Raport Badawczy

Research Report

RB/54/2005

# Analysis of socio-economics factors in polish regional structure

W druku: Kopacek (Ed) ACS, Elsevier, (Pergamon Press), Oxford, London

J. Hołubiec, G. Petriczek

**Instytut Badań Systemowych**
**Polska Akademia Nauk**

**Systems Research Institute**
**Polish Academy of Sciences**

POLSKA AKADEMIA NAUK

**Instytut Badań Systemowych**

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 8373578

fax: (+48) (22) 8372772

Kierownik Pracowni zgłaszający pracę:
Dr inż. Jan Owsiński

Warszawa 2005

**Artykuł przyjęty do druku In: Kopacek P. (Ed): ADVANCED CONTROL STRATEGIES FOR SOCIAL AND ECONOMIC SYSTEMS 2004.** ***Elsevier (Pergamon Press), Oxford, London*** **Seria: IFAC Proceedings Volume.**

# ANALYSIS OF SOCIO-ECONOMICS FACTORS IN POLISH REGIONAL STRUCTURE.

**Jerzy Hołubiec, Grażyna Petriczek**

## SPIS TREŚCI

# ANALYSIS OF SOCIO-ECONOMICS FACTORS IN POLISH REGIONAL STRUCTURE

**Jerzy Holubiec, Grażyna Petriczek**

*Systems Research Institute, Polish Academy of Sciences,*
*01-447 Warsaw, Newelska 6 ,Poland*
*tel. 0-22 36- 44- 14, fax. 022-37-27-72, e-mail: petricz@ibspan.waw.pl*

Abstract: In the paper the concept of partition of Poland territory consisting of vojvodships into group with the similar socio-economic characteristics is presented. This concept is based on the method of determining and selecting the homogeneous groups from the data set describing the analysed phenomenon.
The data set describing vojvodships in Poland has the form of matrix consisting of rows number corresponding to vojvodships number and columns number depended on the number of considered socio-economic factors.
The analysis of partition of Poland territory with respect to various socio-economic characteristics was performed for years 1998, 2000, 2001. In 1999 year new partition of Poland territory was performed. Til 1999 year Poland territory consisted of 49 vojvodships. After administration reform Poland territory contains 16 vojvodships.

KeyWords: Regional modelling, hypotheses, data set homogeneity, statistics , statistically stable boundaries between two sets

## 1. METHOD OF SET DIVISION INTO HOMOGENEOUS SUBSETS

The essential element of the method is a criterion for testing data set homogeneity hypothesis. The criterion function has a form of statistics U, which has the $\chi^2$ distribution. The method consists in iterative partition of nonhomogeneous set into two parts. If number of this division increases, than the process of successive partitions can lead to the existence of statistically unstable boundaries between adjacent homogeneous subsets.
The subsets aggregation is based on adequately formulated hypothesis testing concerning stability of the boundary between the two sets.

Thus, the data set division algorithm contains two basic fundamental steps :
- the set partition procedures into homogeneous, disconnected subsets, that enable the primary set division
- the procedures that examine the stability of boundaries between these subsets.

This two - steps algorithm is as follows:

### *1.1 Set division into homogeneous, separated subsets*

The basic idea underlying the partition is principle of the equivalence of random variable that have the same distribution.
Let $S = \{ s_1 , s_2 , ...., s_n \}$ denotes a data set.

Assume that to each element $s \in S$ corresponds a random variable $\xi_s$ with the distribution function $F_s(x)$.
Let $E^s$ denotes the set of random variables $\xi_s$ and R denotes the set of random variable values x.

Def.1. Random variables $\xi_{s_1}, \xi_{s_2}$ are equivalent if for each two elements $s_1$, $s_2 \in S$ the following relation is fulfilled :

$$F_{s_1}(x) - F_{s_2}(x) = 0 \text{ for any } x \in R \qquad (1)$$

The above condition states that a random variable set $E^s$ can be disconnected into equivalence classes, and subsequently, the set S has the form :

$$S = S_1 \cup S_2 \cup \ldots \cup S_k \; k \geq 1 \qquad (2)$$

If k = 1 then the set S is homogeneous.
If k > 1 then the set S is nonhomogeneous and the relation (2) describes this nonhomogeneity.
On the basis of the conception of random variable equivalence one can formulate the homogeneity definition.

