A program to perform Ward's clustering method on several regionalized variables

Carme Hervada-Sala[a] , **Eusebi Jarauta-Bragulat**[b]

[a] Dept. of Physics and Nuclear Engineering. Universitat Politècnica de Catalunya, Spain. carme.hervada@upc.es

[b] Dept. of Applied Mathematics III. Universitat Politècnica de Catalunya, Spain. eusebi.jarauta@upc.es

Website: ftp://ftp-urgell.upc.es/matematica/ejarauta/C&G2003/

ABSTRACT

Earth science studies deal in general with multivariate and regionalized observations which may be compositional. Sometimes, it is interesting to know whether these data have to be divided into different subpopulations, 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 several 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 produces 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 it can be seen in [2].

Cluster analysis encompasses many diverse techniques for discovering structure within complex sets of data. In a typical example one has a sample of data each described by scores on some variables. The objective of cluster analysis 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 criteria have been described: partitioning methods, arbitrary origin methods, mutual similarity procedures and hierarchical clustering techniques. One of the most widespread hierarchical clustering methods is the *Ward's 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, because variables are not independent. 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 adapted Ward's method, which can be used with spatially dependent variables, but its use is often difficult, because it needs to model variograms and cross-variograms and it has been performed for two variables only. On top of that, variograms and cross-variograms models used must be the same.

The aim of this work is to present a FORTRAN program in GSLIB-style that develops a new methodology to generalise adapted Ward's method to be able in clustering of several regionalized variables. It is based on the use of the correlogram tables calculated using the Fast Fourier Transform (FFT) following [8].

2. Ward's and adapted Ward's methods

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.

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 $C(\mathrm{h})$ 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:

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

Figure 2.1.- Application of adapted Ward's method into four groups.

