The spatial-temporary approach in problems of clusterization

Lyudmyla Sarycheva

Dept. of Geoinformation systems, National Mining University, 49005, K. Marx .av., 19, Dnipropetrovsk, Ukraine

sarychevаl@nmu.org.ua

Abstract. *The problem of cluster analysis of the spatial-temporary data is considered. The new algorithm of GeoTime-clusterization takes into account a temporary neighbourhood of objects features (on the basis of the inductive approach) and absolute concepts of objects location (on the basis of the geoinformation approach). Using the actual data of ecological and socio-economic monitoring of the Europe states the experimental comparison of GeoTime-clusterization with known algorithms is carried out by three clusterization quality criteria and gave the good results.*

Keywords

GeoTime-clusterization, clusterization quality criteria, spatial-temporary data, socio-economic monitoring, proximity measures

1 Introduction

The class of methods and algorithms of cluster analysis is extensive and includes algorithms of hierarchical clusterization, k-means, ISODATA, ОКК, etc. [1 - 4]. In each particular case the methods that take into account features of input data – volume of sampling, number of features, prior information, etc are applied. In many practical problems the input data have spatial-temporary character. For example, in a problem of cluster analysis of ecological and socio- economic (ESE) parameters of regions monitoring the tables containing geographical coordinates and values of several indices for particular number of months (years) are usually used.

In known algorithms the geographical coordinates of objects are used, if used at all, equivalently with other temporary properties, therefore from the informative point of view the interpretation of clusterization results is inconvenient and ambiguous.

The actual problem is clusterization algorithms creation that operate with coordinates of objects and features reflecting temporary properties, not integrating them equivalently in one object-feature table. The expediency of such algorithms is caused by taking into account objects locations topological features and possibility of meaningful interpretation of clusterization results.

The purpose of the article: presentation of the new algorithm of GeoTime-clusterization which takes into account a geographical neighbourhood of objects (on the basis of the geoinformation approach, i.e. consideration of topological concepts of locations) and temporal neighbourhood of their indices (on the basis of the inductive approach).

2 Problem definition

There are n terrain regions as the objects of analysis. They are characterized by the location (defined by two or three geographical coordinates) and indices, measured in the series of instants. It is required to conduct the geographical zoning of terrain by the set of monitoring indices [6].

Mathematical problem definition. Let $x_{ij}(t_s)$ be measurements of features describing given set of objects $Z=\{Z_1, Z_2,\ldots, Z_n\}$ in an instant t_s ($i=1, 2,\ldots, n$ is number of observations, n is a quantity of observations, $j=1, 2, \ldots, m$ is a number of feature, m is a feature index, $s=1, 2, \ldots, L$ is a number of instants); $Q(Z_1), Q(Z_2),\ldots, Q(Z_n)$ are geographical coordinates of objects.
The input data represent a block matrix

$$(\boldsymbol{Q}\mathbb{M}\boldsymbol{X}(t_1)\mathbb{M}\boldsymbol{X}(t_2)\mathbb{M}\ldots\mathbb{M}\boldsymbol{X}(t_L)),$$

where $\boldsymbol{Q}$ is a matrix of objects geographical coordinates with the size $n\times 2$ (or $n\times 3$), $\boldsymbol{X}(t_s)=(\boldsymbol{X}^1(t_s)\mathbb{M}\boldsymbol{X}^2(t_s)\mathbb{M}\ldots\mathbb{M}\boldsymbol{X}^m(t_s))$ is a matrix "object - feature" type, $\boldsymbol{X}^j(t_s)=(x_{1j}(t_s),x_{2j}(t_s),\ldots,x_{nj}(t_s))^T$ is a column vector of j-th feature values for n objects, $\boldsymbol{X}_i(t_s)=(x_{i1}(t_s),x_{i2}(t_s),\ldots,x_{im}(t_s))$ is a row vector of m indices values of i-th object, $j=1,2,\ldots,m;\ i=1,2,\ldots,n;\ s=1,2,\ldots, L$.