Def.2. The set of random variables $E^{S_1} \subset E^S$ is homogeneous if the following condition is fulfilled:

$$F_{s'}(x) - F_{s''}(x) = 0 \quad \text{for any } s', s'' \in S_1 \text{ and } x \in R \qquad (3)$$

By contradiction to the homogeneity condition (3) one obtains the nonhomogeneity definition.
A set of equivalent random variables is homogeneous. In order to formulate homogeneity hypothesis testing criteria one assumes that the random variables are independent and normally distributed with the probability density function :

$$f(x) = \frac{1}{\sqrt{(2\pi)^k}} |\Sigma_s|^{-\frac{1}{2}} \exp\left( -\frac{1}{2}(x - m_s)^T \Sigma_s^{-1} (x - m_s) \right) \qquad (4)$$

where :
$m_s$ - the vector of expected values of the random variable $\xi_s$
$\Sigma_s$ - the covariance matrix with [k × k] dimension
$|\Sigma_s|$ - the determinant of covariance matrix

Moreover, if one assumes that the considered variables have the same covariance matrices then the homogeneity condition takes the following form: ( the $H_0$ hypothesis )

$$H_0 : \quad m_{s'} = m_{s''} \quad \text{for all } s', s'' \in S$$
$$\text{subject to :} \quad \Sigma_{s'} = \Sigma_{s''} \qquad (5)$$

Defining on the set of all partitions of S into two subsets $S_1$ and $S_2$ the function :

$$\delta(S_1, S_2) = \frac{1}{n_1} \sum_{s \in S_1} m_s - \frac{1}{n_2} \sum_{s \in S_2} m_s \qquad (6)$$

where : $n_1$ , $n_2$ - number of elements of $S_1$ and $S_2$ respectively
one obtains a homogeneity index of k - dimensional variable set.
On the basis of index (6) the homogeneity hypothesis can be formulated as follows:

$$H_0 : \quad \delta(S_1, S_2) = 0 \qquad (7)$$

for any pair ( $S_1$, $S_2$ ) belonging to all partitions of S into two subsets.

Let n - be a number of observations and k - be a number of considered characteristics describing the analyzed phenomenon.
The set of all observations of k - dimensional random variable is a matrix

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \ldots, & x_{1k} \\ x_{21}, & x_{22}, & \ldots, & x_{2k} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ x_{n1}, & x_{n2}, & \ldots & x_{nk} \end{bmatrix} \qquad (8)$$

The matrix (8) is a realization of k - dimensional normally distributed random variable $\xi_s$ , with mean value $m_s$ and the same diagonal covariance matrices.
The $H_0$ hypothesis (5) testing is performed as to compare two samples.
That result is the partition of matrix (8) into two various, disconnected parts containing $n_1$ and $n_2$ rows, respectively.
A criterion function for the $H_0$ hypothesis (7) testing is constructed using the maximum likelihood method.
The statistical estimation of k - dimensional partition $\delta$ ( $S_1$, $S_2$ ) is given by the random variable $\tilde{\xi}$ of the form:

$$\tilde{\xi} = \frac{1}{n_1} \sum_{s \in S_1} \xi_s - \frac{1}{n_2} \sum_{s \in S_2} \xi_s \qquad (9)$$

Each component appearing in the expression (9) is a random variable with the following distribution parameters: the expected values $(m'_j, m_j)$ and the variances ( $\sigma_j^2 / n_1$ , $\sigma_j^2 / n_2$ ).
Assuming that the $H_0$ hypothesis given by condition (5) is true, one can formulate the likelihood function of the form:

$$L(x,m) = \frac{1}{\sqrt{(2\pi)^k}} \left( \prod_{j=1}^{k} c_j^2 \right)^{-\frac{1}{2}} \exp\left[ -\frac{1}{2} \sum_{j=1}^{k} \frac{\tilde{x}_j^2}{c_j^2} \right] \qquad (10)$$

where: $c_j^2$ - the variance of the random variable $\xi$ described by the following relation :

$$c_j^2 = \frac{\tilde{\sigma}_j^2 (n_1+n_2)}{n_1 n_2} \qquad (11a)$$

$\tilde{\sigma}_j^2$ - sample variance of the random variable $\xi_{s_j}$

$\tilde{x}_j$ - observation values of j-th component of the random variable $\xi$ determined as follows :

$$\tilde{x}_j = \frac{1}{n_1} \sum_{s \in S_1} x_{sj} - \frac{1}{n_2} \sum_{s \in S_2} x_{sj} \qquad (11b)$$

