STATISTICS IN TRANSITION-new series, December 2012
Vol. 13, No. 3, pp. 581—600

SYMBOLIC APPROACH IN REGIONAL ANALYSES

Justyna Wilk[1]

ABSTRACT

Regional studies cover a spectrum of diversified phenomena and problems including social, economic and environmental ones, which refer to territorial units. Owing to their specific characteristics they are most frequently of both multivariate and complex nature. Conducting regional research is associated with the need to consider such difficulties as large data sets, insufficient precision of phenomena description, disregarding territorial diversification of a given phenomenon, as well as incomplete description of problems.
The objective of the paper is to suggest solutions to these problems by means of symbolic approach application which basically consists in presenting phenomena in the form of symbolic data. The first part of the paper discusses specific nature of symbolic data, methods for collecting symbolic data and methods for these data analysis. The second part presents an empirical example referring to the assessment of labour market situation in Polish regions (NTS-2) using symbolic data and cluster analysis.

Key words: symbolic approach, symbolic data analysis, cluster analysis, regional research, labour market.

1. Introduction

Regional studies cover an overall spectrum of diversified phenomena and problems including social, economic and environmental ones, referring to territorial units (TU) at international, national, regional and local level. Owing to their specific features they are most frequently of multivariate and complex nature.

Conducting regional studies is associated with the need to consider such problems as:

1. Large data set (many objects and variables) or incomplete description of the studied phenomena due to the description reduction requirement. For example, the research of local communities living standard cover a large group of TU and a

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, 58-500 Jelenia Góra, Nowowiejska 3. E-mail: justyna.wilk@ue.wroc.pl.

significant spectrum of problems, including the financial situation of population, public safety level, natural environment condition, etc.

2. Insufficient precision of phenomena description and disregarding territorial diversification of a given phenomenon, in particular at higher territorial division levels. It happens as the result of data generalization, among others, due to descriptive statistics application (e.g. median, mean value). Registered unemployment rate may serve as the example, which in 2011 was specified at the level of 9.9% for Mazowieckie region (NTS-2), while in districts (NTS-4) covered by this region the range of registered values ranged from 3.8 up to 37.2%.

The objective of the paper is to suggest solutions to these problems by means of symbolic approach application which basically consists in presenting phenomena in the form of symbolic data. The application of symbolic data in regional studies allows for:

- description of higher level territorial units (e.g. NTS-1) including the situation of lower level units (e.g. NTS-4),
- characteristics of territorial units (e.g. NTS-2, NTS-5) applying not only metric and non-metric variables but also interval-valued, multivalued and modal variables.

The first part of the paper discusses the specific nature of symbolic data, methods for collecting symbolic data and methods for these data analysis. The second part presents an empirical example referring to the assessment of labour market situation in Polish regions (NTS-2) using symbolic approach and cluster analysis.

2. Symbolic data specification

The approach most often applied in statistical analysis assumes that variables describing the set of objects are of metric nature (represented by a single value for each object) and/or non-metric (represented by a single category for each object) – see e.g. (Walesiak, 1993, Mynarski, 2000).

Symbolic data analysis (SDA) offers the possibility of including symbolic variables which take the form of (see Bock, Diday et al., 2000, Billard and Diday, 2006, Diday, Noirhomme-Fraiture et al., 2008):

- intervals of values, disjoint or non-disjoint (interval-valued variables),
- sets of categories or values (multivalued variables),
- sets of categories (values) with weights, frequencies, probabilities, etc. (modal variables).

The objects described by means of symbolic variables are defined as symbolic objects. With regard to aggregation level the following objects can be specified:

- first order symbolic objects which constitute objects following classical approach, i.e. primary units of the study (e.g. region, country, location)

which are described by means of symbolic and/or classical approach variables,
- second order symbolic objects which result from the aggregation of a set of first order objects (e.g. region made up of districts located in its territory) and they are described by symbolic variables only.

The set of observations referring to symbolic data is entered into a symbolic data table.

3. Symbolic data sources in regional studies

3.1. Aggregation of objects

If symbolic approach is applied the set of classical data (objects described by metric and non-metric data) may be subject to aggregation process as the result of which symbolic data set is obtained. Based on the available literature studies two methods of data aggregation may be distinguished:

1. Aggregation of objects (territorial units) (see Bock, Diday et al., 2000, Diday, Noirhomme-Fraiture et al., 2008, Wilk, 2010b, pp. 86-88).
2. Aggregation of variables (descriptive characteristics for territorial units) (see Wilk, 2010b, pp. 88-91, Wilk, 2011).

The aggregation of objects consists in the representation of lower order objects by means of higher order objects, which results in obtaining second order symbolic objects. Such procedure is carried out to:
- reduce the description due to a very large set of objects, in particular local level territorial units (e.g. NTS-4, NTS-5),
- refine the description of higher order territorial units (e.g. NTS-0, NTS-2), i.e. consider their territorial diversification.

In this case symbolic data for a higher order territorial unit are obtained based on descriptive statistics values (e.g. minimum, maximum, frequency, etc.), assigned to lower order territorial units.

Interval-valued variables for a higher order territorial unit are obtained by specifying minimum and maximum values based on values obtained by lower order units. For example, on the basis of average gross monthly salary earned in sub-regions a synthetic variable may be obtained which illustrates salary spread interval in regions (see Table 1). In a similar way the following may be obtained:
- GDP *per capita* values interval in Polish regions (NTS-2), defined based on values obtained by sub-regions (NTS-3),
- unemployment rate size interval in EU countries (NTS-0), defined based on values obtained by NTS-4 level units.