The clusterization $K=\{K_1, K_2,\ldots,K_k\}$, $1\le k\le n$ of ensemble Z is the set of nonempty, mutually nonintersected subsets (clusters) K_q, $q=1, 2,.., k$, of the ensemble Z, which combination coincides with Z:

$$K_1 \dot{E} K_2 \dot{E}\ldots\dot{E} K_k = Z, \quad K_i \dot{C} K_j = \text{Æ}, \text{ if } i\neq j,\ i,j=1,2,..,k,\ K_q \neq \text{Æ},\ q=1,2,..,k.$$

The clusterization $K^* \in \Phi$ is the best, if

$$K^* = \arg\max_{K\subseteq\Phi} J(K) \quad (\text{or } K^* = \arg\min_{K\subseteq\Phi} J(K)),$$

where Φ is an ensemble of all permissible splittings (clusterizations) of given ensemble Z; $J(K)$ is a criterion of clusterization quality.

It is required to decide the problem of searching the best clusterization K^*.

To solve the clusterization problem it is necessary:
a) to give definition of a cluster - to specify features, general for all objects of a separate cluster (measure of resemblance between objects);
b) to set a way of clusters creation (sorting, re-grouping, affiliation, splitting, addition, searching);
c) to specify J criteria to clusterization quality estimation ($J(K)=J(a(K), b(K), g(K))$; $a(K)$, $b(K)$, $g(K)$ are criterions of accuracy, neighborhood, consistency accordingly);
d) to organize J criterion motion to the maximum (minimum) (so the number of real clusters is also determined).

3 Clusterization GeoTime algorithm of the spatial-temporary data

Offered method and conforming clusterization of the spatially-temporal data algorithm GeoTime are based on the following suppositions.

The supposition 1 (taking into account a temporal neighbourhood). The objects close on their features (included into one cluster) at an instant t can be close also at an instant $(t+1)$ (fig. 1). This supposition is used for computing of clusterization $K(t, t+1)$, that optimizes criterion $g(K)$ of consistency (for example, $g(K)\to min$, where $g(K)$ is a measure of similarity between clusterizations $K(t_s)$ for separate temporary shears $t=t_s$, $s=1, 2,\ldots, L$, and all data sets).

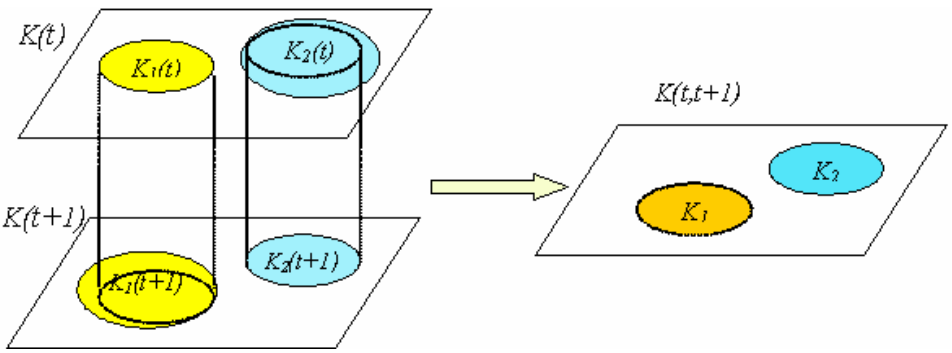


Fig. 1. Succession of clusterizations $K(t)$ and $K(t+1)$ in instants t and $(t+1)$

The supposition 2 (taking into account a geographical adjacency). The objects adjoining in geographical space can organize the homogeneous groups by their features, appropriate to connected areas (complete territorial bands). This supposition is used for a determination of the clusterization that optimizes the criterion $b(K)$ (for example, $b(K)\to max$; $b(K)$ is the sum of elements in a adjacency matrix of separate clusters objects that are considered in the space of geographical coordinates).

The supposition 3 (taking into account the proximity in the space of indications). The objects adjoining in the space of indications X can organize the features homogeneous groups – clusters. This spread for the majority of cluster analysis methods supposition is used for a determination of the clusterization optimizing criteria $a(K)$ (for example, $a(K)\to min$; $a(K)$ is the sum of intracluster distances, where K is a clusterization of ensemble Z without taking into account $\boldsymbol{Q}$).

