Terrain classification by cluster analysis

M. Crespi(.), G. Forlani(.), L. Mussio(.), F. Radicioni(..)
(.) Istituto di Topografia, Fotogrammetria e Geofisica - Politecnico di Milano - Piazza L. da Vinci, 32 - 20133 Milano.
(..) Dipartimento di Scienze dei Materiali e della Terra - Università di Ancona - Via Brecce Bianche - 60100 Ancona.
Italy
Commission No. III

1. Introduction

The digital terrain modelling can be obtained by different methods belonging to two principal categories: deterministic methods (e.g. polinomial and spline functions interpolation, Fourier spectra) and stochastic methods (e.g. least squares collocation and fractals, i.e. the concept of selfsimilarity in probability).
To reach good results, both the first and the second methods need same initial suitable information which can be gained by a preprocessing of data named terrain classification.
In fact, the deterministic methods require to know how is the roughness of the terrain, related to the density of the data (elevations, deformations, etc.) used for the interpolation, and the stochastic methods ask for the knowledge of the autocorrelation function of the data.
Moreover, may be useful or very necessary to split up the area under consideration in subareas homogeneous according to some parameters, because of different kinds of reasons (too much large initial set of data, so that they can't be processed togheter; very important discontinuities or singularities; etc.).
Last but not least, may be remarkable to test the type of distribution (normal or non-normal) of the subsets obtained by the preceding selection, because the statistical properties of the normal distribution are very important (e.g., least squares linear estimations are the same of maximum likelihood and minimum variance ones).
A general way to perform the terrain classification can consists of the following steps:
1 - to split up the original set of data into subsets (may be useful dimensions of these subsets are almost equal);
2 - to compute statistical parameters of each subset;
3 - to cluster these parameters to obtain homogeneous groups (clusters) represented by one value only (cluster point), so that the objects belonging to a cluster show a high degree of similarity while the ones belonging to different clusters are as dissimilar as possible;
4 - to test the significance of each cluster point and to modify the subsets;
5 - to subdivide the test area according to the results of the cluster analysis.
The elevations of the Noiretable area and the deformations of the Ancona '82 landslide are analyzed.
Before showing the results of the research, the algorithm used to perform the cluster analysis is described.

2. The ISODATA algorithm

This algorithm was originally published by G.H. Ball and D.J. Hall in the 1967; some changes were introduced in the steps 3) and 5) as regards to split and to merge the groups (clusters).
The essential steps of the tecnique are:
1 - to choose the initial set of cluster points and the values for the min-

imum norm (e.g. the Euclidean distance) between two cluster points ("dmin") and for the maximum dispersion parameter (e.g. the standard deviation) of each group ("smax");

2 - to group togheter the data closest to the same cluster point according to the norm chosen;

3 - to split into two groups each group whose dispersion parameter is larger than "smax"; the consequence of this operation is the birth of a new cluster point (the modality of this birth is arbitrary; the choice consists in the comparison of dispersions after having split up a group into two ones whose size are balanced according to the ratio 1:3, 2:2, 3:1;

4 - to regroup togheter the data taking into account the new cluster points;

5 - to merge into a group the groups whose cluster points are closer than "dmin"; the consequence of this operation is the death of some cluster points (the modality of this death is arbitrary; the choice consists in the joining of two or three groups, the second possibility occours if the number of groups is odd);

6 - to iterate the procedure from the step 2) (the implemented version of ISODATA algorithm stops if in 3 consecutive iterations the number of cluster points isn't changed and however after 25 iterations).

The program CLUSTER was implemented to perform the algorithm.

3. Subdivision of the original sets of data

The Noiretable test area is a hilly zone of central France, whose shape is a square with the sides about 3.2 km long. A very regular grid of about 6600 measured elevations (average spacing 40 m) was used for computation.

The Ancona '82 landslide area is a zone in the district of Ancona (Italy) interested by a large scale landslide, whose shape is almost rectangular with sides about 2.1 and 3.0 km long. About 150 cross-sections of different lengths and densities about 8900 measured altimetric deformations (average spacing 20 m) were used for computations.

Taking into account the shapes of the two areas and the positions of the measurements, they were subdivided into 64 and 59 subsets respectively, as shown in the figs. 1 and 2.

In such a way, subsets of data very homogeneous for the first area (the size is about 100 points) and homogeneous enough for the second one (the size ranges about 80-220 points) as regards their dimensions were obtained.

