TEKA Kom. Mot. Energ. Roln. – OL PAN, 2008, 8, 35–42

UTILIZATION OF AGGLOMERATION ALGORITHM (TREE CLUSTERING) IN RESEARCH ON CRIMINALITY IN MOTOR TRANSPORT IN POLAND

Zbigniew Burski

University of Life Sciences in Lublin, Faculty of Production Engineering, Department of Power Industry and Vehicles

Summary. In this thesis there are introduced theoretical basics and utilization opportunities of an agglomeration method (tree clustering) in researches on criminality in motor transport in Poland. It concerns a period of intensification of government transitions in the country, in which the problems of property larceny, including cars, have considerably grown. Application of the mathematical, statistical method enabled to acquire an objective image of this research issue in a multidimensional form.

Key words: motor transport, car utilization, mathematical statistics, regionalization of criminality.

INTRODUCTION

The role and meaning of regional development programming and the involved logistic procedures is systematically increasing. The comprehensive regional policy has acquired a special meaning after the day when Poland became a member of the European Union. A significant meaning in every regional policy, no matter of its character and preferable directions of development, is involved in the regions' demarcation – stating a geographical area which this policy would concern. On the other hand, in regional strategic planning the basic thing is to know the demographic, economic, cultural and linguistic potential of the area, which affects the basis of multidimensional economic development and the involved threats.

In the literature on agglomeration issues a special role is played by the issue of infrastructure of technical backup of rural areas. [Gruszczynski et. al. 2000, Sieminski 2000, Kwapisz 2001]. According to these authors, technical infrastructure in its multidimensional character is a potential development basis and plays a great role in the strategic planning of rural areas. There were designated regions of the country with a probability of appearance of such infrastructure, and they have since been an object of regional policy [Wozniak 1997].

PURPOSE OF THE THESIS

The purpose of this thesis is an analysis of motor crime, which is a significant threat for economic regional development [Siemaszko i in. 1999]. In the conducted analysis there was used

an agglomeration method type tree clustering as multidimensional, mathematical criteria of significance of its appearance.

DELIMITATION OF REGIONS AS A RESEARCH ISSUE IN THEORY AND EXPLOITATION PRACTICE

The general problem of regionalization consists in the fact that having some particular number "n" of spatial units with some particular number "m" of features we want to designate smaller number of regions which are constant units and maximally inner uniform in aspect of the set of these features [Analiza skupień, 1999].

The method of resolving the issue of regionalization has a character of some taxonomy – an algorithm of a universal character, which takes into account three essential activities:

– development of an information matrix,
– estimation of probability – taxonomic distances,
– grouping of units – considering condition of continuity and spatial adherence.

The starting point for regionalization procedure is spatial series or geographical matrix – some ordered set of events. In the matrix for each spatial unit there is one line and for each feature – one column. Each element of this matrix is a geographical event. The event in regionalization process is e.g. infrastructure of a particular size (intensity) and localization (e.g. number of households using collective water supply system in particular communities of the region). Such spatial series is a basis information matrix, which is a base of taxonomic procedures of educing similar units (classification).

Region, understood ex definitione as some compact area uniform in some particular criteria, is a class of basic spatial units. In this sense, regionalization, understood as a process of educing classes, is made by grouping or logical division.

Between regionalization and classification there are some significant similarities, however procedures and concepts of two disciplines are not in explicit adequacy. Regionalization, according to its object of research and scientific purpose, is inherently geographical and only in formal aspect identified with classification. Classification, as a more precise, better-known and better-organized activity can be a theoretical model of regionalization. Acceptance of such an attitude entitles to consider regionalization in terms of classification.

An essential difficulty encountered by methods and conceptions of regionalization as classification procedure is a problem of localization (locus) of spatial units. In regionalization, as distinct from classification, apart from the properties of the basic units, their spatial distribution is also considered. Localization (locus) cannot be a distinctive feature in classification because it is a unique feature of each particular spatial unit.

