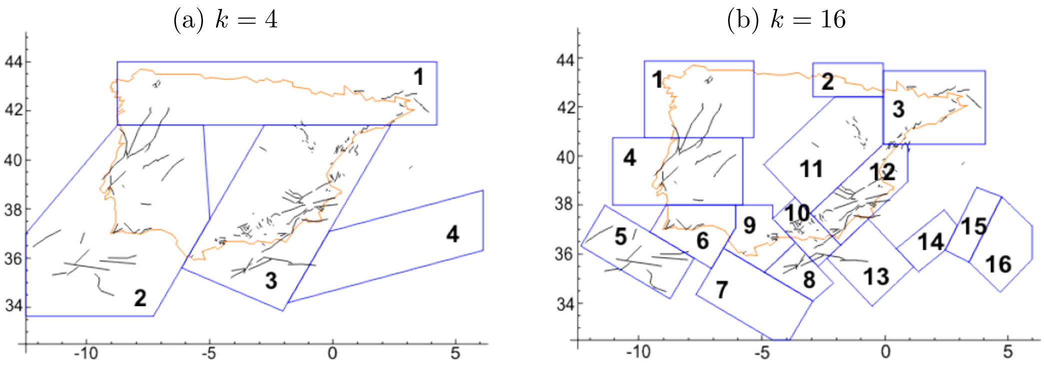
In this paper only the partition with four zones $Z_j^{(4)}$, $j = 1, \ldots, 4$, will be considered. The zones are described in Table 5.

Zone	Description
$Z_1^{(4)}$	Galicia, the Cantabrian mountain mass and the Pyrenees
$Z_2^{(4)}$	The Azores-Gibraltar fault and the south-west of the Iberian Peninsula
$Z_3^{(4)}$	The Betic system, the Alboran Sea, the north of Morocco and the Gibraltar field
$Z_4^{(4)}$	The Tell

Table 5: Earthquake zoning of the Iberian Peninsula (four clusters)

$Z_1^{(4)}$ corresponds to North Spain and includes Galicia, the Cantabrian mountain mass and the Pyrenees. It is a zone of moderate seismic activity. $Z_2^{(4)}$ matches the Azores–Gibraltar fault and the south-west of the Iberian Peninsula. The Azores–Gibraltar fault is characterized by a persistent seismic activity. Large earthquakes are known to have happened within that fault. The south-west of the Iberian Peninsula is mainly affected by the influence of earthquakes of the Azores–Gibraltar fault. $Z_3^{(4)}$ corresponds to the Betic system, the Alboran Sea, the north of Morocco and the Gibraltar field. It is a zone centered in the Alboran Sea. The seismic activity is usual near the Alboran Sea and it decreases for the outlying ones. $Z_4^{(4)}$ is the Tell. The zone presents a high seismic activity and large earthquakes are frequent.

Also the zones for the Iberian Peninsula have been smoothed according to geology. Fig. 12 shows the zones and the active faults for the Iberian Peninsula. These faults have been obtained from the Geological and Mining Institute of Spain (IGME: `www.igme.es`). These are the active faults from the Quaternary and are able to generate seismic activity.



(a) $k = 4$

(b) $k = 16$

Figure 12: Smoothed zones for the Iberian Peninsula

5 Conclusions

In the paper an efficient adaptive Mahalanobis k-means algorithm is constructed and a new efficient algorithm for searching for a globally optimal partition obtained by using the adaptive Mahalanobis distance-like function is proposed. Even if one cannot assert that a globally optimal partition is reached, numerous calculations show that the solution obtained by this algorithm is a satisfactory approximation of the globally optimal solution. Therefore, it is acceptable for applied research.

An important advantage of the proposed algorithm is that it successively gives optimal partitions with $k = 2, 3, \ldots$ clusters. Therefore, for each $k \geq 2$, it is immediately possible to calculate the value of various validity indexes and in this way to estimate the most appropriate number of clusters in a partition. It should also be highlighted that a novel `Area index` to measure the quality of the created partitions has been proposed and used in combination with other well-known indexes.

