Spatial Clustering and the EM Algorithm

Christophe Ambroise et Gérard Govaert
Université de technologie de Compiègne
URA CNRS 817
BP 649 F-60206 Compiègne cedex - France

Abstract

A clustering algorithm for spatial data is presented. It seeks a fuzzy partition which is optimal according to some criterion. We propose to penalize the energy function exhibited by Hathway (1986) with a term taking into account spatial contiguity constraints. The structure of the EM algorithm may be used for maximizing the proposed criterion. The Maximization step is then unchanged and the Expectation step becomes iterative.The efficiency of the new clustering algorithm has been tested with biological images and compared with other clustering techniques.

1 Introduction

When classical clustering techniques are used for partitioning spatial data, the resulting classes will often be geographically very mixed. To avoid this phenomenon, the spatial information of the data has to be taken into account. Very different solutions to this problem have been proposed in the literature.

A natural approach consists in using the geographical coordinates of the individuals, more or less heavily weighted, as an additional pairs of variates (Berry 1966, Jain and Farrokhnia 1991).

Another approach groups individuals which are both similar and contiguous (Legendre 1987, Openshaw 1977). This implies the definition of a neighborhood concept. Defining neighborhood relationship is equivalent to build a graph where each element is represented by a node and each neighborhood relationship is an edge. Clustering with spatial contiguity constraints can be described as the succession of two steps:

1. The definition of a neighborhood graph. This can be done with standard algorithms such as a Delaunay triangulation, a Gabriel graph...

2. Running of a clustering algorithm while respecting the constraints. Many classical clustering algorithms may be modified to take into account the constraints which are summarized by the graph. In Lebart (1978) a classical hierarchical clustering algorithm is adapted.

These procedures produce classes made of adjacent sites. They may separate into different classes individuals which are very similar, if they are geographically far apart.

Oliver and Webster (1989) propose to run clustering algorithms with a modified dissimilarity matrix. This modified matrix is a combination of the matrix of geographical distances and the dissimilarity matrix computed with the non geographical variables. This kind of procedure seems to work well but have no statistical justification.

Other spatial constrained clustering methods and approaches have been developed in the field of unsupervised image segmentation (Ripley 1991). The specificity of these methods is that they deals with pixels and their regular grid structure. An image may well be considered as a regular lattice of pixels. Thus the computation of the neighborhood graph is immediate. The commonest choices for the neighborhood graph are the 4 neighbor graph (horizontal and vertical adjacencies only) and the 8 neighbor (including diagonals).

There are numerous unsupervised segmentation algorithms. In this paper, we consider only some statistically based methods. In this framework, the Bayesian approach proposes some solutions, which may be separated into two families (Masson and Pieczinsky 1993):

1. The local methods make assumptions about the pixel or about small groups of adjacent pixels called "context" (Masson and Pieczinsky 1993).

2. The global methods make assumption about the whole image and generally use a Markov random field model (Geman and Geman 1984, Besag 1974).

Notice that the classes obtained with these methods differ from the classes obtained with hard contiguity constrained algorithms: A class does not necessarily form a single patch on the image. Pixels of the same class may be present in different part of the image. Thus the statistical models used in unsupervised segmentation algorithms take the contiguity constraints into account but do not impose "one region classes".

We develop in this paper a new statistical method for spatial clustering which is based on the EM algorithm (Dempster *et al.* 1977) and takes into account the spatial constraints without requiring a partition made of "one region classes".

2 Introducing Spatial Constraints in the EM Algorithm

2.1 Fuzzy Clustering and the EM Algorithm

In cluster analysis based on Gaussian mixture models (Celeux and Govaert 1995), data are $\mathbb{R}^d$-valued vectors $\mathbf{x}_1, \ldots \mathbf{x}_n$ assumed to be a sample from a mixture of densities:

$$f(\mathbf{x}_i|\Phi) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i|\theta_k), \tag{1}$$

where the p_k are the mixing proportions ($0 < p_k < 1$, for all $k = 1,..,K$ and $\sum_k p_k = 1$) and $f_k(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the density of a Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and variance matrix $\boldsymbol{\Sigma}_k$.