Suppositions 1-3 cause that the offered algorithm implements a method of separation of the given objects ensemble on the groups consisting of interconnected, homogeneous objects. The type of an interior homogeneity is meant here: the aggregation is considered as homogeneous if regularity of objects forming leads to the fact that objects are close to each other in the fixed space of features.

Let's consider succession problem of the clusterizations appropriate to two sequential instants more detailed.

Let $K(t_s)=\{K_1(t_s), K_2(t_s),\ldots,K_k(t_s)\}$ is an object clusterization in an instant $t=t_s$, $s=1,2,\ldots,L$, where k is a number of clusters, $1<k<n$ ($k=1$, $k=n$ are not considered here):

$$\bigcup_{i=1}^{k} K_i(t_s) = X(t_s),\ K_i(t_s)\neq\varnothing,\ K_i(t_s)\cap K_j(t_s)=\varnothing,\ i\neq j;\ i,j=1,2,\ldots,k.$$

Clusterization $K(t_s)$ is obtained on the basis of the indications describing an instant $t=t_s$, i.e. with using of a matrix

$$\boldsymbol{X}(t_s)=(\boldsymbol{X}_1(t_s)\,|\,\boldsymbol{X}_2(t_s)\,|\ldots|\,\boldsymbol{X}_m(t_s))=\begin{pmatrix} x_{11}(t_s) & \ldots & x_{1m}(t_s) \\ \ldots & \ldots & \ldots \\ x_{n1}(t_s) & \ldots & x_{nm}(t_s) \end{pmatrix}.$$

Clusterization $K(t_s)$ puts into correspondence to each object Z_i (identified with a point $((x_{i1}(t_s), x_{i2}(t_s),\ldots, x_{im}(t_s))$, $i=1,2,\ldots,n$, of Euclidean space R^m) the cluster number to which it belongs in an instant t_s.

By results of clusterizations $K(t_s)$ and $K(t_{s+1})$ matrix $\boldsymbol{A}(t_s, t_{s+1})=(a_{ij}(t_s, t_{s+1}))$, $i, j=1, 2,\ldots, k$, is being constructed, where the element $a_{ij}(t_s, t_{s+1})$ defines number of the objects simultaneously entering into cluster $K_i(t_s)$ and cluster $K_j(t_{s+1})$:

$K(t_s)$	$K(t_{s+1})$					
	$K_1(t_{s+1})$	$K_2(t_{s+1})$	…	$K_j(t_{s+1})$	…	$K_k(t_{s+1})$
$K_1(t_s)$	a_{11}	a_{12}	…	a_{1j}	…	a_{1k}
…	…	…	…	…	…	…
$K_i(t_s)$	a_{i1}	a_{i2}	…	a_{ij}	…	a_{ik}
…	…	…	…	…	…	…
$K_k(t_s)$	a_{k1}	a_{k2}	…	a_{kj}	…	a_{kk}

Elements $a_{ij}=a_{ij}(t_s, t_{s+1})$ of matrix $\boldsymbol{A}(t_s, t_{s+1})$ are calculated as follows:

$$a_{ij}=\sum_{e=1}^{n} F_{ij}(X_e),\quad i,j=1,2,\ldots,k, \tag{1}$$

$$F_{ij}(X_e)=\begin{cases} 1, \text{ if } (X_e\in K_i(t_s))\,\&\,(X_e\in K_j(t_{s+1}))=true, \\ 0-otherwise \end{cases}$$

The elements sum of matrix $\boldsymbol{A}(t_s, t_{s+1})$ is equal to number of clustered objects:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}(t_s,t_{s+1})=n,\quad \forall s=1,2,\ldots,L-1.$$

The number of cluster is no more than a mark, i.e. it is possible to interchange indexing of clusters, having maintained a composition of elements entering into them. Therefore for definition of $K(t_{s+1})$

clusterization succession in relation to $K(t_s)$ permutation of columns in matrix $A(t_s, t_{s+1})$is made so that the maximal element of i-th of string can get on the principal diagonal:

$$K_i(t_{s+1}) \leftrightarrow \max_i a_{ii}(t_s, t_{s+1}).$$

Lets mark $P(t_s, t_{s+1}) = \frac{1}{n} A(t_s, t_{s+1}) = (p_{ij}.(t_s, t_{s+1})),\ p_{ij}(t_s, t_{s+1}) = a_{ij}(t_s, t_{s+1})/n,\ i, j = 1,2,\ldots,k.$

The elements sum of matrix $P(t_s, t_{s+1})$ is equal to one:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(t_s, t_{s+1}) = 1, \quad \forall s = 1,2,\ldots, L-1.$$

It is possible to consider an element $p_{ij}(t_s, t_{s+1}))$ as probability of the object having belonged to cluster $K_i(t_s)$ will be into cluster $K_j(t_{s+1})$.

