CLUSTERING GEOSTATISTICAL DATA

D. ALLARD AND G. GUILLOT

Institut National de la Recherche Agronomique,
Unité de Biométrie, Domaine St-Paul, Site Agroparc,
84914 Avignon cedex 9, France

Abstract. We explore and compare different methods for the spatial clustering of geostatistical data. A new methodology based on the likelihood is proposed and compared to the approach by Allard and Monestiez (1999). Both methods are compared on a heavy metal concentration data set in the Swiss Jura.

1. Introduction

Clustering in a spatial context 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 $\mathbf{t}$, is modeled as a Markov random field and the observed scene, say $\mathbf{x}$, is interpreted as a degradation version of $\mathbf{t}$, such that conditionally on $\mathbf{t}$, the values x_i are independent to each other. In this model, label properties and pixel values need only be conditioned on nearest neighbors instead of on all pixels of the map. This was the basis for the works of Geman and Geman (1984), Haslett (1985) and Besag (1986, 1991) in Bayesian image restoration. A review of these methods can be found in Guyon (1995).

Clustering of irregularly spaced data (i.e. geostatistical data) has not been much studied. Oliver and Webster (1989) 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 and to well-behaved maps, the method was not fully statistically grounded. Allard (1998) and Allard and Monestiez (1999) proposed a very different approach, based on the minimization of the within variance/between variance ratio. Ambroise *et al.* (1997) proposed a clustering algorithm for Markov random fields based on the EM algorithm (Dempster *et al.*, 1977) that can be applied to irregular data using a neighbor-

hood defined by the Delaunay graph of the data (i.e. the nearest-neighbor graph based on the Voronoï tessellation).

In this paper, we explore a new clustering method for spatially correlated data. It is not based on the Markov random field paradigm, because Markov random fields rely heavily on the neighborhood structure (the nearest neighbors, for example). For irregularly spaced data, this neighborhood structure does not at all reflect a structure in the data, but rather the structure in the sampling scheme. In the example studied in section 4, the sampling scheme is partly random and is likely to introduce spurious effects in the neighborhood structure. Hence, a random function model with a covariance function was preferred. In this paper we propose a clustering method based on an approximation of the EM algorithm for this model.

The paper is organized as follows. After a short introduction (this section), some simple criteria are presented in section 2. In section 3, the EM algorithm and its application to the clustering of independent data are briefly recalled. Then, our new method for the clustering of geostatistical data using the EC-M algorithm is presented. These different approaches are applied to a heavy metal concentration data set in the Swiss Jura in section 4, before a brief discussion in section 5.

2. A model and two criteria

2.1. THE MODEL

Some region of the plane, say A, is divided in a partition of K unknown regions (i.e. $A = \cup_k A_k$ and $A_k \cap A_l = \emptyset$, if $k \neq l$), where K is known. Some univariate data $\mathbf{x} = (x_1, \ldots, x_n)^t$ are measured at a set of n locations, $\mathbf{s} = (s_1, \ldots, s_n)^t \in A^n$. The data are modeled as a random function $X(s)$, a mosaic of K independent stationary normal (Gaussian) random function on each region A_k, with mean μ_k, variance σ_k^2 and correlation function $\rho_k(\cdot)$:

$$E[X(s)] = \mu_k, \quad Cov(X(s), X(s')) = \sigma_k^2 \rho_k(s' - s), \quad Cov(X(s), X(s'')) = 0, \tag{1}$$

for $s, s' \in A_k$ and $s'' \in A_l,\ k \neq l$. The vector of the n_k data whose co-ordinates are in A_k is denoted $\mathbf{x}_k$, $\mathbf{R}_k$ is the correlation matrix built from the data locations, $\mathbf{R}_k[ij] = \rho_k(s_i - s_j),\ s_i, s_j \in A_k$, and $|\mathbf{R}_k|$ is the determinant of that matrix. The aim is to find the clustering of the data into K groups that best matches the model above. In practice we seek some criterion derived from the model to be minimized. Clusters are described using the $n \times k$ matrix $\mathbf{t}$, with elements $t_{ik} = 1$ if $s_i \in A_k$ and $t_{ik} = 0$ otherwise. The parameters μ_k, σ_k^2 and $\rho_k(\cdot)$ are unknown, and must be estimated along with the groups. The case where K is also unknown and must estimated will not be considered. For notational convenience, the parameters of the random function on A_k will be denoted $\theta_k = (\mu_k, \sigma_k^2, \rho_k(\cdot))$, and $\Theta = (\theta_1, \ldots, \theta_K)$. For this model the