The sample variance can be calculated from the relation:

$$\tilde{\sigma}_j^2 = \frac{1}{n_1+n_2-1}\left[\sum_{s \in S_1} x_{sj}^2 + \sum_{s \in S_2} x_{sj}^2 - \frac{1}{n_1+n_2}\left(\sum_{s \in S_1} x_{sj} + \sum_{s \in S_2} x_{sj}\right)^2\right] \qquad (11c)$$

It can notice that the form of the likelihood function depends on the exponent $\sum_{j=1}^{k} \frac{\tilde{x}_j^2}{c_j^2}$

Substituting the expressions (11a) - (11c) to the exponent one obtains the following criterion function U( $S_1$, $S_2$ ):

$$U(S_1,S_2) = \frac{\frac{1}{(n_1+n_2)n_1 n_2} \sum_{j=1}^{k} \left( n_2 \sum_{s \in S_1} x_{s_j} - n_1 \sum_{s \in S_2} x_{s_j} \right)^2}{\sum_{j=1}^{k} \left( \sum_{s \in S} x_{s_j}^2 - \frac{1}{n_1+n_2} (\sum_{s \in S} x_{s_j})^2 \right)}$$

$$S = S_1 \cup S_2 \qquad (12)$$

The statistics U has the $\chi^2$ - distribution with k degrees of freedom.

From the maximum likelihood principle and the properties of function L( , ) it follows that the $H_0$ hypothesis can be accepted if the following condition holds :

$$U(S_1,S_2) \le \chi^2_{\alpha,k} \qquad (13)$$

for any pair ( $S_1$, $S_2$ ) belonging to all partitions of set S

where : $\alpha$ - significance level
k - degree of freedom

If the above condition is not satisfied then the $H_0$ hypothesis should be rejected; the set is nonhomogenous and can be partitioned into two disconnected subsets.

The determination of statistics U for all pairs ( $S_1$ , $S_2$ ) and examination of inequality (13) for a large number of observations are complicated and causes computational difficulties.

To simplify the $H_0$ hypothesis testing it was assumed, that the observations { $X_1$ , $X_2$ , .... , $X_n$ } are ordered increasingly with regard to most significant characteristics.

In this case it is sufficient to verify the hypothesis $H_0$ only for the pairs ( $S_1$, $S_2$); where the subset $S_1$ consists of l - elements and the subset $S_2$ consists of the remaining n-l elements.

In this case, the statistics U has the form :

$$U(l,n-l) = \frac{\frac{n-l}{n(n-l)l} \sum_{j=1}^{k} \left( (n-l) \sum_{i=1}^{l} x_{ij} - l \sum_{i=l+1}^{n} x_{ij} \right)^2}{\sum_{j=1}^{k} \left( \sum_{i=1}^{n} x_{ij}^2 - \frac{1}{n} (\sum_{i=1}^{n} x_{ij})^2 \right)}$$

$$\text{for } l = 1, 2, \ldots\ldots, n\text{-}1 \qquad (14)$$

From the above consideration, it follows that the homogeneity hypothesis $H_0$ is accepted if the following inequality:

$$U(l, n-l) \le \chi^2_{\alpha,k} \quad \text{for } l = 1, \ldots., n\text{-}l \quad (15)$$

is fulfilled.

If the above inequality is not satisfied for any row l-th then the $H_0$ hypothesis should be rejected ; the set is nonhomogeneous. In this case the alternative hypothesis is assumed:

$$H_1 : \quad \begin{array}{l} \delta(S_1,S_2) \ne 0 \\ U(l,n-l) > \chi^2_{\alpha,k} \end{array} \qquad (16)$$

The partition of nonhomogeneous data set into homogeneous, disconnected subsets is based on the acceptation of the alternative nonhomogeneous hypothesis ( $H_1$ hypothesis ).