Table 1. Average monthly gross wages and salaries in Poland in 2010 (PLN)

NTS-1 (region)		NTS-3 (sub-region)		Descriptive statistics for NTS-3		Symbolic interval-valued variable for NTS-1
Name	Value	Name	Value	Minimum value	Maximum value	
Central	4025.26	Łódzki	2705.93	2666.84	4694.47	[2666.84, 4694.47]
		City of Łódź	3243.15			
		Piotrkowski	3297.76			
		Sieradzki	2666.84			
		Skierniewicki	2810.76			
		Ciechanowsko-Płocki	3366.54			
		Ostrołęcko-Siedlecki	2950.04			
		Radomski	3054.40			
		City of Warsaw	4694.47			
		Warszawski Wschodni	3209.11			
		Warszawski Zachodni	3657.19			
Eastern	2987.67	Bialski	2841.15	2650.41	3410.28	[2650.41, 3410.28]
		Sandomiersko-Jędrzejowski	2982.23			
		. . .	. . .			

Source: Author's estimation based on data provided by Local Data Bank of the Central Statistical Office in Poland.

The construction of symbolic modal variable consists in defining the variable category (e.g. low, medium, high level; very favourable situation, moderately favourable, unfavourable, very unfavourable) and in specifying the percentage of objects meeting the criterion data. For example, in case of unemployment rate in Poland at NTS-4 level the following variable categories were defined:

- less than 10% – low level of unemployment,
- 10-20% – medium level of unemployment,
- more than 20% – high level of unemployment.

Next, the percentage of districts in each region meeting this criteria was calculated (see Table 2). Instead of defining levels the national average may be accepted as the determinant and on this basis it is possible to calculate the percentage of districts in a region ranked above and below national average level. In a similar way other economic aspects may be interpreted (e.g. average gross monthly salary), as well as environmental ones (e.g. the emission of gaseous air pollution from particularly noxious plants per 1 km^2), etc.

Table 2. Registered unemployment rate in Poland in 2010 (%)

NTS-2 (voivodship/region)		NTS-4 (poviat/district)		Fraction of poviats/districts (NTS-4) due to phenomenon level			Symbolic modal variable for NTS-2
Name	Value	Name	Value	low	medium	high	
Mazowieckie	9.7	City of Warszawa	3.5	0.2	0.5	0.3	{low (0.2), medium (0.5), high (0.3)}
		Warszawski Zachodni	5.9				
		. . .	. . .				
		Radomski	30.8				
		Szydłowiecki	36.0				
Kujawsko-Pomorskie	17.0	City of Bydgoszcz	8.0	0.1	0.3	0.6	{low (0.1), medium (0.3), high (0.6)}
		⋮	⋮				
		Lipnowski	28.9				

Source: Author's estimation based on data provided by Local Data Bank of the Central Statistical Office in Poland.

3.2. Aggregation of variables

The second method for obtaining symbolic data is the aggregation based on variables describing the set of objects and, in fact, categories of these variables. Following this approach the territorial level of analyzed units is left unchanged, however, the form of describing phenomena is altered. The procedure consisting in variables aggregation may be used in situations when there is the need for:

- size reduction owing to an extensive set of variables describing territorial units at different level (e.g. counties, NTS-2, NTS-4, cities),
- synthetic description of the analyzed phenomena.

In this case the symbolic variable is constructed as the result of variable category combination following classical approach and, in particular, non-metric variable, e.g. variable presenting the availability of sport objects (swimming pools, tennis courts, sports grounds, etc.), the structure of monthly household expenditure, the occurrence of natural resources (lignite, rock salt, kaolin, etc).

Table 3 illustrates the method for multivalued variable construction which describes the availability of tertiary education institutions in Polish regions by the selected education type. For example, in Podkarpackie region only technical universities are available out of the analyzed educations profiles.

Table 3. Higher education institutions by type, in 2010 (facilities)

NTS-2 (voivodship / region)	Number of institutions (universities, academies) by type					Symbolic multivalued variable for NTS-2
	technical	agricultural	teacher education	medical	fine arts	
Kujawsko-Pomorskie	1	1	0	0	1	{technical, agricultural, fine arts}
Warmińsko-Mazurskie	0	0	2	0	0	{teacher education}
Śląskie	4	0	2	1	2	{technical, pedagogical, medical, fine arts}
Dolnośląskie	1	1	1	1	2	{technical, agricultural, teacher education, medical, fine arts}
Podkarpackie	1	0	0	0	0	{technical}

Source: Author's compilation based on data from Local Data Bank of the Central Statistical Office in Poland.

The combination of variable category illustrating groups of demographic age and their percentage share in Polish cities allowed for obtaining modal symbolic variable (see Table 4). The following factors were considered: pre-working age (up to the age of 17), working age (men aged 18-64, women aged 18-59) and post-working age (men aged 65 and more, women aged 60 and more).

Table 4. Population by economic age groups in Poland in 2010 (%)

NTS-5 (city)	Economic age group			Symbolic modal variable for NTS-5
Name	pre-working	working	post-working	
Łódź	13.9	64.3	21.8	{pre-working (0.139), working (0.643), post-working (0.218)}
Warszawa	15.0	63.9	21.1	{pre-working (0.150), working (0.639), post-working (0.211)}
Kraków	15.5	65.2	19.3	{pre-working (0.155), working (0.652), post-working (0.193)}
Wrocław	14.7	65.7	19.6	{pre-working (0.147), working (0.657), post-working (0.196)}

Source: Author's compilation based on data from Local Data Bank of the Central Statistical Office in Poland.