For regions of moderate seismic activity it is necessary to depict seismogenic zones before conducting a PSHA. Two regions of moderate seismic activity, Croatia and the Iberian Peninsula, have been analyzed. The task of depicting seismogenic zones involves a high degree of subjectivity as it depends on the author's knowledge and criteria. One of the main advantages of the algorithm lies in its ability to depict zones without considering a human decision. Moreover, it also estimates the best number of clusters. Another advantage is that the algorithm is able to plot not only circular but also elliptical zones. Three maps have been proposed for Croatia and two for the Iberian Peninsula. Finally, it must be noted that a satisfactory correlation with findings obtained using tools from geological sciences is obtained.

References

[1] S. Akkar and B. Glavatovic. Harmonization of seismic hazard maps for the western Balkan countries. Technical report, Science for Peace and Security Programme (NATO), 2010.

[2] A. M. Bagirov. Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41:3192–3199, 2008.

[3] E. Buforn, B. Benito, C. Sanz de Galdeano, C. del Fresno, D. Muñoz, and I. Rodríguez. Study of the damaging earthquakes of 1911, 1999 and 2002 in the Murcia, Southeastern Spain, region: seism-tectonic and seismic-risk implications. *Bulletin of the Seismological Society of America*, 95:549–567, 2005.

[4] E. Buforn, A. Udías, and M. A. Colombás. Seismicity, source mechanisms and seismotectonics of the Azores-Gibraltar plate boundary. *Tectonophysics*, 152:89–118, 1988.

[5] Senior Seismic Hazard Analysis Committee. Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on uncertainty and use of experts. Technical Report CR-6372, US Nuclear Regulatory Commission Report, 1997.

[6] C. A. Cornell. Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58(2):729–754, 1968.

[7] B. Durak. *A Classification Algorithm Using Mahalanobis Distances Clustering of Data with Applications on Biomedical Data Set.* PhD thesis, The Graduate School of Natural and Applied Sciences of Middle East Technical University, 2011.

[8] D. E. Finkel. *DIRECT Optimization Algorithm User Guide.* Center for Research in Scientific Computation. North Carolina State University, 2003. `http://www4.ncsu.edu/definkel/research/index.html`.

[9] J. M. Gablonsky. DIRECT version 2.0. Technical report, Center for Research in Scientific Computation. North Carolina State University, 2001.

[10] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications.* SIAM, Philadelphia, 2007.

[11] J. García-Mayordomo. Considering geological data and geologically based criteria in seismic hazard analysis of moderate activity regions: I. Definition and characterization of seismogenic sources. *Geogaceta*, 41:87–90, 2007.

[12] I. Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:773–781, 1989.

[13] J. J. Giner, S. Molina, J. Delgado, and P. J. Jáuregui. Mixing methodologies in seismic hazard assessment via a logic tree procedure: an application for Eastern Spain. *Natural Hazards*, 25:59–81, 2003.

[14] E. Gracia, R. Pallas, J. I. Soto, M. Comas, X. Moreno, and E. Masana. Active faulting offshore SE Spain (Alboran Sea): implications for earthquake hazard assessment in the Southern Iberian Margin. *Earth and Planetary Science Letters*, 241(3):734–749, 2006.

[15] R. Grbić, E. K. Nyarko, and R. Scitovski. A modification of the DIRECT method for Lipschitz global optimization for a symmetric function. *Journal of Global Optimization*, 57(4):1193–1212, 2013.

[16] D. Herak, M. Herak, and B. Tomljenović. Seismicity and earthquake focal mechanisms in North-Western Croatia. *Tectonophysics*, 465:212–220, 2009.

[17] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79:157–181, 1993.

[18] A. A. Kiratzi and C. B. Papazachos. Active crustal deformation from the Azores triple junction to the Middle East. *Tectonophysics*, 152:1–24, 1995.

[19] J. Kogan. *Introduction to Clustering Large and High-dimensional Data.* Cambridge University Press, 2007.

[20] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36:451–461, 2003.

