STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 243–264

USING SYMBOLIC DATA IN GRAVITY MODEL OF POPULATION MIGRATION TO REDUCE MODIFIABLE AREAL UNIT PROBLEM (MAUP)

Justyna Wilk[1]

ABSTRACT

Spatial analyses suffer from modifiable areal unit problem (MAUP). This occurs while operating on aggregated data determined for high-level territorial units, e.g. official statistics for countries. Generalization process deprives the data of variation. Carrying out research excluding territorial distribution of a phenomenon affects the analysis results and reduces their reliability. The paper proposes to use symbolic data analysis (SDA) to reduce MAUP. SDA proposes an alternative form of individual data aggregation and deals with multivariate analysis of interval-valued, multi-valued and histogram data.

The paper discusses the scale effect of MAUP which occurs in a gravity model of population migrations and shows how SDA can deal with this problem. Symbolic interval-valued data was used to determine the economic distance between regions which served as a separation function in the model. The proposed approach revealed that economic disparities in Poland are lower than official statistics show but they are still one of the most important factors of domestic migration flows.

Key words: modifiable areal unit problem (MAUP), symbolic data analysis (SDA), gravity model, population migration, economic distance.

1. **Introduction**

A large number of spatial analyses suffer from modifiable areal unit problem (MAUP), regardless of the research field (e.g. economics, biology, sociology, finance, medicine, etc.) (see Openshaw, 1984; Arbia, 1989, pp. 7-21; King, Tanner and Rosen (Eds.), 2004; Wong 2009). MAUP occurs while operating on aggregated data which is a procedure frequently used to describe higher-level territorial units, e.g. countries, metropolitan areas, regional labour markets, etc. Generalization process deprives the data of variation. Carrying out research excluding territorial

[1] Wrocław University of Economics, Department of Econometrics and Computer Science, 58-500 Jelenia Góra (Poland), Nowowiejska 3 Street. E-mail: Justyna.Wilk@ue.wroc.pl.

distribution and spatial features of a phenomenon affects the analysis results and reduces their reliability.

This problem is mostly seen in socio-economic studies in which a territorial unit is a result of an administrative division or territorial division for statistical purposes (see, e.g. NUTS classification). For example, Poland is administratively divided into 2479 municipalities (LAU 2 units). Each of them is located in the territory of one of 314 districts (LAU 1 units). A set of bordered districts is assigned to one of 16 provinces (NUTS 2). Official (economic, social, environmental, demographical etc.) statistics present aggregated values which generalize the situations of territorial units. They do not show the ranges, densities, distributions, outlier values or spatial variation in data. Then, we cannot infer results from one scale to the other scales of territorial division due to ecological fallacy. The paper deals with MAUP which occurs while modelling of population migrations.

In the era of market economy, domestic population migrations represent an integral part underlying the functioning of societies and economies. They regulate the size and structure of human resources, as well as job market situation, the consumption of goods and services, etc. Thus, an integral part of developing the policy of sustainable regional development is carrying out research studies regarding not only the results of migration flows (e.g. an amount of inflow) but, first of all, the conditions and causes of people’s decisions why and where to migrate (see Bunea, 2012; Lucas, 1997; White and Lindstrom, 2006).

The intensity of domestic migration flows is strongly determined by the macroeconomic trends which affect people’s propensity to move. But the directions of migrations depend on regional factors such as an economic, social, political, environmental situation, etc., as well as spatial and relational factors, e.g. ethnic differences (Van der Gaag (Ed.), 2003, pp. 1-141). These factors can be examined using an econometric gravity model.

The aim of the empirical study is to examine the determinants of domestic migrations in Poland. The research covers migration flows between 16 Polish NUTS 2 units (provinces) in the years 2011-2013, in which the world economy was overcoming the economic crisis. In terms of relatively stable political and cultural terms, the strongest determinants of migration processes are economic motives, e.g. improving the standard of living (Todaro, 1980; Lucas, 1997; Kupiszewski, Durham and Rees, 1999; Holzer, 2003; White, Lindstrom, 2006; Ghatak, Mulhern and Watson, 2008). Cohesion policy of the European Union directs national policies of regional development to convergence processes. Thus, in this study, the crucial issue is to identify the economic disparities in Poland and examine their impact on domestic migration flows.

The preliminary data analysis showed the occurrence of the scale effect of MAUP. Therefore the objective of this paper is to propose a solution to MAUP. The proposed approach employs symbolic data analysis (SDA) to construct the gravity model of population migration. SDA covers multivariate analysis of interval-valued, multi-valued, modal and dependent data. It is a support to manage data structure and reduction problems (see Bock and Diday (Eds.), 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture (Ed.), 2008; Gatnar and Walesiak, 2011).

