Domaining by clustering multivariate geostatistical data

Thomas Romary*[1], Jacques Rivoirard[1], Jacques Deraisme[2], Cristian Quinones[3] and Xavier Freulon[3]

Abstract

Domaining is very often a complex and time-consuming process in mining assessment. Apart from the delineation of envelopes, a significant number of parameters (lithology, alteration, grades) are to be combined in order to characterize domains or subdomains within the envelopes. This rapidly leads to a huge combinatorial problem. Hopefully the number of domains should be limited, while ensuring their connectivity as well as the stationarity of the variables within each domain. In order to achieve this, different methods for the spatial clustering of multivariate data are explored and compared. A particular emphasis is placed on the ways to modify existing procedures of clustering in non spatial settings to enforce the spatial connectivity of the resulting clusters. K-means, hierarchical methods and model based algorithms are reviewed. The methods are illustrated on a simple example and on mining data.

1 Introduction

In mining assessment, once the delineation of mineralization envelopes has been performed, it is often necessary to partition the area inside this envelope into several homogeneous subdomains. This is particularly the case when the extracted materials have to be subsequently chemically processed. It is also helpful to assess the viability of a mining project for its planning optimization. A significant number of parameters (lithology, alteration, grades...) are to be combined in order to characterize domains or subdomains. This rapidly leads to a huge combinatory. Methods to automatize this task are therefore

* thomas.romary@mines-paristech.fr,[1] Mines Paristech,[2] Géovariances,[3] AREVA

necessary. Almost no method has been proposed in the literature expect from an univariate approach based on grade domaining ([5]). Consequently, we focus here on a sensible approach that consists in adapting statistical clustering procedures.

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval and bioinformatics ([7]).

In the settings of independent observations, no particular structure is expected among the data. In a geostatistical context however, one expects to obtain a classification of the data that presents some spatial connexity.

Clustering in a spatial framework has been mainly studied in the image analysis and remote sensing context where the model is usually the following: the true but unknown scene, say t, is modeled as a Markov random field and the observed scene, say x, is interpreted as a degradation version of t, such that conditionally on t, the values x_i are independent to each other. In this model, label properties and pixel values need only to be conditioned on nearest neighbors instead of on all pixels of the map, see e.g. [6] for a review.

Clustering of irregularly spaced data (i.e. geostatistical data) has not been much studied. Oliver and Webster [8] proposed a method for clustering multivariate non-lattice data. They proposed to modify the dissimilarity matrix of the data by multiplying it by a variogram. Although this approach leads to a sensible algorithm, the method was not fully statistically grounded. Indeed, it terms to smooth the dissimilarity matrix for pairs of points at short distances but will not enforce the connexity of the resulting clusters. Contrarily, this tends to mitigate the borders between geologically different areas, making it difficult to differentiate between them.

Ambroise et al. [2] proposed a clustering algorithm for Markov random fields based on the expectation-maximization algorithm (EM, see [4]) that can be applied to irregular data using a neighborhood defined by the Delaunay graph of the data (i.e. the nearest-neighbor graph based on the Vorono tessellation). However this neighborhood structure does not reflect a structure in the data, but rather the structure in the sampling scheme. A Gaussian Markov random field model, while adapted to lattice data, is not natural on such a graph. Furthermore, this method does not ensure the connexity of the resulting clusters either.

Finally, Allard and Guillot [1] proposed a clustering method based on an approximation of the EM algorithm for a mixture of Gaussian random functions model. However this method relies on strong assumptions that are not likely to be encountered in practice and particularly with mining deposit data: the data are assumed to be Gaussian and data belonging to different clusters are assumed independent. Moreover, this last method is not suitable to large multivariate datasets as it computes the maximum likelihood estimator of

the covariance matrix at each iteration of the EM algorithm. Thus, a single iteration requires several inversions of a $(N \times P) \times (N \times P)$ matrix, where N is the number of data and P is the number of variables. This becomes quickly intractable as N and P increase.
In this paper, we first review existing procedures in an independent context. In section 2, we describe a novel geostatistical clustering algorithm that ensures the connexity of the resulting clusters. It is based on a slight modification of the hierarchical clustering algorithm. We compare its performances with other methods on a toy example. Finally, an application on mining data is exposed.

