Systems Analysis Modelling Simulation
Vol. 43, No. 10, October 2003, pp. 1409-1414

Taylor & Francis
Taylor & Francis Group

USING GMDH IN ECOLOGICAL AND SOCIO-ECONOMICAL MONITORING PROBLEMS

LYUDMILA SARYCHEVA*

Institute of Geoinformatics, National Mining University of Ukraine, Dnepropetrovsk, K. Marx av. 19, Ukraine

(Received 13 August 2001)

Geographically distributed and time-phased ecological and socio-economical indices are treated as source data for the analysis, carried out for taking reasoned management decision. Analysis complexity, related to multi-dimensional data is overcome thanks to using group methods of data handling (GMDH). This work shows the results of Ukraine's regions cluster analysis by the totality of ecological and socio-economical indices. These results are visualized as homogeneity classes' map in geographic information system.

Keywords: Ecological and socio-economical monitoring (ESEM); Socio-economical index; Objective clusterization

1. INTRODUCTION

A system of ecological and socio-economical monitoring (ESEM) provides the system approach to exploration of human society's interaction with environment. ESEM includes: (a) keeping unified monitoring database; (b) data visualization; (c) data analysis; (d) building up ecological and socio-economical model; (e) forecasting ecological, economical and social situations development; (f) forming management decisions, based on modeling and forecasting.

For solving (c), (d) and (e) problems of ESEM, GMDG are suitable. These problems are based on joint analysis of three informational blocks: ecological, economical and social. Analyzed objects are n regions of the territory, characterized by m ecological and socio-economical indices (attributes). Examination results are represented as an "object-attribute" matrix of dimension $m \times n$. The number of attributes is greater than the number of objects: $n < m$.

GMDH expediency can be explained by the fact that these methods use iterative schemes of models complication. GMDH are based on the analogy with biological selection of organisms. Models complication during selection takes place due to "cross-breeding" of the best models of the previous selection row [1,2]. Iterative methods are efficient not only when m is little, but also when $m > 100$ and even

*E-mail: Sarycheva@prognoz.dp.ua; gis@nmuu.dp.ua

ISSN 0232-9298 print: ISSN 1029-4902 online
DOI: 10.1080/02329290290024925

$n < m$. GMDH fundamental principles are: (a) intermediate solution is not final; (b) external compliment (using verifying sampling); (c) self-selection of intermediate solution; (d) uniqueness of final solution. This work solves the problem of ESEM data clusterization according to GMDH ideology.

2. PROBLEM STATEMENT

Let the investigated territory be divided into regions $\mathbf{X}_i$, $i = 1,2,\ldots,n$, (region is a structural spatial unit of territory lay-out, usually a city district or industrial domain), n is the number of regions. According to ESEM data, each region is characterized by the totally of $m = \lambda + \mu + \nu$ indices: λ – ecological indices (atmospheric emission of contaminants, square of land resources, waster product of mining operations and mineral processing etc.), μ - economical indices (make quantity per head, consumption of electrical energy, gross output of agriculture, etc.) and ν social indices (level of unemployment, cash income of population, demographical indices, etc.). The territory has to be divided into homogeneous zones by the totality of ecological, eco-economical and eco-socio-economical indices.

In mathematical problem statement, regions are represented as objects (vectors), ecological and socio-economical indices are represented as these objects' attributes (vector coordinates). An "object-attribute" matrix $(\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_n)^{\mathrm{T}}$ of dimension $m \times n$ is analyzed, which has the following structure:

$$\begin{array}{c} \\ \mathbf{X}_1 \\ \\ \mathbf{X}_n \end{array} \begin{pmatrix} \overbrace{\phantom{X_{ij}}}^{\lambda} & \overbrace{\phantom{X_{ij}}}^{\mu} & \overbrace{\phantom{X_{ij}}}^{\nu} \\ & X_{ij} & \\ & & \end{pmatrix}$$

$i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m$. Each row of the matrix describes one object. Clusterization problem has to be solved. Clusterization $\mathbf{K}(X) = \{\mathbf{K}_1(X), \mathbf{K}_2(X), \ldots, \mathbf{K}_k(X)\}$, $n \geq k \geq 1$, of set X is a collection of nonempty pairwise disjoint subsets (clusters) $\mathbf{K}_q(X)$, $q = 1,2,\ldots,k$ of the X set, whose union coincides with $X : \mathbf{K}_1 \cup \mathbf{K}_2 \cup \cdots \cup \mathbf{K}_k = X$, $\mathbf{K}_1 \cap \mathbf{K}_2 \cap \cdots \cap \mathbf{K}_k = \emptyset$, $\mathbf{K}_q \neq \emptyset$, $q = 1,2,\ldots,k$.