The EM algorithm is often used for finding the unknown parameters of the mixture. It maximizes (it finds often a local maximum) the likelihood of the sample. Hathaway (Hathway 1986) has shown that the EM algorithm is an alternate optimization algorithm which considers the following decomposition of the likelihood:

$$\mathrm{L}(\mathbf{c}, \theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} \log(p_k f_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} - \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} \log(c_{ik}) \tag{2}$$

where c_{ik} may be considered as the probability of $\mathbf{x}_i$ belonging to class k ($0 \leq c_{ik} \leq 1, \sum_{k=1}^{K} c_{ik} = 1, \sum_{i=1}^{n} c_{ik} > 0, 1 \leq i \leq n, 1 \leq k \leq K$).

Two kinds of parameters appear in this criterion: the parameters of the mixture, $\Phi = \{p_1, \cdots, p_{K-1}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K\}$ and the fuzzy classification matrix, $\mathbf{c} = [c_{ik}]$. Each iteration of the EM algorithm optimizes alternatively the criterion according to the parameters of the mixture with a fixed classification matrix and then according to the classification matrix with the parameters fixed. Thus each EM iteration comprises two steps:

- **E-step:** Maximization of $L(\mathbf{c}, \Phi)$ according to $\mathbf{c}$. Let us write the Lagrange's equation:

$$\mathcal{L}(\mathbf{c}) = L(\mathbf{c}, \Phi) + \sum_{i=1}^{n} \lambda_i((\sum_{k=1}^{K} c_{ik}) - 1), \tag{3}$$

  where the λ_i are the Lagrange coefficients corresponding to the constraints $\sum_{k=1}^{K} u_{ik} = 1, \forall i$. Solving the system, we may write:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial c_{ik}} = \log\left(p_k f_k(\mathbf{x}_i|\theta_k)\right) - 1 - \log c_{ik} + \lambda_i = 0 \\ \sum_{k=1}^{K} u_{ik} = 1 \end{cases}$$

  which gives us,

$$\begin{cases} c_{ik} = \exp\left\{-1 + \lambda_i + \log\left(p_k f_k(\mathbf{x}_i|\theta_k)\right)\right\} \\ \sum_{k=1}^{K} \exp\left\{-1 + \lambda_i + \log\left(p_k f_k(\mathbf{x}_i|\theta_k)\right)\right\} = 1 \end{cases}$$

  Thus we get the following values for the c_{ik}:

$$c_{ik} = \frac{p_k f_k(\mathbf{x}_i|\theta_k)}{f(\mathbf{x}_i|\Phi)} \tag{4}$$

- **M-step:** Maximization of $L(\mathbf{c}, \Phi)$ according to Φ. The maximum likelihood estimates are computed. In the most general case we get:

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} c_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} c_{ik}} \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_k = \sum_{k=1}^{K}\sum_{i=1}^{n} c_{ik}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^t \tag{6}$$

$$\hat{p}_k = \frac{\sum_{i=1}^{n} c_{ik}}{n} \tag{7}$$

2.2 Penalization of the Likelihood

We propose to regularize the likelihood with a term, which takes into account the spatial information relative to the data. Let $\mathbf{V}$ be the "neighborhood matrix":

$$v_{ij} = \begin{cases} \alpha > 0 & \text{if } \mathbf{x}_i \text{ et } \mathbf{x}_j \text{ are neighbors} \\ 0 & \text{if } \mathbf{x}_i \text{ et } \mathbf{x}_j \text{ are not neighbors} \end{cases}$$

One can use the graph structure presented in the introduction for computing the matrix $\mathbf{V}$ or use a function of the spatial distance between the elements of the data set.