2 Review of some methods for independent observations

The goal of cluster analysis is to partition the observations into clusters such that those within each cluster are more closely related to one another than variables assigned to different clusters. A central notion for clustering is the degree of similarity (or dissimilarity) between the individual observations being clustered. A clustering method attempts generally to group the observations based on the definition of dissimilarity supplied to it.

2.1 Dissimilarity matrix

Most of the clustering algorithms take a dissimilarity matrix as their input, the first step is to construct pairwise dissimilarities between the observations. For quantitative variables, one can choose among euclidean, squared euclidean, 1-norm (sum of absolute differences), ∞-norm (maximum over absolute differences). For ordinal variables, where the values are represented as contiguous integers (*e.g.* alteration degree), error measures are generally defined by replacing their N original values with

$$\frac{i - 1/2}{N}, \quad i = 1 \dots N$$

in the prescribed order of their original values. They are then treated as quantitative variable on this scale. For unordered categorical variables however the degree of difference between pairs of values must be delineated explicitly (*e.g.* for geological factors). The most common choice is to take the distance between two observations to be 0 when they belong to different categories, 1 otherwise.
In a multivariate context, the next step is to define a procedure for combining the individual variable dissimilarities into a single overall measure of dissimilarity. This is done by means of a weighted average, where weights

are assigned to regulate the relative influence of each variable. In general, setting the weight as the inverse of the average individual dissimilarity for all variables will cause each one of them to equally influence the overall dissimilarity between pairs of observations. Variable that are more relevant in separating the groups should be assigned a higher influence in defining object dissimilarity.

2.2 Partitioning clustering

The most popular clustering algorithms directly assign each observation to a group or cluster without regard to a probability model describing the data. A prespecified number of clusters $K < N$ is postulated, and each one is labeled by an integer $k \in 1, \ldots, K$. Each observation is assigned to one and only one cluster. The individual cluster assignments for each of the N observations are adjusted so as to minimize a cost function that characterizes the degree to which the clustering goal is not met. A natural cost function is the sum over the clusters of the average distance between observations within each cluster. Cluster analysis by combinatorial optimization is straightforward in principle. As the amount of data increases however, one has to rely on algorithms that are able to examine only a very small fraction of all possible assignments. Such feasible strategies are based on iterative greedy descent. An initial partition is specified. At each iterative step, the cluster assignments are changed in such a way that the value of the criterion is improved from its previous value. The popular $K-means$ algorithm and its variant K-medoids are built upon that principle. In order to apply K-means or K-medoids one must select the number of clusters K and an initialization, see [11] for a review. The number of clusters may be part of the problem. A solution for estimating K typically examine the within-cluster dissimilarity as a function of the number of clusters K, see [7], chapter 14, for more details.

2.3 Hierarchical clustering

In contrast to K-means or K-medoids clustering algorithms, (agglomerative) hierarchical clustering methods do not require the choice for the number of clusters to be searched and a starting configuration assignment. Instead, they require the user to specify a measure of dissimilarity between (disjoint) groups of observations, based on the pairwise dissimilarities among the observations in the two groups. As the name suggests, they produce hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level.
Agglomerative clustering algorithms begin with every observation represent-

ing a singleton cluster. At each of the N-1 steps the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level. Therefore, a measure of dissimilarity between two clusters must be defined. *Single linkage* agglomerative clustering takes the intergroup dissimilarity to be that of the closest (least dissimilar pair). This is also often called the nearest-neighbor technique. *Complete linkage* agglomerative clustering (furthest-neighbor technique) takes the intergroup dissimilarity to be that of the furthest (most dissimilar) pair. *Group average* clustering uses the average dissimilarity between the groups. Although there have been many other proposals for defining intergroup dissimilarity in the context of agglomerative clustering (see *e.g.* [10]), the above three are the ones most commonly used.

2.4 Model-based clustering

Contrarily to the two previous methods, model-based clustering methods rely on the assumption that the data are drawn from a particular distribution. Generally, this distribution is a Gaussian mixture model, *i.e* a weighted sum of Gaussian distributions each with a different mean (which corresponds to the centroid in K-means) and covariance. Each component of the mixture defines a cluster, *i.e.* each observation will be considered to have been drawn from one particular component of the mixture.
The estimation of the parameters and the assignment of each observation to a cluster is conducted through an expectation-maximization (EM) algorithm [4]. The two steps of the alternating EM algorithm are very similar to the two steps in K- means. There exists a different version of the EM algorithm called classification EM (CEM, [3]) that may be more adapted to classification problems.

