Statistical Methods & Applications (2001) 10: 237–256

SMA

A constrained k-means clustering algorithm for classifying spatial units

G. Damiana Costanzo

Dipartimento di Economia e Statistica, Università della Calabria, Arcavacata di Rende 87036 (CS), Italy
(e-mail: dm.costanzo@unical.it)

Abstract. In some classification problems it may be important to impose constraints on the set of allowable solutions. In particular, in regional taxonomy, urban and regional studies often try to segment a set of territorial data in homogenous groups with respect to a set of socio-economic variables taking into account, at the same time, contiguous neighbourhoods. The objects in a class are thus required not only to be similar to one another but also to be part of a spatially contiguous set. The rationale behind this is that if a spatially varying phenomenon influences the objects, as could occur in the case of geographical units, and this spatial information were ignored in constructing the classes then it would be less likely to be detected. In this paper a constrained version of the k-means clustering method (MacQueen, 1967; Ball and Hall, 1967) and a new algorithm for devising such a procedure are proposed; the latter is based on the efficient algorithm proposed by Hartigan and Wong (1979). This algorithm has proved its usefulness in zoning two large regions in Italy (Calabria and Puglia).

Key words: k-means clustering, constrained optimisation, contiguity matrix, spatial data, regional taxonomy, segmentation.

1. Introduction

In certain classification problems it may be important to impose constraints on the set of allowable solutions. For example, restrictions can be placed on the properties of classes in a partition (number of groups, maximum number of members in a cluster, shape of the clusters) or on the topology of a tree diagram specifying a hierarchical classification. The reasons for imposing such additional conditions on a classification depend on the application and include, for example, the need to match a sales area to resource constraints in market segmentation (DeSarbo and Mahajan, 1984) or to obtain enumeration districts for administrative or electoral

purposes (Mills, 1967; Openshaw, 1977; Taylor, 1973). Besides these external reasons, one might want to constrain the classification in order to test a specific theory or hypothesis (e.g. De Soete et al., 1984).

In the analysis of spatially located data, constrained classifications are usually obtained by using relational constraints, that is symmetric and reflexive relations in addition to proximities between objects. Depending on the application relational constraints may be defined in different ways; for example, in image processing where the image consists of pixels characterised by different grey-level intensities, the eight neighbouring pixels (west, south-west, south, etc.) may define the contiguity relationships for any given pixel. Another example comes from soil science: scientists study the properties of soil profiles at many different sites aiming, for convenience of soil management, to create parcels of land with similar soil properties (Webster,1977). The terrain may be subdivided into square parcels and the neighbours of a parcel may be defined as its eight adjacent parcels. Lebart (1978), Lechevallier (1980), Batagelj (1984), Ferligoj and Batagelj (1982, 1983, 1992, 1998, 2000) among others, deal with clustering problems with relational constraints.

In regional taxonomy, urban and regional studies often try to divide (regionalize or zone) a set of territorial data (administrative units, regions, countries, etc.) into homogenous groups with respect to a set of socio-economic variables, while taking into account contiguous relationships. The rationale behind this is that if a spatially varying phenomenon influences the objects and the spatial information were ignored in constructing the classes then it would be less likely to be detected (Monestiez, 1977; Legendre, 1987; Gordon, 1996).

In the case of territorial objects, the most common and easiest way to express the neighbour relationships is by a binary matrix, with a contiguity value $c_{ij} \in \{0, 1\}$ defined for each pair of objects. The specification of a contiguity matrix on a set of geographic objects is not a simple matter owing to surface irregularities such as rivers or lakes that make definition difficult. This problem has been directly addressed by some authors (Gordon, 1973, 1999) and usually areas sharing a common boundary are regarded as contiguous.

A number of approaches for overcoming contiguity constrained clustering problems has been proposed; extensive reviews are provided by Murtagh (1985), Zani (1993) and, more recently, by Gordon (1996). One approach has been to adopt traditional procedures which have proved their worth in standard problems and to enforce geographic connectivity by incorporating contiguity constraints, e.g. the use of agglomerative hierarchical algorithms (e.g. Lebart, 1978; Lefkovitch, 1980; Gordon, 1973, 1980, 1987; Legendre, 1987) and local optimisation clustering procedures (Ferligoj and Batagelj, 1982, 1983).

However, as far as we know, despite the number of constrained clustering algorithms proposed, there have been a few practical applications of such procedures in real situations.

In this paper a constrained version of the k-means (Mac Queen, 1967; Ball and Hall, 1967) clustering method and a new algorithm for devising such a procedure are proposed; the latter is based on the efficient algorithm proposed by Hartigan and Wong (1979). It has proved its usefulness in the zoning of two large regions in

the south of Italy (Calabria and Puglia). The aim was to detect sub areas in the two Italian regions in order to target development policies more efficiently.

The paper is organised as follows. In Section 2, the model on which the new algorithm is based is formally introduced; next, in Section 3, the algorithm developed for the model is described and, finally, case study applications are presented in Section 4.

2. The model

The k-means clustering problem can be posed in the following terms (De Soete and Carroll, 1994). Let $\mathbf{X} = [x_{ik}]$ be the matrix containing measurements of K variables ($k = 1, \ldots, K$) on N individuals or objects ($i = 1, \ldots, N$). If the aim of the analysis of the data matrix $\mathbf{X}$ entails clustering the N objects into L ($L < N$) homogenous (non overlapping) clusters starting from $\mathbf{X}$, let $\mathbf{M}= [m_{lk}]$ be the $L \times K$ matrix specifying the centroids of the L clusters and let $\mathbf{U}$ be an $N \times L$ binary indicator matrix designating the cluster membership for each object

$$u_{il} = \begin{cases} 1 \text{ if object } i \text{ belongs to cluster } l \\ 0 \text{ otherwise} \end{cases}$$

The k-means clustering requires that each object be assigned to one and only one cluster, that is $\sum_{l=1}^{L} u_{il} = 1$ holds for $i = 1, \ldots, N$. The matrices $\mathbf{M}$ and $\mathbf{U}$ are determined so that the sum of the squared Euclidean distances between the object and the centroids of the clusters to which they belong is minimal; that is $\mathbf{M}$ and $\mathbf{U}$ are determined so that the least-squares loss function

$$\begin{aligned} \mathrm{F}(\mathbf{M}, \mathbf{U}) &= \|\mathbf{X} - \mathbf{UM}\|^2 = \mathrm{tr}(\mathbf{X} - \mathbf{UM})'(\mathbf{X} - \mathbf{UM}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}(x_{ik} - \sum_{l=1}^{L} u_{il}m_{lk})^2 = \sum_{i=1}^{N}\sum_{l=1}^{L} u_{il}\sum_{k=1}^{K}(x_{ik} - m_{lk})^2 \end{aligned}$$

is minimal.

In a contiguity-constrained formulation of this problem, it is required that the objects in a class are not only similar to one another but also that the classes comprise a spatially contiguous set of objects.[1] The contiguity relationship between spatial objects may be defined in different ways depending on the application (cfr. Murtagh, 1985); geographical contiguity is a special case of contiguity constraint and rests on the physical closeness of bordering geographical areas.