Note that in the matrices $\{A(t_s, t_{s+1})\}$, $s=1,2,\ldots, L-1$, having built in such way following rations are fulfilled:

$$\forall\, i \in \{1,2,\ldots,k\} \quad a_{ii}(t_s, t_{s+1}) \neq 0,\ a_{ii}(t_{s-1}, t_s) \geq a_{ii}(t_s, t_{s+1}),\ s=2,\ldots,L-1, \tag{2}$$

$$\{ j_1^L, j_2^L, \ldots, j_{n_i(L)}^L \} \subset \{ j_1^{L-1}, j_2^{L-1}, \ldots, j_{n_i(L-1)}^{L-1} \} \subset \ldots \subset \{ j_1^1, j_2^1, \ldots, j_{n_i(1)}^1 \}, \tag{3}$$

where $j_1^s, j_2^s, \ldots, j_{n_i(s)}^s$ are numbers of the objects which have entered into cluster $K_i(t_s)$; $n_i(s) = |K_i(t_s)|$ is the number of objects in $K_i(t_s)$, $\sum_{i=1}^{k} n_i(s) = n$, $s=1,2,\ldots,L$.

The ration (3) shows that for any $i \in \{1,2,\ldots,k\}$ sequences of objects numbers $j_1^s, j_2^s, \ldots, j_{n_i(s)}^s$ of clusters $K_i(t_s)$ $s=1,2,\ldots,L$, are enclosed.

The theorem. For any $i \in \{1,2,\ldots,k\}$ there is at least one point that belongs to all clusters $K_i(t_s)$, $s=1,2,\ldots,L$, simultaneously, i.e.:

$$\{ j_1^L, j_2^L, \ldots, j_{n_i(L)}^L \} \cap \{ j_1^{L-1}, j_2^{L-1}, \ldots, j_{n_i(L-1)}^{L-1} \} \cap \ldots \cap \{ j_1^1, j_2^1, \ldots, j_{n_i(1)}^1 \} \neq \varnothing.$$

The proof goes by the contradiction method.
From the theorem it follows there are k points $X_{i_1}, X_{i2}, \ldots, X_{i_k}$:

$X_{i_1} \in K_1(t_1) \cap K_1(t_2) \cap \ldots \cap K_1(t_L)$, $X_{i_2} \in K_2(t_1) \cap K_2(t_2) \cap \ldots \cap K_2(t_L), \ldots, X_{i_k} \in K_k(t_1) \cap K_k(t_2) \cap \ldots \cap K_k(t_L)$.
While searching of the best clusterization in GeoTime method such points are taken for initial centers (kernels) of clusters (identifying object Z_i with point $X_i = (x_{i1}(t_1), x_{i2}(t_1), \ldots, x_{im}(t_1), x_{i1}(t_2), x_{i2}(t_2), \ldots, x_{im}(t_2), \ldots, x_{i1}(t_L), x_{i2}(t_L), \ldots, x_{im}(t_L)) \in R^p$, $p = mL$).

Proximity measures between two objects, between object and cluster, between two clusters, applied in GeoTime-clusterization, and presented in tab. 1. To estimate proximity between two various clusterizations K and Q of an output set of objects the proximity measure is used

$$d(K,Q) = \frac{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right) - \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} |K_i \cap Q_j|^2}{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right)}, \tag{4}$$

where k_1, k_2 is a number of clusters (subsets of the initial ensemble) in clusterizations K and Q accordingly; $|K_i|, |Q_j|, i = 1,2,\ldots,k_1;\ j = 1,2,\ldots,k_2$, – potencies of appropriate subsets, i.e. elements number in clusters.