likelihood of the data is

$$\begin{aligned} L(\Theta, \mathbf{t}; \mathbf{x}) &= \prod_{k=1}^{K} f_k(\mathbf{x}_k; \theta_k) \\ &= \frac{1}{(2\pi)^{n/2}} \prod_{k=1}^{K} \frac{1}{\sigma_k^{n_k}} \frac{1}{|\mathbf{R}_k|^{1/2}} \exp\left\{ -\frac{(\mathbf{x}_k - \mu_k \mathbf{1})^t \mathbf{R}_k^{-1} (\mathbf{x}_k - \mu_k \mathbf{1})}{2\sigma_k^2} \right\}. \end{aligned} \tag{2}$$

Here, and in all the paper, $f_k(\,\cdot\,;\theta_k)$ is the Gaussian density in group k, with parameters θ_k and appropriate dimension n_k. Three different clustering methods are now presented: two methods based on the minimization of a variance and likelihood criterion respectively, and one method using an approximation of the EM algorithm.

2.2. VARIANCE CRITERION

For fixed groups and fixed $\rho_k(\cdot)$s, the unbiased maximum likelihood estimators (MLEs) for μ_k and σ_k^2 are

$$\hat{\mu}_k = \frac{\mathbf{x}_k^t \mathbf{R}_k^{-1} \mathbf{1}_k}{\mathbf{1}_k^t \mathbf{R}_k^{-1} \mathbf{1}_k}, \quad \hat{\sigma}_k^2 = \frac{(\mathbf{x}_k - \hat{\mu}_k \mathbf{1}_k)^t \mathbf{R}_k^{-1} (\mathbf{x}_k - \hat{\mu}_k \mathbf{1}_k)}{n_k - 1}, \tag{3}$$

where $\mathbf{1}_k$ is a vector of length n_k. In geostatistical terms, $\hat{\mu}_k$ is the kriging of the mean of the random function X_k (Cressie 1993, Wackernagel 1995). It is well known that it is the best linear unbiased predictor of μ_k with variance $\sigma_k^2 / \mathbf{1}_k^t \mathbf{R}_k^{-1} \mathbf{1}_k$. Allard and Monestiez (1999) sought a clustering minimizing the variance in each group and maximizing the contrast between groups. They proposed to minimize the ratio W/B where $W = \sum_{k=1}^{K} (n_k - 1)\hat{\sigma}_k^2$ is the within variance and $B = \sum_{k=1}^{K} n_k (\hat{\mu}_k - \hat{\mu})^2$ is the between variance, computed according to (3). The average of X over A is $\mu = \int_A X(s)ds/|A|$. Hence, $\mu = \sum_k \mu_k |A_k|/|A|$ and $\hat{\mu} = \sum_k \hat{\mu}_k n_k / n$. Unlike the variance decomposition of independent data, $W + B$ is not constant in a spatially correlated context and minimizing W is not equivalent to maximizing B. Minimizing the ratio W/B was shown to be a reasonable choice.

2.3. LIKELIHOOD CRITERION

For the Gaussian model, a maximum likelihood criterion can alternatively be derived. Replacing μ_k and σ_k^2 by their MLEs in (2) yields (up to a constant term) the negative concentrated log-likelihood

$$2\ell(\Theta, \mathbf{t}; \mathbf{x}) = \sum_{k=1}^{K} (n_k \ln \hat{\sigma}_k^2 + \ln |\mathbf{R}_k|), \tag{4}$$

to be minimized. The first term favors groups with low variance, i.e. with similar values. Since $\mathbf{R}_k$ is a correlation matrix, the logarithm of its determinant is always negative (Searle 1982, p.260)

and is equal to 0 if there are no correlation. Thus, the second term acts as a regularization term, and it tends to favor groups with high spatial correlations, i.e. that are not scattered. This criterion is different from the previous one and leads to a different clustering.