The regularizing term we propose is the following:

$$\mathrm{G}(\mathbf{c}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ik}.c_{jk}.v_(i,j) \tag{8}$$

The more the classes contain adjacent elements, the more this term is important. Notice that this term is only function of the classification matrix. The new criterion we consider is then a penalized likelihood:

$$\mathrm{U}(\mathbf{c},\Phi) = \mathrm{L}(\mathbf{c},\Phi) + \beta.\mathrm{G}(\mathbf{c}) \tag{9}$$

A modified version of the EM algorithm can be used for maximizing this criterion. We named this version *Neighborhood EM algorithm* (NEM). As the new regularizing term does not contain any parameter of the mixture, the Maximization step remains unchanged. The Estimation step (maximization of the criterion relatively to the classification matrix with fixed mixture parameters) changes:

1. The classification matrix $\mathbf{c}^0$ is randomly initialized (n lines, K columns). A neighborhood matrix $\mathbf{V}$ is computed according to the spatial relationship.

2. At each iteration the followings steps are realized until convergence:

   (a) **E-step:** The necessary conditions of optimality take the following form:

   $$\begin{cases} \frac{\partial U}{\partial c_{ik}} = \log\left(p_k f_k(\mathbf{x}_i|\theta_k)\right) + 1 - \log c_{ik} + \lambda_i + \beta\sum_{j=1}^{n} c_{jk}v_{ij} \\ \sum_{k=1}^{K} c_{ik} = 1 \end{cases}$$

   which may be written as,

   $$\begin{cases} c_{ik} = \exp\left\{\log(p_k f_k(\mathbf{x}_i|\theta_k)) + 1 + \lambda_i + \beta\sum_{j=1}^{n} c_{jk}v_{ij}\right\} \\ \sum_{k=1}^{K} \exp\left\{\log(p_k f_k(\mathbf{x}_i|\theta_k)) + 1 + \lambda_i + \beta\sum_{j=1}^{n} c_{jk}v_{ij}\right\} = 1 \end{cases}$$

Finally we get the following equation:

$$c_{ik} = \frac{p_k f_k(\mathbf{x}_i|\theta_k) \cdot \exp\{\beta \sum_{j=1}^{n} c_{jk} v_{ij}\}}{\sum_{\ell=1}^{K} p_\ell f_\ell(\mathbf{x}_i|\theta_\ell) \cdot \exp\{\beta \sum_{j=1}^{n} c_{j\ell} v_{ij}\}} \tag{10}$$

which suggests an iterative computing algorithm of the form $\mathbf{c} = g(\tilde{\mathbf{c}})$, where $\tilde{\mathbf{c}}$ is the old classification matrix. From a practical point of view, a few iterations produce a reasonable new classification matrix $\mathbf{c}$, which can be used for the next *M-step*.

(b) **M-step:** The maximum likelihood estimates are computed.

Empirically the NEM algorithm converges with all tested problems, but no proof of local convergence is available yet.

At each iteration the spatial information modify the partition according the importance of the β coefficient. The Estimation step smoothes the map of the labels.

The preceding algorithm solves two different tasks which are highly dependent: it estimates the parameters of the mixture and finds a partition. The two tasks are performed at the same time.

2.3 Hard Partition

The NEM algorithm may be easily adapted for seeking a hard partition ($c_{ik} \in \{0, 1\}$):

- The easiest way is to consider the resulting fuzzy partition at the convergence and to assigned each individual to the most probable class according to the a posteriori probabilities.

- As in the CEM algorithm (Celeux and Govaert 1992), it is possible to add an intermediate classification step between the E and the M step.

3 A Bayesian Interpretation

Notice that it is possible to have a Bayesian interpretation of the NEM algorithm. Maximizing the criterion $U(\mathbf{c}, \Phi)$ is equivalent to maximizing

$$\exp\{U(\mathbf{c}, \Phi)\} = \exp\{L(\mathbf{c}, \Phi)\} \cdot \exp\{\beta \cdot G(\mathbf{c})\} \tag{11}$$

We may write,

$$\exp\{G(\mathbf{c})\} \propto P(\mathbf{c}) \tag{12}$$

where $P(\mathbf{c})$ is a Gibbs distribution with energy function $-G(\mathbf{c})$.