3 Geostatistical hierarchical clustering

In this section, we describe a novel geostatistical clustering algorithm that ensures the spatial connexity of resulting clusters. It is based on a slight modification of the hierarchical clustering algorithm described above. It practically consists of two steps: first, the data are structured on a graph according to their location; second, a hierarchical clustering algorithm is conducted where the merging of two clusters is conditioned by their connection in the graph structure. This enforce the spatial connexity of the clusters while respecting the dissimilarities between pairs of observation.

3.1 Algorithm

The first step of the proposed algorithm consists in building a graph over the data to structure them with respect to their proximity. In two dimensions, this task is straightforward as we can consider the Delaunay triangulation, associated to the sampling scheme of the data. Powerful algorithms exist to carry out this task efficiently. Figure 1 *b.* presents an example of a Delaunay triangulation associated to the sampling performed for the next section example. We can see that each point is connected to surrounding points, not necessarily to its nearest neighbours. Some branches of that graph may seem too long, particularly on the borders of the field. The graph can be post processed by pruning the longest edges so as to avoid spurious connections.
In a geological 3-dimensional context however, the Delaunay tetrahedralization, apart from being tricky to compute, may not be relevant for the purpose of domaining. As an example, consider a vein-type deposit with non horizontal veins. We would like one vein to belong to a unique cluster, which implies the samples belonging to the vein to be connected. Suppose that samples are located along parallel cores. The tetrahedralization will produce flat horizontal tetrahedra and the subsequent connections between points will be irrelevant with the geological configuration. Therefore, we propose to proceed in two steps to build the connections between sample points:

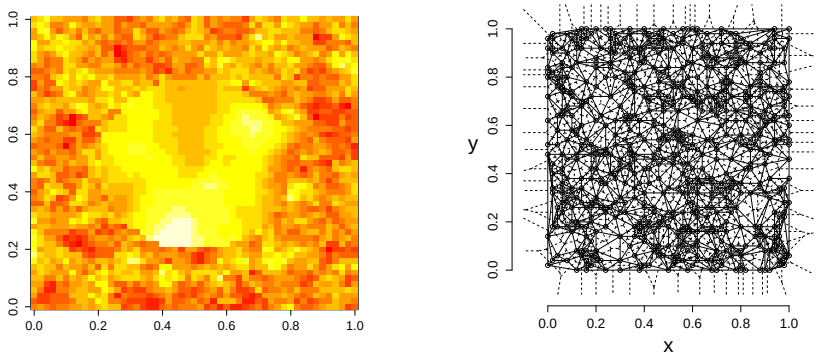
1. compute the Delaunay graph for one or several 2D surrogates of the deposit (linking the cores), possibly post process it,
2. extend the connections in the third dimension along the cores and between the cores by taking into account the geology (*e.g.* orientation), as far as possible.

Once the graph has been built, the second step of our method consists in running a slightly modified version of the hierarchical clustering algorithm (see section 2.3), the trick being to authorize two clusters to merge only if they are connected (two clusters are considered connected if there exists a connected pair of points between the two clusters). This will ensure the connexity of the resulting clusters. We chose to perform complete linkage clustering upon numerical experiments results, as it tends to produce more compact clusters. Finally, the user can choose the hierarchical level of clustering to be considered in the final classification.

3.2 Example

Here, we describe a 2D example on which we have evaluated the performances of some previously exposed methods. We consider a random function on the unit square which is made of a Gaussian random function with mean 2 and a cubic covariance with range 0.3 and sill 1 on the disk of radius 0.3

and center (0.5,0.5) and a Gaussian random function with mean 0 and an exponential covariance with range 0.1 and sill 1. A realization is shown in figure 1 *a.* while figure 1 *b.* corresponds to Delaunay graph associated to the sampling performed by picking 650 points out of the 2601 points of the complete realization.



a. *b.*

Fig. 1 Example dataset: full realization *a.* and Delaunay triangulation corresponding to the sampling performed *b.*

While we can clearly see a smooth surface with high values in the central disk in figure 1 *a.*, it is much more difficult to distinguish between the two areas in figure 1 *b.*, which makes this example challenging. We now test the performances of four different methods for this task: the K-means algorithm, the complete linkage hierarchical clustering algorithm (HC), Oliver and Webster's method (O&W) and our geostatistical hierarchical clustering (GHC) algorithm.