The difficulty with the clustering methods based on these two criteria is that the correlation functions $\rho_k(h)$ must be known for computing W/B or ℓ and for finding the clusters minimizing these criteria. But on the other hand, the groups must be known for estimating the parameters, in particular the $\rho_k(h)$s. Readers are referred to Allard and Monestiez (1999) for computational details on an iterative algorithm estimating the clusters and the correlation functions in turn for the variance criteria. In this iterative algorithm, the clustering minimizing W/B for a given correlation function is found using a simulated annealing procedure. The correlation functions are estimated with the new clustering. In practice, a few iterations only are necessary because often the new experimental variograms can be fitted with the old model.

These methods will be illustrated on the Swiss Jura data set. They are very computer intensive and the minima are not easily found. Hence, a different approach is now proposed.

3. EM algorithm for spatial clustering

Instead of considering the indicator matrix $\mathbf{t}$ as an unknown parameter to be estimated as in the previous section, it will now be considered as some missing data in a maximum likelihood setting, with a probability model. A well known iterative algorithm for estimating missing data is the EM algorithm, which is now presented and applied to the spatial clustering problem.

3.1. THE EM ALGORITHM

The Expectation-Maximization (EM) algorithm is a general iterative algorithm that computes maximum likelihood estimates in the presence of missing data. A detailed presentation can be found in the seminal paper by Dempster, Laird and Rubin (1977); McLachlan and Krishnan (1997) provide a review with many examples and extensions. The general idea of the EM algorithm is to iteratively compute the augmented likelihood of the data (the augmented likelihood is the likelihood containing both the measured data and the missing data). Starting with some initial values, the expectation step (E-step) computes the conditional expectation of the augmented log-likelihood, given the current value of the parameters. The maximization step (M-step) computes the MLEs of the parameters, given the measured data and the updated missing data from the E-step. Dempster, Laird and Rubin (1977) showed that each step increases the augmented likelihood. These two steps are iterated until convergence occurs at a local maximum of the likelihood surface.

Clustering methods of independent data based on mixtures of normal (Gaussian) distributions coupled with an EM algorithm have been shown to be powerful, see for example McLachlan

and Basford (1988) and Banfield and Raftery (1993). The data are supposed to be originated from a mixture $f(x) = \sum_k p_k f_k(x)$ where the p_ks are the mixing proportions. In this case, because the data are independent, it is easily shown that the E-step is equivalent to estimating the conditional probabilities

$$\hat{t}_{ik} = P(x_i \in \text{group } k \mid (\hat{\theta}_1, \ldots, \hat{\theta}_K)) = \frac{\hat{p}_k f_k(x_i; \hat{\theta}_k)}{\sum_{l=1}^{K} \hat{p}_l f_l(x_i; \hat{\theta}_l)}. \tag{5}$$

After an E-step, the classification matrix is not a 0/1 matrix but each row of the matrix still adds up to one. Celeux and Govaert (1992) proposed a Classification EM algorithm (CEM) as an extension of the EM algorithm. In the CEM algorithm, a classification step is added between the E-step and the M-step in which the missing cluster indicators t_{ik} are replaced with their maximum a posteriori (m.a.p.) estimates:

$$\tilde{t}_{ik} = 1 \text{ if } \arg\max_l \hat{t}_{il} = k, \qquad \tilde{t}_{ik} = 0 \text{ otherwise}, \tag{6}$$

which is also denoted $\tilde{t}_{ik} = \text{m.a.p.}(\hat{t}_{ik})$. Celeux and Govaert showed that any sequence of the CEM algorithm increases the classification likelihood and that it converges to a stationary value. Moreover, if the iterates get close enough to a point that produces a local optimum, the CEM sequence will converge to it.