The expression $\exp\{L(\mathbf{c}, \Phi)\}$ may be interpreted as $P(\mathbf{x}|\mathbf{c}, \Phi)$ the conditional density of the sample, $\mathbf{x}$, knowing the classification matrix, $\mathbf{c}$, and the parameters of the Gaussian mixture. Thus,

$$\begin{aligned} \exp\{U(\mathbf{c}, \Phi)\} &\propto P(\mathbf{c}) \cdot P(\mathbf{x}|\mathbf{c}, \Phi) \\ &\propto P(\mathbf{c}|\mathbf{x}, \Phi) \end{aligned}$$

where $P(\mathbf{c}|\mathbf{x},\Phi)$ is the posterior distribution of the classification matrix $\mathbf{c}$.

Maximizing $U(\mathbf{c},\Phi)$ is equivalent to finding the matrix $\mathbf{c}$ which maximizes the posterior distribution $P(\mathbf{c}|\mathbf{x},\Phi)$, which is a Gibbs distribution with energy function $-U$. Thus, the NEM algorithm may be interpreted as an algorithm for searching the MAP estimate of the classification matrix $\mathbf{c}$.

This Bayesian interpretation assumes that there are two random field: $\mathbf{X} = \{X_s, s \in S\}$, which is the observed random field and $\mathbf{C} = \{C_s, s \in S\}$, which is the random field corresponding to the classification matrix (S is the set of sites). The X_s take their values in $\mathbb{R}^d$ and the $C_s = (C_{s1}, \cdots, C_{sK})$ in a subset of $[0,1]^K$. Both field $\mathbf{X}$ et $\mathbf{C}$ follow Gibbs Distribution.

This approach may be used in image segmentation and provides an alternative to the existing fuzzy non supervised segmentation algorithms (Kent and Mardia 1991, Caillol *et al.* 1993).

3.1 Link with Markov Random Fields

If we consider a hard partition (the $C_s = (C_{s1}, \cdots, C_{sK})$ are random variables which take their values in $\{0,1\}^K$) and the following neighborhood matrix:

$$v_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ et } \mathbf{x}_j \text{ are neighbors} \\ 0 & \text{if } \mathbf{x}_i \text{ et } \mathbf{x}_j \text{ are not neighbors} \end{cases}$$

the a priori distribution of the classification matrix becomes:

$$\begin{aligned} P(\mathbf{C} = \mathbf{c}) &= \frac{1}{Z} exp\{\beta G(\mathbf{c})\} \\ &= \frac{1}{Z} exp\{\frac{\beta}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{K} c_{ik} \cdot c_{kj} \cdot v_{ij}\} \\ &= \frac{1}{Z} exp\{\beta \sum_{i<j} v_{ij} (\mathbf{c}_i \cdot \mathbf{c}_j)\} \\ &= \frac{1}{Z} exp\{\beta (\#\text{neighbors belonging to the same class})\} \end{aligned}$$

where Z is a normalizing constant. This Gibbs distribution defines a Markov random field which is a model proposed by Strauss (1977) for image segmentation.

4 An Application to biological Images

A sample of living cells is laid on a nutritive substance. After a few days new living cells appear and form a thin but visible layer around the original sample. Biologists are interested in determining the surface of the new layer. We have a collection of images of size 512 by 512 [1] of such experimental results and present here the analysis of a representative image.