4. Statistical analysis of the subsets

The following statistical elaborations were carried out:

- computation of average, standard deviation, asymmetry, kurtosis and first autocorrelation coefficient for each subset;
- construction of the histograms of these statistical parameters;
- construction of the histograms of the data for each subset (some examples in figs. 3 and 4);
- computation of the empirical autocorrelation functions of the data for each subset;
- best fitting of these functions with teorical ones (some examples in figs. 3 and 4);
- construction of the histrograms of the noise standard deviation and of the theorical autocorrelation function first zero.

Note that the histograms of the signal standard deviation are the same of the histograms of the first autocorrelation coefficient.

Two conclusions were reached by this statistical analysis:

- in general, the subsets haven't a normal distribution;
- all the theorical autocorrelation functions which best fit the empirical

ones belong to the family of the J_o Bessel functions which have the following expression:

$$f(x) = a\ J_o(cx),$$

where a and c are two parameters related to the variance of the signal and the first zero.

Note that the subsets of Ancona'82 landslide area were divided into two groups according to the result of comparison between the values of the noise and the signal standard deviation as an altimetric deformation didn't happen in all the grid-points, so that some autocorrelation functions don't represent a stochastic process but a white noise only.

These deductions are important because the first establishes the least squares method can't reach the best results in this case, and the second gives a very exact information about the most suitable theorical autocorrelation function to use in the stocastic approach to the digital terrain modelling.

5. Cluster analysis of statistical parameters

The cluster analysis was applied to all the statistical parameter computed before. The results regarding kurtosis, noise standard deviation, first autocorrelation coefficient and theorical autocorrelation function first zero are shown only for the sake of brevity. On the other hand, these last are the most meaningful results, because the kurtosis is generally discriminant about the normality of the distribution and the other parameters characterize completely the autocorrelation function, then the stochastic process.

The choices regarding the initial set of cluster points and the values for "smax" and "dmin" were made in the following way for each statistical parameter:

- the initial cluster points were the averages of the modal classes found by the histogram;
- the values for "smax" and "dmin" were assumed equal to the double of the class interval of the histogram.

At the present a significance test for of the cluster points isn't available. Indeed to test them, distribution-free tests concerning non-random samples are necessary.

So the figures show the results of the ISODATA algorithm without subsequent refinings; however some observations can be developed.

The grey tonality of the fig. 5 is uniform: the kurtoses of the subsets belonging to the Noiretable area are grouped togheter around one cluster point only, whose value is remarkably lower than 3 (the significance is evalueted by the D'Agostino-Pearson test) according to the powerful Kolmogorov-Smirnov test. It means not only the most part of subsets haven't a normal distribution (same of them could have it, but they are so few they can't cause the birth of an other cluster point), but also the global set of data hasn't a normal distribution.

Also the most part of the subsets of data of the Ancona '82 landslide has a kurtosis remarkably greater than 3. Indeed principal clusters show a behaviour apart from the normality, but the geomorphological interpretation of each cluster isn't immediate (see fig. 6).

The grey tonality of the fig. 7 is uniform: the noise standard deviation of the subsets belonging to the Noiretable area are grouped togheter around one cluster point only whose square value is remarkably lower than 0.5 times the standard deviation of the process. It means there is a stochastic process as regards the elevations in each part of the area.

The clearest grey tonality of the fig. 8 corresponds to the cluster point of the lowest value for the noise standard deviation computed on Ancona '82

landslide area data; its square value is remarkably lower than 0.5 times the standard deviation of the process. It means there is a stochastic process as regards the altimetric deformations in some parts of the area only, that is true. Moreover, the comparison between figs. 2 and 8 shows the classification by cluster analysis is able to distinguish the parts interested or no by landslide deformation exactly enough, i.e. the parts with and without the stochastic signal as regards the altimetric deformations. Some subsets belonging to the middle grey tonality are located along the border of the landslide; a more refined partition could solve the ambiguity.
Note that the cluster analysis applied to the averages can be useful to recognize very important discontinuities or singularities. Besides the cluster analysis applied to all the theorical autocorrelation function parameters (see figs. 9, 10, 11 and 12) individuates connected homogenous subareas useful for a successive data filtering.

Appendix

There are two fundamental types of algorithms for cluster analysis.
The "iterative" algorithm, like ISODATA, in which one must choose the initial set of cluster points and the values of two parameters related to the relative positions of the cluster points and to the dispersion of each cluster before starting the procedure.
The algorithm with an objective function, like MEDOIDS and FUZZY ISODATA, in which one must choose the number of cluster points only before starting the procedure. In general, the objective function has the following expression:

$$F_{p,q}(C_1,\ldots,C_n) = \sum_{1}^{n} {}_j\; e_{p,q}\,(C_j),$$

where:

$$e_{p,q}(C_j) = \min_{y_j} \sum_{i\varepsilon C_j} ||x_i - y_j||_p^q \quad (1 \leq p \leq \infty,\ q \geq 1),$$

C_j is the j-th cluster $(1 \leq j \leq m)$,

x_i is the i-th object $(1 \leq i \leq n,\ n \geq m)$,

y_j is the j-th cluster point (representative value for the j-th cluster C_j),

i.e. the sum (on all the clusters) of the sum (on all the objects belonging to each cluster) of the L_p distances to the q-th power of the x_i $(i \ \varepsilon\ C_j)$ from the unknown cluster point y_j. At the present, it is possible to compute the cluster points y_j, then to determinate C_j for p=q=1, p=q=2 and p=∞, q=1 only.