However, the above is not the only possible definition of contiguity for spatial objects. For example, in spatial statistics where data are frequently organized as contiguous *quadrats* the following definitions of contiguity can be considered (Upton and Fingleton, 1985): a) the *rook's* case where quadrats abut only if they

[1] Actually, the main idea of a local optimisation procedure for clustering with relational constraints has been discussed by Ferligoj and Batagely (1982). In their procedure the local optimisation neighbourhood of clustering is based on the transformation which transfers an object from one cluster to another one.

have touching edges; *b*) when interest is focused on diagonal spatial trends a more relevant definition of contiguity is the *bishop's* case, where quadrats abut if they have touching corners; *c*) another composite definition of contiguity is the *queen's* case where quadrats abut if they have either touching corners or touching edges.

Neighbourhood information based on distance, e.g. Euclidean, may be also specified; for example, any site which is within a specified distance value ρ of the i^{th} site can be considered a neighbour of this site. Other definitions may be based on certain probabilistic rules; for example, site k can be defined as neighbour of site i if the conditional distribution of the spatial process $Z(i)$, given all other site values, depends functionally on $z(k)$, for $k \neq i$ (cfr. Cressie, 1993, pp. 414–416).

Whatever the contiguity definition one considers, let $\mathbf{C}$ be an $N \times N$ binary matrix that expresses the contiguity relationship among the objects to be clustered, with a contiguity value

$$c_{ij} = \begin{cases} 1 \text{ if object } i \text{ is contiguous to object } j \\ 0 \text{ otherwise} \end{cases}$$

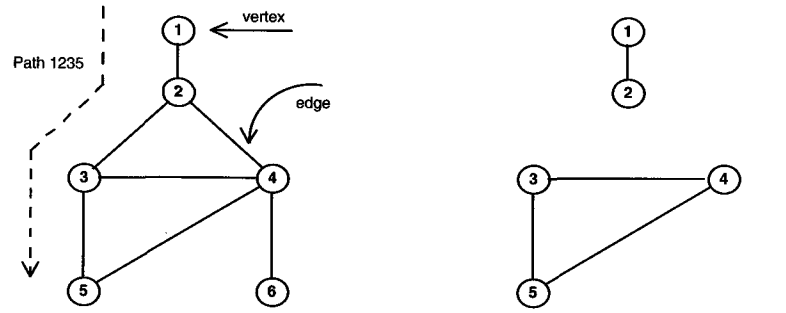
defined for each pair of objects.

In a constrained partitioning object i belonging to cluster l is required to be contiguous to, at least, one of the other contiguous objects in the same class. That is, we require that each object i belongs to a "contiguous" cluster. The idea of contiguous clusters can be conveniently handled by borrowing some basic principles from graph theory, namely those of *connectivity* and *reachability* (see Christofides, 1975, Ch. 2).

Various authors investigated the relationships between graph theory and the clustering of a set of objects (see for example Hubert, 1974 for a review), while others approached their classification problems directly through graph-theoretic models (e.g. recently Maravalle et al., 1997, on relational constraints).

Actually, matrix $\mathbf{C}$ can be considered as an adjacency matrix, say $\mathbf{A}$, of a graph with all its diagonal elements set to 1 (assuming that each object is contiguous to itself), that is $\mathbf{C} = \mathbf{I} + \mathbf{A}$, where $\mathbf{I}$ is the $N \times N$ identity matrix. Since the adjacency matrix defines the structure of the graph completely, contiguity matrix $\mathbf{C}$ induces a (non directed) graph G(N,$\mathbf{C}$) where the set of the objects $\{1, 2, \ldots, N\}$ is the set of the *vertex* or knots of the graph, while $c_{ij} = 1$ means that there is an *edge* or link between vertex i and vertex j. For graph-theory terminology we refer to Wilson (1996) or Harary (1969), while some basic terms are illustrated in Fig. 1.

Let $\mathbf{C}_l$ be the contiguity sub-matrix of the matrix $\mathbf{C}$ for the lth cluster, for the above contiguity constraint to be satisfied, the sub-graph G($N(l)$,$\mathbf{C}_l$) induced by $\mathbf{C}_l$ needs to be connected. That is, every pair of vertices in the graph must be joined by a *path*. Another way to define the connectivity of a graph is in matrix terms, if there is no labeling of the knots of the graph such that its adjacency matrix can be reduced to a block diagonal matrix (see Harary, 1969 pp. 150–159). The latter condition could be used to ascertain the contiguity of the clusters. However, a computationally simpler and more direct way to establish the connectivity of the clusters of a classification which also allow us to render the contiguity constraint in formal terms is based on the *cardinality* of a path, that is the number of edges appearing in the path.



$$\mathbf{C = A + I} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

R(1) = {1} ∪ {2} ∪ {3,4} ∪ {5,6} ∪ {3,4}
= {1,2,3,4,5,6}

Fig. 1. A graph, its contiguity matrix and a reachable set

For a given object i in a cluster l consider the following sets

$\Gamma_l(i)=\{j \in l : c_{ij} =1; j \neq i\}$;
$\Gamma_l(\Gamma_l(i)) = \Gamma_l^2(i) = \{k \in l : c_{jk} = 1; \forall j \in \Gamma_l(i); k \neq j\}$;
$\vdots$
$\Gamma_l^p(i) = \{r \in l : c_{rs} = 1; \forall s \in \Gamma_l^{(p-1)}(i); r \neq j\}$.

That is the sets of objects which are reachable from i along a path of cardinality $1, 2, \ldots, p$ respectively, where p is bounded by the number of vertices in the graph minus one, $N(l) - 1$.

Let us assume that $\Gamma_l^0(i) = \{i\}$ and let $|S|$ denote the number of elements in S, that is the *cardinality* of the set S; if the cluster l is connected

$$|\bigcup_{\gamma=0}^{p} \Gamma_l^{\gamma}(i)| = N(l)$$

with $p \leq N(l) - 1$. That is, *reachable set* $R(i) = \bigcup_{\gamma=0}^{p} \Gamma_l^{\gamma}(i)$ is obtained by performing the union operations from left to right until the current total set is no longer increased in size by the next union.[2] When this occurs any subsequent union will not add new members to the set. Thus if i is a member of a contiguous cluster l

[2] Note that there are some slight differences between our definition of a reachable set and the one given by Christofides (1975) mainly due to the use of the contiguity matrix instead of the adjacency one.

then all the objects in the cluster must be reachable from i along (at least) one path, that is $R(i)$ must be equal to the size of the cluster. An example of a reachable set is given for the graph in Fig. 1.

Let us indicate $\bigoplus$ the Boolean arithmetic sum (i.e. $0+0=0, 0+1=1+1=1$), the above contiguity constraint can be expressed in an equivalent form involving the columns of the matrices **C** and **U**.

Let $\mathbf{u}_l$ be the column vector of **U** so that $u_{il}=1$, denoting the cluster l to which object i belongs, and let $\mathbf{c}_i$ be the column vector of **C**, the contiguity-constrained k-means problem can be posed in the following terms:

$$minimize\ \mathrm{F}(\mathbf{M},\mathbf{U}) = \sum_{i=1}^{N}\sum_{l=1}^{L} u_{il} \sum_{k=1}^{K}(x_{ik}-m_{lk})^2 \tag{1}$$

with constraints

$$sub \begin{cases} \mathbf{u}_l' \left(\bigoplus_{j \in R_l(i)} \mathbf{c}_j \right) = \mathbf{u}_l'\mathbf{u}_l \\ \sum_{l=1}^{L} u_{il} = 1 \\ u_{il} \in \{0,1\} \end{cases} \qquad i=1,\ldots,N; l=1,\ldots,L \tag{2}$$

The k-means clustering problem then entails determining **U** and **M** so that the loss function (1) subject to constraints (2) is minimized.