[1]We are indebted to Eugenio Grapa, from the Université de Technologie de Compiègne, for the biological pictures

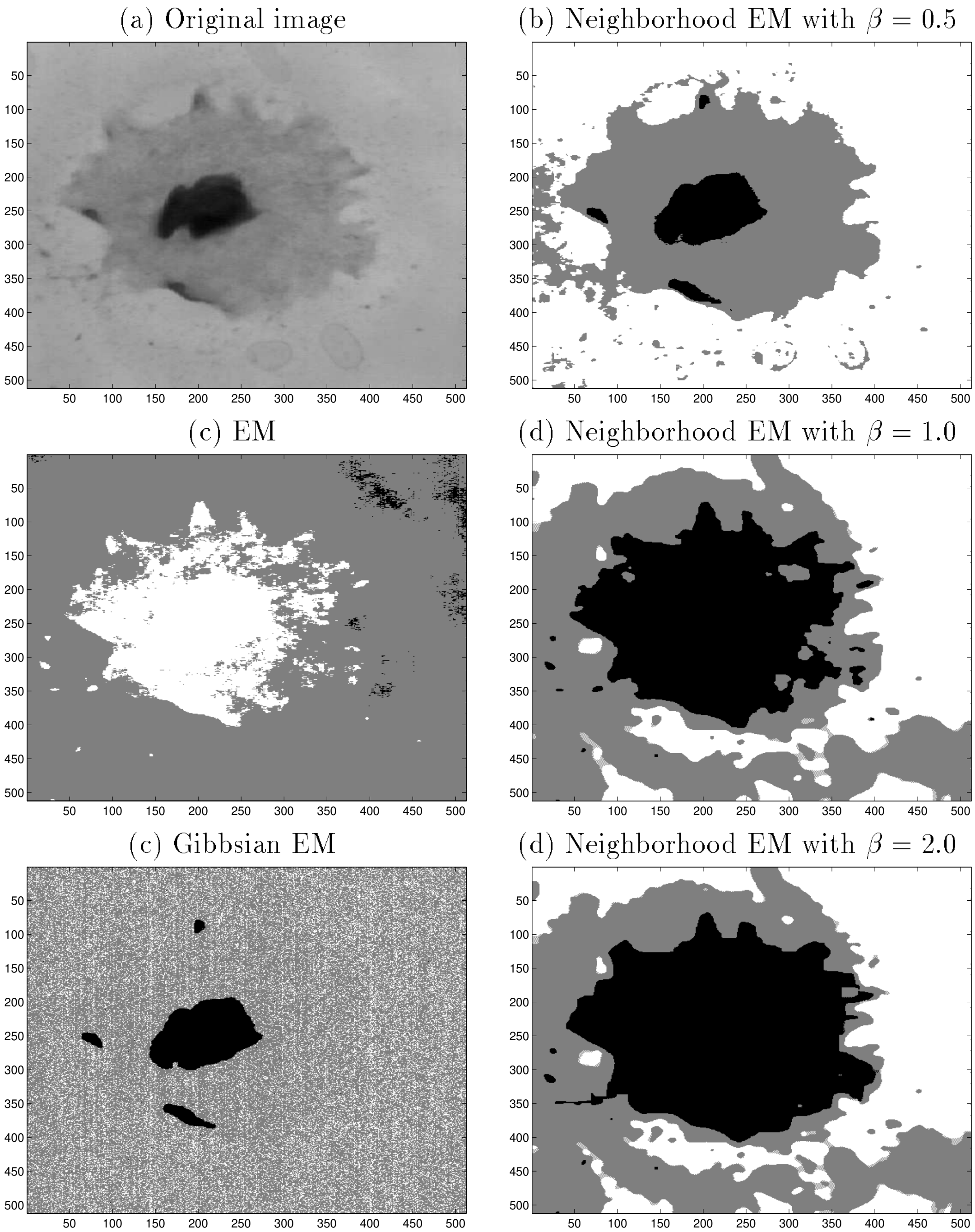

Figure 1: Segmentation of a biological image by different algorithms

We aim to distinguish three different kind of patterns, in order to determine automatically the size of the area covered by the new cells. A "good" segmentation from the biologist point of view should separate the image in three different areas representing:

- The original sample
- The nutritive substance
- The new cells

We have tested three different algorithms:

- The classical EM algorithm, which does not take into account the spatial information.
- An unsupervised segmentation algorithm, called Gibbsian EM ((Chalmond 1989)), which is based on Markov random field model.
- The proposed version of the EM algorithm, which minimizes the previous defined penalized likelihood.