Magnitude $d(K, Q)$ accepts values from 0 up to 1: 0 - at completely coincident partitions in clusterizations K and Q, 1 - at completely distinct, when $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} |K_i \cap Q_j|^2 = 0$.

The supposition 2 causes application of geomatic algorithm. Paper [4] describes it.

Tab. **1**. Proximity measures.

<table>
<tr><th colspan="2">Proximity measures</th></tr>
<tr><th>between objects, $d(X_i, X_j)$</th><th>between clusters, $d(K_i, K_j)$</th></tr>
<tr><td>Euclidean distance
$d_E(X_i, X_j) = \left[\sum_{l=1}^{m} (x_{il} - x_{jl})^2\right]^{1/2}$</td><td>nearest neighborhood
$d_{min}(K_i, K_j) = \min_{X_l \in K_i, X_m \in K_j} d(X_l, X_m)$</td></tr>
<tr><td>weighted Euclidean distance
$d_B(X_i, X_j) = \left[\sum_{l=1}^{m} w_l (x_{il} - x_{jl})^2\right]^{1/2}$</td><td>remote neighborhood
$d_{max}(K_i, K_j) = \max_{X_l \in K_i, X_m \in K_j} d(X_l, X_m)$</td></tr>
<tr><td>potential function
$d_{П}(X_i, X_j) = \left[1 + a\, d_E^2(X_i, X_j)\right]^{-1}, a > 0,$</td><td>average neighborhood
$d_{mean}(K_i, K_j) = \frac{1}{n_i \cdot n_j} \sum_{X_l \in K_i} \sum_{X_m \in K_j} d(X_l, X_m)$</td></tr>
<tr><td>eigther
$d^*_{П}(X_i, X_j) = exp(-a\, d_E^2((X_i, X_j))$
or
$d_{Пф}(X_i, X_j) = \left\lvert \frac{\sin a\, d_E^2(X_i, X_j)}{a\, d_E^2(X_i, X_j)} \right\rvert$</td><td>distance between centers of gravity
$d_c(K_i, K_j) = d_E(m_i, m_j)$
potential function
$d_P(K_i, K_j) = \frac{1}{n_i \cdot n_j} \sum_{X_l \in K_i} \sum_{X_m \in K_j} d_P(X_l, X_m)$</td></tr>
<tr><td>angular measure
$d(X_i, X_j) = arccos((X_i^{\chi} X_j)/(\lvert X_i \rvert^{\chi} \lvert X_j \rvert))$</td><td>Mahalonobis distance
$d_M(K_i, K_j) = (m_i - m_j)^T C^{-1} (m_i - m_j)$</td></tr>
<tr><th colspan="2">Proximity measures between object and cluster, $d(X, K_i)$</th></tr>
<tr><td colspan="2">Mahalonobis distance
$d_M(X, K_i) = (X - m_i)^T C_i^{-1} (X - m_i)$</td></tr>
<tr><td colspan="2">proximity measures function
$d_{PMF}(X, K_i) = \left(\prod_{X_j \in K_i} d(X, X_j)\right)^{1/n_i}$ or $d^*_{ФМБ}(X, K_i) = \frac{1}{n_i} \sum_{X_j \in K_i} ln\, d(X, X_j)$</td></tr>
<tr><td colspan="2">potential function
$d_P(X, K_i) = \frac{1}{n_i} \sum_{X_j \in K_i} d_P(X, X_j)$</td></tr>
<tr><td colspan="2">similarity angular measure
$d_{PSI}(X, K_i) = \left[\prod_{X_j \in K_i} \sin(X, \hat{} X_j)\right]^{1/n_i}$</td></tr>
<tr><td colspan="2">distance for centers of gravity of cluster
$d_c(X, K_i) = d_E(X, m_i)$</td></tr>
<tr><td colspan="2">where $\mu_j = \frac{1}{n_j} \sum_{X_i \in K_j} X_i$ is a vector of cluster averages K_j, w_l is a weight coefficient , C_i is a covariance matrix of cluster K_i, $C = C_i = C_j$</td></tr>
</table>

GeoTime includes the basic steps or geomatic algorithm and complicates it with procedure of kernels selection (based on succession of clusterizations in sequential instants) - the clusters forming a complete geographical region.