The first section of the paper discusses the essence of the modifiable areal unit problem. The second part concerns the scale effect of MAUP occurring in the gravity model of population migration. The third section employs symbolic data analysis to reduce this problem. The fourth part discusses the results of the study and shows the influence of MAUP on the spatial interaction analysis results.

2. Modifiable areal unit problem (MAUP) of spatial data analyses

Yule and Kendall (1950) introduced a fundamental distinction between two different kinds of analysed units: the non-modifiable and modifiable units. Modifiable units differ from non-modifiable units because they can be further decomposed into smaller units and, moreover, this decomposition can be done in a few ways. The relevance of this distinction is that the value of any statistical measure "will, in general, depend on the unit chosen if that unit is modifiable" (Yule and Kendall, 1950).

This problem is known in the literature as the modifiable areal unit problem (MAUP). MAUP results from data generalization and multiscaling of spatial phenomena (see Openshaw, 1984; Arbia, 1989, pp. 7-21; Anselin, 1988, p. 26-28; Suchecka (Ed.), 2014, pp. 56-60; Gotway, Crawford and Young, 2004; Wong, 2009). The problem arises from the fact that areal units are usually arbitrarily determined and modifiable in the sense that they can be aggregated to form units of different sizes or spatial arrangements (Jelinski and Wu, 1996, p. 130).

Openshaw and Taylor (1979) distinguished two aspects of MAUP: the scale effect and zonation effect. The scale (aggregation) aspect refers to different results which can be achieved in statistical analysis with the same set of data grouped at different scale levels (e.g. countries or regions). Thus, the scale effect occurs if a set of areas is considered from the point of view of larger areal units, with each combination leading to different data values and inferences. The problem is "the variation in results that may be obtained when the same data are combined into sets of increasingly larger areal units of analysis" (Openshaw and Taylor, 1979).

The scale effect of MAUP results from a few of reasons, e.g. human society is organized in territorial units usually arranged into nested hierarchies, e.g. town, regions, states, countries (see Moellering and Tobler, 1972; Cliff and Ord, 1981, p. 133). The scale effect of MAUP was proved by Gehlke and Biehl, 1934, pp. 169-170; Jelinski and Wu, 1996; Dark and Bram, 2007, pp. 471-479. Parts a-c of Figure 1 illustrate the scale effect of MAUP.

a

2	4	6	1
3	6	3	5
1	5	4	2
5	4	5	4

Ar. mean = 3.75
St. dev. = 2.60

b

3.0	3.5
4.5	4.0
3.0	3.0
4.5	4.5

Ar. mean = 3.75
St. dev. = 0.50

c

3.75	3.75
3.75	3.75

Ar. mean = 3.75
St. dev. = 0.00

d

2.5	5.0	4.5	3.0
3.0	4.5	4.5	3.0

Ar. mean = 3.75
St. dev. = 0.93

e

2.75	4.75	4.5	3.0

Ar. mean = 3.75
St. dev. = 1.04

f

4.0	1.0
4.0	3.67

Ar. mean = 3.17
St. dev. = 2.11

Figure 1. Examples of the scale effect (a-c) and zoning effect (d-f) of MAUP

Source: Jelinski D. E., Wu J., 1996, The modifiable areal unit problem and implications for landscape ecology, Landscape Ecology, Vol. 11, No. 3, pp. 129−140.

The operation of “averaging” data results in smoothing the data and losing information. For example, the disposable income in Swedish NUTS 2 units was between 155 and 168 SEK, whereas the values recorded by 284 Swedish municipalities (LAU 2) are held in [137,000 – 352,000] SEK *per capita* in 2002 (see parts a, b, d of Figure 2). The scale effect has at least two consequences. The data aggregation (shifting from a finer to a coarse scale) results in decreasing the variance (see Moellering and Tobler, 1972), as well as the statistical correlation tends to increase with increasing the size of the areas considered (see Yule and Kendall, 1950).

The zonation (grouping, delimitation) effect concerns the spatial arrangement in zones. It considers the variability of results not due to variations in the size of the areas but rather to their shapes, e.g. metropolitan areas, local labour markets, urban areas, tourist regions, etc.

When dealing with the aggregation problem, no loss of information occurs if we shift from one boundary system to another, rather there is an alternation of information (see Arbia, 1989, p. 18). For example, in parts c, e and f of Figure 1 one can see that even when the number of zones is held constant (N = 4) the mean and variance is affected. A comparison of parts b and d of Figure 2 shows a change in variance when the orientation is altered but the size of the units remains fixed (Jelinski and Wu, 1996). For example, depending on the zone boundaries, the interpretation of disposable income in Sweden changes (see parts c and d of Figure 2).