From the maximum likelihood method follows that the $H_1$ hypothesis can be accepted if the L function achieves its maximum value. Thus, the exponent occurring in this function should be minimized.

By appropriate transformations the minimization problem leads to the maximization of statistics U( l, n-l ) and it can be written as follows:

$$\max_{l} U(l,n-l) \qquad (17)$$

From the condition (17) it follows that the likelihood function achieves the maximum for such a partition of nonhomogeneous data set into two parts, that gives rise to maximum of statistics U( l, n-l ).

The condition (17) establishes the theoretical base of the method for set partitioning into two homogeneous, separated subsets.

### *1.2 The subsets aggregation - hypothesis on unstable intergroup boundaries*

The presented method of set division consists in the iterative partitions of nonhomogeneous set into two parts. If number of these partitions ( iteration number ) increase that successive divisions can cause the existence of statistically unstable boundaries between adjacent homogeneous subsets.

The statistic stability of boundaries between subsets can be examined by comparing multi- dimensional means. If multi- dimensional means of two compared groups are statistically equivalent than it can be assumed, that the statistically unstable boundary exists, so the boundary can be removed and the sets can be aggregated into one homogeneous set.

The examination of the intergroup boundaries stability leads to verification of suitably formulated hypothesis.

Let $m_i$ - denote multi- dimensional mean of i-th subset.( i = 1,...., M )

The hypothesis $H_0$ on statistic unstable intergroup boundary has the form :

$$H_0 : \quad m_i - m_{i+1} = \{ 0, 0, ......, 0 \} \quad (18)$$

If the $H_0$ hypothesis holds, than the boundary between two sets is statistically unstable and the sets can be connected.

The rejection of the $H_0$ hypothesis of the form (18) enables acceptation of the alternative hypothesis ( on boundaries stability ) and it means that there exist essential stable intergroup boundaries.

The examination of boundaries can be performed successively for all considered subsets.

A criterion function for the $H_0$ hypothesis testing is constructed using the maximum likelihood method. As the result, one obtains the U statistics of the form :

$$U(S_i, S_{i+1}) = \frac{\frac{n_i + n_{i+1} - 1}{(n_i + n_{i+1}) n_i n_{i+1}} \sum_{j=1}^{k} \left( n_{i+1} \sum_{s \in S_i} x_{s_j} - n_i \sum_{s \in S_{i+1}} x_{s_j} \right)^2}{\sum_{j=1}^{k} \left( \sum_{s \in S} x_{s_j}^2 - \frac{1}{n_i + n_{i+1}} \left( \sum_{s \in S} x_{s_j} \right)^2 \right)}$$

$$S = S_i \cup S_{i+1} , \quad i = 1, ...., M \quad (19)$$

where :

M - the number of subsets

k - the number of considered characteristics

$n_i$ , $n_{i+1}$ - number of elements of $S_i$ , $S_{i+1}$ , respectively

It can be proved that under some assumptions the $H_0$ hypothesis testing leads to inequality of the form:

$$U(S_i, S_{i+1}) \leq \chi^2_{\alpha,k} \quad i = 1, ...., M \quad (20)$$

If the condition (20) holds, than one can assume that the boundary between subsets $S_i$ and $S_{i+1}$ is unstable.

In this situation these subsets can be interconnected and the hypothesis of the stability between subsets ( $S_i \cup S_{i+1}$ ) and $S_{i+2}$ is tested.

If the condition (20) is not satisfied then the boundary between sets $S_i$ and $S_{i+1}$ is stable ( essential ) and these sets can not be connected.

## 2. MODELLING REGIONAL STRUCTURE

The presented algorithm consists of two steps:

a) The initial partition of input set into separate ( disconnected ) homogeneous subsets. For this purpose one tests the homogeneity hypothesis of the form (15) ; if $H_0$ is rejected, then the nonhomogeneity hypothesis (16) is examined. The $H_1$ hypothesis testing leads to the U statistics maximization problem (17).

b) The examination of stability of the boundaries between subsets obtained by the initial partition. It results in testing hypothesis on the boundary stability. Depending on the case whether the $H_0$ hypothesis of the form (20) is true or not, the two adjacent subsets can be connected or remain disjoined.