The number of cluster k can be unknown beforehand. For solving the clusterization problem, it is necessary to: (a) give definition of a cluster, i.e. indicate the properties, common for all the objects of a certain cluster (measure of resemblance between objects); (b) specify the method of partitioning objects into clusters; (c) specify clusterization quality criterion J (measure of resemblance between classes); (d) organize this criterion's movement to minimum (maximum) (during this process the actual number of clusters is defined).

3. OBJECTIVE CLUSTERIZATION ALGORITHM

Clusterization is carried out under the multiple-state scheme. On the first stage Euclidean distance (or generalized Euclidean distance) is taken as a measure of resemblance between objects and then clusterizations are analyzed, received by three methods (realizing different methods of fragmentation), under the assumption that the number of really existing clusters $k = 2$.

In order to evaluate clusterization quality, initial set of objects is divided in a certain way into two equinumerous subsets A and B. Clusterization is considered to be objective if separately carried out for each equinumerous subset clusterization give the greatest congruence (thus, GMDH principle of external compliment is realized).

On the following stages the assumed number of cluster is increased: $k = k + \backslash(k < n)$. The best clusterizations of the previous stage are stored and then compared with newly obtained ones.

If during an iteration the number of clusters increases to the number of objects, but the desired criterion of clusterization objectivity is not obtained, consecutive replacement takes place: first, valuation method is replaced, then, measure of resemblance between classes is replaced, and finally, other clusterization methods are taken.

The scheme of one stage of the algorithm is the following:

1. The data are valuated.
2. Initial set X, containing n objects, is divided into two non-overlapping equinumerous subsets A and B ($\mathbf{A} \cap \mathbf{B} = \emptyset, \mathbf{A} \cup \mathbf{B} = X$):
   (a) $n(n-1)/2$ distances r_{ij} between objects X_i and X_j are calculated, where $i = 1, 2, \ldots, n-1, j = i+1, i+2, \ldots, n$;
   (b) $r_{qs} = \min_{i,j} r_{ij}$ is determined;
   (c) qth object is assigned to subset A, and the closest to it sth object is assigned to subset B;
   (d) steps (b)–(c) are repeated for the remaining objects until all the objects are assigned either to A or B subset. Subset A contains objects with number $q_1, q_2, \ldots, q_{n/2}$, and subset B contains objects with numbers $s_1, s_2, \ldots, s_{n/2}$ (n is assumed to be an even number, otherwise some of the objects are regarded twice).
3. Subsets A and B are clusterized (separately) by k average method (k is the expected number of clusters), closest neighbor method or ISODATA method [3]. Number k changes from iteration to iteration (usually $k = 2, 3, \ldots, n$).

The method of k average minimizes internal clusterization quality criterion:

$$J(\mathbf{K}_i) = \sum_{\mathbf{X} \in \mathbf{K}_i} d^2(\mathbf{X}, \mathbf{z}_i), \quad i = 1, 2, \ldots, k,$$

where $\mathbf{z}_i$ is a center of cluster $\mathbf{K}_i$, $d(\mathbf{X}, \mathbf{z}_i)$ is the distance between object X and center $\mathbf{z}_i$.

The ISODATA algorithm performs the following steps:

- given objects are distributed among k clusters, corresponding to initial centers $\mathbf{z}_i$, $j = 1, 2, \ldots, k$, by the following rule: $\mathbf{X} \in \mathbf{K}_j$, if $d(\mathbf{X}, \mathbf{z}_j) < d(\mathbf{X}, \mathbf{z}_i)$, $i \neq j$;
- if $n_j < \Theta_N$, cluster $\mathbf{K}_j$ is excluded (n_j is the number of objects in $\mathbf{K}j$; Θ_N is a parameter to compare the number of objects in a cluster with);
- clusters' centers $\mathbf{z}_i$ are corrected: $\mathbf{z}_i = 1/n_i \sum_{\mathbf{X} \in \mathbf{K}_i} \mathbf{X}$;
- the mean distance D_i is calculated between the objects in cluster $\mathbf{K}_i$ and the corresponding center $\mathbf{z}_i$; generalized mean distance D is calculated between objects in clusters and corresponding centers;
- for each cluster $\mathbf{K}_i$ the vector of mean square deviation σ_i is determined, each component of which characterizes mean square deviation of an object of cluster $\mathbf{K}_i$ by one of coordinate axis;