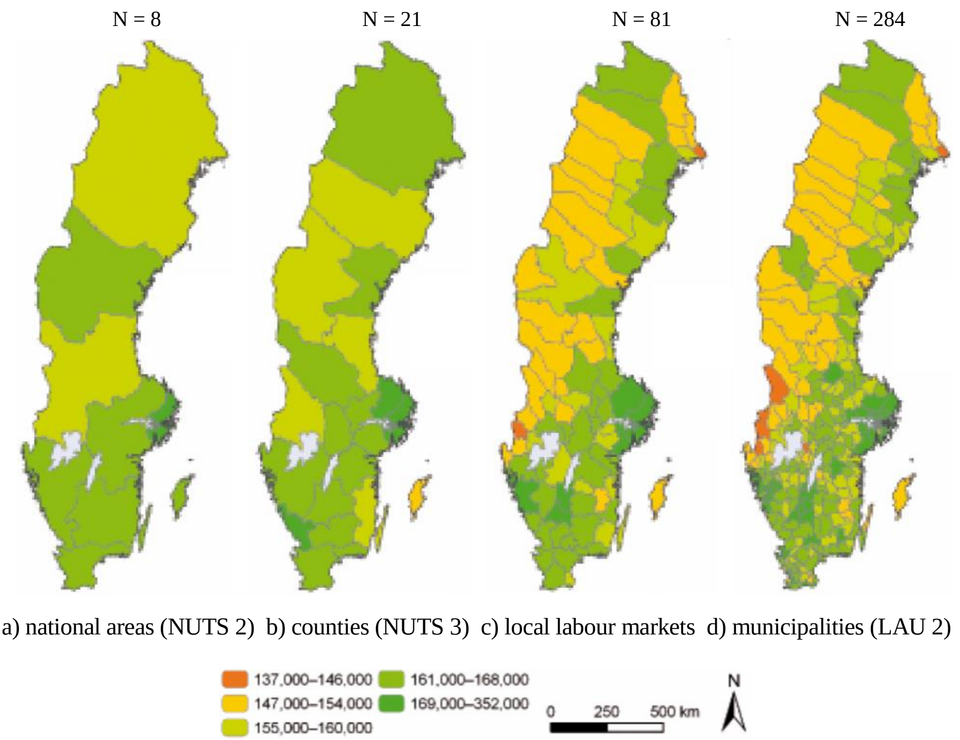


Figure 2. Disposable income *per capita* (20-64 years old people) in Sweden in 2002 (SEK)

Source: the modifiable areas unit problem, European Observation Network, Territorial Development and Cohesion, The ESPON Monitoring Committee, Luxembourg 2006, p. 47.

In regional studies based on a set of units resulted from an administrative or statistical division of a territory the zonation problem exists but is omitted due to operating on territorial units which are defined in advance (e.g. NUTS classification of territorial units) and function independently. The empirical study presented in this paper is based on NUTS 2 Polish units (provinces) which function as self-government territorial units. Therefore, in the article, special attention is paid to the scale effect of MAUP and methods which deal with it.

Some solutions to this case are discussed in the literature, for example King (1997) proposed error-bound approach, Tobler (1979) formed scale-insensitive migration model, Tate and Atkinson (2001) proposed to use fractal analysis and geostatistics (kriging and related methods such as variograms), Benali and Escoffier (1990) proposed smooth factorial analysis, Fotheringham, Charlton and Brunsdon (2001) proposed the geographically weighted regression. However, none of these solutions is sufficient and universal. The scale effect of MAUP is still an open issue.

3. The scale effect of MAUP in gravity model of population migration

Migrations occur in territorial space as flows from one area to another. An econometric gravity model is a tool which examines the internal and external conditions of flows, by analogy with Newton's (1687) concept of gravity (see Isard, 1960; Chojnicki, 1966; Anderson, 1979; Fotheringham and O'Kelly, 1989; Grabiński, Malina, Wydymus and Zeliaś, 1988; Sen and Smith, 1995; Zeliaś, 1999, pp. 172-175; Roy, 2004; LeSage and Pace, 2008; Suchecki (Ed.), 2010, pp. 226-230; Chojnicki, Czyż and Ratajczak, 2011, Shepherd, 2013, Beine, Bertoli and Fernández-Huertas Moraga, 2015).

The model typically examines three types of factors to explain mean interaction frequencies (Fischer and Wang, 2011):

a) factors pushing flows from the origin location (outflows) which indicate the ability of the origin location to produce or generate flows,
b) factors pulling flows to the destination location (inflows) which show the attractiveness of the destination location,
c) separation function that reflects the way spatial separation of origins from destinations constrains or impedes interaction such as geographical, time, economic, social, political, cultural, technological distance between locations etc.