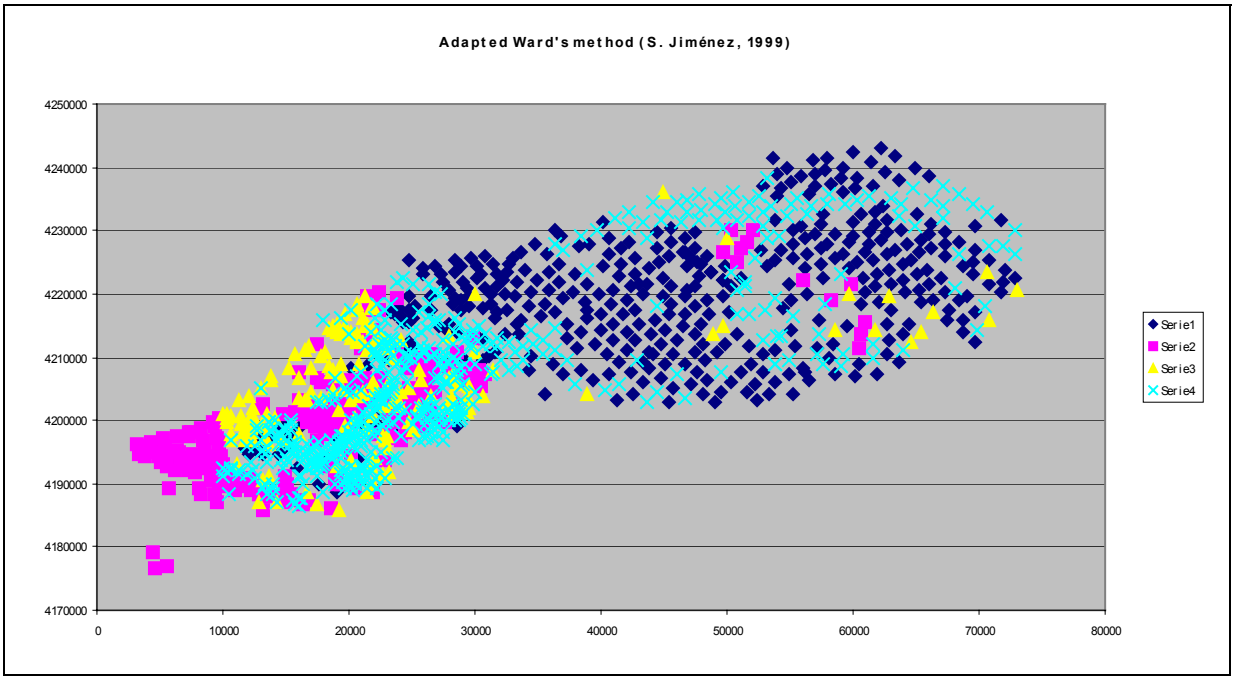
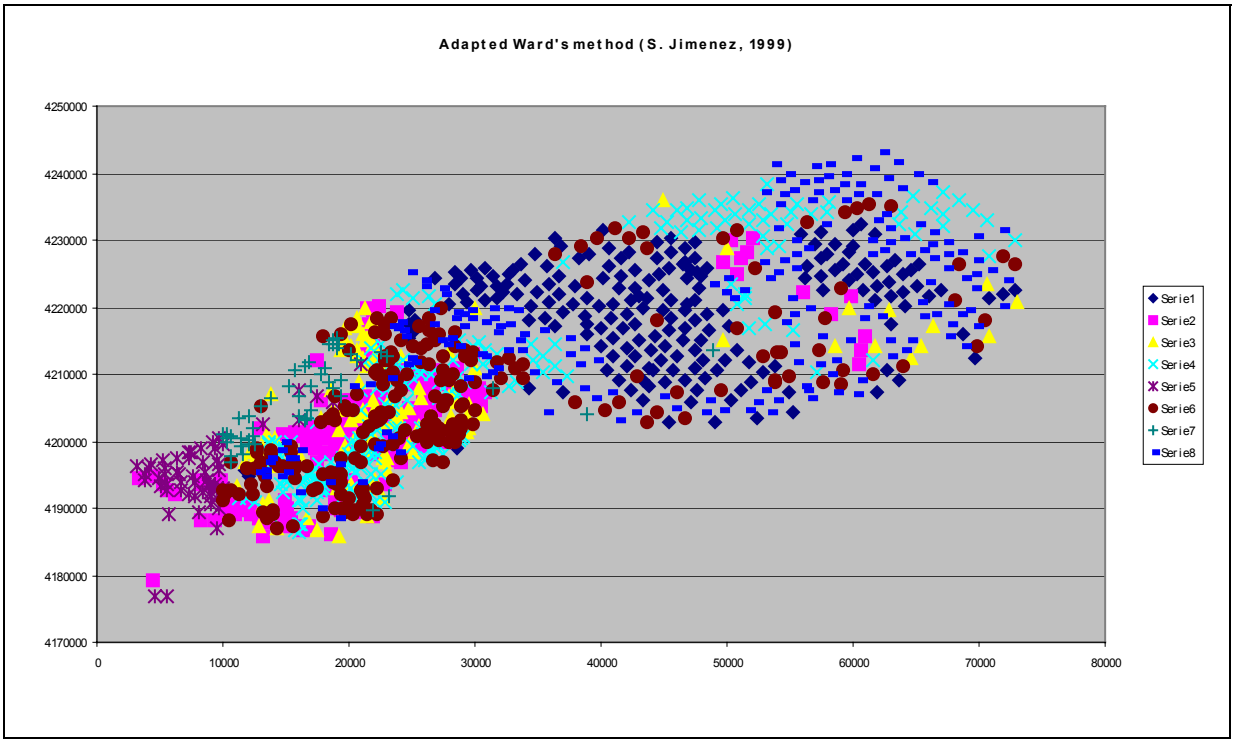


Figure 2.2.- Application of adapted Ward's method into eight groups.



Susana Jimenez (1999) published the adapted Ward's method and wrote a program to apply it to a set of two regionalized variables [4]. The data she used is called Darss Sill data. 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, 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.