In a similar way the share structure may be obtained, which illustrates an economic profile of a territorial unit based on employment size in economic sectors (agricultural, industrial and service sector), or added value produced by a given sector, and also monthly household expenditure on food, clothes, footwear, house maintenance, etc.

4. The symbolic data analysis methods

The complex structure of symbolic data prevents direct application of statistical methods dedicated to data following classical approach (see e.g. Hair et al., 2006, Everitt and Dunn, 2001). Two approaches are suggested in the professional literature for symbolic data analysis (see e.g. Bock, Diday et al., 2000, Wilk, 2010a).

The first approach consists in adapting algorithms originally designed for the analysis of data in classical approach and in particular distance matrix based methods. In such case distance measurement has to be performed using distance measures of symbolic objects (see e.g. Malerba et al., 2001, Malerba, Esposito, Monopoli, 2002, Wilk 2006a, 2006b).

The second approach is based on the application of methods dedicated to symbolic data, e.g. TREE algorithm from the family of classification trees, or Interscal in order to perform multivariate scaling. Selected methods, applied in symbolic data analysis, are presented in Table 5.

Table 5. Selected methods applied in symbolic data analysis

Name	Group	Selected algorithms	
		adopted	original
Decision trees	classification	–	TREE, BDT, SDT
	regression	–	–
Neural networks	supervised	multilayer perceptron	–
	unsupervised	self-organizing maps	–
Multidimensional scaling	based on distance matrix	bi-dimensional mapping	–
	based on symbolic data table	–	InterScal, SymScal, I-Scal
Cluster analysis	hierarchical	Ward's, complete linkage	Brito's, Chavent's
	non-hierarchical	*k*-medoids	SCLUST, DCLUST

Source: Author's compilation based on (Bock, Diday et al., 2000, Billard and Diday, 2006, Diday, Noirhomme-Fraiture et al., 2008, Wilk, 2009, 2010a).

5. An empirical example

The objective of the study is to define labour market situation in Polish regions (NTS-2) in 2010. Objects (regions) were described by means of variables characterizing the employment level and structure, as well as average wages and salaries. Symbolic approach was applied following which data aggregation was performed in line with classical approach. The approach based on objects aggregation as well as the one based on variables aggregation were considered.

As the result of aggregation 5 symbolic variables were obtained, including 3 interval-valued variables and 2 modal variables (see Table 6). Observations referring to variables are presented in the symbolic data table (see Appendix 1).

Table 6. The set of symbolic variables

No.	Variable name	Aggregation method	Type of symbolic variable	Set of variable implementation	Measure unit
1	Employment rate in sub-regions (NTS-3)	aggregation of objects	interval-valued	[36.4, 92.3]	%
2	Employment by economic sectors	aggregation of variables	modal	{agricultural [2.9, 28.2], industrial [20.9, 38.0], services [48.3, 66.2]}	%
3	Employment by education level	aggregation of variables	modal	{tertiary [21.8, 37.1], post-secondary and vocational secondary [26.3, 31.2], general secondary [6.8, 10.5], basic vocational [20.5, 34.0], lower secondary, primary and lower [4.0, 11.9]}	%
4	Job vacancy rate*	aggregation of variables	interval-valued	[0.29, 0.89]	%
5	Average monthly gross wages and salaries in poviats/districts (NTS-4)	aggregation of objects	interval-valued	[2106.84, 6012.95]	PLN

* Job vacancy rate is the ratio of vacancies to occupied and unoccupied jobs (quarterly data).

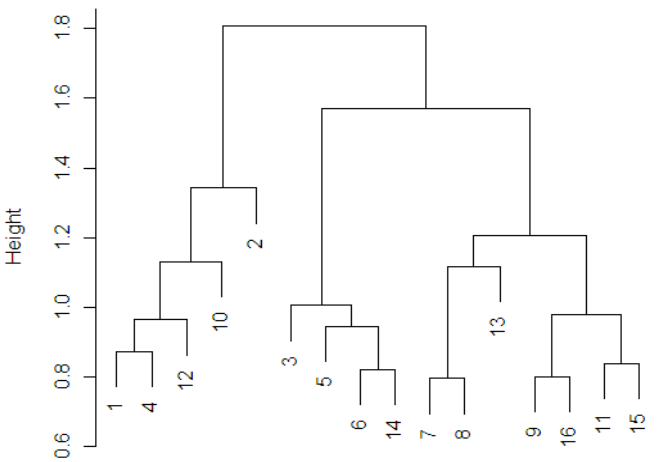
Source: Author's compilation based on data provided by the Central Statistical Office in Poland (Local Data Bank, Labour Force Survey and elaboration on The demand for labour in 2010).

Cluster analysis was applied in order to distinguish homogenous classes of regions. The first step of classification procedure focused on measuring symbolic objects distance using two componentwise distance measures, i.e. U_3 Ichino-Yaguchi normalized measure for interval-valued variables and PU_2 normalized measure for modal variables, as well as Euclidean distance as objectwise distance measure (see e.g. Bock, Diday et al., 2000). Distance measurement results are presented in Appendix 2.

The initial distant matrix analysis indicates that Mazowieckie region significantly differs from others. Measure values are higher than 1 for all regions and in particular for Lubelskie and Podkarpackie. The most similar ones, with regard to the accepted variables set, are Warmińsko-Mazurskie and Lubuskie regions.