[21] P. J. G. Lisboa, T. E. Etchells, I. H. Jarman, and S. J. Chambers. Finding reproducible cluster partitions for the k-means algorithm. *BMC Bioinformatics*, 14(Suppl. 1)(S8):1–19, 2013.

[22] C. López-Casado, C. Sanz de Galdeano, J. Delgado, and M. A. Peinado. The b parameter in the Betic Cordillera, Rif and nearby sectors. Relations with the tectonics of the region. *Tectonophysics*, 248:277–292, 1995.

[23] C. López-Fernández, J. A. Pulgar, J. Gallart, J. M. González-Cortina, J. Díaz, and M. Ruiz. Zonación sismotectónica del NO de la Península Ibérica. *Geo-Temas*, 10:1031–1034, 2008.

[24] S. Markušić. Seismicity of Croatia. *NATO Science Series: IV: Earth and Environmental Sciences*, 81:81–98, 2008.

[25] S. Markušić and M. Herak. Seismic Zoning of Croatia. *Natural Hazards*, 18:169–285, 1999.

[26] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar-Ruiz. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1230–1243, 2011.

[27] A. J. Martín. *Riesgo sísmico en la Península Ibérica.* PhD thesis, Polytechnical University of Madrid, 1984.

[28] J. Mezcua, J. Rueda, and R. M. García-Blanco. A new probabilistic seismic hazard study of Spain. *Natural Hazards*, 59:1087–1108, 2011.

[29] W. G. Milne and A. G. Davenport. Distribution of earthquake risk in Canada. *Bulletin of the Seismological Society of America*, 59(2):729–754, 1969.

[30] A. Morales-Esteban, J. L. de Justo, F. Martínez-Álvarez, and J. M. Azañón. Probabilistic method to select calculation accelerograms based on uniform seismic hazard acceleration response spectra. *Soil Dynamics and Earthquake Engineering*, 43:174–185, 2012.

[31] A. Neumaier. Complete search in continuous global optimization and constraint satisfaction. *Acta Numerica*, 13:271–369, 2004.

[32] J. M. Nocquet and E. Calais. Geodetic measurements of crustal deformation in the western Mediterranean and Europe. *Pure and Applied Geophysics*, 161:661–681, 2004.

[33] J. D. Pintér. *Global Optimization in Action (Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications).* Kluwer Academic Publishers, Dordrecht, 1996.

[34] K. Sabo, R. Scitovski, I. Vazler, and M. Zekić-Sušac. Mathematical models of natural gas consumption. *Energy Conversion and Management*, 52(3):1721–1727, 2011.

[35] R. Scitovski and K. Sabo. Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters. *Knowledge-Based Systems*, 57:1–7, 2014.

[36] R. Scitovski and S. Scitovski. A fast partitioning algorithm and its application to earthquake investigation. *Computers & Geosciences*, 59:124–131, 2013.

[37] H. Späth. *Cluster-Formation und Analyse.* R. Oldenburg Verlag, Munich, 1983.

[38] M. Teboulle. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8:65–102, 2007.

[39] S. Theodoridis and K. Koutroumbas. *Pattern Recognition.* Academic Press, Burlington, 4th edition, 2009.

[40] B. Tomljenović, D. Herak, M. Herak, and K. Kralj. Seismogenic zones of northwestern Croatia. In *Proceedings of the Convegno Nazionale di Geofisica della Tierra Solida*, pp. 46–47, 2008.

[41] G. Vanucci and P. Gasperini. The new release of the database of Earthquake Mechanisms of the Mediterranean Area (EMMA Version 2). *Annals of Geophysics*, 47:307–334, 2009.

[42] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. On the comparison of relative clustering validity criteria. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 733–744, 2009.

[43] L. Ye, C. Qiuru, X. Haixu, L. Yijun, and Z. Guangping. Customer segmentation for telecom with the k-means clustering method. *Information Technology Journal*, 12(3):409–413, 2013.

[44] K. S. Younis, *Weighted Mahalanobis distance for hyper-ellipsoidal clustering*, Ph.D. thesis, Air Force Institute of Technology, Ohio, 1999.