However, the most important issue in regionalization – grouping of spatial units, is estimation of their multidimensional similarity.

Searching for similarity between units from the viewpoint of many descriptive features and their grouping is a classic issue of numeric taxonomy, especially its part concerning cluster analysis.

Most of taxonomic procedures are based on adequately defined measures of distance (or similarity) between classified units (objects). Stating of measures, which are needed to set distances (similarities) between units in multi-dimensional spaces as well as methods and procedures of consolidating similar units in groups, is an area of interest for many researches from various disciplines [Woźniak 1997, Siemaszko i in. 1999, Burski i in. 2003].

The choice of classification procedure implicates the way of defining taxonomic similarity. A well-developed theory from the area of taxometry presents a wide choice of classification

methods of multidimensional objects. The most frequently applied are: agglomeration methods of hierarchical grouping (the nearest neighborhood, median, centre of gravity, group mean or J.H.Ward method) and from the group of iteration-optimization methods – method of k-means [Ward 1963, Analiza skupień 1999].

CONCEPT OF UTILIZATION OF AGGLOMERATION METHODS IN REGIONALIZATION

Aforementioned agglomeration procedures, especially dendrite Ward's method, became a basis of an attempt of creating own classification concept of motor crime based on these methods as well as on practical insights of other authors. Regionalization as a procedure of educing homogeneous and compact regions consists in making classification and grouping of objects.

A starting point in the assumed concept is two matrixes of distance. The first matrix, built on the basis of spatial series is matrix $\left|D^e\right|_{nxn}$. Elements of this matrix are taxonomic distances of grouped items in multidimensional space of their properties. In other words, it is a matrix of mutual similarities between objects.

These similarities can be specified by an ordinary Euclidean metric designate from the formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{r}(x_k - x_{jk})^2}\text{ , dla } (i,j = 1,2 \ldots n) \text{ i } (k = 1,2 \ldots r), \tag{1}$$

where:
d_{ij} – measure of similarity of taxonomic distance from space item i to item j,
x_{ik}, x_{jk} – realization of k feature for items i and j.

Elements of the second matrix $\left|D^e\right|_{nxn}$ are geodetic distances between the spatial items d_{ij}^g analysed by regionalization method, which can be stated as voivodships, counties or communes, so their geographical distances can be designated straight from the map. This approach was used e.g. in researches presented in the thesis [Wozniak 1997].

Therefore, on the basis of digitalization of the Malopolskie voivodship map, to each object (commune) there were allocated geographical coordinates X_i and Y_i. On the basis of these coordinates geometric distances were designated between communes, using a general formula:

$$d_{ij}^g = \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)}\text{ , dla } i < j, \tag{2}$$

where:
d_{ij}^g – measure of geographical distance from spatial item i to item j,
X_i, Y_i – geographical coordinates of spatial item i.

The matrix of geographical distance $|D^g|$ constructed in this way makes a matrix of n verses and n columns, in which on the main diagonal there are only zeros and above these, diagonal elements of matrix d_{ij}^g which convey a geometric (geographical) distance between i and j commune.

These two matrixes $|D^e|$ and $|D^g|$ are similar, both have similar positions of verses and columns as they refer to the same spatial units – they have some special properties.

In matrix $|D^e|$ the lower absolute value of an element d_{ij}^e, the more similar the objects are. And vice versa, the higher the absolute value, the less similar the objects are.

Also in matrix $|D^g|$ elements d_{ij}^{e} refer to physical nearness between objects – the lower the value of means, the closer the neighborhood is.