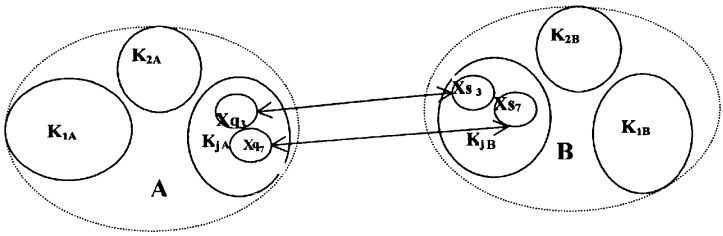
- in each of vectors σ_i the maximal component $\sigma_{i\,\max}$ is determined;
- distances D_{ij}, $i = 1, 2, \ldots, k-1, j = i+1, \ldots, k$, between all pairs of clusters' centers are calculated and compared with parameters Θ_c, characterizing the compactness; Θ_L distances which are less than Θ_c are ranged in ascending order; clusters' pairs, corresponding to the lowest values of distances D_{ij} are merged.

Described algorithm, implemented by authors, maximizes internal clusterization quality criterion:

$$J = \ln \frac{k^{k-1} \cdot \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(\mathbf{z}_i, \mathbf{z}_j) \cdot \prod_{i=1}^{k} n_i / n}{(k-1)(1+\sigma_{\max})(1 + 1/n \sum_{i=1}^{k} \bar{D}_i / n_i)},$$

where $\sigma_{\max} = \max_i(\sigma_{i\max})$, $\bar{D}_i$ is mean distance between objects of cluster $\mathbf{K}_i$.

4. Among the obtained clusterizations the best one is selected. Selection of best clusterization is realized according to the external criterion of "objectivity", evaluating the correspondence of clusterizations, obtained separately from excerpts A and B. Clusterization is considered best if it has greatest number of objects pairs (q_l, s_l), $l = 1, 2, \ldots, n/2$, located in the corresponding clusters of subsets A and B. E.g. if objects with numbers q_3, q_7, q_{10} are in the same class of subset A and objects with numbers s_3, s_7, s_{10} are in the same class of subset B, the clusterization is objective:



In conclusion we will note the distinctive features of the algorithm: multiple-stage search of the best clusterization, clusterization quality evaluation by means of external and internal criteria, usage of collections of measures of resemblance and clusterization methods.

4. OBJECTIVE CLUSTERIZATION METHOD APPLICATION FOR ESEM DATA ANALYSIS

Department of regional policy of the Ministry of Economy of Ukraine places on its official web-site (www.me.gov.ua) materials [4], characterizing the main indices of socio-economical conditions in Ukraine: industry growth rate by the main branches of economy, indices characterizing industry, agriculture, building and investment of capital, goods turnover, export and import, state of backlogs analysis, financial state, social state, privatization and private enterprise, and ecology. These materials for 1999 and January-July 2000 were taken as source data for the cluster analysis of Ukraine's regions by the totality of indices.

"Object-attribute" tables of dimension m x n are the subject to multi-dimensional analysis, $n = 27$ is the number of initial set objects (regions of Ukraine), m is the

number of attributes, describing these objects (ESE indices, e.g. $m = 5$ is the number of ecological indices, $m = 47$ is the number of social sphere indices, $m = 94$ is the total number of indices, etc.). Let us briefly describe some obtained results.