Acknowledgements

The authors wish to thank very much prof. B. Benciolini and Mr. L. Pallottino, for the use of their programs to plot images: they were very helpful to draw all the figures.

References

Ammannati F., Benciolini B., Mussio L., Sansò F. (1983): "An Experiment of Collocation Applied to Digital Height Model Analysis". Proceedings of the Int. Colloquium on Mathematical Aspects of Digital Elevation Models, Stockholm 19-20 April 1983, Dept. of Photogrammetry, Royal Istitute of Technology, pp. 1.1 -1.9, Stockholm, 1983.

Ammannati F., Betti B., Mussio L. (1984): "Statistical Analysis of Relevant Terrain Features Through Digital Elevation Models". Int. Archives of Photogrammetry and Remote Sensing, vol. XXV, part A3A/B, Rio de Janeiro 17-29 June 1984, pp. 774-780, Rio de Janeiro, 1984.

Ball G.H., Hall D.J. (1967): "A Clustering Tecnique for Summarizing Multivarite Data". Behavioral Science, 12 (1967), pp. 153-155, 1967.

Colombo L., Fangi G., Mussio L., Radicioni F. (1986): "Further Development on Digital Models in a Control Problem for the Ancona '82 Landslide". Proceedings of the Int. Symp from Analytical to Digital of the ISPRS Comm. III, Rovaniemi 19-22 August 1986, Int. Archives of Photogrammetry and Remote Sensing, vol. XXVI, parte 3.1, pp. 134-147, Rovaniemi, 1986.

Crippa B., Mussio L. (1987): "The new ITM System of Programs MODEL for Digital Modelling". Proceedings of the Int. Colloquium on Progress in Terrain Modelling, Copenhagen 20-22 May 1987, Institute of Surveying and Photogrammetry, Technical University of Denmark, pp. 77-88, Lyngby, 1987.

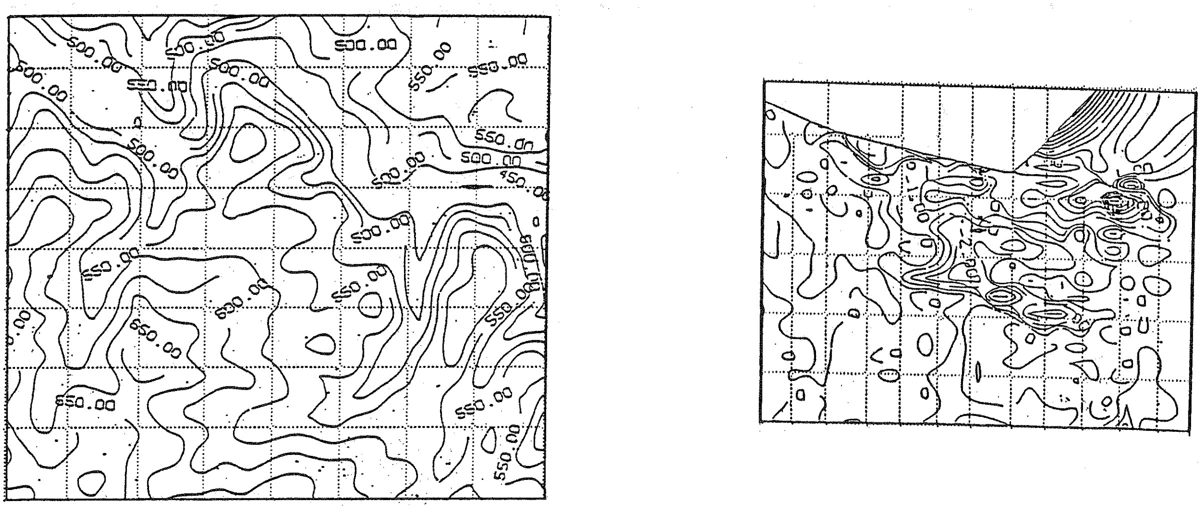
Cunietti M., Fangi G., Mussio L., Radicioni F. (1984): "Block Adjustment and Digital Model of Photogrammetric Data in a Control Problem for the Ancona 82 Landslide". Int. Archives of Photogrannetry and Remote Sensing, vol. XXV, part A3A/B, Rio de Janeiro 17-29 June 1984, pp. 774-780, Rio de Janeiro, 1984.