The model can also examine the determinants of migration flows within locations (intra-regional flows). In its extended version, the model identifies the nature of spatial dependences between locations (see LeSage and Pace, 2008; Griffith and Fischer, 2013).

A researcher should also consider some problems in the construction and estimation of gravity modelling. Bertoli and Fernández-Huertas Moraga, 2013, pay a special attention to multilateral resistance in a gravity model. Santos Silva and Tenreyro, 2006, consider econometric problems resulting from heteroscedastic residuals, variables bias and the zero problem. This paper discusses the scale effect of modifiable areal unit problem which affects the results of a gravity model.

The following study concerns the economic determinants of population migrations in Poland in the years 2011-2013. The study examines factors pushing and pulling migration flows and the role of distance. The paper intentionally uses a relatively simple version of a gravity model and ignores any other problems with the construction of a gravity model to consider the scale effect of MAUP.

The gravity model used in the study (after logarithmic linearization) takes the form of:

$$\overline{Y}^* = \beta_0^* + X_o\overline{\beta}_o + X_d\overline{\beta}_d + \gamma\overline{d}^* + \overline{\varepsilon} \quad (1)$$

where: $\overline{Y}^* = \ln\overline{Y}$, $\overline{Y}$ – vector of flows from origin to destination locations,

$\boldsymbol{X}_o$ ($\boldsymbol{X}_d$) – matrices of explanatory variables realizations in the origin (destination) locations,

$X_o = [\ln\overline{x}_{o1}, \ln\overline{x}_{o2},..., \ln\overline{x}_{ok}]$, $X_d = [\ln\overline{x}_{d1}, \ln\overline{x}_{d2},..., \ln\overline{x}_{dk}]$,

$\bar{d}$ – vector containing distances between each pair of locations,

$\bar{\beta}_o, \bar{\beta}_d \;\; \gamma$ – structural parameters,

β_o^* – constant,

$\bar{\beta}_o = [\beta_{o1}, \beta_{o2}, ..., \beta_{ok}]'$, $\bar{\beta}_d = [\beta_{d1}, \beta_{d2}, ..., \beta_{dk}]'$,

$\bar{\varepsilon}$ – vector of disturbances.

The intensity of domestic migrations is strongly affected by macroeconomic trends. In respect of the registered migrations for permanent residence, the biggest domestic migration flows occurred just before Poland's accession to the European Union (2001-2004) and in the first years of accession (2005-2007) in which Polish economy was in the economic upturn. A big decrease in migration flows in 2008 was a reaction to the world financial and economic crisis. In subsequent years, the intensities of internal migration flows did not fluctuated. The following study covers the years 2011-2013 in which the economic situation in Poland was going to stabilize and the intensity of domestic migration flows was not changing rapidly.

The aggregated number of migration flows for permanent residence from an origin to destination province (NUTS 2 unit) in the years 2011-2013 in relation to 100 thousand inhabitants of the destination province in these years defines the dependent variable. Statistical data was collected from the Demography Database of the Central Statistical Office of Poland. Migration flows occur in territorial space and each origin is also a destination, thus we form a non-symmetric squared data matrix. This matrix is transformed into a data vector according to the approach presented in LeSage and Pace, 2008. An alternative approach is to use a panel gravity model (see Parikh and Van Leuvensteijn, 2002; Bunea, 2012; Pietrzak, Drzewoszewska and Wilk, 2012). This will allow for including provincial fixed effects and considering the issue of multilateral resistance to migration.

A set of explanatory variables was used to explain the changes of the dependent variable. The first subgroup refers to the factors which push and pull migration flows. People usually migrate to improve their living and working conditions. But their migration decisions are frequently affected by their economic and socio-economic situation and environment. In this paper we use Gross Domestic Product *per capita* (in PLN), which is a popular indicator of regional development level, as an explanatory variable of people propensity to migrate.

In an origin province, the level of regional development indicates the factor pushing migration flows to the other provinces, e.g. a weak access to education in a province may provoke massive emigrations. But for a destination province, the level of regional development is a factor pulling migration flows, e.g. relatively low costs of living may attract people to come and live in the province. The values of GDP *per capita* in 16 Polish provinces refers to 2011, which is the year of opening the studied period (2011-2013). Migrations are a long-term reaction to previous economic situation. Statistical data was provided by the Local Data Bank of the Central Statistical Office of Poland.

Other economic features (e.g. investment outlays, salary and wages, etc.) can be also used in the gravity model. But they were statistically correlated with GDP *per capita* and were excluded from the analysis to avoid multicollinearity. The alternative solution is to employ structural equation modelling in the construction of the gravity model (see, e.g. Pietrzak, Żurek, Matusik and Wilk, 2012).

