ADAPTED WARD'S CLUSTERING METHOD: GENERALIZATION TO SEVERAL VARIABLES USING THE FAST FOURIER TRANSFORM

Carme Hervada-Sala[a] , **Eusebi Jarauta-Bragulat**[b], **Ángela M. Diblasi**[c]

[a] carme.hervada@upc.es Dept. de Física i Enginyeria Nuclear, EUETIT (UPC), Spain.
[b] eusebi.jarauta@upc.es Dept. de Matemàtica Aplicada III, ETSECCPB (UPC), Spain.
[c] angelad@uncu.edu.ar Universidad Nacional de Cuyo, Argentina.
Website: ftp://ftp-urgell.upc.es/matematica/ejarauta/iamg2002/article2

ABSTRACT

Earth science studies deal in general with multivariate, regionalized, observations, which may be compositional, or not. Frequently, it is of interest to know whether those data have to be divided into different populations, a task usually performed by cluster analysis. This problem cannot be studied with traditional methods because samples are not independent. In that case, an extension of Ward's clustering method to spatially dependent samples can be used. This methodology is based on a generalized Mahalanobis distance, which uses the covariance and cross covariance (or variogram and cross-variogram) matrices. In its original version, the method was iterative and tedious, as it was necessary to re-estimate the spatial covariance structure at each step. In this work, we stay within the same theoretical framework, but we improve the methodology using the Fast Fourier Transform (FFT) method to find the covariance structure. Thus, we obtain a generalization to many compositional variables of adapted Ward's clustering method.

1. Introduction

Earth sciences deal with great amounts of data which have to be analysed, organised and also cleaned up to obtain information about a given problem. There are many statistical techniques that allow finding similarities or differences among data and variables. Multivariate methods allow us to consider changes in several properties simultaneously. One of the most widely used multivariate procedures in Earth science is the discriminant function. The aim of discriminant analysis is to find a linear combination of the variables, which produce the maximum difference among the previously defined groups. However, when classifications of objects have to be done, cluster analysis is used. Cluster analysis is the name given to a bewildering assortment of techniques designed to perform classification by assigning observations to groups so each group is more or less homogeneous and distinct from other groups. There is no analytical solution to this problem, as can be seen in [2].

Cluster analysis encompasses many diverse techniques for discovering structure within complex bodies of data. In a typical example one has a sample of data each described by scores on some variables. The objective is to group either the data or the variables into clusters such that the elements within a cluster have a high degree of "natural association" among themselves while clusters are "relatively distinct" from one another. The approach to this problem and the results achieved depend on how the investigator chooses to give operational meaning to the phrases "natural association" and "relatively distinct". To do so, many criterions have been described: partitioning methods, arbitrary origin methods, mutual similarity procedures and hierarchical clustering techniques. One of the most widespread

hierarchical clustering methods is Ward method, described in [7], also known as method of the minimum variance. However, when this method is used to classify regionalized variables some difficulties arise. A first step to solve them is to use a generalization of Ward's method proposed in [3] and developed in [4]. This generalization is known as Ward's adapted method. This new method can be used with spatially dependent variables, but its use is often boring because it needs to model all variograms and cross-variograms. In this work this new methodology is enlarged with the use of the FFT to calculate the correlogram tables [8] and, a `FORTRAN` program to generalise [4] when there are many variables to cluster.

Ward's clustering method is a hierarchical agglomerative method whose philosophy can be summarized as follows. Assuming that there are N elements to cluster, begin with N clusters consisting exactly of one entity, search the similarity matrix for the most similar pair of clusters and reduce the number of clusters by one through merger the most similar pair of clusters. Perform those steps until all clusters are merged. The Ward objective is to find at each stage those two clusters whose merger gives the minimum increase in the total within group error sum of squares (or distances between the centroids of the merged clusters).

Total within group error sum of squares $V_T(K)$ in one stage with K groups, J variables and there are N_i elements in each group, is defined as:

$$V_T(K) = \sum_{k=1}^{K}\left(\sum_{j=1}^{J}\left(\sum_{i=1}^{N_i}(x_{ijk} - \overline{x}_{jk}(i))^2\right)\right);\ \ \overline{x}_{jk}(i) = \frac{1}{N_i}\sum_{i=1}^{N_i} x_{ijk};\ \ \sum_{i=1}^{K} N_i = N \tag{1}$$

Where x_{ijk} is the value of j-th variable, from i-th observations in k-th group, and $\overline{x}_{jk}(i)$ is the average value inside this group. So, following summation order, the first sum corresponds to variability inside a group for a given variable, the second one summing up all variables and the last one is the total variability.