In the next step hierarchical classification of objects set, using Ward's method, was performed. The obtained dendrogram is presented in Figure 1. Next, the number of classes was defined based on Hubert and Levine internal cluster quality index (see Figure 2). Finally, the 5 class structure was accepted for which index received the lowest value (0.396).

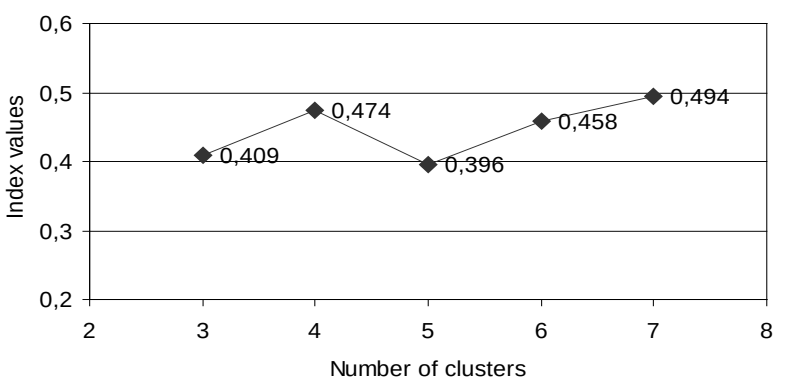
Figure 1. Dendrogram obtained using Ward's method



Explanations: 1 – Łódzkie, 2 – Mazowieckie, 3 – Małopolskie, 4 – Śląskie, 5 – Lubelskie, 6 – Podkarpackie, 7 – Podlaskie, 8 – Świętokrzyskie, 9 – Lubuskie, 10 – Wielkopolskie, 11- Zachodniopomorskie, 12 – Dolnośląskie, 13 – Opolskie, 14 – Kujawsko-Pomorskie, 15 – Pomorskie, 16 – Warmińsko-Mazurskie.

Source: Author's estimation in `clusterSim` *and* `cluster` *package of* `R` *software v. 2.13.0.*

Figure 2. Values of Hubert & Levine index for different number of clusters



Source: Author's estimation in `clusterSim` *and* `cluster` *package of* `R` *software v. 2.13.0.*

The final classification procedure step consisted in the description (interpretation) of the obtained classes using CLINT technique (see Brito, 2004). For interval-valued variables minimum and maximum variable value was defined as obtained by the given class objects (see Table 7). On the other hand, in case of modal variables the average weight obtained by objects included in the class for every variable category.

Table 7. Cluster description

No	Cluster members	Employment rate in NTS-3 (%)	Employment by economic sector (%)	Employment by education level (%)	Job vacancy rate (%)	Average monthly gross wages and salaries in NTS-4 (PLN)
1	Łódzkie, Śląskie, Dolnośląskie, Wielkopolskie	[40.0, 79.8]	{agricultural (9.3*), industrial (34.5), services (56.1)}	{tertiary (26.2), post-secondary and vocational secondary (28.7), general secondary (9.1), basic vocational (29.0), lower secondary, primary and lower (6.6)}	[0.48, 0.80]	[2106.84, 6012.95]
2	Mazowieckie	[43.2, 92.4]	{agricultural (11.3), industrial (22.3), services (66.2)}	{tertiary (37.1), post-secondary and vocational secondary (26.4), general secondary (9.7), basic vocational (20.5), lower secondary, primary and lower (6.3)}	[0.79, 0.89]	[2410.05, 4694.47]
3	Małopolskie, Lubelskie, Podkarpackie, Kujawsko-Pomorskie	[42.7, 73.4]	{agricultural (19.6*), industrial (27.8*), services (52.6)}	{tertiary (24.6), post-secondary and vocational secondary (28.5), general secondary (8.1), basic vocational (29.9), lower secondary, primary and lower (8.9)}	[0.29, 0.66]	[2500.93, 4428.15]

Table 7. Cluster description (cont.)

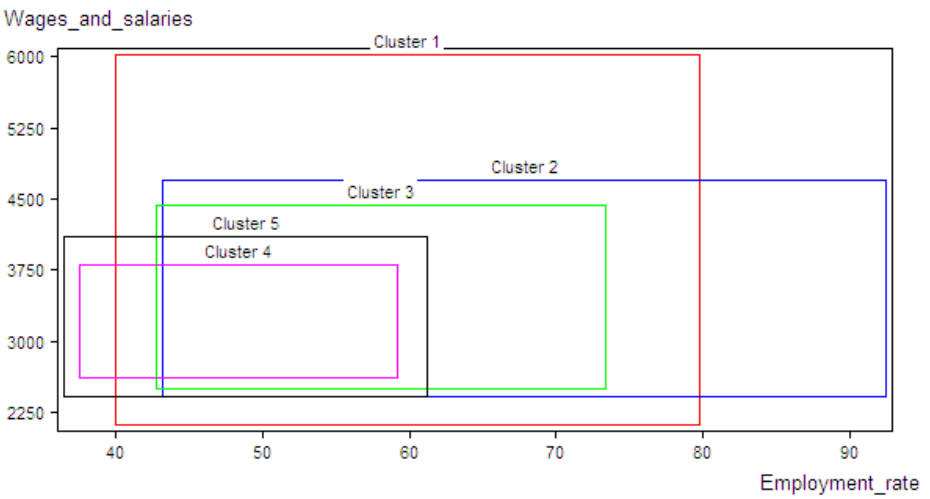
No	Cluster members	Employment rate in NTS-3 (%)	Employment by economic sector (%)	Employment by education level (%)	Job vacancy rate (%)	Average monthly gross wages and salaries in NTS-4 (PLN)
4	Podlaskie, Świętokrzyskie, Opolskie	[37.5, 59.2]	{agricultural (19.3*), industrial (29.6*), services (51.2)}	{tertiary (24.7), post-secondary and vocational secondary (28.7), general secondary (7.5), basic vocational (29.6*), lower secondary, primary and lower (9.4)}	[0.44, 0.86]	[2613.35, 3798.54]
5	Lubuskie, Zachodnio-Pomorskie, Pomorskie, Warmińsko-Mazurskie	[36.4, 61.2]	{agricultural (8.7), industrial (31.8), services (59.5)}	{tertiary (26.2), post-secondary and vocational secondary (27.6), general secondary (9.1), basic vocational (29.0), lower secondary, primary and lower (8.1)}	[0.45, 0.72]	[2414.19, 4108.37]