The minimization procedure for solving such a constrained optimisation problem involves alternating between minimizing the loss function F(**M**,**U**) with respect to **U** subject to constraints (2) given the current estimates of the cluster centroids **M**, and minimizing F(**M**,**U**) with respect to **M** given the current cluster assignment **U**.

In other words, the minimization of F(**M**,**U**) can be achieved by means of an alternating procedure that we can roughly summarize in the following two steps:

a) given current estimates of the cluster centroids **M** updating **U** through new estimate cluster assignment $\hat{\mathbf{U}}$ so that:

$$\hat{u}_{il} = \begin{cases} 1 \text{ if } \begin{cases} \sum_{k=1}^{K} (x_{ik}-m_{lk})^2 < \sum_{k=1}^{K} (x_{ik}-m_{l'k})^2 \\ \mathbf{u}_l' \left(\bigoplus_{j \in R_l(i)} \mathbf{c}_j \right) = \mathbf{u}_l'\mathbf{u}_l \end{cases} \qquad l'=1,\ldots,L; l' \neq l \\ 0 \text{ otherwise} \end{cases}$$

b) given the new cluster $\hat{\mathbf{U}}$ assignment, updating **M** by computing new cluster centroids $m_{lk} = \frac{1}{N(l)} \sum_{i=1}^{N} \hat{u}_{il} x_{ik}$, where $N(l) = \sum_{i=1}^{N} \hat{u}_{il}$ is the number of objects in cluster l until, step by step, no more changes occur in **U** and hence in **M**.

In the next section we present an algorithm to implement this basic two step procedure.

3. A k-means constrained clustering algorithm

Actually, there are a number of versions of the (standard) k-means algorithm. They essentially differ with respect to (i) how the clusters are initiated, (ii) how objects are allocated to clusters (that is the *movement rule*), (iii) how some or all of the already clustered objects are reallocated to other clusters (that is the *updating rule*). The different rules have an effect on calculation time, storage requirements, convergence to at least a local optimum, expected difference between the local optimum and the global optimum. They are also usually employed in evaluating and comparing the different algorithms (see Seber, 1984, Hartigan, 1975 and Späth, 1980 for a discussion on the k-means algorithms).

The algorithm developed for the model outlined in the previous section is based on the algorithm **KMNS** proposed by Hartigan and Wong (1979, *Algorithm AS 136*, pp. 100–108) for performing a standard k-means cluster analysis.

The **KMNS** algorithm aims at finding a k-partition with a local optimal within-cluster sum of squares which cannot be reduced by moving points from one cluster to the other.

One attractive feature of this algorithm is the reduced number of optimal transfer iterations considerably reducing calculation time through the use of a "live set" of clusters, an optimal transfer stage (*OPTRA*) and a quick-transfer stage (*QTRAN*). Clusters involved in a transfer at a given step are, in fact, memorised (in the "live set") and only these updated clusters enter the calculations at the subsequent step. The iterative relocation procedure stops when the live set is empty. Moreover, at the stage where each point is examined in turn to see if it should be reassigned to a different cluster, the algorithm searches for two cluster centres, the closest one and the second closest one. In fact, the use of the closest centre alone to check for possible reallocation of a point does not guarantee an optimal solution since a cluster centre other than the closest (in a Euclidean sense) one, may result in a lower value of the objective function (cfr. Hartigan and Wong, 1979).

The constrained k-means algorithm proposed in this paper, uses a three stage procedure: a constrained optimal assignment (*COPASS*) stage, a constrained optimal-transfer (*COPTRA*) stage and a constrained quick-transfer (*CQTRAN*) stage. The latter two stages being obtained by modifying the *OPTRA* and *QTRAN* stages of **KMNS** algorithm by allowing for contiguity constraints.

The general procedure for solving the constrained problem outlined in the previous section involves finding a partition by moving points from one cluster to another in such a way that in transferring a point the following conditions are met:

i) with respect to the cluster the point has to be reassigned to: object is contiguous to at least one of the other objects in the target cluster and its transfer results in a smaller value of the objective function. Given current cluster assignment **U**, if $d(i,l)$ denote the Euclidean distance between the point i and the centroid of cluster l, this is attained by considering object i for transfer from cluster l to cluster l' if

$$\{N(l')/[N(l')+1]\}\, d^2(i,l') < \{N(l)/[(N(l)-1])\}\, d^2(i,l)$$

and

$$\sum_{j=1}^{N} c_{ij} u_{jl'} \geq 1$$

that is if i is contiguous to, at least, one other object belonging to cluster l'.

ii) with respect to the cluster the point has to be transferred from: the transfer of point i does not cause the cluster l to which it belongs to be divided into more than one contiguous cluster, that is if u_{il} =0 we should have

$$\mathbf{u}_l' \left(\bigoplus_{j \in R_l(i)} \mathbf{c}_j \right) = \mathbf{u}_l' \mathbf{u}_l = N(l) - 1 \; \forall k \in l$$

Similar conditions on the reassignment of the points are given by Ferligoj and Batagelj (1982).

Given a matrix **X** of N points in K dimensions, a matrix **S** of L initial cluster seeds in K dimensions and an $N \times N$ contiguity matrix **C**, define IC1 and IC2 two N-dimensional column vectors representing cluster membership variables. They denote for each point i, the closest and the second closest contiguous cluster centres respectively. Further, we assume that IC2(i)=0 means that the second closest contiguous cluster does not exist. The general structure of the constrained k-means algorithm **CKMNS** is shown in Table 1.

To start the algorithm an initial estimate of **M** is required; in *Step 0a* and *Step 0b* initial estimates of cluster centres are formed starting from L seeds, randomly chosen or selected on the basis of some *a priori* information about the objects to be classified.

Note that the **CKMNS** algorithm differs from the **KMNS** algorithm in that some objects at a particular step can not have a second closest contiguous cluster. This is the case, for example, with those objects that, in the initial cluster formation process of aggregation to the starting nucleus L, fall "inside" the clusters: at the *COPASS Step 1* it can happen that they belong to just one cluster. The subsequent allocation-reallocation iterative procedure of the algorithm avoids objects being forced to stay in the starting cluster. However, in the case of geographical contiguity, some spatial units, for example, those on the border of the area to be segmented, can be neighbours to just one cluster. A control structure for such units has been considered in the algorithm.

The check before the transfer of objects in *Step 3a* is actually performed in two separate steps *3a.1* and *3a.2* to save calculation time while taking into account complex cluster shapes. In fact, in our experience in real and simulated applications, during the formation process clusters are in general more likely to grow into spherical or chain like structures. In the latter case problems can arise in the relocation process if the candidates for transfer represent the joining ring points of the chains. *Step 3a.1* on the contiguous points allow us to check this kind of situation. At the same time, owing to the complexities of the links between the units in an area, more complicated structures, such as one cluster encapsulated inside another, can develop (this was the case, for example, with one of the applications considered

Table 1. The **CKMNS** algorithm

Step 0 Initialisation

Step 0a Given L seeds aggregate their contiguous points i $(i \neq l; l = 1, \ldots, L; i = 1, \ldots, N)$. If object i is contiguous to more than one seed aggregate to the closest (in a Euclidean sense) one.