Results obtained using those data are shown in figures 2.1 and 2.2. Figure 2.1 shows the classification into four groups and figure 2.2 into eight groups. Both of them show a strong weight of spatial coordinates.

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

From our experience, we think that 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, Jiménez 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 needs the variogram models for the initial step. Instead of using variogram models, we propose to calculate the correlogram tables using the FFT approach according to [8]. Using it, we are able to find C(h) and C(0) matrices, which will be summed and inverted. Using so, we are able to implement adapted Ward's 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. The program to perform adapted Ward's method is written in FORTRAN language, with the same structure of GSLIB programs (see [1]); so, it is a modular structured program, which may be compiled and executed in any machine. The full program can be obtained from our website. A parameter file, which contains the information on input and output information, is needed to be able to run; such file is shown in Table 3.1. 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.

Table 3.1.- The parameter file.

```
Parameters for FFT Ward
**********************
START OF PARAMETERS:
darss.dat                   \file with data
1 2                         \x column, y column
3                           \debugging level
darss.dbg                   \file for debug
darss.out                   \file for output
1.67E-3 1.67E-3             \inverse distance x nodes, y nodes
2                           \number of variables
31 31                       \number of x nodes, y nodes
mapdarss                    \file for correlogram
```

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.

To compare results obtained using the methodology proposed by Jimenez [4], we have used the same data set with the same variables. Results are shown in Figures 3.1 and 3.2 for four and eight groups, respectively. In both figures, we can see that there is not the strong correspondence among groups and position pointed out with adapted Ward's method application, because in this case a variogram spherical model for all variables is forced, and even all variograms are forced to have same range and does not allow for nested structures.

Figure 3.1.- Application of generalized Ward's method into four groups.