Figure 2 shows the results obtained by each four methods. Each subpicture represents the dataset on scatterplots with respect to the coordinates (X and Y) and the sampled value (Z). K-means (*a.*) identifies well the central area but the result lacks of connexity. It can be seen that the method discriminates between low and high values: the limiting value between the two clusters can be read as 0.5. HC (*b.*) also discriminates between low and high value but the limiting value is lower. To sum up, those two classical methods in an independent observations context fail to produce spatially connected clusters. O & W's approach has been tested with various variograms and variogram parameter values but it never showed any structured result (*c.*). The interpretation that we give is that multiplying the dissimilarity matrix by a variogram may erase some dissimilarities, inducing a loss in the structure of the data. The GHC algorithm succeeded in providing a clustering with spatial

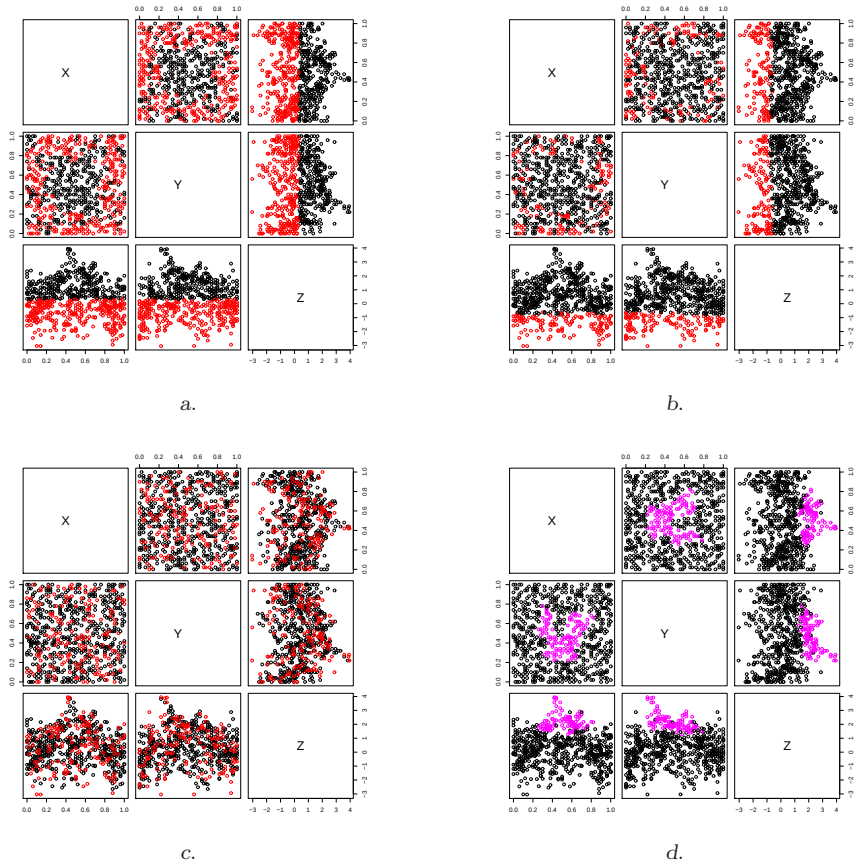


Fig. 2 Results of K-means *a.*, hierarchical clustering *b.*, Oliver and Webster's method *c.* and geostatistical hierarchical clustering *d.*

connexity (*d.*). A part of the disk is misclassified however. If we turn back to the complete realization in figure 1 *a*, we can see that the misclassified area corresponds to the low values of the realization around the border of the disk that are very close to the values taken outside the disk and are thus difficult to classify well.

We applied each four algorithms to 100 realizations of the same geostatistical model each with a different uniform random sampling. Then we computed the mean, median and 10% percentile of the rate of misclassified points. Results are summarized in table 1.