Frederiksen P., Jacobi O., Kubik K. (1984): "Modelling and Classifying Terrain". Int. Archives of Photogrannetry and Remote Sensing, vol. XXV, part A3A/B, Rio de Janeiro 17-29 June 1984, pp. 256-267, Rio de Janeiro, 1984.

Kaufman L., Rousseeuw P.J. (1987): "Clustering by means of Medoids". In Dodge Y. (Ed.), Statistical Data Analysis Based on the L1-norm and Related Methods, pp. 405-416, North-Holland, 1987.

Spath H. (1987): "Using the L1 norm within cluster analysis". In Dodge Y. (Ed.), Statistical Data Analysis Based on the L1-norm and Related Methods, pp. 427-433, North-Holland, 1987.

Trauvert E. (1987): "L1 in Fuzzy Clustering". In Dodge Y. (Ed.), Statistical Data Analysis Based on the L1-norm and Related Methods, pp. 417-426, North-Holland, 1987.



Figs. 1-2 - Contour lines of data and borders of the 64 subsets of the Noiretable area and of the 59 subsets of the Ancona '82 landslide area.

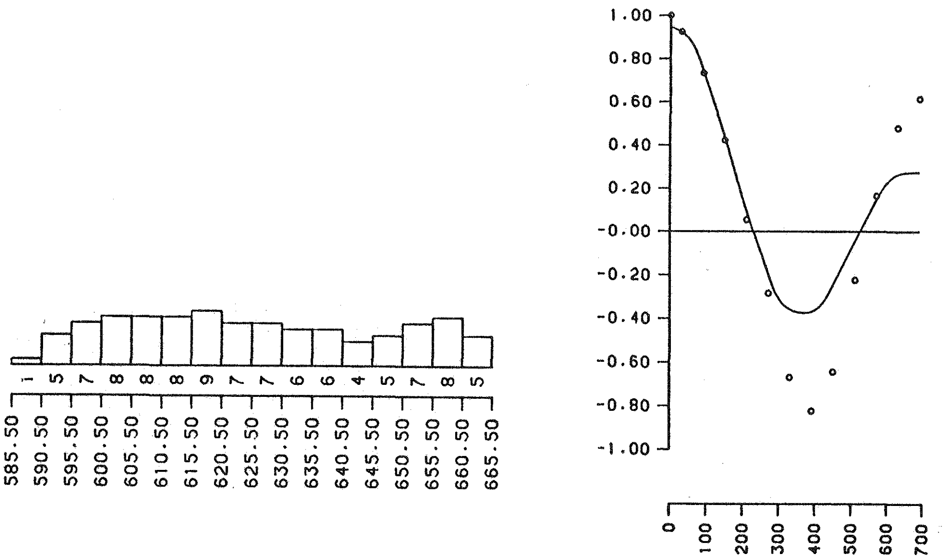
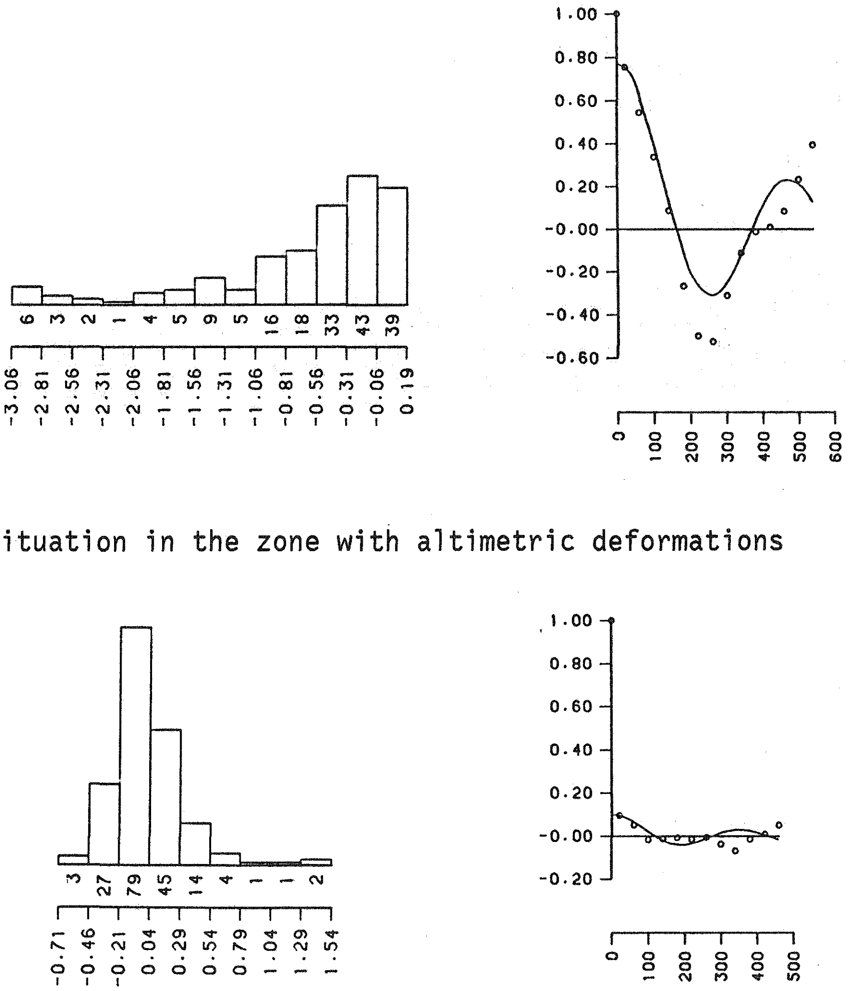