3.2. THE EC-M ALGORITHM FOR SPATIAL CLUSTERING

We now see how the EM algorithm can be applied to the problem of clustering spatially dependent data. The missing data $\mathbf{t}$ are modeled with a probability distribution $p(\mathbf{t})$ defined on $\Omega = \{1, \ldots, K\}^n$. As a first attempt, we will consider in this paper that each data location belongs to A_k with probability p_k, independently on all other data location. Hence, $p(\mathbf{t}) = \prod_k p_k^{n_k}$. This is for example the case if the spatial structure of the segmentation is of a smaller range than the distances between data location. Thus all the spatial information is supposed to be contained in the covariance structure of the Gaussian random functions, and in particular, in the absence of correlation between groups. This is certainly an oversimplification and applications of this algorithm with a more appropriate model remain to be developed. The complete log-likelihood of the augmented data is then

$$\ln L_c(\Theta; \mathbf{x}, \mathbf{t}) = \ln\{L(\Theta, \mathbf{t}; \mathbf{x}) p(\mathbf{t})\} = \ln L(\Theta, \mathbf{t}; \mathbf{x}) + \sum_k n_k \ln p_k. \tag{7}$$

Applying directly the EM algorithm to the likelihood (7) is not possible for several reasons:

1. In a cluster, data are not independent. Unlike usual model based clustering for independent data, all computations need to take into account the spatial auto-correlation of data

belonging to the same cluster. Thus, for being able to compute the M-step, it is necessary to work with a hard classification matrix instead of a fuzzy classification matrix $\hat{\mathbf{t}}$. This can done by using a classification version of the EM algorithm, thereby defining a classification matrix $\tilde{\mathbf{t}}$ at each iteration according to (6).

2. The E-step amounts to computing the conditional expectation of the log-likelihood (7), which is intractable. Hence approximations of the E-step must the found.

To solve these problems, we propose an algorithm for approximating the EM algorithm in the case of spatial clustering that addresses the two problems mentioned above. The EC-M algorithm, as its name indicates, is decomposed into two steps, the Expectation-Classification step considered as a whole, and the more usual Maximization step.

The EC step

The logarithm of the likelihood (7) can be rewritten using the trivial conditional decomposition:

$$\ln L_c(\Theta; \mathbf{x}, \mathbf{t}) = \sum_{k=1}^{K} t_{1k} \ln f_k(x_1) p_k + \sum_{i=2}^{n} \sum_{k=1}^{K} t_{ik} \ln f_k(\mathbf{x}_{<i}, \mathbf{t}_{<i}) p_k, \tag{8}$$

where $\mathbf{x}_{<i}$ and $\mathbf{t}_{<i}$ denote the data and the classification matrix for indices less than i. As mentioned above, the conditional expectation of this expression is not tractable. However, it can be approximated by the linear expression in which the conditioning at site x_i is computed using the m.a.p estimates at the previous sites:

$$\sum_{k=1}^{K} t_{1k} \ln f_k(x_1) p_k + \sum_{i=2}^{n} \sum_{k=1}^{K} t_{ik} \ln f_k(x_i \mid \mathbf{x}_{<i}, \tilde{\mathbf{t}}_{<i}) p_k. \tag{9}$$

This approximation leads to the following estimate for t_{ik}, $i = 1, \ldots, n$:

$$\hat{t}_{ik} = \frac{f_k(x_i \mid \mathbf{x}_{<i}, \tilde{\mathbf{t}}_{<i}) p_k}{\sum_l f_l(x_i \mid \mathbf{x}_{<i}, \tilde{\mathbf{t}}_{<i}) p_l}, \qquad \tilde{t}_{ik} = \text{m.a.p.}(\hat{t}_{ik}) \tag{10}$$

The algorithm is then to compute $\tilde{t}_{ik}$ sequentially for $i = 1, \ldots, n$ using equations and (10). Equation (8) is true for all ordering of the sites, but the approximation (9) depends very much upon the order of visit of the sites.