Step 0b Determine L centres to start the algorithm: compute the L initial centres to be the averages of the points added to the L seeds.

Step 1 The constrained optimal-assignment stage ($COPASS$): objects are progressively assigned to the L clusters. For each object $i (i = 1, ..., N)$, find its closest and, if it exists, the second closest contiguous cluster centre, IC1(i) and IC2(i) respectively. Assign point i to cluster IC1(i). At each assignment cluster centres, IC1(i) and IC2(i) are updated. Repeat until all objects are allocated.

Step 2 All clusters belong to the live set.

Step 3 Consider each of those points for which there is a second closest contiguous cluster in turn. Let point i be in cluster $l1$. If cluster $l1$ is updated in the last constrained quick-transfer ($CQTRAN$) step, then it belongs to the live set through this stage. Otherwise, at each step, it is not in the live set if it has not been updated in the last constrained optimal transfer step.

Step 3a Before considering object i for transfer check if it can moved from its starting cluster $l1$.

Step 3a.1 Consider the subset of cluster $l1$ of the objects contiguous to the candidate moving point i. Check if its relocation causes this subset to be divided into two or more contiguous subsets. If this is not the case, point i can be considered for transfer at this stage and go to *Step 3b*. Otherwise do *Step 3a.2*.

Step 3a.2 This step is the same as *Step 3a.1* except that the check is extended to consider all the objects belonging to $l1$. If at this step transfer of point i does not cause cluster $l1$ to be divided into more than one contiguous clusters it can be considered for the subsequent relocation procedure.

Step 3b The constrained optimal-transfer stage ($COPTRA$): if $l1$ is in the live set, do *Step 3b.1*; otherwise, do *Step 3b.2*.

Step 3b.1 Compute the minimum of quantity $R2 = N(l) \times d^2(i, l)/(N(l)+1)$, over all contiguous clusters $l (l \neq l1, l = 1, 2, \ldots, L)$. Let $l2$ be the contiguous cluster with the smallest $R2$. If this value is greater than or equal to $R1 = N(l1) \times d^2(i, l1)/(N(l1)-1)$ no reallocation is necessary and $l2$ is the new IC2(i)*. Otherwise, point i is allocated to cluster $l2$ and $l1$ is the new IC2(i). Update cluster centres to be the mean of the points assigned to them, if reallocation has taken place. The two clusters that are involved in the transfer of point i at this particular step are now in the live set.

Step 3b.2 The same as *Step 3b.1*, except that here the minimum $R2$ is computed only over clusters in the live set.

Step 4 If the live set is empty then stop. Otherwise go to *Step 5*.

Step 5 Constrained quick-transfer ($CQTRAN$): each point i for which $l1 =$ IC1(i) and $l2 =$ IC2(i) have changed in *Step 3* is considered in turn and the following values are computed:

$$R1 = N(l1) \times d^2(i, l1)/(N(l) - 1) \text{ and } R2 = N(l2) \times d^2(i, l2)/(N(l2) + 1)$$

If $R1$ is less than $R2$, point i remains in cluster $l1$. Otherwise, IC1(i) and IC2(i) are switched and the centers of clusters $l1$ and $l2$ are updated. The clusters involved in a transfer at this step are in the live set.

Step 6 If no transfer has taken place in the *Step 5* go to *Step 1* to optimal assign (after one pass thorough the data set) to the (new) cluster centers those objects for which in the last step has been IC2(i) = 0. Otherwise go to *Step 5*.

* As in Hartigan's **KMNS** algorithm, $R1$ is remembered and will remain the same for object i until cluster $l1$ is updated.

in the following section). In this instance, if a candidate for transfer in *Step 3a.1*, cannot be moved from its starting cluster because it causes the cluster to which it belongs to divide, in *Step 3a.2* we need to check if it is lying on an outer cluster of an encapsulated structure. In this case, in fact, its transfer will not cause the cluster to which it belongs to be divided into more than one contiguous clusters.

As in Hartigan's algorithm the constrained optimal-transfer *Step 3b* has been split into two separate steps *3b.1* and *3b.2* through the use of a "live set" of clusters. That is, if cluster l $(l = 1, \ldots, L)$ has been updated in the last constrained quick-transfer stage, then it belongs to the live set throughout this stage and step *3b.1* is considered. Instead, cluster l is not in the live set, if it has not been updated in the last constrained optimal transfer step; in this case step *3b.2* is considered.

Finally note that when in *Step 6* the algorithm goes back to *Step 1* the objects which have not been involved in the last relocation stage because they were contiguous to just one cluster, may change cluster and/or be contiguous to more than one contiguous cluster after this step. They may now enter the subsequent relocation stage.

The algorithm described has been written in SAS-IML language. It comprises three main modules, *COPASS*, *COPTRA* and *CQTRAN*, and two sub-modules *CAT* and *CER* to check if an object can be moved from the cluster it belongs to.

As is usual for all k-means clustering, **CKMNS** algorithm produces a clustering which can be locally optimal only. Subject to contiguity constraints, the within cluster sum of squares may not be decreased by transferring an object from one cluster to another; however, different partitions may have the same or smaller within cluster sum of squares. It is advisable in that case to use several initial partitions and to use those final partitions with minimal value of the objective function.

4. Case-study

POM (*Programma Operativo Multiregionale*) is an acronym for a series of policies by which the Italian government financially supports research projects investigating less developed areas of the peninsula where defining regional development programmes is a priority, especially with respect to the traditional agricultural sector. Since regional development is related to a series of spatially varying attributes and since aid policies can be more efficient if they are tailored for the specific needs of an area, the aim of the research project was to segment basic territorial units to establish regional sub areas, which were both homogeneous with respect to the selected attributes and, in addition, spatially connected, so that planning would be more efficient.

The research studied two large regions in the south of Italy, in particular Puglia and Calabria; the algorithm discussed in the previous section was applied to zoning their administrative units (basic spatial units), 256 and 409 respectively. Each administrative unit (*comune*) in the two regions is described by a descriptor vector comprising eighteen socio-economic indicators:[3]

[3] Sources of the data: *ISTAT: censimento generale della Popolazione e delle Abitazioni 1991; ISTAT: censimento generale dell'Industria e dei servizi 1991; ISTAT: Censimento Generale dell'Agricoltura*

earth surface for agriculture use per capita (SAU);
SAU for farms larger than 50 hectares per capita;
number of farms;
population density;
natural population turnover rate (SALDO);
population not of working age;
population enrolled in higher education;
working population;
population working in agriculture;
population working in industry;
population working in public services;
working women;
number of employees in firms;
electricity consumption per capita,
number of powerful cars (greater than 2000 c.c.) per capita;
number of houses built after 1981 per capita;
per capita income;
IRPEF tax (local rates).