Fig. 3 - Remarkable histogram and autocorrelation function for the Noiretable area.



a) situation in the zone with altimetric deformations

b) situation in the zone without altimetric deformations

Fig. 4 - Remarkable histograms and autocorrelation functions for the Ancona '82 landslide area.

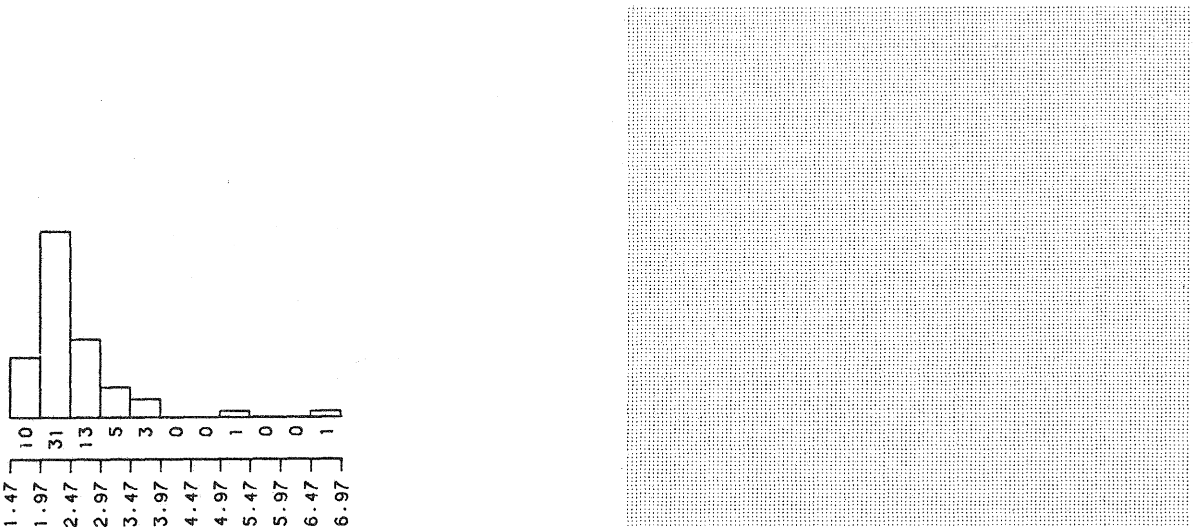


Fig. 5 - Histogram and representation of cluster analysis for the kurtoses of the subsets of the Noiretable area.