The human vision distinguish clearly the three classes. In fact each class is far from having the same gray level: Some pixels representing the new cells have exactly the same color as the nutritive substance pixels, but the human eyes makes automatically some adjustments. Thus the unsupervised segmentation of this kind of images is harder than one could think at first.

The EM algorithm isolates the nutritive substance, but tends to empty the class corresponding to the original sample. Numerous runs give always the same kind of results.

The Gibbsian EM algorithm does not make any distinction between the new cells and the nutritive substance.

The NEM algorithm was tested with different values for the parameter β. When the parameter is two small the same results as the EM algorithm are obtained. This seems to be logical, because if we set β to zero, NEM becomes the classical EM algorithm. When $\beta = 0.5$, the result is very satisfying and allows the automatic computation of the surface covered by the new cells. If we try to run NEM with greater values of the parameter, the spatial information becomes preponderant and the segmentation does not make a lot of sense.

5 Discussion

The preceding example showed the practical efficiency of the proposed algorithm on this particular problem. We tested the NEM algorithm on other segmentation tasks and get regularly good results once the β parameter was tuned. Further experiments should test the performance of the algorithm with spatial data which have an irregular lattice structure.

The main problem of the NEM algorithm consists in deciding on β. In this problem we validate the value using our knowledge of what the "good" solution should look like. It would be easier to have an automatic estimation of this parameter. This problem should be overcome by considering the Bayesian interpretation of the algorithm.

References

Berry, B. (1966). Essay on commodity flows and the spatial structure of the indian economy. Research paper 111. University of Chicago, departement of geography.

Besag, J. (1974). 'Spatial interaction and the statistical analysis of lattice systems'. *Journal of the Royal Statistical Society* **35**, 192–236.

Caillol, H., A. Hillion and W. Pieczinsky (1993). 'Fuzzy random fields and unsupervised image segmentation'. *IEEE transactions on geoscience and remote sensing* **31**(4), 801–810.

Celeux, G. and G. Govaert (1992). 'A classification em algorithm for clustering and two stochastic versions'. *Computational statistics and data analysis* **14**, 315–332.

Celeux, G. and G. Govaert (1995). "gaussian parsimonious clustering models'. *Pattern Recognition* **28**, 781–793.

Chalmond, B. (1989). 'An iterative gibbsian technique for reconstruction of m-ary images'. *Pattern Recognition* **22**(6), 747–761.

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

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the baysian restoration of images. In 'IEE Transactions on Pattern Analysis and Machine Intelligence'. Vol. PAMI-6. pp. 721–741.

Hathway, R. (1986). 'Another interpretation of the em algorithm for mixture distributions'. *Journal of Statistics & Probability Letters* **4**, 53–56.

Jain, A. and F. Farrokhnia (1991). 'Unsupervised texture segmentation using gabor filters'. *Pattern Recognition* **24**(12), 1167–1186.

Kent, J. and K. Mardia (1991). 'Spatial classification using fuzzy membership'. *IEER Trans. Patt. Anal. Machine Intell.* **10**(5), 135–143.

Lebart, L. (1978). 'Programme d'agrégation avec contraintes (c.a.h. contiguité'. *Cahier de l'analyse des données* **3**, 275–287.

Legendre, P. (1987). 'Constrained clustering'. *developpments in numerical ecology. NATO ASI Series* **G 14**, 289–307.

Masson, P. and W. Pieczinsky (1993). 'Sem algorithm and unsupervised statistical segmentation of satellite images'. *IEEE transactions on geoscience and remote sensing* **31**(3), 618–633.

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

Openshaw, S. (1977). 'A geographical solution to scale and aggregation problems in region-building, partitionning and spatial modelling'. *Transactions of the institute of british geographers* **2**, 459–472.

Ripley, B. (1991). *Statistical inference for spatial processes.* Cambridge university press.

Strauss, D. (1977). 'Clustering on coloured lattices'. *Journal of appl. probab.* **14**, 135–143.