The common scheme of GeoTime algorithm is as follows.

1. To discover an adjacency matrix $G=(g_{ij}), i, j=1,2,\ldots,n,$ between objects $Z_1, Z_2,\ldots, Z_n$ in geographical space; g_{ij} is being defined by one of the following:

a) $g_{ij}=\begin{cases} 1, if\ \ r_{ij} < c, \\ 0-otherwise, \end{cases}$ b) $g_{ij}=\begin{cases} 1, if\ Z_i\ and\ \ Z_j-neighbors, \\ 0-otherwise, \end{cases}$

where r_{ij} is Euclidean distance between objects Z_i and Z_j in geographical space, ($\boldsymbol{Q}$ - the initial matrix for its evaluation), c is a threshold.

2. To compute matrix $D=(d_{ij}), i, j=1,2,\ldots,n,$ where $d_{ij}=d(Z_i,Z_j)$ is a similarity measure between Z_i and Z_j in indications space (tab. 1) ($\boldsymbol{X}=(\boldsymbol{X}(t_1)\mathbb{M}\boldsymbol{X}(t_2)\mathbb{M}\ldots\mathbb{M}\boldsymbol{X}(t_L))$ is the initial matrix for d_{ij} evaluation). To compute matrix $D(t_s)=(d_{ij})|_{ts}$ analogous to D (using $\boldsymbol{X}(t_s)$) for each instant t_s ,$s=1,2,\ldots,L$.

3. To define clusters centers of an initial partition (kernel) $X_{i_1}, X_{i2},\ldots, X_{i_k}$, using the described algorithm of clusterizations succession definition in sequential instants. To accept

$$K_1^1=\{X_{i_1}\}, K_2^1=\{X_{i_2}\},\ldots, K_k^1=\{X_{i_k}\}.$$

4. To suppose $r=1$.

5. Let on step $r \hat{I} \{1,\ldots,n-k\}$ classes $K_1^r,\ldots,K_k^r$ are gained.

6. To discover $j \hat{I} \{1,\ldots,n\}$, $v \hat{I} \{1,\ldots,k\}$: $d_{ji_v}=min\{d_{ji_q} \mid \exists X_m \in K_q : g_{jm}=1, q \in \{1,2,\ldots,k\}\}$.

7. To suppose $\forall\, i \hat{I} \{1,\ldots,k\}$ $K_i^{r+1}=\begin{cases} K_i^r \cup \{X_j\}, i=v; \\ K_i^r, i \neq v. \end{cases}$

(The supposition 2 is considered on this step).

8. To suppose $d_{ji_v}=\infty$.

9. Algorithm is finished if $r=n-k$, otherwise suppose $r=r+1$ and go to step 5.

Thus selecting some areas in each cluster forming a complete geographical region (spatial volume), one suppose them as kernels. The selected clusters kernels are dilated by additional classification of remained objects on the algorithm above.

Advantages of offered clusterization are:

- taking into account of topographic features of object locations for selection of complete geographical bands made by separate clusters;
- taking into account of temporal trends of indices modification for selection of initial clusters kernels;
- a possibility of informative interpretation of the selected clusters.