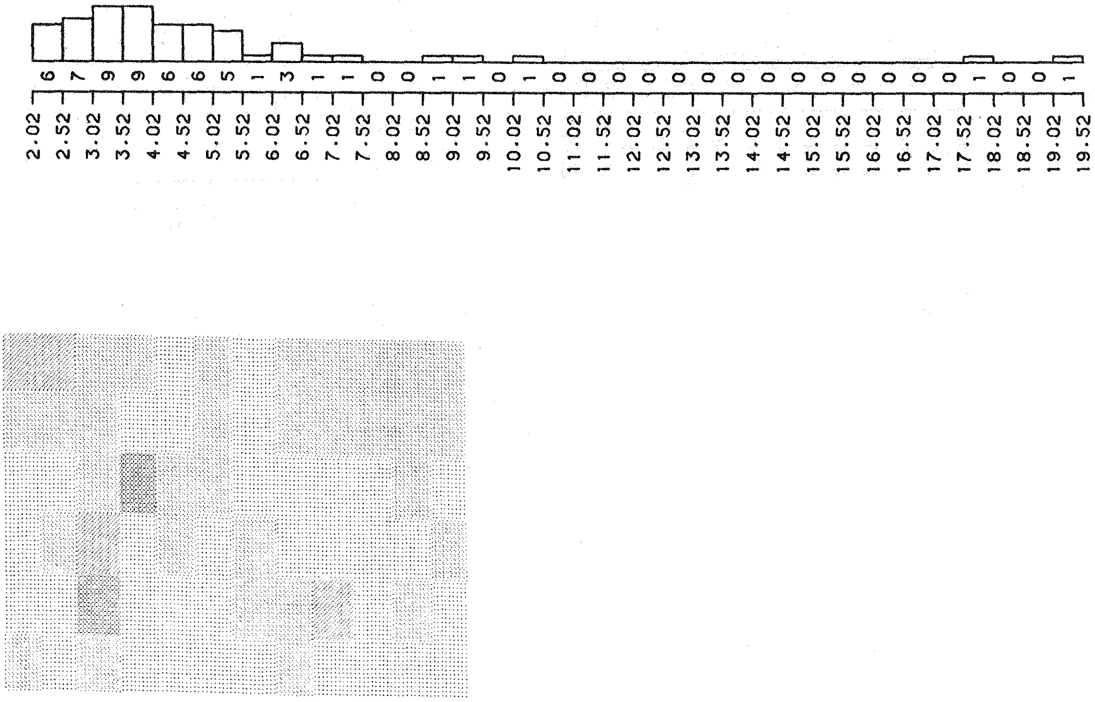


Fig. 6 - Histogram and representation of cluster analysis for the kurtoses of the subsets of the Ancona'82 landslide area.

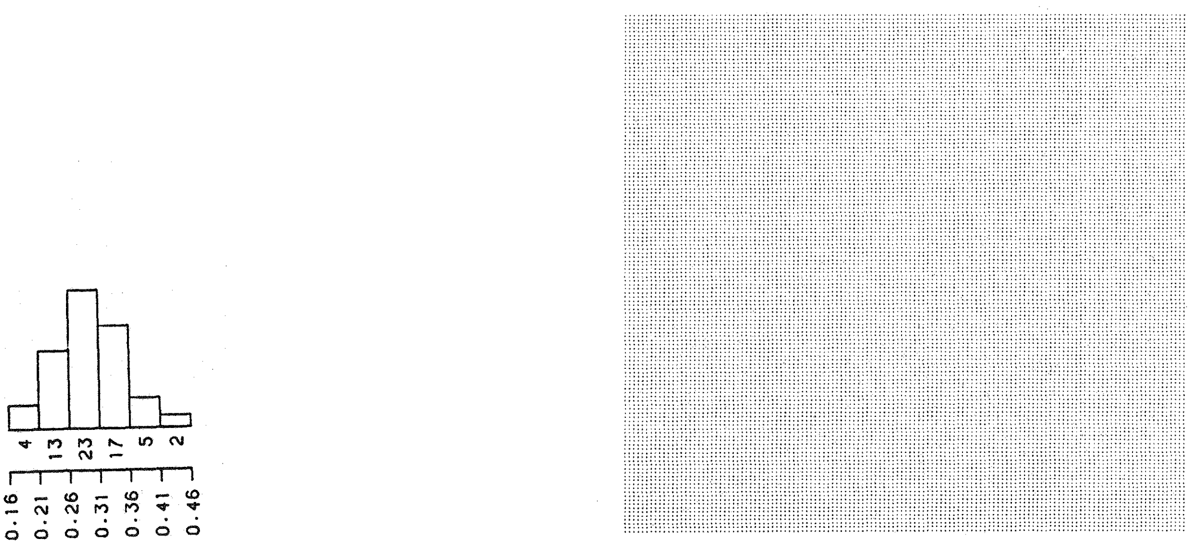


Fig. 7 - Histogram and representation of cluster analysis for the noise standard deviations of the subsets of the Noiretable area.