To circumvent this problem, an iterative procedure can be suggested for the EC step where the $\tilde{t}_{ik}$'s are not computed sequentially, but are rather computed in parallel, using all other $\tilde{\mathbf{t}}_{-i}$ computed at the previous iteration. In the first iteration, $\tilde{\mathbf{t}}^{(1)}$ is estimated ignoring the spatial correlation

$$\hat{t}_{ik}^{(1)} = \frac{f_k(x_i) p_k}{\sum_l f_l(x_i) p_l}, \tag{11}$$

where $f_k(x)$ is the marginal distribution of group k and $\tilde{\mathbf{t}}^{(1)} = \text{m.a.p.}(\hat{\mathbf{t}}^{(1)})$. Then, $\tilde{\mathbf{t}}^{(m+1)}$ is iteratively re-estimated until convergence occurs by computing:

$$\hat{t}_{ik}^{(m+1)} = \frac{f_k(x_i \mid \mathbf{x}_{-i}, \tilde{\mathbf{t}}_{-i}^{(m)}) p_k}{\sum_l f_l(x_i \mid \mathbf{x}_{-i}, \tilde{\mathbf{t}}_{-i}^{(m)}) p_l}, \qquad \tilde{\mathbf{t}}_i^{(m+1)} = \text{m.a.p.}(\hat{\mathbf{t}}_i^{(m+1)}) \tag{12}$$

where $\mathbf{x}_{-i}$ and $\tilde{\mathbf{t}}_{-i}^{(m)}$ are the vector $\mathbf{x}$ and the matrix $\tilde{\mathbf{t}}^{(m)}$ without index i. The conditional distributions f_k in (12) are easily computed: in the Gaussian case, they simply are Gaussian densities with means and variances given by simple kriging. We did not succeed in finding formal proof of the convergence of the parallel procedure, but in practice convergence was always reached in few iterations.

Note that the EC-step proposed above is in some respect similar to the Iterated Conditional Mode proposed by Besag (1986, 1991). The differences are in the spatial structure modeling. While Besag considered a Markov random field model for the unknown clusters and conditionally independent observed data, we consider independent and identically distributed prior probabilities p_k for group k and correlated observed data.

The M step

In the M-step, new parameters are estimated, given the classification matrix defined at the previous EC-step. For known correlation functions, means and variances can easily be estimated using equations (3), and $\hat{p}_k = \sum_i \tilde{t}_{ik}/n$. If the correlation function is not known, it must be estimated during the M-step. In general, we will assume a common parametric form for all correlation functions, with some parameters ψ_k different for each cluster. A simple example is the decreasing exponential function, $\rho_k(h) = \exp(-|h|/a_k)$. Plugging the unbiased estimates $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ in (7) yields (up to a constant term) the negative log profile likelihood for the cluster k:

$$2\ell_k(\psi_k; \mathbf{x}) = \ln |\mathbf{R}_k(\psi_k)| + n_k \ln \hat{\sigma}_k^2, \tag{13}$$

which will be minimized with respect to ψ_k.

Finally, the complete negative log-likelihood that is minimized is

$$\ell_c(\Theta; \mathbf{x}, \mathbf{t}) = \sum_{k=1}^{K} (\ell_k(\hat{\psi}_k; \mathbf{x}) - n_k \ln \hat{p}_k) = \frac{1}{2} \sum_{k=1}^{K} (n_k \ln \hat{\sigma}_k^2 + \ln |\mathbf{R}_k(\hat{\psi}_k)| - 2n_k \ln \hat{p}_k). \tag{14}$$

The difference between this expression and (4) is the last term, related to the entropy of the distribution $\hat{p}_k$. It favors groups of very different size (it is maximized when one group is of size n). The solution given by the EC-M algorithm leads to a local maximum of the complete log-likelihood (14).