* There is relatively high diversity of values (max – min = 10-15 percentage points).

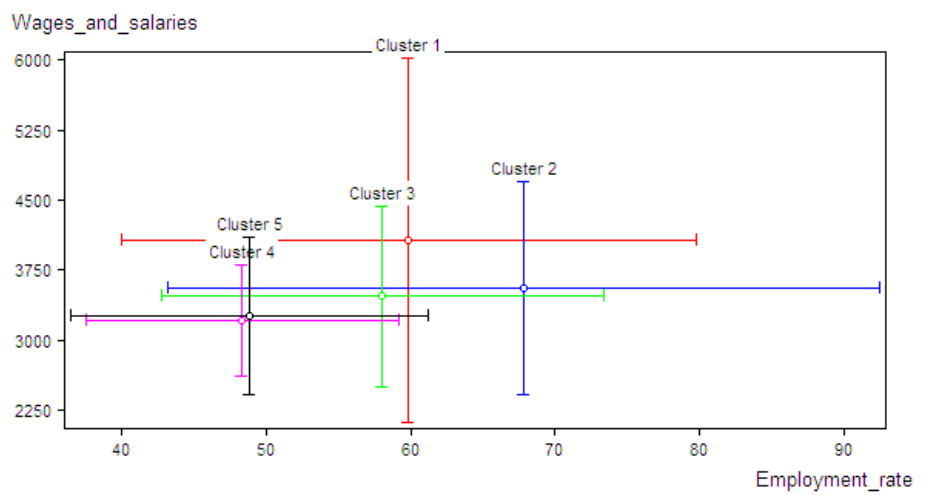
Source: Author's estimation in `clusterSim` *and* `cluster` *package of* `R` *software v. 2.13.0.*

Picture 3 presents class characteristics with reference to employment level and average earnings. Squares in picture 4a show intervals of variables values, whereas perpendicular, intersecting lines in picture 4b illustrate centres and ranges of intervals. The smaller the squares the lower the territorial diversification of the analyzed qualities. As both pictures show the relatively low employment rate and low salaries, they are characteristic to regions included in classes 4 and 5. These are eastern and southern regions of the country, with the exception of Śląskie voivodship.

Figure 3. Cluster description regarding wages and salaries and employment rate



a) boxes

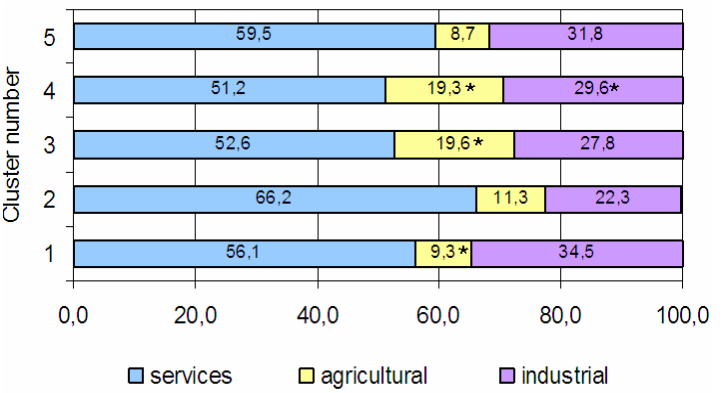


b) crosses

Source: Author's estimation in Sodas software v. 2.0.

Service sector is the dominating one in all regions. In the second class represented by Mazowieckie region as many as 2/3 of labour force is employed in this sector and slightly more then 50% in the regions covered by classes 3 and 4 (see Figure 4). The regions listed in classes 4 and 5 may be referred to as the industry oriented since every third person is employed in industry sector.

Figure 4. Cluster description regarding employment by economic sector

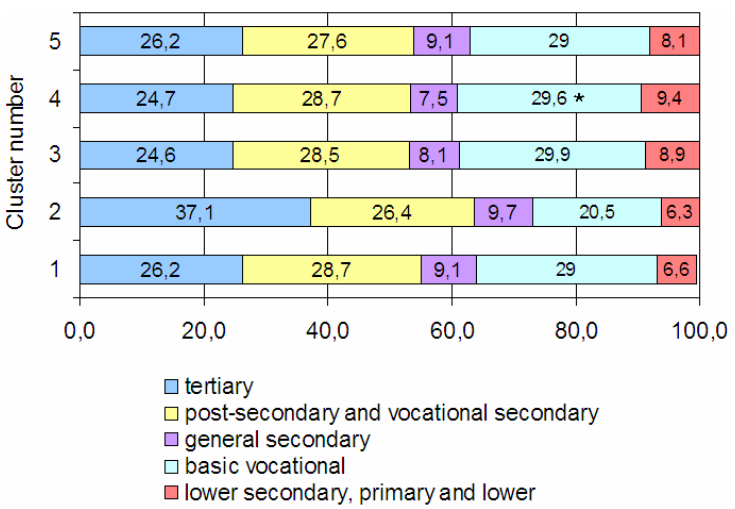


* There is relatively high diversity of values (max – min = 10-15 percentage points).