The presented algorithm was used for selecting the homogeneous groups of vojvodships in Poland described by a set of characteristics.

It means, that the vojvodships belonging to the same group have similar feature; for example the socio-economic characteristics or development dynamic characteristics. The selection of the characteristics describing vojvodships depends on applications.

The algorithm was used for analyzing the Polish regional structure for 3 years - 1998, 2000, 2001.

In 1999 administration reform was performed and the new regional structure was introduced.

Until 1999 the Polish regional structure consisted of 49 vojvodships and after new partition there are 16 vojvodships.

The following cases were considered :

a) the vojvodships described by 3 characteristics (population, employment, investments )

b) the vojvodships described by 5 characteristics (population, employment, investments, industry production, number of flats )

c) the vojvodships described by 9 characteristics (population, employment, investments, industry production, number of flats, agricultural grounds , construction production, unemployment, accommodation )

The analyzed statistical data concern 3 years: 1998 – when Poland territory consisted of 49 vojvodships and 2000, 2001 – when Poland territory contained 16 vojvodships.

In all cases the number of people is assumed to be the most significant characteristic. The $\chi^2$ distribution value for 3 degrees of freedom equals 7.815 ; for 5 degrees of freedom it is equal to 11.07 and for 9 degrees of freedom equals 16.919. The significante level $\alpha = 0.05$.

The input data set is ordered increasingly with regard to the significant feature assumed.
The input data set is represented as matrix with: the number of rows corresponding to the number of vojvodships and the number of columns corresponding to the number of characteristics considered.
For each year the partition of input data set into homogeneous, separable groups of vojvodships were obtained.

° for 1998 for two cases ( 3 characteristics and 5 characteristics ) after 10 iterations one obtains the complete partition of input data set into 11-th homogeneous groups of vojvodships. The subsets obtained from 1-st step procedure have stable boundaries.

For 9 characteristics after 11 iterations one obtains the initial partition of input data set into 12-th homogeneous, disconnected subsets ( 1-st step of algorithm )
From the analysis of stability of the boundaries between subsets ( step 2 -nd algorithm) follows that the subsets number 2, 3 can be connected ; the boundaries between subsets are unstable. Finally one obtains 11-th groups of vojvodships.

From the analysis of obtained subsets results that number of characteristics has not caused any significant change in number of subsets ; only the attachment of particular vojvodships to groups has varied
Tables given below present partition of Poland territory for 1998 year into homogeneous, separable groups of vojvodships described by 3, 5 and 9 characteristics respectively.

Table 1. Partition of 49 vojvodships into groups 1998- 3 characteristics

| **1- subset** | **2- subset** | **3- subset** |
|---|---|---|
| Chelmskie | Leszczynkie | Koninskie |
| BialskoPodlaskie | Sieradzkie | Suwalskie |
| Lomzynskie | Ostroleckie | Zamojskie |
|  | Przemyskie | Elblaskie |
|  | Skierniewickie | Pilskie |
|  | Slupskie | Krosnienskie |
|  | Wloclawskie |  |
|  | Ciechanowskie |  |

| **7 - subset** | **8 - subset** | **9 - subset** |
|---|---|---|
| Nowosadeckie | Bielskie | Lubelskie |
| Rzeszowskie | Szczecinskie | Lodzkie |
| Radomskie | Opolskie | Kieleckie |
| Olsztynskie |  | Wroclawskie |
| Czestochowskie |  | Bydgoskie |

| **10 - subset** | **11 -subset** |
|---|---|
| Krakowskie | Warszawskie |
| Poznanskie | Katowickie |
| Gdanskie |  |

From the analysis of obtained results for 3 and 5 characteristics one can perform following conclusion:

- for 16 vojvodships belonging to subsets 1, 8 9, 10, 11 increasing number of characteristics has not caused any changes in the attachment of particular vojvodships to groups.
- for 2 – subset in the case of 5 characteristics follows partition of this group into two subsets, which are given in below:

Table 2. 1998 - 5 characteristics. The partition subset number 2 into two groups

| **2 - subset** | **3 - subset** |
|---|---|
| Leszczynki | Skierniewickie |
| Sieradzkie | Slupskie |
| Ostroleckie | Wloclawskie |
| Przemyskie | Ciechanowskie |

- for the 5 characteristics Gorzowskie vojvodship belong to 3-group
- for 5 characteristics the changes occur in subsets 5, 6, 7 and it is presented in Table 3.