The undertaken effort [3] of regionalization of rural regions of Malopolskie voivodship on account of infrastructure, according to the definition of a region takes into account also taxometric similarity and geographical nearness of units in order to gain consistency and continuity of regions. It is needed to find such a matrix $|D^*|$ which will be in the middle between two matrixes $|D^e|$ and $|D^g|$, treating matrix of physical distances as a system of weights for taxonomic distances falling within agglomeration. Symmetry of matrixes $|D^e|$ and $|D^g|$ enables to make simple algebraic operations in order to transform these matrixes in a matrix $|D^*|$ following the pattern:

$$|D^e| \cdot |D^g| = |D^*| \tag{3}$$

It means that elements of a matrix $|D^*|$ are designated by multiplying each element d_{ij}^{e} of matrix $|D^e|$ by responding to its element d_{ij}^{e} in matrix $|D^g|$.

The property of $|D^*|$ is then strengthening the similarity of units adhering spatially against the remaining units.

The transformed matrix of taxonomic distances $|D^*|$ obtained in this way is introduced as an input matrix of data for the procedure of hierarchic grouping by any agglomeration procedure.

From the group of hierarchic methods of motor crime analysis there was chosen the Ward method, well-described theoretically and providing the best agglomeration results of the analysed objects of research [Burski et.al. 2003].

CHARACTERISTICS OF THE ASSUMED METHOD OF CALCULATIONS OF REGIONALIZATION OF MOTOR CRIME IN POLAND

From the presented agglomeration methods of hierarchical grouping and methods of iteration-optimization for the purposes of this thesis the method of full bond-furthest neighbour was used.

In the method of full bond the distance between clusters is determined by the largest of the distances between any two objects depending on different clusters (that is the furthest neighbors). It was used because of a possibility of having a great diversification between groups of objects (voivodships or macroregions). In the assumed agglomeration method the starting point is a matrix of distances built on the basis of spatial series, the elements of which are taxometric distances of grouped units in the multidimensional space of their properties. In other words, it is a matrix of mutual similarities between objects. These similarities can be stated by an ordinary Euclidean metric designated from the formula (1).

From the point of view of algorithm of merging it does not matter if distances which are set are real or only derived from the measure important for the analysis.

RESULTS OF RESEARCHES ON CLUSTERING OF SAFETY OF VEHICLES' EXPLOITATION AND UTILIZATION IN MACROREGIONS OF POLAND

In order to state clustering of safety and utilization of vehicles' according to EU standards, there were made calculations of indicators of larceny growth as two- and three-sort classification and quotient of larceny for 9 macroregions of Poland (according to GUS). In table 2 there are presented statistical data of police Headquarters and GUS, calculations of quotients and classifications

of macroregions according to safety of vehicles' utilization – in accordance to art. 9, ust. 1 of Act from 21 of June 1996 about agency of Home and Administration Secretary (Dz. U. nr 106, poz. 491 and Dz. U. from 1997 Nr 70, poz. 443).

Table 1. Results of classification of macroregions of Poland according to standards of two- and three-sort classification

Macro region	Classification					
	two-sort			three-sort		
	1995	1996	1997	1995	1996	1997
south	B	B	B	B	B	B
south-east	B	B	B	B	B	B
south-west	B	A	A	C	C	C
north-east	A	A	A	C	C	C
north	A	A	A	D	D	D
capital	A	A	A	C	C	C
central	B	B	B	B	B	B
central-east	B	B	B	B	B	B
central-west	A	A	A	C	C	C

Table 2. a i b. Results of calculations of car larceny quotients and classification of macroregions of Poland

a)

Macro region	lost				cancelled			
	1995	1996	1997	1998	1995	1996	1997	1998
south	6269	6702	8127	7973	2456	2270	3062	3317
south-east	4283	4347	5008	5582	1807	2039	2233	2255
south-west	5273	6036	6719	6261	2187	3318	3499	2995
north-east	2629	3279	4036	3610	824	1443	2046	1946
north	8283	9504	11553	10982	2527	3252	4027	4134
capital	13967	12287	13671	14792	5827	4818	4443	3959
central	6750	6458	7577	9230	1868	2203	3047	3202
central-east	1024	961	1399	1339	576	785	992	1170
central-west	6072	5640	6890	7964	574	912	2437	3264
total	54550	55214	64980	67733	18646	21540	25786	26242

b)