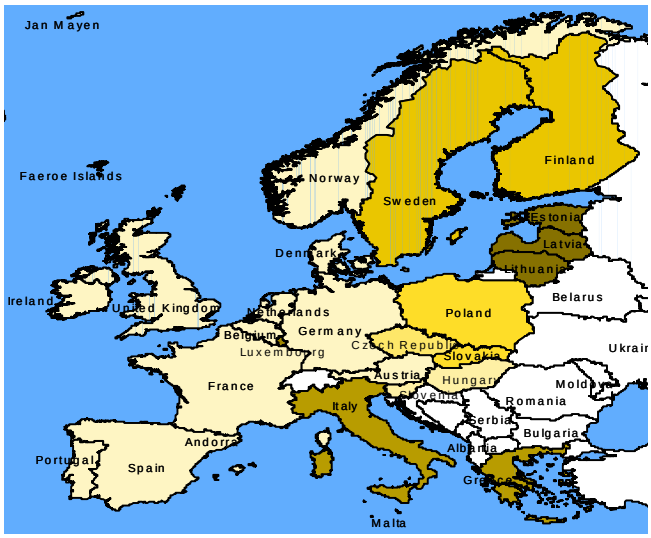
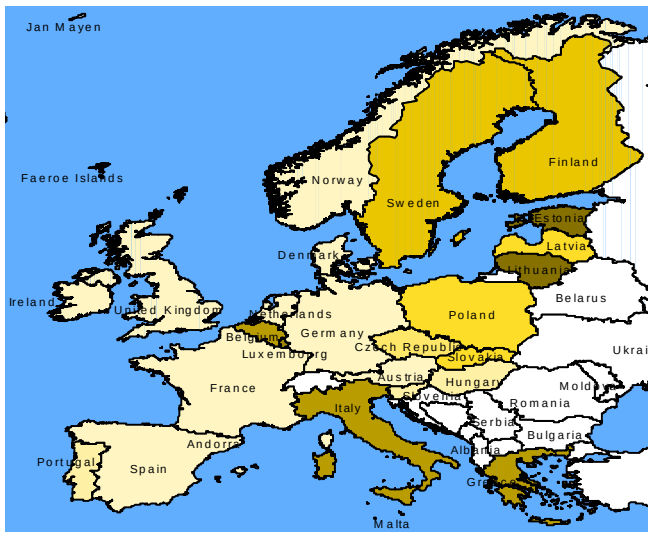
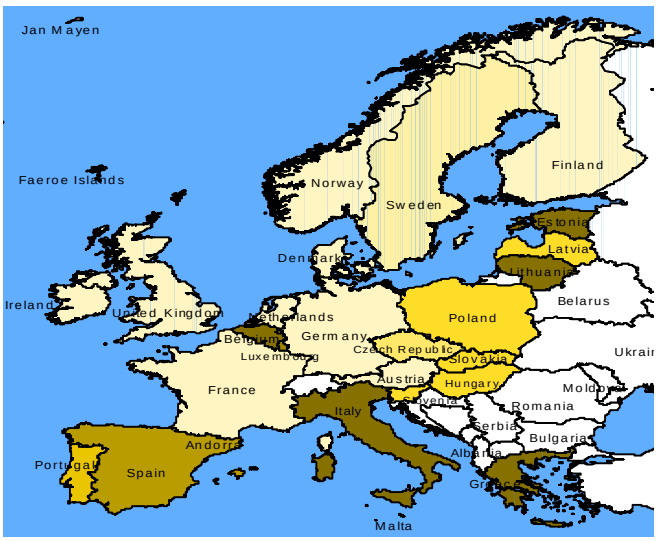
4 Experimentation

The experimental matching GeoTime with known algorithms was spent on the accessible actual data of ESE-monitoring published on the site [5]. Sixteen ESE-indices of twenty four states of Europe appeared as the initial indications for nine years (1996-2004):

- economic indices – Business investment, % of GDP; Consumption expenditure at constant prices, index 1995 = 100; GDP per capita in PPS, index EU-25 = 100; General goverment gross debt, % of GDP; Inflation rate, %; Net saving: Public sector, % of GDP; Public investment, % of GDP; Real GDP growth rate, %; Total investment, % of GDP; Unit labour cost growth: total economy, %;
- social indices – Total employment growth, %; Total employment rate, %; Electricity consumption by households, Index 1995 = 100;
- ecological indices – CO_2 emissions per capita, tonnes; Forest trees damaged by defoliation, %; Municipal waste generated, kg per capita.

In tab.2 some outcomes of the experimental comparison GeoTime with known algorithms of clusterization (number of clusters $k=8$, $m=16$ indications, $n=24$ objects) are presented.

Tab. 2. Outcomes of the experimental matching of clusterization methods.