Table 3. 1998- 5 characteristics. The subset in which occur the changes

| **5 - subset** | **6 - subset** | **7 - subset** |
|---|---|---|
| Plockie | Tarnobrzeskie | Bialostockie |
| Jeleniogorskie | Piotrkowskie | Kaliskie |
| Legnickie | Siedleckie | Walbrzyskie |
| Koszalinskie | Torunskie | Nowosadeckie |
|  | Zielonogorskie | Rzeszowskie |
|  | Tarnowskie | Radomskie |
|  |  | Olsztynskie |
|  |  | Czestochowskie |

- for 9 characteristics the changes occur in subsets number 1, 2, 3, 4, 5, 6, 7 and it is presented in Table 4.

Table 4. 1998- 9 characteristics. The subset in which occur the changes

| **1 - subset** | **2 - subset** | **3 -subset** |
|---|---|---|
| Chelmskie | Lomzynskie | Koninskie |
| BialskoPodlaskie | Leszczynskie | Suwalskie |
| | Sieradzkie | Zamojskie |
| | Ostroleckie | Elblaskie |
| | Przemyskie | Pilskie |
| | Skierniewickie | Krosnienskie |
| | Slupskie | Gorzowskie |
| | Wloclawskie | |
| | Ciechanowskie | |

| **4 - subset** | **5 - subset** | **6 - subset** |
|---|---|---|
| Plockie | Tarnobrzeskie | Walbrzyskie |
| Jeleniogorskie | Piotrkowskie | Nowosadeckie |
| Legnickie | Siedleckie | |
| Koszalinskie | Torunskie | |
| | Zielonogorskie | |
| | Tarnowskie | |
| | Bialostockie | |
| | Kaliskie | |

| **7 - subset** |
|---|
| Nowosadeckie |
| Rzeszowskie |
| Radomskie |
| Olsztynskie |
| Czestochowskie |

° for the years 2000 and 2001 ( after administration reform ) homogeneous groups with 3, 5 and 9 characteristics considered were taken into account. The of characteristics did not influence on the number of homogeneous groups.

For the years 2000 and 2001 4 homogeneous groups with stable boundaries between subsets were obtained. Table 5 present partition of Poland territory for years 2000 and 2001 into homogeneous, separable groups of vojvodships described by appropriate 3, 5 and 9 characteristics.

Table 5. Partition of 16 vojvodships into groups – for the years 2000, 2001 for 3 , 5 and 9 characteristics

| **1 - subset** | **2 - subset** | **3 - subset** | **4 - subset** |
|---|---|---|---|
| Lubuskie | Zachodnio-Pomorskie | Lodzkie | Slaskie |
| Opolskie | Kujawsko-Pomorskie | Dolnoslaskie | Mazowieckie |
| Podlaskie | Podkarpackie | Malopolskie | |
| Świetokrzyskie | Pomorskie | Wielkopolskie | |
| Warminsko-Mazurskie | Lubelskie | | |

It can be seen that increasing number of characteristics has not caused any significant change in number of subsets ; only the attachment of particular vojvodships to groups has varied.

The determined vojvodships groups form regional structures with similar characteristics.

## 3. REFERENCES

Fisz, M. (1967). Rachunek prawdopodobieństwa i statystyka matematyczna. PWN, Warszawa.

Kildyshev, G.S., and J.A. Abolentzev (1978). Mnogomernye gruppirovki. Izd. Statistika , Moskva

Mardia, K.V., J.T. Kent and J.M Bibby (1979). Multivariate Analysis. Academic Press,, London

Hołubiec, J. and G. Petriczek (1997). Homogeneity Algorithm For Modeling Regional Structure. In: *Proceedings of the International Conference on Methods and Models in Automation and Robotics MMAR* (S. Domek, Z. Emirsajłow, R. Kaszyński, Ed.). Vol.1, pp. 397-402. Technical University of Szczecin, Szczecin

Statistical Year-book 1998, 2000, 2001. Główny Urząd Statystyczny, Warsaw