In order to eliminate the different average dimension and the different kinds of measurements all the variables were divided by their respective mean, while the two geographical contiguity matrices were defined on the assumption that administrative units sharing a common boundary were contiguous.

Table 2 presents the results for the Puglia region of the criterion function $F(\mathbf{M}, \mathbf{U})$[4] for ordinary **KMNS** and constrained **CKMNS** algorithms in 10 groups for 20 initial random configurations and for three configurations obtained by considering the ten administrative units furthest apart with respect to the indicators SAU, SALDO and IRPEF. The values of the criterion function for the initial configurations (after the *OPASS* step in the **CKMNS** case) are termed F_0 (**M**, **U**), while the values for corresponding (local) minimum are denoted $F_{10}(\mathbf{M}, \mathbf{U})$.

Note how the decrease of the criterion is somewhat larger in the case of the ordinary clustering **KMNS** than in the constrained case. This could be due to the fact that in the constrained case the initial step of assignment of the objects to the L initial centres is optimised; in fact in the *COPASS* step each time an object is assigned to a cluster, the cluster centroid is updated. The evident divergence between the values of the criterion in the two cases proves the effectiveness of the **CKMNS** algorithm in discovering spatially connected groups. In fact, given the same initial configurations, in the constrained situation objects are not free to associate on the basis of the minimum of their distances from the cluster centroids only, but they must also satisfy various contiguity constraints. On the other hand, from Table 2 we can see that the constrained situation exhibits more variability in the values

1991; ANCITEL-ENEL 1991; ANCITEL-ACI 1991; ANCITEL-Ministero delle Finanze 1991. The selection of the 18 attributes for the analysis was highly influenced by the data availability as well as by the knowledge of the socio-economic structure of the two regions.

[4] The criterion function has been normalized to express the goodness of fit of the final solution. For variables column-centered, the normalized version of the criterion function in Section 2 is $F^*(\mathbf{M}, \mathbf{U}) = F(\mathbf{M}, \mathbf{U}) / \| \mathbf{X} \|^2$ (De Soete and Carrol, 1994).

Table 2. Criterion function for 20 different initial configurations (Puglia)

Initial configurations	Ordinary **KMNS**			Constrained **CKMNS**		
	F_0 (**M**, **U**)	F_{10} (**M**, **U**)	F_{21} (**M**, **U**)	F_0 (**M**, **U**)	F_{10} (**M**, **U**)	F_{21} (**M**, **U**)
1	0,601	0,370	0,229	0,781	0,594	0,448
2	0,577	0,391	0,234	0,775	0,629	0,406
3	0,553	0,374	0,230	0,772	0,592	0,430
4	0,620	0,367	0,227	0,770	0,571	0,413
5	0,547	0,378	0,233	0,753	0,622	0,478
6	0,493	0,392	0,229	0,612	0,572	0,422
7	0,529	0,358	0,234	0,648	0,619	0,426
8	0,614	0,378	0,229	0,764	0,635	0,425
9	0,638	0,361	0,231	0,806	0,630	0,441
10	0,619	0,361	0,233	0,798	0,631	0,415
11	0,584	0,371	0,230	0,622	0,564	0,414
12	0,609	0,363	0,234	0,788	0,646	0,420
13	0,580	0,345	0,230	0,646	0,599	0,416
14	0,629	0,361	0,224	0,795	0,639	0.491
15	0,551	0,355	0,229	0,783	0,649	0,423
16	0,536	0,361	0,232	0,749	0,573	0,429
17	0,531	0,365	0,251	0,622	0,564	0,419
18	0,574	0,357	0,231	0,793	0,642	0,425
19	0,651	0,371	0,241	0,798	0,634	0,422
20	0,549	0,379	0,230	0,646	0,599	0,416
SAU	0,586	0,357	0,232	0,769	0,627	0,405
SALDO	0,618	0,345	0,237	0,770	0,563	0,421
IRPEF	0,571	0,403	0,229	0,652	0,578	0,498

of F_{10}(**M**, **U**) compared to the ordinary case. One reason for this is that, as one would expect, the **CKMNS** algorithm is likely to be more dependent on the starting configuration than the ordinary algorithm. Nevertheless, there is agreement among the values of the criterion function when the number of clusters increases, as can be observed from Table 2 where we also present values F_{21} (**M**, **U**) of the criterion function for an "optimal" clustering in 21 groups obtained starting from the 20 corresponding initial configurations in ten clusters. Figure 2(a) plots the minimum value of criterion function in Table 2 for any given number of clusters, from ten to thirty.

In fact to determine the unknown number of clusters, we increased their number in a sequential way by picking the group with the greatest error (variance within) at the end of the procedure (**KMNS** and **CKMNS**) for a given starting number of clusters L ($L = 10$ for these applications) and by adding the object with the largest distance from the corresponding centroid as the new seed for a new run of

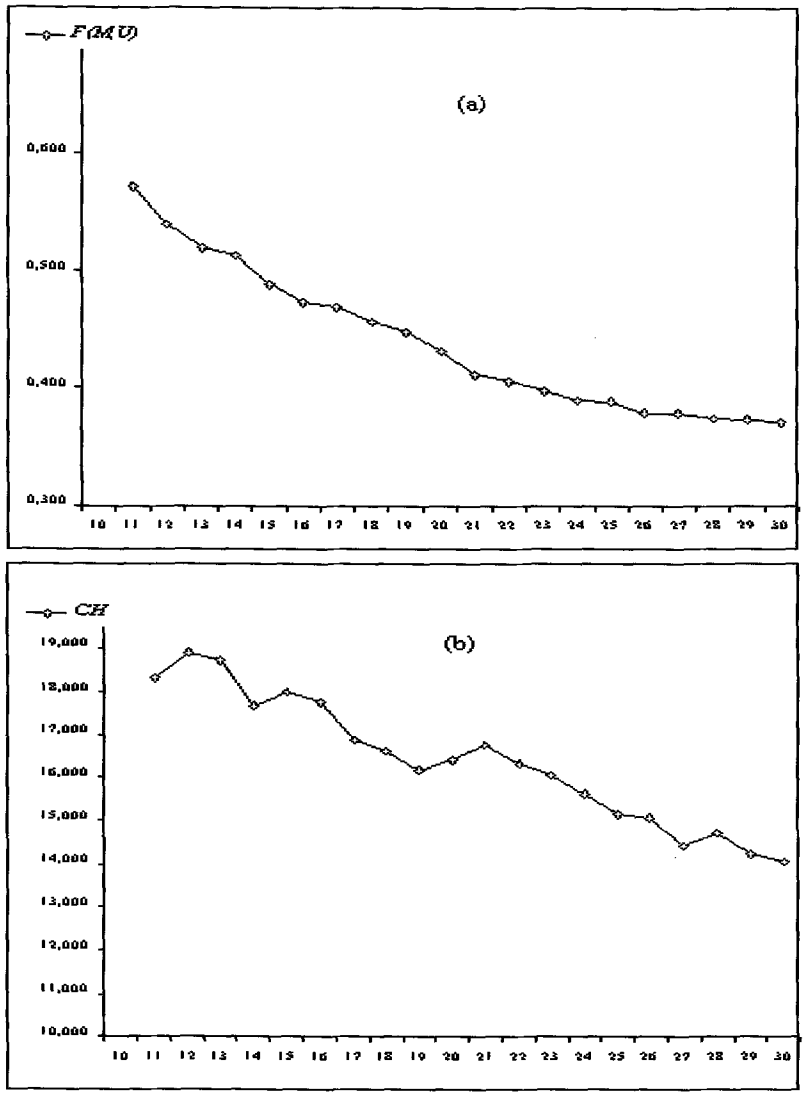


Fig. 2. Criterion function (**a**) and *CH* index (**b**) against the number of clusters (Puglia)

the algorithm with a number of clusters equal to $L + 1$.[5] When a selected seed was already included in previous stages, it was (temporary) deleted and the variances within clusters were recomputed; again the object with the largest distance from the centroid is considered as a new seed for a new run of the algorithm. In the applications considered in this paper we started from ten clusters and stopped to "add seeds" according to the outlined above procedure when we reached thirty