GHC exhibits the best performances overall whith 11% misclassified points in average while K-means is not so far, O & W performing the worst with the HC in between. If we look at the median however, GHC has the lowest one with a larger margin. The 10% percentile indicates that in the 10% most

	K-means	HC	O & W	GHC
Mean	0.13	0.23	0.35	0.11
Median	0.12	0.20	0.34	0.09
10% percentile	0.08	0.08	0.28	0.01

Table 1 Rates of misclassified points for the 4 algorithms

favorables cases, GHC misclassified only 0.01% of the points, while all the other algorithms performs a largely worse job. It can also be seen that the 10% percentile are similar for the K-means and the HC. This can be explain by the fact that the HC, and GHC (its worse result in this task was a misclassification of almost 50%), can sometimes perform really bad, whereas the K-means algorithm gives more stable results. In the favorable cases however, this algorithm works as well as the K-means. Concerning GHC, it performed worse than the K-means in less than 10% of the cases.

4 Application to an ore deposit

In this section, we present a preliminary study for the application of statistical clustering methods on a ore deposit. We describe the different steps and exhibits some results.
The first step has been to select the data that will be used for the domaining. The following variables have been chosen:

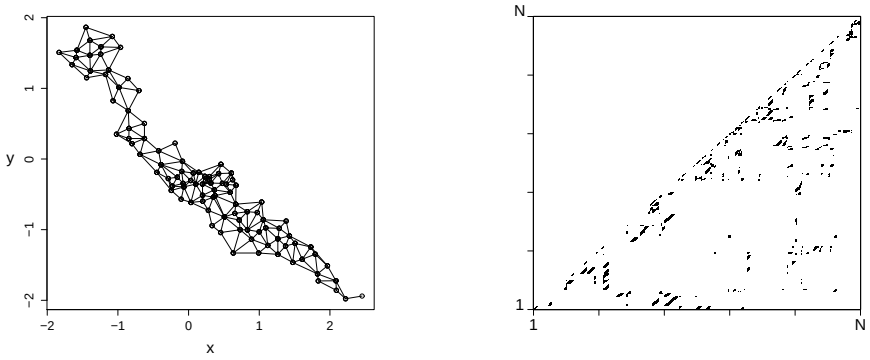
- coordinates, X, Y and Z
- uranium grades
- a geological factor describing the socle
- the hematization degree

This choice has been made upon an exploratory analysis of the data and discussion with geologists. Some transformations of the data have been performed:

- the coordinates have been normalized,
- uranium grades have been log-transformed and normalized,
- the degree of hematization has been transformed into a continuous variable, then normalized.

Then, the connections between close samples have been designed. As the mineralization envelop of the deposit exhibit a horizontal shape, we chose to build a 2D Delaunay triangulation first. To do that, we have selected for each core the observation closest to the median altitude of the whole sample. Then we performed the Delaunay triangulation on that subsample. Some branches of the graph were very long and did not correspond to vicinity. Consequently, the longest branches have been pruned and we obtained the graph pictured

in figure 3 *a*. Then, for each core, the adjacent observations are connected



a. *b.*

Fig. 3 Pruned Delaunay triangulation at median altitude *a.* and adjacency matrix *b.*

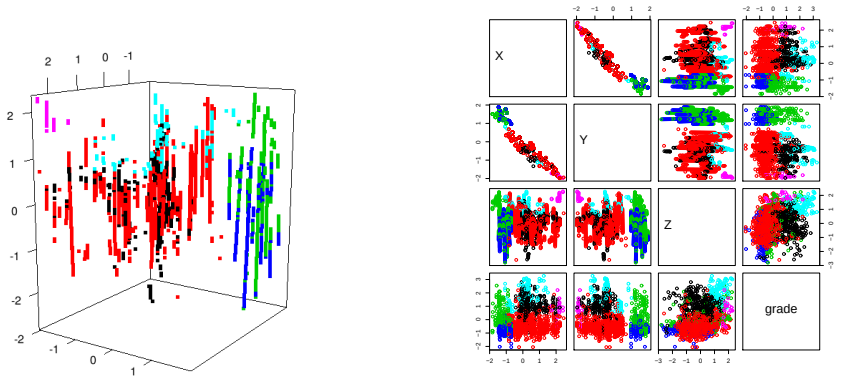
along the core. Close observations from two connected cores are connected as well. In this way, we built the adjacency matrix, an $n \times n$ lower diagonal matrix whose entries are one if the points are connected and zero otherwise. This matrix is plotted in figure 3 *b*. Note that the black points near the diagonal line corresponds to the connections along the cores. The diagonal of this matrix contains only zero.