The second subgroup of explanatory variables includes factors which show statistical distances between provinces. In a typical version of a gravity model of migration flows, the geographical distance is used as a separation function. However, Greenwood (1997, pp. 648-720) noticed that geographical distance elasticity of migration declines over time due to modern information, communication and transport technologies. Therefore, the economic distance is an important area of interests. In gravity models the economic distance is defined in a few ways, e.g. transportation costs, economic disproportions between units, e.g. countries, companies (see Conley and Topa, 2002, Horning and Dziadek, 1987, *Reshaping*…., 2009, p. 75, Pietrzak and Wilk, 2014).

In the following study, the economic distance will indicate the scale of the economic disparities between 16 Polish provinces and serve as the last explanatory variable in the gravity model. Because the economic disproportions result from many issues such as the level of economic activity, economic profile, attractiveness of foreign capital, local society's purchasing power and propensity to invest, labour market absorption, entrepreneurship, productivity, capacity of industry, etc., we determined a set of variables to define it (see Table 1).

Table 1. Set of variables defining the economic distances between 16 Polish provinces in 2011

No	Abbreviation	Definition	Unit
1	Investments	Investment outlays in companies per working-age people	PLN
2	Wages and salaries	Average monthly gross wages and salaries	PLN
3	Unemployment	Registered unemployment rate	%
4	Foreign capital	Companies with foreign capital per 10 thousand people	entity
5	Individual businesses	Natural persons conducting economic activity per 100 working-age people	entity
6	Employment in T&S	People employed in trade and service sectors (PKD 2007 classification) per 1 thousand working-age people	person
7	New entities	New entities of the national economy registered in REGON register per 10 thousand people	entity

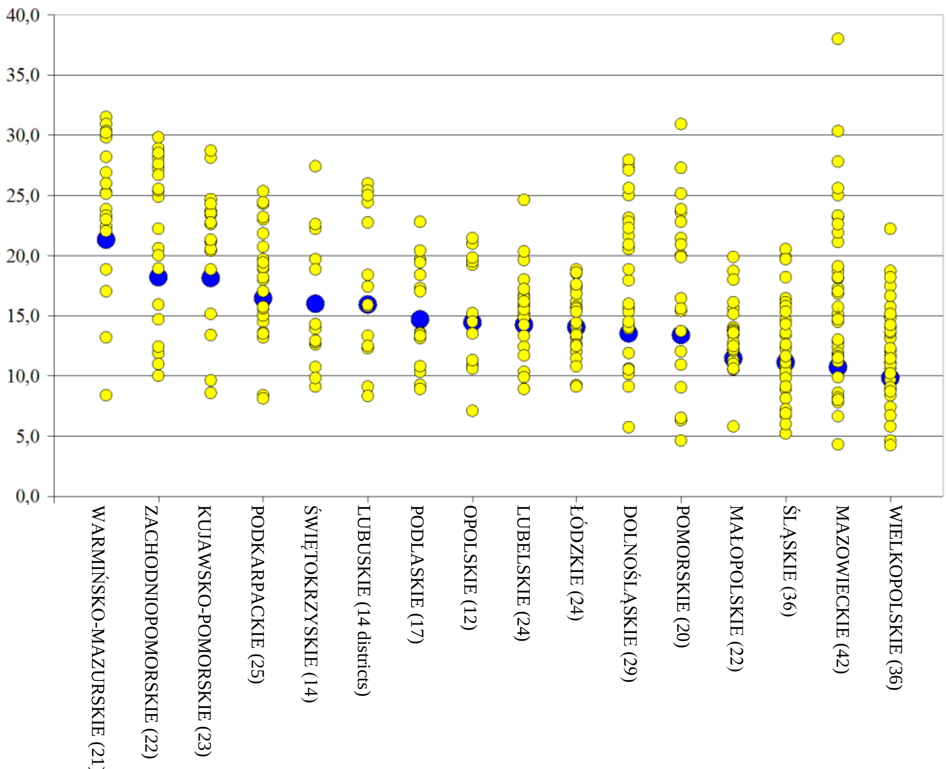
The set of diagnostic variables meet the following application criteria: statistical data availability , comparability, clear definition of the research problem and measurability. High statistical variation and low statistical correlation were also required.

The preliminary data analysis is carried out to examine if the scale effect of MAUP exists. A situation of each province was separately examined according to each variable based on statistical data for its districts (LAU 1 units).

An empirical example of the scale effect of MAUP will be presented based on the Unemployment variable. Figure 3 shows the values of the registered unemployment rate for 16 Polish provinces and 379 Polish districts assigned to provinces they are located. Dark circle tags indicate the values of official statistics for provinces, while grey circle tags show the values of official statistics for districts.