4. Application to the Swiss Jura Data Set

We now apply these three clustering methods to a soil data set in the Swiss Jura. These data were first analyzed by Atteia, Dubois and Webster (1994). They are used by Goovaerts (1998) as illustration of his geostatistics textbook. A detailed presentation of these data can be found in these two references. A total of 359 data were collected on an area of 14.5 km^2. A first set of 214 regularly spaced data were collected. Among these, 29 data points were used as starting points for a nested sampling with distances 100 m, 40 m, 15 m and 6 m. The concentrations of Co, Cr, Cu, Ni, Pb and Zn were measured. The region embraces outcrops of several geological formations of Jurassic limestone. The geology seems to be the factor the most related to heavy metal concentration. In particular a one-way analysis of variance showed that most of the effect is attributed to the difference between the Argovian and the other formations data points. The average value of Co and Ni is about twice in Non Argovian (nA) than in Argovian (Ar).

Figure 1 shows the Ar/nA classification (first row) and the non spatial clustering (second row). Argovian data points are depicted with proportional circles and non Argovian data points are located with a dot. The Argovian rock type (76 data points) is situated on the Northern edge and in the South-East corner of the studied region. It includes mainly low Co values, with a few spots of higher values, specially in the extreme North-East. The non spatial clustering is defined by the cutoff maximizing the contrast between the two groups, without accounting for any spatial correlation. The group of lower values contains many more points (135) than the Argovian group. In addition to the two groups of the Argovian classification, it contains a scattered group of low values in the central-Eastern corner. Isotropic variograms were computed in each group. The Argovian group and the group of lower values have longer range than the non Argovian/higher values groups. Exponential variograms were fitted with the following parameters: $a = 0.3$ in the group of lower values, and $a = 0.15$ in the group of higher values. The sills of the variograms depicted Figure 1 are equal to the variances $\hat{\sigma}_k^2$ computed according to (3).

We will now apply the clustering methods presented in Section 2 and 3 using these parameters as starting values. Table 1 shows the main statistics of the different clusterings, compared to the Ar/nA statistics. These statistics are computed using the exponential variograms fitted above. For comparison purposes, the same variograms were used for computing the statistics of the Argovian classification and the non spatial clustering.

Figure 2 shows the clusterings of the minimization method using the simulated annealing algorithm for the criteria W/B and ℓ. The experimental variograms computed on the new groups showed good agreement to the variogram models in terms of the range parameters. Hence, there was no need to re-estimate new theoretical variograms and to re-run the simulated annealing procedure.

Minimizing the variance criterion leads to a reasonable clustering (Figure 2, first row). The

TABLE 1. Statistics computed on the different clusterings, including the Argovian classification. The range parameters used for computing $\hat{\mu}$ and $\hat{\sigma}^2$ are $a = 0.3$ and $a = 0.15$ in groups 1 and 2, respectively.

Clustering	*Group 1*			*Group 2*			*Global*	
	n	$\hat{\mu}$	$\hat{\sigma}^2$	n	$\hat{\mu}$	$\hat{\sigma}^2$	W/B	ℓ
Argovian	76	6.43	7.51	283	10.3	11.5	4.25	580
Non spatial	135	6.42	6.73	224	11.5	5.53	0.968	377
Variance crit.	112	5.42	4.29	247	11.1	6.14	0.804	355
Likelihood crit.	162	7.23	4.88	197	11.8	4.55	0.919	307
EC-M algorithm	135	6.54	4.69	224	11.4	5.09	0.890	327

two Argovian areas are very well recovered: a total of 58 Argovian data points (out of 76) are in the group of the lower values, along with 54 non Argovian values. There are two main differences with the Argovian map: the first one is in the North-East corner, where high values are excluded but many low values are included. The second one is the existence of a small and scattered area in the eastern part of the studied area. The statistics show a high contrast between the groups ($\hat{\mu}_1 = 5.42$ and $\hat{\mu}_2 = 11.1$) and a low variance in the groups. The estimated variances are in accordance with the variogram sills.