The next step has been to build the dissimilarity matrix. This has been done using all the variables listed above and considering a particular distance for the geological factor: it has been chosen to be 1 when the samples have different factor values and 0 otherwise. Weights have been set by trial and error: we finally set the weights to 1 for the coordinates, 4 for the grade, 2 for the hematization degree and 10 for the geological factor.

Finally, we were able to run the geostatistical hierarchical algorithm described in section 3.1. We retained six clusters after the visualization of different hierarchical configurations. The results are depicted in figure 4 and described in table 2.

	Cyan	**Purple**	**Red Black**	**Blue Green**
(X,Y)	middle	S-E	middle to S-E	N-W
Z	top	top	middle to bottom	all along
Grade	high	middle	middle high	middle high
Geol.	sand	sand	socle	sand & socle
Hemat.	middle to large	low	no role	no role

Table 2 Description of the resulting clustering



a. *b.*

Fig. 4 Results of the algorithm with 6 clusters: in 3d *a.* and scatterplots *b.*

5 Conclusions

In this paper, we presented an insight towards geostatistically adapted clustering procedures. We presented an hierarchical algorithm conditioned to a connexity structure imposed on the data. Two applications have been provided, the first one on a toy example and the second on the deposit data.
The results shown on the toy example clearly assess the superiority of our method over tested ones as it is able to produce compact, connected clusters. The results obtained for the application where also satisfactory as they depicted a synthesised description of the deposit. Moreover, thanks to the sequential nature of the algorithm, our method generates a whole ensemble of clusterings that can be useful to the user: he can visualize the results at different hierarchical levels which leads to different interpretation levels. Furthermore, the user can also play with the weights of each variable to produce different clusterings, according to its knowledge of the geology.
Still, some improvements can be done. The first point is the way how we connect the observations in 3D. Performing the Delaunay to a 2D surrogate can be extended to non horizontal deposits by e.g. transforming the mineralization envelop into a flat manifold. Then the observations between connected cores should be connected if and only if they are not to distant away. Second, ways to define properly the weights associated to each variable according to the desired results should be investigated. Then, we could think of a more adapted linkage criterion than the complete linkage in the hierarchical algorithm. This new criterion would account for instance for the homogeneity of the cluster.
Finally, implementing a K-medoids algorithm based on the connection rela-

tions may be an interesting perspective, as it presents more appealing theoretical properties than hierarchical algorithms and is much faster.

References

[1] Allard, D., and Guillot, G. Clustering geostatistical data. In *Proceedings of the sixth geostatistical conference* (2000).

[2] Ambroise, C., Dang, M., and Govaert, G. Clustering of spatial data by the EM algorithm. In *geoENV I Geostatistics for Environmental Applications* (1995), A. S. et al., Ed., Kluwer Academic Publishers, p. pp. 493504.

[3] Celeux, G., and Govaert, G. A Classification EM Algorithm for Clustering and Two Stochastic versions. *Computational Statistics and Data Analysis*, 14 (1992), 315–332.

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

[5] Emery, X., and Ortiz, J. M. Defining geological units by grade domaining. Technical report, Universidad de Chile (2004).

[6] Guyon, X. *Random fields on a network*. Springer, 1995.

[7] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning*, 2nd edition, ed. Springer, 2009.

[8] Oliver, M., and Webster, R. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology 21* (1989), 15–35. 10.1007/BF00897238.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[10] Saporta, G. *Probabilités, analyses des données et statistiques*, 2nd edition ed. Technip, 2006.

[11] Steinley, D., and Brusco, M. J. Initializing k-means batch clustering: A critical evaluation of several techniques. *Journal of Classification 24* (2007), 99–121. 10.1007/s00357-007-0003-0.