2. Ward's adapted method to regionalized variables.

To be able to apply Ward's method to spatially dependent data, equation (1) has to be corrected. To do so, Mahalanobis distance defined in [5] has to be used. Euclidean distance between points z_i, z_j in a sample with N elements, is:

$$d_M^2(z_i, z_j) = (z_i - z_j)^T C^{-1}(z_i - z_j);\ \ C = \frac{1}{N}\sum_{i=1}^{N}(z_i - \overline{z})(z_i - \overline{z})^T,\ i, j = 1, 2, ..., N \tag{2}$$

Where $\overline{z}$ means the average. Then, for regionalized variables distance has to be generalized following [5] so it is:

$$d_{MG}^2(z_i, z_j) = (z_i - z_j)^T \left[C(0) + C(h)\right](z_i - z_j);\ \ i, j = 1, 2, ..., N, (i \neq j) \tag{3}$$

where $C(0)$ is the sample variance-covariance matrix, and $C(h)$ is the auto-covariance matrix at lag $h = \| z_i - z_j \|$. Using this approach total variability is, now

$$V_T(K) = \sum_{k=1}^{K} \left(\sum_{i,j \in C_k} d_{MG}^2(z_i, z_j) \right) \tag{4}$$

At each stage those two clusters that minimise the loss of variability are merged. Then, the total variability after having merged two groups is measured with:

$$V_T(K-1) = \sum_{k=1}^{K-1} \left(\sum_{i,j \in C_k} d_{MG}^2(z_i, z_j) \right) + \sum_{i,j \in C_{(k_1,k_2)}} d_{MG}^2(z_i, z_j);\ k \neq k_1, k_2 \tag{5}$$

So the loss of variability can be calculated as:

$$\Delta V(K, K-1) = V_T(K-1) - V_T(K) \tag{6}$$

3. Database.

The data set used to illustrate this methodology and compare it to methodology described in [4] has been given for Susana Jiménez. Full data base description can be found in [4]. The data set consists of 1250 sandy surface samples taken from Darss Sill in the Baltic Sea. The sediments where dried and sieved into eight-weight percent size fractions. So, for each sample ten variables are known: two UTM coordinates (Easting and Northing) and eight fractions.

Jiménez [4], transforms the eight percent variables into two, median and sorting, following Tauber's methodology [6]. Those two variables describe fairly well and easily granulometric parameters. Median is a variable that measures the average size of the sample and sorting reflects the dispersion of grain size, it is quite similar to standard deviation. Those variables can be obtained adjusting the accumulated distribution function to a logistic distribution in median and sorting. To compare results obtained using the methodology proposed by Jimenez [4], we have used the same data set with the same variables.

4. Generalization of Ward's adapted method. Application and discussion.

We think it is not much operative to calculate and model variograms and cross-variograms for more than two variables. To avoid modelling them at each stage [4] uses the recurrent Lance and Williams's formula. However, this is still too difficult to implement when there are more than two variables because it still needs the variogram models for the initial step. Instead of using variogram models we propose to calculate the correlogram tables using the FFT approach [8]. Using it, we are able to find C(h) and C(0) matrices. They will be summed and inverted. Using so, we will be able to implement Ward's adapted method given in [3].

Once C(h) and C(0) are known, distances matrix can be defined using equation (3). At each stage the groups that merge are those which total variability increment is minimum. Our program, which can be found in our website, is structured like GSLIB [1] programs, it is also written in `FORTRAN`. The corresponding parameter needs the following information:

- Number of groups (or steps to compute)
- Name of the data file (Geo-EAS format)
- Number of cases to study.

- Name of the initial correlogram file
- Number of grids in the correlogram matrix
- Distance between points in the correlogram matrix
- Output file name

The output of the program is an ASCII file with a matrix where rows are the cases and columns are the groups which they belong to. It is similar to the output in [4] but more specified.

To compare outputs from [4] and from our program figures 1 and 2 can be compared. Figure 1 shows classification using [4] and figure 2 shows the results using our program. In figure 1 there is a stronger correspondence among groups and position than in figure 2; we think that this is because [4] forces a variogram spherical model for all variables, it also forces all of them to have same range and does not allow for nested structures.

5. Conclusions

Those are the most remarkable conclusions from this work:

a) Ward's adapted method described in [4] allows the use of spatially dependent data, or regionalized variables.
b) The drawbacks in the program described in [4] are: the restriction to the variogram model, the restriction of their parameters, the impossibility to assume nested variogram structures, the limitation to two variables and so the high dependence of the resulting groups to physical space.
c) The best advantages of the program presented in this work are: its structure in the same form as GSLIB programs, the better specification of the parameter file and the output file, and that the drawbacks described in b) are avoided. At last, as we use FFT instead of modelling variograms it is less subjective.