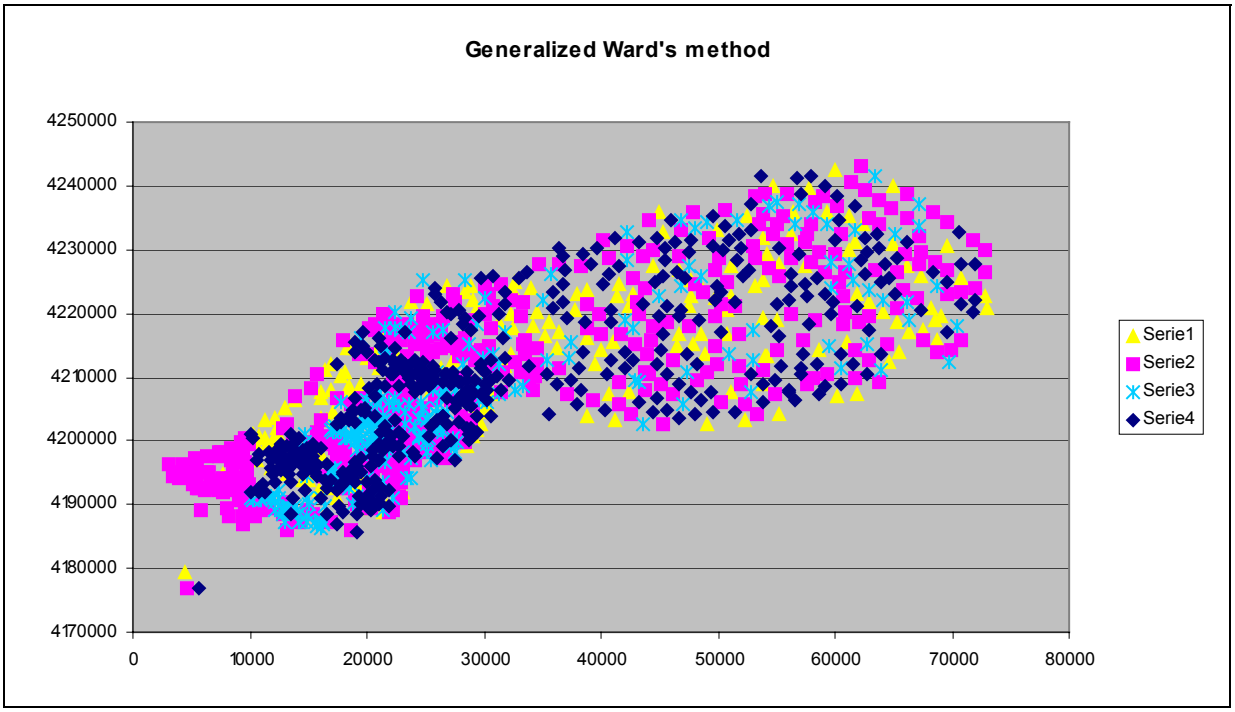
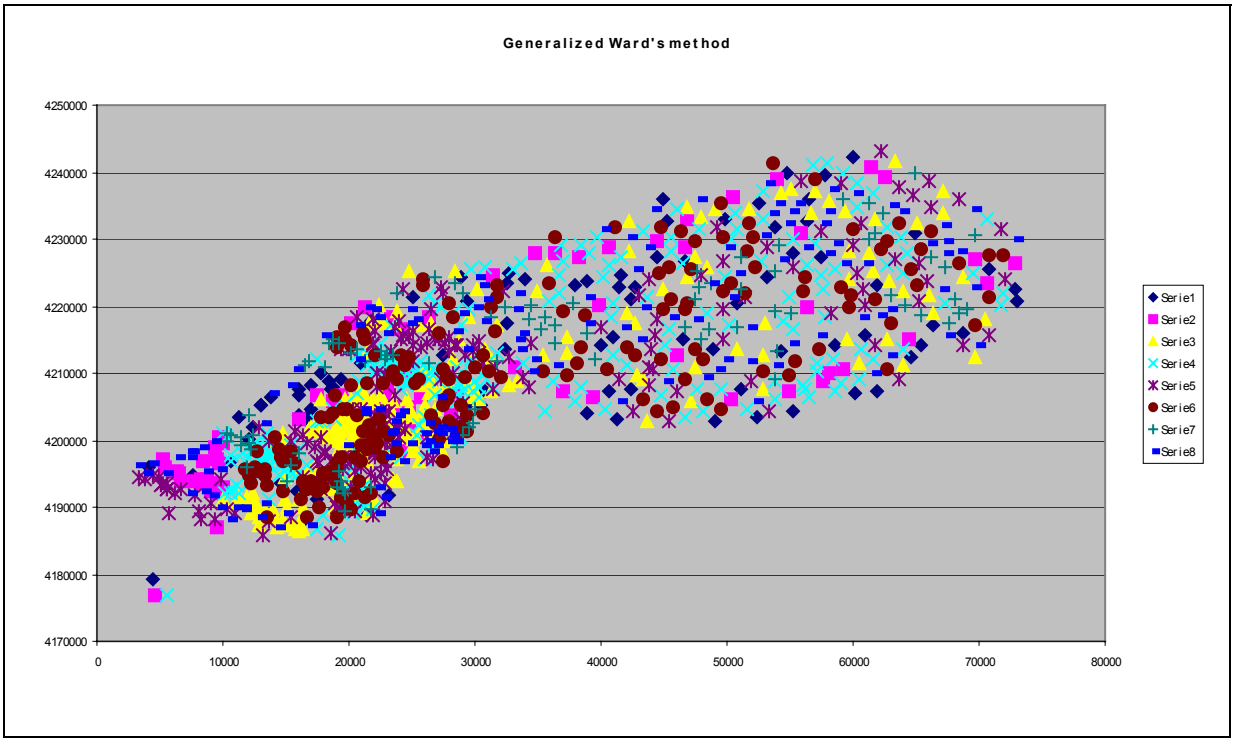


Figure 3.2.- Application of generalized Ward's method into eight groups.



4. Conclusions

As the most remarkable conclusions from this work, we can point out the following:

a) Adapted Ward's method described in [4] allows for 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 significant advantages of the methodology and 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.

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 set.