Macro region	cancelled			larceny quotient (%)		
	1995	1996	1997	1995	1996	1997
south	1508100	1600100	1684800	0,416	0,419	0,882
south-east	1265600	1339900	1430000	0,338	0,325	0,350
south-west	879200	956400	1020200	0,600	0,631	0,659
north-east	405600	430100	460100	0,648	0,762	0,877

north	824100	876100	952900	1,005	1,085	1,212
capital	1458600	1560600	1585900	0,958	0,787	0,862
central	1465800	1566800	1654400	0,460	0,412	0,458
central-east	410500	447300	483700	0,249	0,215	0,289
central-west	739000	7940800	830300	0,822	0,710	0,830
total	8956500	9571100	10102300	0,609	0,577	0,843

Larceny quotient:

$$w.k. = \frac{P_u(szt.)}{P_r(szt.)}, \tag{4}$$

where:
P_u – number of cars lost (pcs),
P_r – number of cars registered (pcs).

In Table 1 there are presented calculations of larceny quotient and preliminary classification according to the rule*: w.k.* < 0,5 % => B, *w.k.* 0,5%-1% => C, *w.k.*> 1% => D and classification according to the rule: *w.k.* above mean for all country => A, *w.k.* below mean for all country => B.

Preliminary classification enabled educing macroregions of Poland of decreased safety of utilization and exploitation of vehicles (A) in a global perspctive, which are the following: capital, north, north-east, middle-west and south-west and regions of decreased larceny threat (B) i.e. regions: central, central-east, south and south-east.

A more detailed classification (3 grades – B, C, D) has shown that the north macroregion is the most threatened by car larceny region.

Table 3. Matrix of Euclidean distance of car larceny quotient

	south	south-east	south-west	north-east	north	capital	central	central-east	central-west
south	0,000000	0,001797	0,003316	0,005727	0,011506	0,007573	0,000514	0,003264	0,006083
south-east	0,001797	0,000000	0,005073	0,007518	0,013288	0,009270	0,001849	0,001539	0,007821
south-west	0,003316	0,005073	0,000000	0,002595	0,008227	0,004402	0,003280	0,006576	0,002911
north-east	0,005727	0,007518	0,002595	0,000000	0,005863	0,003108	0,005776	0,008969	0,001874
north	0,011506	0,013288	0,008227	0,005863	0,000000	0,004621	0,011481	0,014765	0,005664
capital	0,007573	0,009270	0,004402	0,003108	0,004621	0,000000	0,007423	0,010758	0,001598
central	0,000514	0,001849	0,003280	0,005776	0,011481	0,007423	0,000000	0,003346	0,005976
central-east	0,003264	0,001539	0,006576	0,008969	0,014765	0,010758	0,003346	0,000000	0,009298
central-west	0,006083	0,007821	0,002911	0,001874	0,005664	0,001598	0,005976	0,009298	0,000000