Ranges and spacing show the differences and similarities between provinces in densities, variation and reveal outlier values. For example the Zachodniopomorskie and Kujawsko-pomorskie provinces present the same level of the unemployment rate (approximately 18 %) according to provincial statistics. But in Zachodniopomorskie province the situation is much more serious. Half of its districts note at least 25 % of the unemployment rate, while majority of Kujawsko-pomorskie province's districts record less than 25 % of the unemployment rate.

One of the lowest values of the unemployment rate is presented in the Mazowieckie province (10.7 %), while above 80 % of its districts note higher level of the unemployment rate. In the Podkarpackie and Warmińsko-mazurskie provinces, the outlier values make the official statistics much lower than they would really be.



Explanation: ● province (NUTS 2 unit) official statistics, ● district (LAU 1 unit) official statistics.
WARMIŃSKO-MAZURSKIE (21) the name of a province (the number of districts located in the province)

Figure 3. Registered unemployment rate in Polish provinces and districts in 2011 (%)

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

Table 2 presents province official statistics (POS) of the unemployment rate and basic statistics for provinces based on district data in 2011. In a vast majority of provinces, the coefficient of variation is higher than 20%, which proves that there is a relatively high internal diversification of the unemployment rate. Province official statistics are close to median values. Normal distributions do not exist for any of provinces.

Table 2. Province official statistics (POS) of registered unemployment rate and basic statistics for provinces based on district data in 2011

Name of province (NUTS 2 unit)	POS	Min	Max	Range	Mean	Median	POS per mean	POS per median	Stand. dev.	Coef. of variation (%)	Kurto-sis	Skew-ness
Łódzkie	14.00	9.10	18.80	9.7	14.39	14.00	0.97	1.00	2.78	19.30	-0.82	-0.11
Mazowieckie	10.70	4.30	38.00	33.7	16.56	15.70	0.65	0.68	6.78	40.93	1.13	0.82
Małopolskie	11.40	5.80	19.90	14.1	13.39	13.10	0.85	0.87	3.02	22.59	1.12	0.07
Śląskie	11.10	5.20	20.50	15.3	12.11	11.60	0.92	0.96	4.09	33.80	-0.66	0.37
Lubelskie	14.20	8.90	24.60	15.7	15.46	15.90	0.92	0.89	3.36	21.71	1.34	0.34
Podkarpackie	16.40	8.10	25.30	17.2	17.98	18.90	0.91	0.87	4.49	25.00	-0.11	-0.44
Podlaskie	14.70	8.90	22.80	13.9	15.18	13.60	0.97	1.08	4.05	26.71	-1.03	0.11
Świętokrzyskie	16.00	9.10	27.40	18.3	16.39	14.10	0.98	1.13	5.53	33.74	-0.98	0.49
Lubuskie	15.90	8.30	26.00	17.7	17.61	16.65	0.90	0.95	5.96	33.82	-1.41	0.05
Wielkopolskie	9.80	4.20	22.20	18.0	11.97	11.70	0.82	0.84	4.00	33.39	0.11	0.25
Zachodnio-Pomorskie	18.20	10.00	29.80	19.8	21.84	24.90	0.83	0.73	6.57	30.10	-1.21	-0.56
Dolnośląskie	13.50	5.70	27.90	22.2	17.66	17.90	0.76	0.75	6.35	35.95	-0.94	-0.16
Opolskie	14.40	7.10	21.40	14.3	15.33	14.85	0.94	0.97	4.58	29.85	-1.30	-0.20
Kujawsko-Pomorskie	18.10	8.60	28.70	20.1	21.36	22.60	0.85	0.80	5.19	24.30	1.02	-1.07
Pomorskie	13.40	4.60	30.90	26.3	17.31	18.15	0.77	0.74	7.28	42.05	-0.88	-0.12
Warmińsko-Mazurskie	21.30	8.40	31.50	23.1	24.05	25.10	0.89	0.85	5.76	23.96	1.32	-1.10

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

In view of the preliminary analysis results, we can conclude that official statistics poorly illustrate the real situations of provinces. Similar situation occurs in the rest of variables from Table 1. Therefore the scale effect of modifiable areal unit problem has been detected and requires a special attention.

4. Symbolic data analysis as a tool to reduce the scale effect of MAUP

As Tobler (1989) and Openshaw (1984) advocated: data aggregation is not only a quantitative process; it changes dramatically the main point of the units as well as the variables. The question of MAUP must be posed before the treatment rather than during it. The approach proposed in this paper deals with this standpoint and employs symbolic data analysis to reduce the scale effect of MAUP.