Source: Author's estimation.

Education structure among labour force is similar in all classes (see Figure 5). The occurring differences amount to about 5 percentage points. Class 2 (Mazowieckie region) is the only exception in which over 1/3 of labour force are tertiary education graduates which is responsible for 10 percentage points more than in other classes.

Figure 5. Cluster description regarding employment by education level



* There is relatively high diversity of values (max-min = 10-15 percentage points).

Source: Author's estimation.

The first class includes the following regions: Łódzkie, Śląskie, Dolnośląskie and Wielkopolskie (see Figure 6). They are characterized by a very high rate of employment in industry sector at the background of other regions. Districts of these regions feature relatively high diversification regarding wages and salaries level. This group registers the lowest percentage of labour force characterized by lower secondary, primary and lower education.

Figure 6. Classes of regions

Source: Author's compilation.

The second cluster represents Mazowieckie region characterized by the highest employment rate level in services (over 66%) and the highest percentage of labour force recruiting from tertiary education graduates (over 37%). At the same time the region presents the highest diversification of districts (NTS-4) with regard to wages and salaries level, as well as many unoccupied jobs, which is indicated by high values of job vacancy rate.

The third class lists regions situated mainly in south-eastern part of the country, i.e. Małopolskie, Lubelskie, Podkarpackie and also Kujawsko-Pomorskie. They are featured by relatively low employment rate in services at sub-regional level. It has to be emphasized that labour supply in these regions corresponds to demand (job vacancy rate obtains relatively low values).

The fourth class is made up of the following regions: Podlaskie, Świętokrzyskie and Opolskie. Their characteristic feature is low employment rate at sub-regional level, as well as extensive seasonal fluctuations in this matter.

Very low earnings were registered in districts of these regions, however, at the same time the lowest territorial diversification was observed with regard to wages and salaries level.

The final, fifth class is composed of regions situated in northern (Zachodnio-Pomorskie, Pomorskie and Warmińsko-Mazurskie) and western (Lubuskie) part of Poland. The employment rate level in sub-regions of these regions is relatively low and additionally the lowest territorial diversification was also observed in this respect. Relatively high employment rate is typical for industry sector and low for agricultural sector. Districts in these regions feature relatively low level of wages and salaries.

6. Conclusions

The objective of the paper is to present the proposal for symbolic approach application in regional studies. The suggested approach usefulness depends, mainly, on the goal underlying the performed study and the accepted research procedure. Symbolic approach is adequate in case of both small and large data sets when the need occurs for synthetic and detailed description of the analyzed phenomena or it is necessary to reduce the description.

Among difficulties resulting from this approach application of more laborious and time consuming research related to classical approach may be indicated, as well as more complicated interpretation of both data and the obtained results. The application of the procedure based on aggregation of objects is to a large extent determined by the availability of statistical data.

REFERENCES

BILLARD, L., DIDAY, E. (2006). Symbolic Data Analysis. Conceptual Statistics and Data Mining, Chichester: Wiley.

BOCK, H. H., DIDAY, E. (Eds.) (2000). Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data, Berlin-Heidelberg: Springer-Verlag.

BRITO, P. (2004). Clustering Interpretation. Interpreting Clusters by using the module CLINT, In: M. Noirhomme-Fraiture (Ed.), User manual for SODAS 2 software, Software Report, Analysis System of Symbolic Official Data, Project no. IST-2000-25161.

DIDAY, E., NOIRHOMME-FRAITURE, M. (Eds.) (2008). Symbolic data analysis and the Sodas software, Chichester: John Wiley & Sons.

EVERITT, B. S., DUNN, G. (2001). Applied Multivariate Data Analysis, London: Arnold.

HAIR, J. F., BLACK, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. (2006). Multivariate Data Analysis, New Jersey: Pearson Prentice Hall.

MALERBA, D., ESPOSITO, F., GIOVALLE, V., TAMMA, V. (2001). Comparing Dissimilarity Measures for Symbolic Data Analysis, In: P. Nanopoulos (Ed.), New Techniques and Technologies for Statistics and Exchange of Technology and Know-how (ETK-NTTS'01), conference materials, pp. 473-481.

MALERBA, D., ESPOSITO, F., MONOPOLI, M. (2002). Comparing dissimilarity measures for probabilistic symbolic objects, In: A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.), Data Mining III, Series Management Information Systems, vol. 6, Southampton: WIT Press, pp. 31-40.

MYNARSKI, S. (2000). Praktyczne metody analizy danych rynkowych i marketingowych [Practical methods for market and marketing data analysis], Kraków: Zakamycze.

Popyt na pracę w 2010 r. [The demand for labour in 2010] (2011). Statistical Information and Studies, Central Statistical Office, Warsaw: Central Statistical Office Publishing House.

WALESIAK, M. (1993). Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu [Procedures in statistical research in case of variables set measured in different types of scales], Badania Operacyjne i Decyzje [Operational Studies and Decisions] no. 1, pp. 71-77.

WILK, J. (2010a). Metody analizy danych symbolicznych [Symbolic data analysis methods], In: J. Dziechciarz (Ed.), Econometrics 29. The Application of quantitative methods, Research Studies of Wrocław University of Economics no. 141, Wrocław: Wrocław University of Economics Publishing House, pp. 29-38.