Figure 1 shows results of cluster analysis of $n = 27$ Ukraine regions by the totality of: (a) $m = 47$ social sphere indices; (b) $m = 8$ investment indices; (c) $m = 94$ ecological and

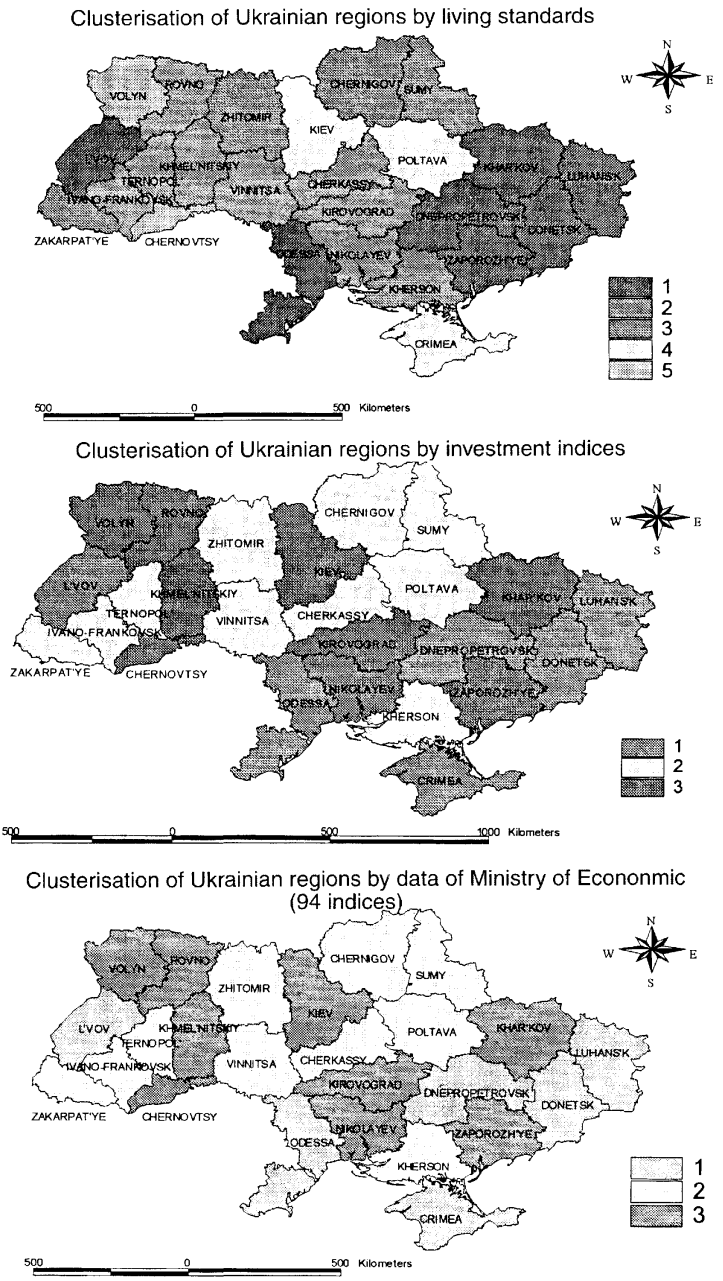


FIGURE 1 Results of cluster analysis.

socio-economical indices. Resulting maps analysis allows to extract regions, homogeneous by the totality of indices. The suggested method gives a convenient and effective tool for taking management decisions. Instead of operating with multi-dimensional tables, a person, taking a decision can analyze visual informative maps.

CONCLUSIONS

A solution on of one of the mathematical problems of ESEM is considered (the problem of territory regions cluster analysis by the totality of its indices) based on the GMDH ideology. Features of the developed method: multiple-stage search of the best clusterization, clusterization quality evaluation by means of verifying subsample, usage of collection of decision functions and measures of resemblance between two objects.

The work gives the results of Ukraine regions cluster analysis by the totality of ecological and socio-economical indices. Obtained results are visualized as homogeneity classes' maps in geographic information system. The method is actual for ESEM data analysis and for taking reasoned management decisions.

References

[1] A.G. Ivakhnenko and V.S. Stepashko (1985). Pomekhoustoychivost' modelirovanya (Noicc Immunity of Modeling). Naukova dumka, Kiev.
[2] A.G. Ivakhnenko and Y.P. Yurachkovsky (1987). Complex system modeling by the experimental data. Radio i Svyaz., 1-20.
[3] J. Tou and R. Gonzales (1974). *Pattern Recognition Principles.* Addison-Wesley Publishing Company.
[4] www.me.gov.ua - Regional policy department of Ministry of Economy of Ukraine official web site.