Symbolic data differs from classical data situation. Conventional data set includes the observations of variables which realizations are in the form of real values or single categories. Symbolic data analysis (SDA) deals with another data form to solve data structure and data reduction problems. The first issue refers to the analysis of fuzzy (e.g. freezing interval of liquids), imprecise (e.g. income intervals of respondents), multinominal data (e.g. foreign languages known by job recruits), data fluctuated in time (e.g. investment outlays). The second problem occurs if a set of data is too large or too complicated to be analysed, e.g. a big set of units, a set of correlated variables, a big set of variables and time series.

SDA distinguishes two types of objects (units) regarding the level of data aggregation: first order and second order symbolic objects. First order objects are indivisible units (e.g. respondents, products, patients, etc.) which cannot result from data aggregation, and are described by symbolic data. Second order objects are the result of an aggregation of first order objects described by classical data. These objects can be seen as more or less homogeneous classes of individuals described by symbolic data.

The observations on symbolic variables take the form of intervals of values (interval-valued variables), sets of categories, values or intervals (multivalued variables), values or intervals with associated weights, frequencies, probabilities, etc., (modal variables), and also taxonomic, logical or hierarchical structures (dependent variables). Furthermore, classical data is a special case of symbolic data, e.g. a classical ratio data is equivalent to the point which is a special case of the interval-valued variable realization, a classical categorical data is equivalent to a single realization of a multivalued variable (see Bock and Diday (Eds.), 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture (Ed.), 2008; Wilk, 2011).

Symbolic data occurs in a natural form or results from classical data aggregation. The ways of symbolic data construction in presented in Bock and Diday (Eds.), 2000; Wilk, 2012. In regional research, higher level units (e.g. regions) can be characterized based on situations of lower level units (e.g. towns) using symbolic data.

For example, we gathered a classical data set regarding 16 Polish NUTS 2 units (provinces) and the values of unemployment rate recorded by all Polish LAU 1 units (districts). Each district is located in the territory of one of the provinces. We can distinguish a set of unemployment categories such as low (unemployment rate of less than 10%), medium (unemployment rate of 10-20%) and high unemployment (unemployment rate higher than 20%) levels in sub-regions. Then, for each province we calculate the frequencies of districts which satisfy each category. In this way we have already constructed a set of second order symbolic objects (provinces) described by a symbolic modal variable (see Table 3).

Table 3. Registered unemployment rate in Polish regions in 2010 (%)

NUTS 2 unit (province)		LAU 1 unit (district)		Fraction of LAU 1 units satisfying each level of unemployment*			Symbolic modal variable realizations
Name	Value	Name	Value	low	medium	high	
Mazowieckie	9.7	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	Bydgoszcz	8.0	0.1	0.3	0.6	{low (0.1), medium (0.3), high (0.6)}
		⋮	⋮				
		Lipnowski	28.9				
⋮	…	…	…	…	…	...	…

* low (under 10%), medium (10-20%], high (over 20%) level of unemployment rate

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

The aim of the following procedure is to reduce the scale effect of MAUP in the examination of the economic distance between provinces using symbolic data analysis. In the first step of the procedure we construct a set of second order symbolic objects. These objects represent 16 Polish NUTS 2 units (provinces) which are described by 7 symbolic interval-valued variables. We use a set of indicators presented in Table 1.

In this case, the constructions of variables result from the aggregation of district data. For each province we construct an interval of values which consists of minimum and maximum values recorded by the districts located in a province as regards each variable. But the construction of intervals of values required to detect and remove outlier values. We removed the values which were much higher than 80% of observations for districts in a province but not higher than 10% of observations for districts in the province.

In the next step we determine the statistical distances between symbolic objects. We use Ichino-Yaguchi's normalized distance measure (Ichino and Yaguchi, 1994):

$$d_{ijk} = \mu(v_{ik} \oplus v_{jk}) - \mu(v_{ik} \otimes v_{jk}) + \gamma v(v_{ik}, v_{jk}) \qquad (2)$$

where: i, j – the number of an object, $i \in [1, m]$,

k – the number of a variable, $k \in [1, p]$,

$v_{ik} \oplus v_{jk} \equiv [\min\{\underline{v_{ik}}, \underline{v_{jk}}\}, \max\{\overline{v_{ik}}, \overline{v_{jk}}\}]$,

$v_{ik} \otimes v_{jk} \equiv v_{ik} \cap v_{jk}$,

$\nu(v_{ik}, v_{jk}) \equiv 2\mu(v_{ik} \otimes v_{jk}) - \mu(v_{ik}) - \mu(v_{jk})$,

$\underline{v_{ik}}, \underline{v_{jk}}$ ($\overline{v_{ik}}, \overline{v_{jk}}$) – the start point and end point of interval of values observed by *k* variable, accordingly, for *i* and *j* objects,

γ – parameter, $\gamma \in [0.0, 0.5]$.