WILK, J. (2006b). Miary odległości obiektów opisanych zmiennymi symbolicznymi z wagami [Distance measures for objects described by symbolic variables with weights], In: K. Jajuga, M. Walesiak (Eds.), Taxonomy 13. Data classification and analysis – theory and applications, Research Studies of University of Economics in Wrocław no. 1126, Wrocław: University of Economics in Wrocław Publishing House, pp. 224-236.

WILK, J. (2006a). Problemy klasyfikacji obiektów symbolicznych. Symboliczne miary odległości [Problems of symbolic objects classification. Symbolic distance measures], In: J. Garczarczyk (Ed.), Quantitative and qualitative methods for market research. Measurement and its efficiency, Research Bulletins University of Economics in Poznań no. 71, Poznań: University of Economics in Poznań Publishing House, pp. 69-83.

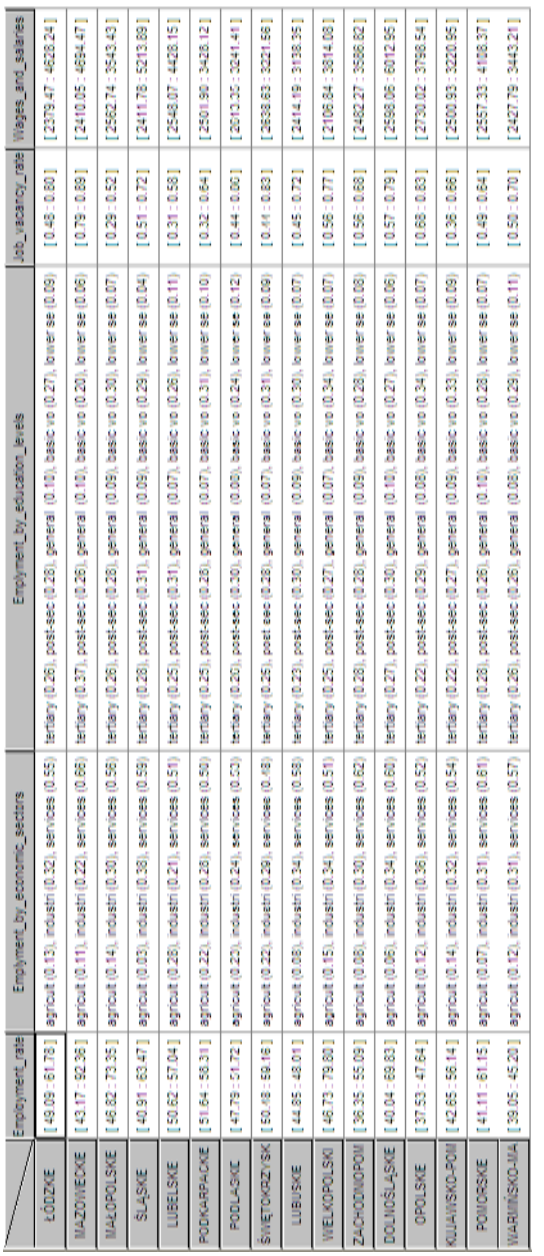
WILK, J. (2010b.) Problemy segmentacji rynku z wykorzystaniem metod klasyfikacji i danych symbolicznych [Market segmentation problems applying classification methods and symbolic data], Wrocław University of Economics, doctoral dissertation (unpublished), Jelenia Góra.

WILK, J. (2009). Przegląd metod wielowymiarowej analizy statystycznej wykorzystywanych w badaniach segmentacyjnych [The review of multivariate statistical analysis applied in segmentation studies], In: J. Dziechciarz (Ed.), Econometrics 23. The Application of quantitative methods, Research Studies of Wrocław University of Economics no. 37, Wrocław: University of Economics in Wrocław Publishing House, pp. 59-70.

WILK, J. (2011). Taksonomiczna analiza rynku pracy województw Polski – podejście symboliczne [Taxonomic analysis of labour market in Polish regions – symbolic approach], In: J. Dziechciarz (Ed.), Econometrics 34. The Application of quantitative methods, Research Studies of Wrocław University of Economics no. 200, Wrocław: Wrocław University of Economics Publishing House, pp. 26-37.