[5] There are, in fact, several suggestions (Ball and Hall, 1967; MacQueen, 1967) which warn against rigid adherence to the assumed number of clusters, but to split up into two clusters the group with the greatest value of the sum of squares after each "step".

clusters. For economic reasons, the two extremes of ten and thirty clusters were chosen as the smallest and the largest number of classes to segment the two regions.

One significant problem that need to be addressed was the determination of the proper number of clusters in the final solution. Apart from the relative value of the objective function, there is no strong evidence about this number (i.e. an elbow point) in the "scree plot" in Fig. 2a owing to the smooth monotone tendency of the curve.

Numerous strategies have been proposed for selecting the number of clusters in the "standard" clustering procedures (and for hierarchical methods in particular). We refer to Vicari (1990) and Gordon (1996*a*) for a review. However, few constrained classification studies have addressed this problem. Amongst the few formal tests are some stopping rules proposed in the case of constrained agglomerative methods (references are given by Gordon (1996)). Anyway, if no suitable "rule" for the situation at hand exists the number of clusters can be determined solely by the interpretability of the partitions (i.e. Gordon and Vichi, 2001).

In our case we investigated the behaviour of the index proposed by Calinsky and Harabasz (1974)

$$CH = [B/(L-1)]/[W/(N-L)]$$

where W and B denote the total within-cluster sum of squared distance (about the centroids), and the total between-cluster sum of squared distances respectively. In this index the same quantities as in the objective function and also the weights which take into account the different number of clusters (L) are involved. Thus it can provide guidance in the selection of the number of clusters. In fact, a value of CH increasing monotonically with L suggests no cluster structure, whereas CH decreasing monotonically with L suggests a hierarchical structure. However CH rising to a maximum at L suggests the presence of L clusters (see Milligan and Cooper (1985) regarding the use of CH in estimating the number of clusters).

Figure 2b plots the values of the index for the corresponding partitions in Fig. 2a. The behaviour of the index suggests twelve as a suitable number of groups because in this partition it gets its maximum value.

However, the economic need to have a more detailed segmentation for this region together with information from previous economic research led us to consider the relative maximum points in Fig. 2b as candidate partitions as well. The relative value of objective function and the interpretability of the results indicated the classification of twenty-one groups as the proper segmentation of the Puglia region. A map of this classification is shown in Fig. 4.

The analysis of the composition of the clusters on the other hand confirmed the suitability of this choice. Actually, the classification in twenty-one groups is, roughly speaking, a "segmentation" of that partition in twelve clusters. In fact, the small groups are almost the same in size and composition for both partitions. For example, group 6 we will illustrate later, "looses" only two units passing from 12 to 21 clusters, while the group surrounding the city of Taranto (group 12 on the map in Fig. 4), composed of nine administrative units, stays exactly the same. Further, in the classification of 21 clusters some groups are clearly originated from the division of some big groups in the classification of 12 clusters; for example, the

Table 3. Criterion function for 10 initial random configurations (Calabria)

Initial configurations	*Ordinary KMNS*			*Constrained CKMNS*		
	F_0 (M, U)	F_{18} (M, U)	F_{30} (M, U)	F_0 (M, U)	F_{18} (M, U)	F_{30} (M, U)
1	0,626	0,336	0,249	0,879	0,783	0,697
2	0,499	0,331	0,259	0,852	0,730	0,645
3	0,739	0,330	0,259	0,871	0,720	0,642
4	0,667	0,348	0,253	0,877	0,774	0,686
5	0,597	0,354	0,255	0,857	0,695	0,626
6	0,651	0,324	0,258	0,889	0,780	0,660
7	0,761	0,336	0,258	0,844	0,717	0,657
8	0,657	0,330	0,258	0,880	0,786	0,613
9	0,582	0,338	0,254	0,897	0,734	0,621
10	0,612	0,338	0,258	0,803	0,723	0,643

very large cluster comprising the city of Bari and its surrounding district gave rise to approximately three clusters, one of which is a singleton in the classification of 21 clusters (the three clusters are the groups 2, 10 and 17 on the map).

The segmentation in twenty-one groups has allowed for a detailed territorial analysis which enabled us to identify sub-areas within the region with similar developmental contexts, which are also geographically connected. The latter aspect, not necessarily emerging in an ordinary clustering, together with the knowledge of the group profiles, can facilitate aid development policies because they can be directed more effectively towards the specific socio-economic characteristics of the sub-areas. For example, cluster 6 in Fig. 4 emerged from the zoning comprising 11 administrative units whose mountainous terrain forms the Gargano National Park. Also given the orographical features, this area shows (as emerged from the cluster profile interpretation) very little tendency towards industrial and agricultural activities but a very positive attitude towards tourism. Another compact sub-area emerging from the analysis is group 5 comprising 19 administrative units covering a flat area called Tavoliere di Puglia. Also in this case, the spatial characteristics of the area influence economic attitudes, which are mainly connected to the agriculture. Whereas a group which shows a very positive attitude towards the industrial activity is the city of Taranto with its surrounding district.

Finally, observe that the partition selected as the optimal one is the partition in 21 clusters from the starting configuration in ten clusters selected on the basis of the indicator SAU.

Table 3 presents the results of the criterion function for ordinary and constrained algorithms for 10 random initial configurations and a number of clusters raising from ten to thirty for the Calabria region. Figure 3(a) and (b) plot the minimum value of criterion function in Table 3 for any given number of clusters and the corresponding values of the index CH respectively. Also in this case, there was not any clear indication about the number of groups in the "scree plot", while the plot