Algorithm, Clusterization, quality Criterions	Geoiconik model	Region, cluster
Geomatic J_1=415,8 J_2=171,1 J_3=0,78		Belgium 1 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 6 Spain 1 France 1 Ireland 1 Italy 6 Latvia 8 Lithuania 8 Luxembourg 7 Hungary 2 Netherlands 1 Austria 1 Poland 4 Portugal 1 Slovenia 2 Slovakia 4 Finland 5 Sweden 5 United Kingdom 1 Norway 1
GeoTime J_1=415,9 J_2=179,4 J_3=0,81		Belgium 6 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 6 Spain 1 France 1 Ireland 1 Italy 6 Latvia 4 Lithuania 8 Luxembourg 7 Hungary 2 Netherlands 1 Austria 1 Poland 4 Portugal 2 Slovenia 2 Slovakia 4 Finland 5 Sweden 5 United Kingdom 1 Norway 1
k-means J_1=417,3 J_2=178,2 J_3=0,75 1 2 3 4 5 6 7 8 No Data		Belgium 7 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 7 Spain 6 France 1 Ireland 1 Italy 7 Latvia 4 Lithuania 8 Luxembourg 7 Hungary 4 Netherlands 1 Austria 1 Poland 4 Portugal 5 Slovenia 4 Slovakia 4 Finland 1 Sweden 2 United Kingdom 1 Norway 1

As a similarity measure between objects in 16-dimensional indices space the angle between vectors X_i and X_j was applied:

$$d(X_i, X_j)=\arccos((X_i \times X_j)/(|X_i| \times |X_j|)).$$

For matching algorithms *J1, J2, J3* are enumerated in tab. 2 – they are the values of following formal criterions of an clusterization quality estimation [1, 2]:

1) criterion of intracluster dispersions

$$J_1 = \sum_{j=1}^{k} \sum_{X_i \in K_j} d_E^2 \left(X_i, \mu_j\right),$$

where $m_j = \frac{1}{n_j} \sum_{X_i \in K_j} X_i$ - is a barycenter of cluster K_j; n_j is a number of objects in it;

2) criterion of paired intracluster distances between objects:

$$J_1 = \sum_{j=1}^{k} \sum_{X_i \in K_j} d_E^2 \left(X_i, \mu_j\right),$$

3) criterion of an intercluster scatter of objects (the more magnitude of J_3 $(0<J_3<1)$, the greater portion of a common objects scatter is illustrated by the intergroup scatter and the quality of a partition is better):

$$J_3 = 1 - \frac{W}{S},$$

where $W = \sum_{j=1}^{k} W_j$; $W_j = \sum_{X_i \in K_j} d^2\left(X_i, \mu_j\right)$ is an intracluster scatter; $S = \sum_{i=1}^{n} d^2\left(X_i, \overline{X}\right)$ is a common dispersion.

Criteria J_1, J_2, J_3 shown in table 2, gave the best results for GeoTime-clusterisation comparison with Geomatic and k-means algorithms.

5 Conclusions

New GeoTime algorithm is developed for clusterization of the spatial-temporary data, considering the geographical neighbourhood of objects (on the basis of the geoinformational approach, i.e. reviewing of topological features of regions locations) and the temporal neighborhood of their indices (on the basis of the inductive approach).

Advantages of the offered algorithm: 1) aiming at selection of complete geographical bands formed by separated clusters; 2) the registration of temporal trends of indices modification for selection of initial clusters kernels; 3) a possibility of informative interpretation of the selected clusters.

Numerous experiments on actual data of region ESE-monitoring have proved that geomatic and GeoTime-clusterizations have slightly better indices of known formal quality criterions compared to clusterization gained by the k-means method (outcomes are worse for hierarchical methods, than for k-means).

The outcomes considered above show application expediency of the offered spatial-temporary approach to regions clusterization according to ESE-monitoring data.

References

[1] Classification and reduction of dimensionality. Edited by prof. S.A.Aivazyan, 1989. –608 p.
[2] Tou J.T., Gonzalez R.C. Pattern recognition principles.– Addison-Wesley Publishing Company, 1974.– 411 p.
[3] Duran B., Odell P. Cluster Analysis. – M.: Statistica, 1977. – 128 p.
[4] Sarycheva L., Boyko A. Geomatic clusterization based on ecological-social-economical monitoring rates // Scientific Bulletin of NMU, 2006, №5.– P.76-81.
[5] Eurostat // Эл. ресурс. URL: http://www.eurostat.com/
[6] Sarycheva L., Zhuykov A. Cluster Analysis of Territories by the Totality of Ecological and Socio-Economic Indices // IEEE 2001 International Geoscience and Remote Sensing Symposium, UNSW. – P.663-664.