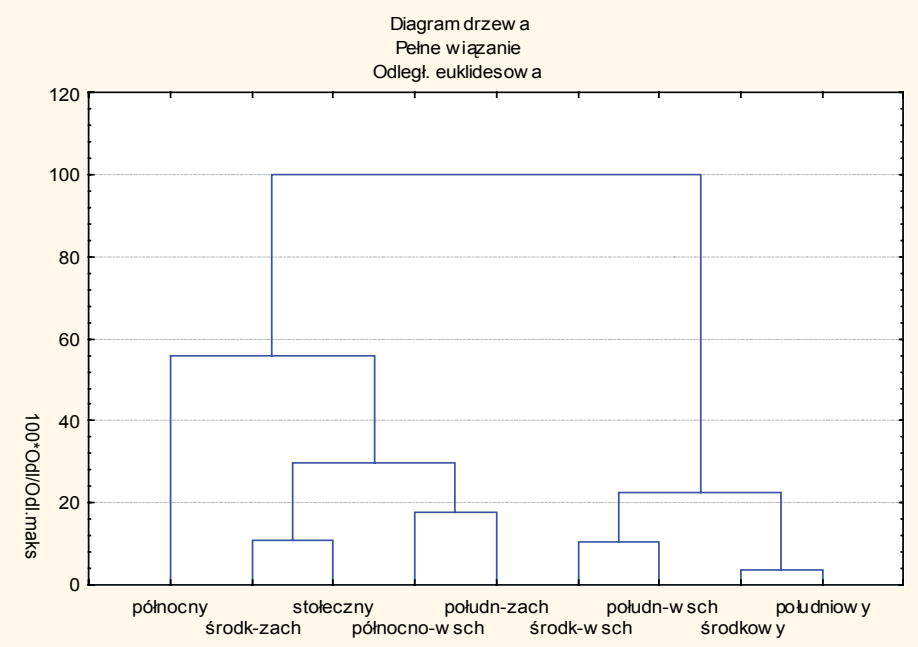

Fig. 1. Diagram of full bond of standardized Euclidean distance of car larceny quotient in macroregion apprehension of EU

Results of calculations of macroregions clustering on the basis of standardized Euclidean distance are presented in Table 3. In Fig. 1 is presented a diagram of full bond of Euclidean distance of car larceny quotient in macroregion apprehension. According to preliminary classification, the method of standard Euclidean distance has affirmed that there are two clusters of macroregions and that there was educed the north macroregion.

CONCLUSION

1. According to standards of EU regarding estimation of steady development of macroregions in this thesis there were used the agglomeration method tree clustering occurring in cluster analysis. Methods of cluster analysis were applied in exploration phase of these researches, when testing of statistical significance is not actually necessary.
2. Preliminary two-grade classification (traditional) has proved to be too general (global). Application of three-grade classification has shown the region which is the most threatened by motor crime – north macroregion.
3. According to preliminary (traditional) classification the method of standardized Euclidean distance has affirmed the existence of two clusters of macroregions of less and more crime and visible educing of north macroregion.
4. Further researches will be continued in the direction of an influence of social engineering and cultural factors on the phenomenon of car larceny in particular macroregions of Poland.

REFERENCES

Analiza skupień. 1999. Copyright StatSoft, SKU 3159.

Burski Z., Burski P. 2003. *Karnoprawna analiza kradzieży pojazdów samochodowych o charakterze zbójniczym, ze wspólnym uwzględnieniem obszarów wiejskich*. Wyd. KMiER, PAN Oddział w Lublinie, s. 11-23.

Gruszczyński J., Kwapisz J. 2000. *Stan infrastruktury technicznej Małopolski na tle kraju*. Zeszyty Naukowe AR w Krakowie, z. 72, nr 365. Kraków.

Kwapisz J. 2001. *Mierniki lokalizacji i koncentracji infrastruktury obszarów wiejskich w badaniach przestrzennych*. Inżynieria Rolnicza, nr 9, Warszawa, s. 329-356.

Siemaszko A., Gruszczyńska B., Marczewski M. 1999. *Atlas przestępczości w Polsce*. Warszawa.

Siemiński J. 2000. *Infrastruktura techniczna obszarów wiejskich w koncepcji rozwoju zrównoważonego i trwałego kraju*. MKN-T Mat. Konf. Kielce

Ward J.H. 1963. *Hierarchical grouping to optimize and objective function*. Journal, The America Statistical Association, nº 58.

Woźniak A. 1997. *Infrastruktura techniczna w lokalnych strategiach rozwoju regionalnego*. Zeszyty Problemowe Postępów Nauk Rolniczych, z. 445, s. 181-194.