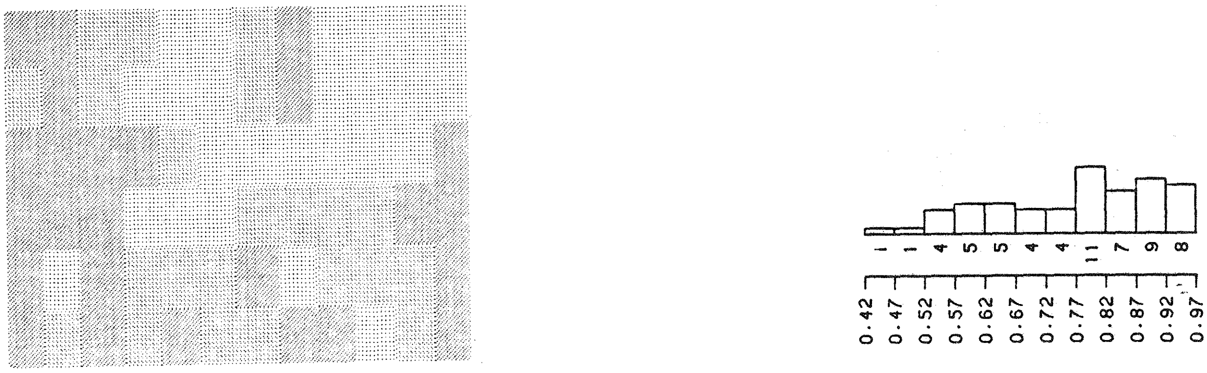


Fig. 8 - Histogram and representation of cluster analysis for the noise standard deviations of the subsets of the Ancona '82 landslide area.

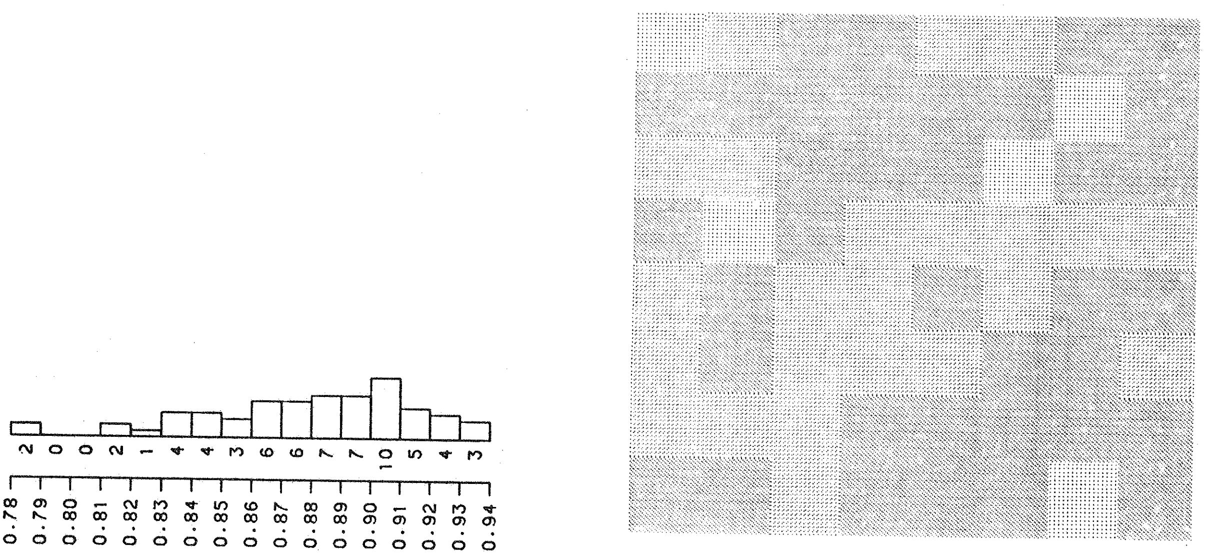


Fig. 9 - Histogram and representation of cluster analysis for the first autocorrelation coefficient of the subsets of the Noiretable area.

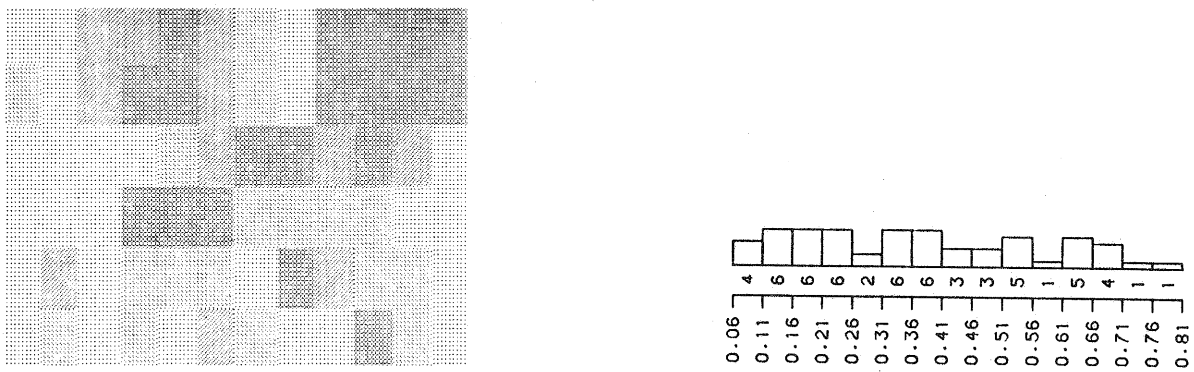


Fig. 10 - Histogram and representation of cluster analysis for the first autocorrelation coefficient of the subsets of the Ancona '82 landslide area.