Minimizing the likelihood criterion turned out to be unsuccessful. The group of lower values is very large (162 points), and very scattered throughout the whole area. The contrast between the groups is low and the variances are underestimated compared to the sill of the variograms computed in the groups (Figure 2, second row). This can be explained as follows. The likelihood (4) involves only within-group variances and correlation matrices, but does not explicitly involve any means. Hence, minimizing the variance criterion is mainly minimizing the variance of the groups. The problem is that the maximum likelihood estimator (3) of the variance relies heavily on the Gaussian assumption. When the histogram of the data are close the a Gaussian histogram, the MLE performs pretty well. This is the case for the group of higher values (Figure 2b), in which the variogram sill is in very good accordance with the estimated variance. When the histogram is non-Gaussian, as it is the case in the group of lower values (Figure 2b), the MLE performs poorly, and the variance is underestimated when compared to the variogram sill. Hence, the solution minimizing the likelihood criterion may very well be a clustering with bi-modal histograms (hence, a low contrast between groups) underestimating the variances. Such a situation did not happen with the variance criterion because it is not only based on the variances, but also on the contrast B between the groups.

The clustering given by the EC-M algorithm is depicted Figure 3. It is intermediate between

the previous solutions. The group of lower values contains 135 data points. It includes 62 Argovian points (out of 76). It has 110 points in common with the clustering given by the variance criterion (out of 112). The contrast between the groups is high. An interesting fact is that the variances of the groups are well estimated.

5. Discussion

Three methods for clustering geostatistical data have been tested on a data set. None of these methods is fully satisfactory. The main reason is the absence of model (or more precisely the implicit assumption of an independence model) for the underlying Ar/nA process. The usual methods that can be found in the literature assume a Markov random field model for the underlying process, and conditional independence of the measurements. In this paper, the opposite model was tested: independence for the underlying process, and correlated measurements. The conclusion of this work is that using only the correlation of the measured data is not strong enough for estimating correctly the underlying process, ie. for clustering the data; a model describing the cluster process $\mathbf{t}$ must also be provided. Ideally this model should be defined on the plane, and not only on the nearest neighbor graph.

Having said that, the method minimizing the variance criterion W/B leads to a sensible clustering, but this method can not be easily generalized to multivariate data, nor can it account for a prior model on $\mathbf{t}$. The EC-M algorithm is the best candidate for both a multivariate generalization and a model on $\mathbf{t}$. Equation (12) needs only to be slightly changed, in order to include the conditional probabilities $P(t_{ik} = 1 \mid \mathbf{t}_{-i})$. The model on $\mathbf{t}$ must be such that these conditional probabilities are computable.

References

Allard, D. (1998) Geostatistical Classification and Class Kriging, *Journal of Geographical Information and Decision Analysis*, **2**, 87-101.

Allard D. and Monestiez P. (1999) Geostatistical Segmentation of Rainfall Data, in *geoENV II: Geostatistics for Environmental Applications*, eds. Gomez-Hernandez J., Soares A. and Froidevaux R., Kluwer Academic Publishers, Dordrecht, pp. 139-150.

Ambroise, C., Dang, M. and Govaert, G. (1995) Clustering of spatial data by the EM algorithm, in A. Soares *et al.* (eds.), *geoENV I – Geostatistics for Environmental Applications.* Kluwer Academic Publishers, Dordrecht, pp. 493–504.

Atteia, O., Dubois, J.-P. and Webster, R. (1994) Geostatistical Analysis of soil contamination in the Swiss Jura, *Environmental Pollution* **86**, 315–327.

Besag, J. E. (1986) On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society, Ser. B* **48**, 259–302.

Besag, J. E., York J. and Mollie A. (1991) Bayesian image restoration with two applications in spatial statistics (with discussion), Annals of the Institue for Statistical Mathematics, **43**, 1-59

Cressie, N. (1993) *Statistics for Spatial Data*, Wiley-Interscience, New-York.

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

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Goovaerts, P. (1997) *Geostatistics for natural resources evaluations*, Oxford University Press.

Guyon, X. (1995) *Random Field on a network*, Springer verlag.

Haslett, J. (1985) Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context, *Pattern Recognition* **18**, 287–296.

Oliver, M. A. and Webster, R. (1989) A Geostatistical basis for spatial weighting in multivariate classification, *Mathematical Geology* **21**, 15–35.

Searle, R. S. (1982), *Matrix algebra useful for statistics*, John Wiley & Sons, New-York.

Wackernagel, H. (1995), *Multivariate Geostatistics*, Springer-Verlag, Berlin.