6. References

[1] Deutsch, C.V. and A.G. Journel (1998). *Geostatistical software library and user's guide – GSLIB*. Oxford University Press, 1 CD + 369 pp.
[2] Everitt, B.S. (1993). *Cluster analysis*. Edward Arnold, Cambridge, 170 pp.
[3] Jiménez-González, S. et al. (1998). *Ward's method adapted to spatially dependent data*. Conference of the International Association for Mathematical Geology, 445-450.
[4] Jiménez-González, S. (1999). *Agrupación de datos con dependencia espacial. El método de Ward adaptado y la estimación iterativa de semivariogramas.* (UPC), 151 pp.
[5] Pawlowsky-Glahn, Vera et al. (1997). *Spatial cluster analysis using a generalized Mahalanobis distance*. Conference of the International Association for Mathematical Geology, 175-180.
[6] Tauber, F. (1997). *Treating grain-size data as continuous functions.* Proceedings of the Conference of the International Association for Mathematical Geology, 169-174.
[7] Ward, J.H. (1963). *Hierarchical grouping to optimise an objective function.* Journal of the American Statistics Association. Vol. 58, 236-244.
[8] Yao, T. and A.G. Journel (1998). *Automatic modelling of (cross)covariance tables using fast Fourier transform.* Mathematical Geology, 30(6), 589−615.

Acknowledgements
Authors want to thank specially:

- To Susana Jiménez, to have given her programs and Darss Sill data to the authors.
- To Department of Applied Mathematics III UPC.
- To the School of Civil Engineering (ETSECCPB) in UPC.
- To the "Consorci Escola Industrial de Barcelona" (CEIB).

Figure 1. Results from Ward's adapted method using [4] with Darss Sill data.

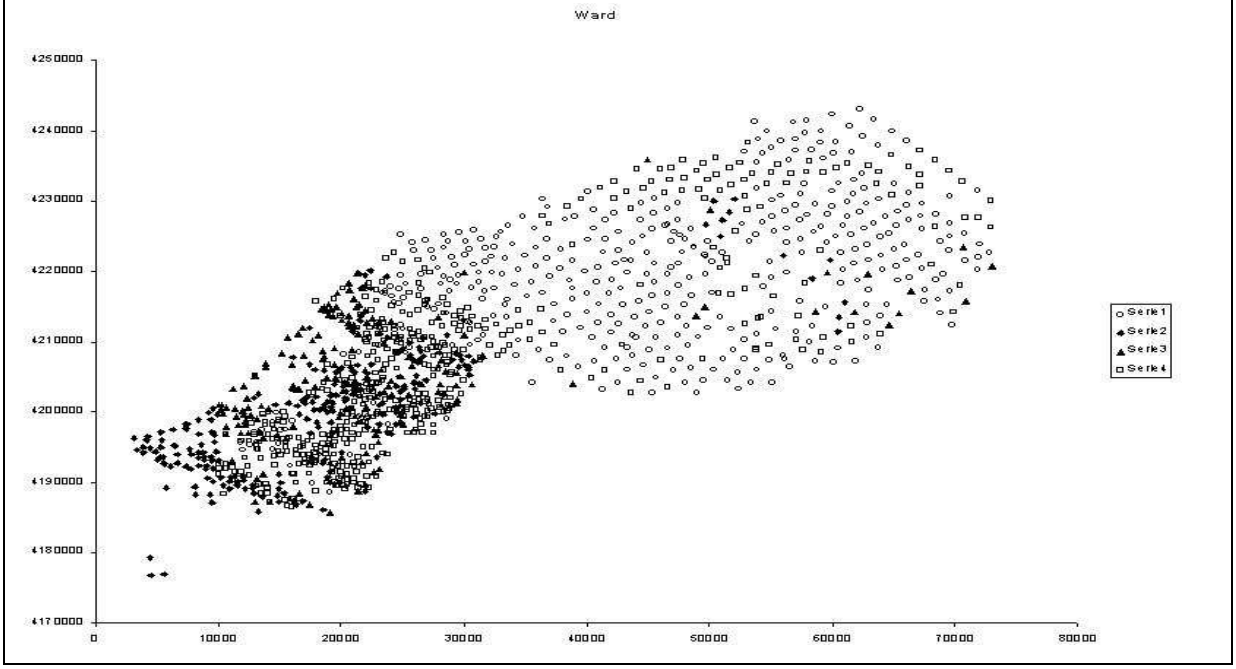

Figure 2. Results from Ward's adapted method using the FFT approach.