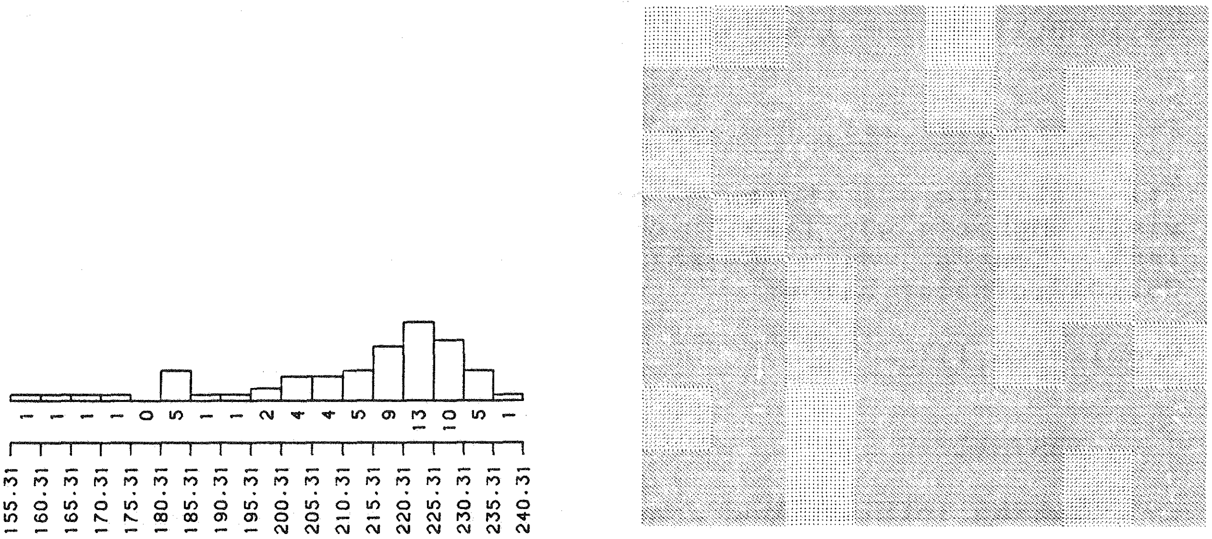


Fig. 11 - Histogram and representation of cluster analysis for the theorical autocorrelation functions first zero of the subsets of the Noiretable area.

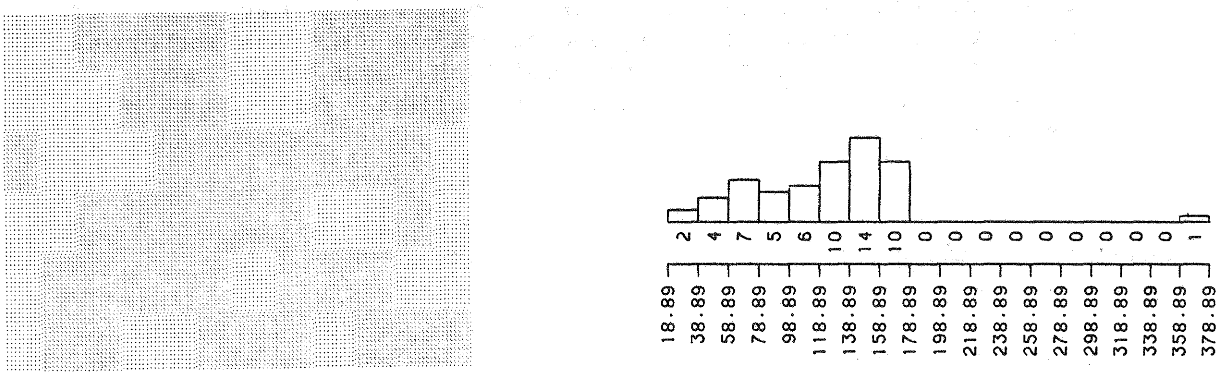


Fig. 12 - Histogram and representation of cluster analisys for the theorical autocorrelation functions first zero of the subsets of the Ancona '82 landslide area.