APPENDIX

Appendix 1. Symbolic data table*



	Employment_rate	Employment_by_economic_sectors	Employment_by_education_levels	Job_vacancy_rate	Wages_and_salaries
ŁÓDZKIE	[49.09 : 61.78]	agricult (0.13), industri (0.32), services (0.55)	tertiary (0.26), post-sec (0.28), general (0.19), basic vo (0.27), lower se (0.09)	[0.48 : 0.80]	[2379.47 : 4628.24]
MAZOWIECKIE	[43.17 : 62.38]	agricult (0.11), industri (0.22), services (0.66)	tertiary (0.37), post-sec (0.26), general (0.19), basic vo (0.20), lower se (0.06)	[0.79 : 0.89]	[2410.05 : 4684.47]
MAŁOPOLSKIE	[46.82 : 73.35]	agricult (0.14), industri (0.30), services (0.56)	tertiary (0.28), post-sec (0.28), general (0.09), basic vo (0.30), lower se (0.07)	[0.29 : 0.52]	[2562.74 : 3543.43]
ŚLĄSKIE	[40.91 : 63.47]	agricult (0.03), industri (0.38), services (0.59)	tertiary (0.28), post-sec (0.31), general (0.09), basic vo (0.29), lower se (0.04)	[0.51 : 0.72]	[2411.78 : 5213.89]
LUBELSKIE	[50.62 : 57.04]	agricult (0.28), industri (0.21), services (0.51)	tertiary (0.25), post-sec (0.31), general (0.07), basic vo (0.26), lower se (0.11)	[0.31 : 0.58]	[2548.07 : 4428.15]
PODKARPACKIE	[51.64 : 58.31]	agricult (0.22), industri (0.28), services (0.50)	tertiary (0.25), post-sec (0.28), general (0.07), basic vo (0.31), lower se (0.10)	[0.32 : 0.64]	[2501.96 : 3428.12]
PODLASKIE	[47.79 : 51.72]	agricult (0.22), industri (0.24), services (0.53)	tertiary (0.20), post-sec (0.30), general (0.08), basic vo (0.24), lower se (0.12)	[0.44 : 0.86]	[2613.35 : 3241.41]
ŚWIĘTOKRZYSKIE	[50.48 : 59.16]	agricult (0.22), industri (0.29), services (0.48)	tertiary (0.25), post-sec (0.28), general (0.07), basic vo (0.35), lower se (0.09)	[0.44 : 0.83]	[2638.63 : 3221.58]
LUBUSKIE	[44.85 : 48.01]	agricult (0.08), industri (0.34), services (0.58)	tertiary (0.23), post-sec (0.30), general (0.08), basic vo (0.30), lower se (0.07)	[0.45 : 0.72]	[2414.19 : 3138.35]
WIELKOPOLSKI	[46.73 : 79.80]	agricult (0.15), industri (0.34), services (0.51)	tertiary (0.25), post-sec (0.27), general (0.07), basic vo (0.34), lower se (0.07)	[0.56 : 0.77]	[2106.84 : 3814.08]
ZACHODNIOPOM	[36.35 : 55.09]	agricult (0.08), industri (0.30), services (0.62)	tertiary (0.28), post-sec (0.28), general (0.09), basic vo (0.28), lower se (0.08)	[0.56 : 0.68]	[2482.27 : 3556.82]
DOLNOŚLĄSKIE	[40.04 : 69.83]	agricult (0.06), industri (0.34), services (0.60)	tertiary (0.27), post-sec (0.30), general (0.10), basic vo (0.27), lower se (0.06)	[0.57 : 0.79]	[2598.06 : 6012.95]
OPOLSKIE	[37.53 : 47.64]	agricult (0.12), industri (0.36), services (0.52)	tertiary (0.22), post-sec (0.29), general (0.08), basic vo (0.34), lower se (0.07)	[0.68 : 0.83]	[2730.02 : 3798.54]
KUJAWSKO-POM	[42.65 : 58.14]	agricult (0.14), industri (0.32), services (0.54)	tertiary (0.22), post-sec (0.27), general (0.09), basic vo (0.33), lower se (0.09)	[0.36 : 0.68]	[2500.93 : 3226.95]
POMORSKIE	[41.11 : 61.15]	agricult (0.07), industri (0.31), services (0.61)	tertiary (0.28), post-sec (0.26), general (0.10), basic vo (0.28), lower se (0.07)	[0.49 : 0.64]	[2557.33 : 4108.37]
WARMIŃSKO-MA	[39.05 : 45.20]	agricult (0.12), industri (0.31), services (0.57)	tertiary (0.26), post-sec (0.26), general (0.08), basic vo (0.29), lower se (0.11)	[0.59 : 0.70]	[2427.79 : 3443.21]

* The names of the selected objects, variables and variables categories were shortened (for full names see Table 6 and Figure 6).

Source: Author's compilation.

Appendix 2. Dissimilarity matrix

Object (territorial unit) number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1.273														
3	1.193	1.553													
4	0.873	1.270	1.226												
5	1.106	**1.590**	0.973	1.148											
6	1.072	1.571	0.943	1.152	0.862										
7	0.986	1.387	1.242	1.172	1.130	1.040									
8	0.958	1.388	1.184	1.134	1.118	0.968	0.796								
9	0.996	1.455	1.148	1.062	1.109	0.977	0.886	0.916							
10	0.974	1.160	1.197	1.070	1.264	1.160	1.117	1.039	1.107						
11	1.013	1.347	1.193	0.990	1.136	1.049	1.056	1.048	0.910	1.092					
12	1.000	1.191	1.347	0.882	1.315	1.336	1.261	1.230	1.246	1.147	1.119				
13	1.078	1.334	1.430	1.134	1.365	1.289	1.041	1.052	1.057	1.151	0.988	1.163			
14	1.042	1.500	0.985	1.097	0.966	0.821	1.003	0.962	0.865	1.121	0.930	1.277	1.185		
15	0.897	1.327	1.059	0.898	1.012	0.958	1.064	1.017	0.948	1.043	0.839	1.055	1.111	0.909	
16	1.014	1.450	1.209	1.045	1.132	1.033	0.971	1.009	**0.800**	1.154	0.833	1.208	1.008	0.911	0.915

low similarity (above 1.4) high similarity (below 0.9)

Explanations: 1 – Łódzkie, 2 – Mazowieckie, 3 – Małopolskie, 4 – Śląskie, 5 – Lubelskie, 6 – Podkarpackie, 7 – Podlaskie, 8 – Świętokrzyskie, 9 – Lubuskie, 10 – Wielkopolskie, 11- Zachodniopomorskie, 12 – Dolnośląskie, 13 – Opolskie, 14 – Kujawsko-Pomorskie, 15 – Pomorskie, 16 – Warmińsko-Mazurskie.

Source: Author's compilation.