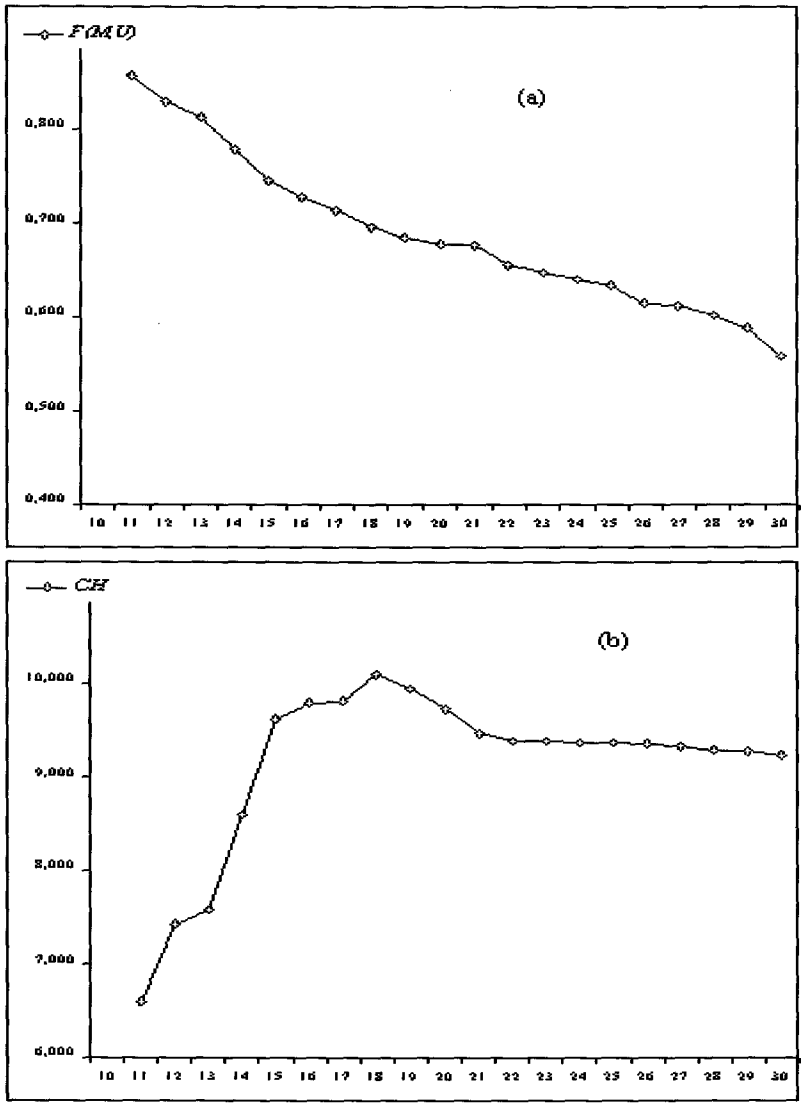


Fig. 3. Criterion function (**a**) and *CH* index (**b**) against the number of clusters (Calabria)

of the index *CH* clearly suggests a suitable candidate number. Figure 5 shows the "optimal" segmentation in eighteen sub-areas chosen as indicated by the maximum value of the index. Here too from the interpretation of this classification a strong link emerged between the segmentation and the territorial characteristics of the region. For example, from the resulting segmentation a cluster emerged of 32 administrative units covering approximately the flat area called Piana of Sibari (group 2 on the map) and for which one dominant common characteristic is the "strong presence" of technologically advanced agricultural activities compared with the rest of the region. Another compact cluster is group 18 comprising 40 administrative units

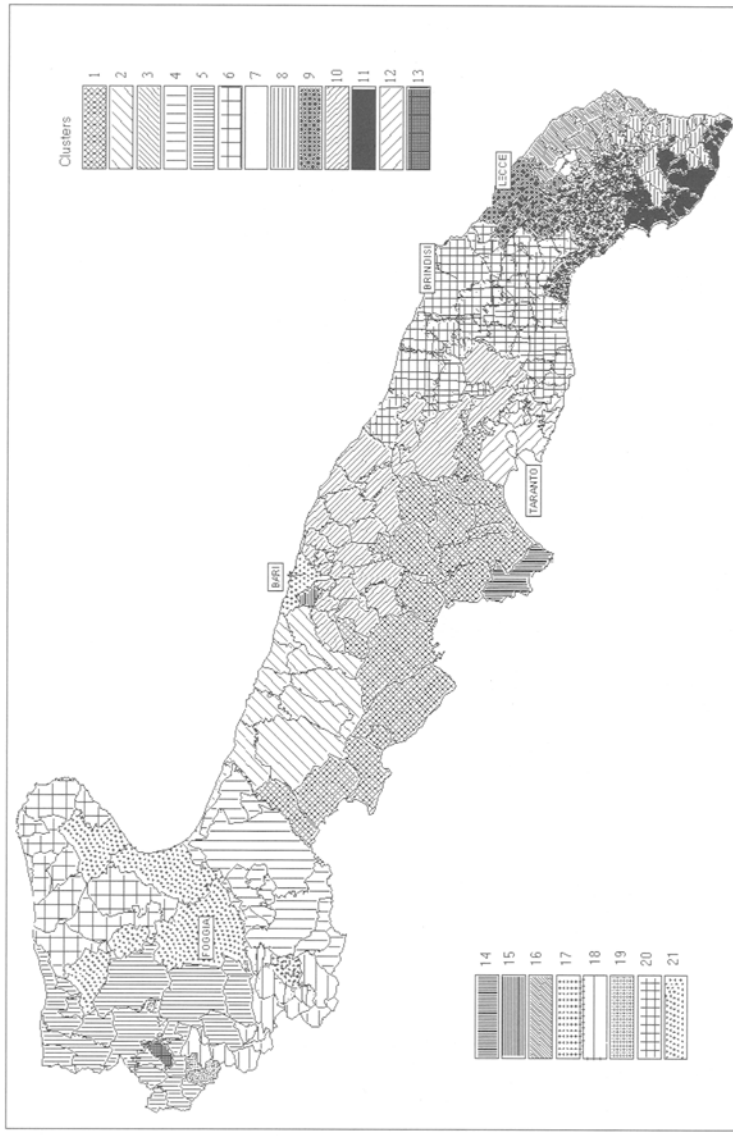


Fig. 4. Segmentation of Puglia in 21 contiguous subareas

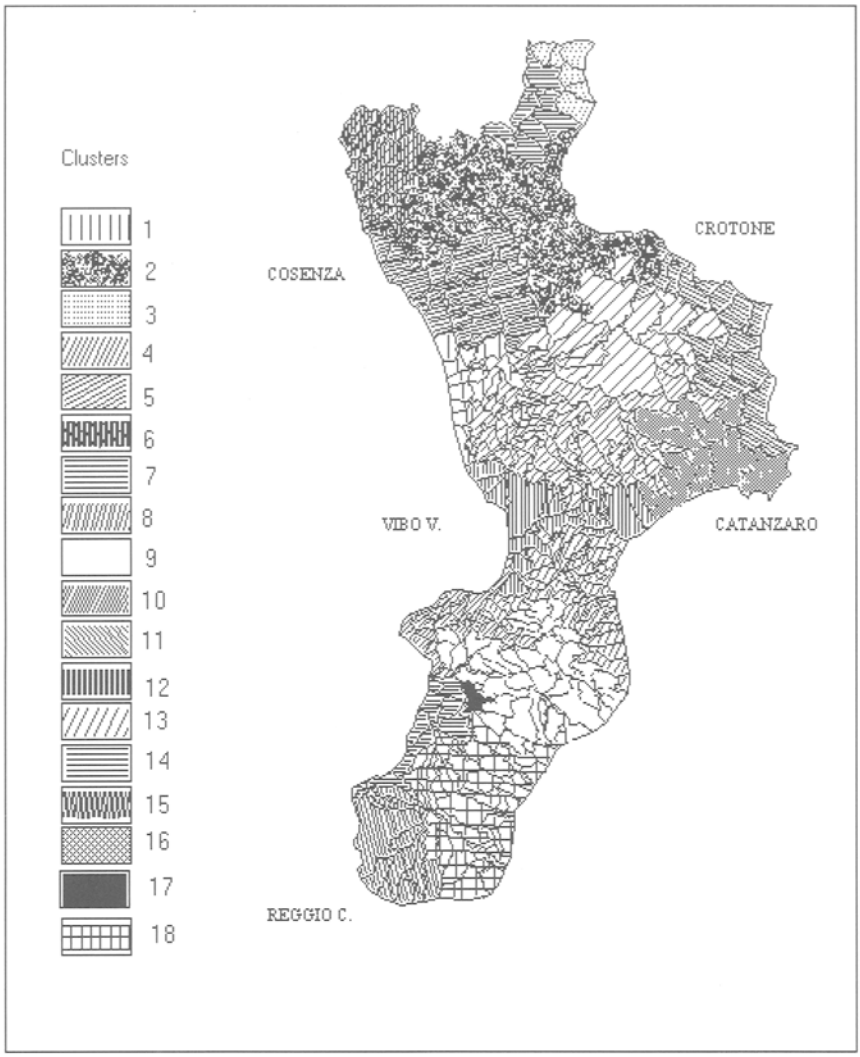


Fig. 5. Segmentation of Calabria in 18 contiguous subareas

most of which belonging to the mountainous area of the Aspromonte Park. Once again, the spatial characteristics of the area influence economic attitudes, which are mainly connected to traditional agricultural activity.

For a deeper analysis of the composition of the classes and a detailed interpretation and discussion of the sub-area profiles of the segmentations illustrated in this paper we refer to Anania et al. (2001).

Acknowledgements. The author wishes to thank G. Anania and S. Ingrassia for their helpful comments on an earlier draft of this paper and the editor and referees who helped to improve it. This research was

financially supported by "Misura 2 (Innovazioni tecnologiche e trasferimento dei risultati della ricerca) of POM" research project funds.

References

Anania G, Cersosimo D, Costanzo GD (2001) Le Calabrie contemporanee. Un'analisi delle caratteristiche degli ambiti economico produttivi sub-regionali. In: Scelte pubbliche, strategie private e sviluppo economico in Calabria. Conoscere per Decidere, Rubbettino, Soveria Mannelli, 333–380

Ball GH, Hall DJ (1967) A clustering technique for summarizing multivariate data. Behavioural Science **12**, 153–155

Batagelj V (1984) Agglomerative methods in clustering with constraints. Preprint Series Dept. Math. Univ. Ljublijana **22** (102), 5–19

Caliñski T, Harabasz J (1974) A dendrite method for cluster analysis. Communications in Statistics **3**, 1–27

Christofides N (1975) Graph Theory. Academic Press, London.

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

De Soete G, DeSarbo WS, Furnas GW, Carrol JD (1984) The estimation of ultrametric and path trees from rectangular proximity data. Psychometrika **49**, 289–310

De Soete G, Carrol JD (1994) K-means clustering in a low-dimensional Euclidean space. In: Diday E et al. (eds.) New approaches in classification and data analysis, pp. 212–219. Springer, Berlin Heidelberg New York

DeSarbo WS, Mahajan V (1984) Constrained classification: the use of a priori information in cluster analysis. Psychometrika **49**, 187–215

Ferligoj A, Batagelj V (1982) Clustering with relational constraint. Psychometrika **47**, 413–426

Ferligoj A, Batagelj V (1983) Some types of clustering with relational constraint. Psychometrika **48**, 541–522

Ferligoj A, Batagelj V (1992) Direct multicriteria clustering algorithms. Journal of Classification **9** (1), 43–61

Ferligoj A, Batagelj V (1998) Constrained clustering problems. In: Proceedings of IFCS '98, Rome, 541–522

Ferligoj A, Batagelj V (2000). Clustering relational data. In: Gaul W, Opitz O., Schader M (eds.) Data analysis. Springer, Berlin heidelberg New York, 3–15

Gordon AD (1973) Classifications in the presence of constraints. Biometrics 29, 821–827

Gordon AD (1980) Methods of constrained classification. In: Tomassone R (ed.) Analyse de données et informatique.(INRIA, Le Chesnay), 149–160

Gordon AD (1999) Classification. Chapmann & Hall, London

Gordon AD (1987) Parsimonious trees. Journal of Classification **4**, 85–101

Gordon AD (1996) A survey of constrained classification. Computational Statistics & Data Analysis **21**, 17–29

Gordon AD (1996) (a). How many clusters? An Investigation of five procedures for detecting nested cluster structure. In: Hayashi C et al. (eds.) Data science, classification, and related methods. Berlin Heidelberg New York, Springer, 109–116

Gordon AD, Vichi M (2001) Fuzzy partition models for fitting a set of partitions. Psychometrika **66**, 229–248

Harary F (1969) Graph theory. Addison-Wesley, Reading, MA

Hartigan JA (1975) Clustering algorithms. Wiley, New York

Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. Applied Statistics **28** (1), 100–108

Hubert LJ (1974) Some applications of graph theory to clustering. Psychometrika **39** (3), 283–308

Lebart L (1978) Programme d'agrégation avec contraintes. Le Cahiers de l'Analyse des Données **3**, 275–287

Lechevallier Y (1980) Classification sous contraintes. In: Diday E et al. (eds.) Optimisation en classification automatique. INRIA, Paris, 677–696

Lefkovitch LP (1980) Conditional clustering. Biometrics **36**, 43–58

Legendre P (1987) Constrained clustering. In: Legendre P et al. (eds.) Developments in numerical ecology. Springer, Berlin Heidelberg New York

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: LeCam LM et al. (eds.) Proceedings ot the Fifth Berkeley Symposium on Mathematic, Statistics and Probability, vol. 1, Statistics, University of California Press, Berkeley, CA, 281–298

Maravalle M, Simeone B, Naldini, R (1997). Clustering on trees. Computational Statistics & Data Analysis **24**, 217–234

Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**, 159–179

Mills G (1967) The determination of local government boundaries. Operational Research Quarterly **18**, 243–255

Monestiez P (1977) Méthode de classification automatique sous contraintes spatiales. Statistique et Analyse des Données **3**, 75–84

Murtagh F (1985) A survey of algorithms for contiguity-constrained clustering and related problems. Computer Journal **28**, 82–88

Openshaw S (1977) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. Transaction of the Institute of British Geographers **52**, 247–258

Seber GAF (1984) Multivariate observations. Wiley, New York

Späth H (1980) Cluster analysis algorithms. Ellis Horwood, Chichester

Taylor PJ (1973) Some implications of the spatial organizations of elections. Transaction of the Institute of British Geographers **60**, 121–136

Upton G, Fingleton B (1985) Spatial data analysis by example, vol. 1. Wiley, New York

Vicari D (1990) Indici per la scelta del numero dei gruppi. Metron **49**, 473–492

Webster R (1977) Quantitative and numerical methods in soil classification and survey. Clarendon Press, Oxford New York

Wilson RJ (1996) Introduction to graph theory. Addison Wesley Longman, England

Zani S (1993) Classificazione di unità territoriali e spaziali. In: Zani S (ed.) Metodi statistici per le analisi territoriali. Franco Angeli, Milano, 93–121