The distance measure (Equation 2) examines a dissimilarity between *i* and *j* objects regarding *k* variable. It calculates Cartesian meet ($v_{ik} \otimes v_{jk}$) and Cartesian join ($v_{ik} \oplus v_{jk}$). If the intersection $v_{ik} \otimes v_{jk}$ takes the empty value, both intervals of values observed for *i* and *j* objects have no common part. If both intervals of values observed for *i* and *j* objects have the same minimum and maximum values, the Cartesian join $v_{ik} \oplus v_{jk}$ is an interval with minimum and maximum values observed for *i* or *j* unit (see Ichino and Yaguchi, 1994, Wilk, 2006). Parameter γ takes values from 0.0 to 0.5 but for 0.5 the $\nu(v_{ik}, v_{jk})$ is equal to 0 and no intersection is included. Then, in the following study, the parameter γ is equal to 0.4.

We use Minkowski's metric (with λ equal to 2) to determine the total distance between *i* and *j* objects:

$$d_{ij} = \left[\sum_{k=1}^{p} (d_{ijk})^{\lambda} \right]^{1/\lambda} \tag{3}$$

where: λ – parameter, $\lambda \geq 1$.

Minkowski's metric takes the values in $[0.0, \infty]$. If all variables take the same observations for *i* and *j* objects, then the value of the measure is equal to 0. The higher the values, the longer the distance between *i* and *j* objects. If the distance is short, then the similarity of two compared provinces is high. If all pairs of provinces present very short distances then the economic disparities would be very low as well.

The same distance measure was also implemented to determine dissimilarities between 16 Polish provinces described by province official statistics (ratio data). For ratio data, the intersection $v_{ik} \otimes v_{jk}$ takes the values of 0 if $v_{ik} \neq v_{jk}$. If $v_{ik} = v_{jk}$, then $v_{ik} \otimes v_{jk} \equiv v_{ik} \oplus v_{jk}$.

Table 4 presents the comparison of results of economic distance measurements based on province official statistics and symbolic interval-valued data. It includes basic descriptive statistics for both the distance matrices and the sets of pairs of provinces with very long and very short distances.

Table 4. Comparison of the economic distance measurement results based on two data sets (official statistics and symbolic data)

Data set	Descriptive statistics			Coef. of var. (%)	Economic distance	
	Min	Max	Median		Very short	Very long
Province official statistics	0.19	2.46	0.83	52.77	Lubelskie vs. Podlaskie (0.19) Podlaskie vs. Świętokrzyskie (0.24) Dolnośląskie vs. Pomorskie (0.31) Lubelskie vs. Podkarpackie (0.33)	Mazowieckie vs. Warmińsko-Mazurskie (2.46) Mazowieckie vs. Podkarpackie (2.38) Mazowieckie vs. Lubelskie (2.29) Mazowieckie vs. Świętokrzyskie (2.23) Mazowieckie vs. Podlaskie (2.20)
Symbolic interval-valued data	0.16	0.91	0.53	34.20	Lubelskie vs. Podlaskie (0.16) Lubelskie vs. Podkarpackie (0.16) Lubelskie vs. Łódzkie (0.20)	Mazowieckie vs. Świętokrzyskie (0.91) Świętokrzyskie vs. Wielkopolskie (0.88) Świętokrzyskie vs. Pomorskie (0.87) Mazowieckie vs. Podlaskie (0.87)

Source: own estimation in symbolicDA package (Dudek, Pełka and Wilk, 2013) of R-CRAN, based on Local Data Bank of the Central Statistical Office of Poland.

In both cases distances between provinces are higher than 0 which means that there are no two provinces with the same values of all variables and the economic disproportions in Poland exist. The comparison of distance ranges proves that the scale of the economic disparities in Poland is smaller than official statistics show. Moreover, the economic relations between provinces based on symbolic interval-valued data set differ from those presented by province official statistics.

5. Conditions of domestic population migration in Poland

Three gravity models of population migration in Poland in 2011 were used in the following study (Table 5). The aggregated number of migration flows for permanent residence from an origin to destination province (NUTS 2 unit) in the years 2011-2013 in relation to 100 thousand inhabitants of the destination province in the same period defines the dependent variable.

Table 5. Gravity models of domestic population migration in Poland in 2011

Specification	Model 1	Model 2	Model 3
Pulling and pushing factor	GDP *per capita*	GDP *per capita*	GDP *per capita*
Separation function	Geographical distance	Economic distance	Economic distance
Data set	Province official statistics	Province official statistics	Symbolic interval-valued data

Source: own elaboration.

All models include two types of dependent variables: a pushing and pulling factor and a separation function as well. All models employ GDP *per capita* to measure the impact of regional development level in an origin to push migration flows and attractiveness of a destination to pull migration flows.

But the first model examines geographical distance as well. In this study, the number of kilometres in a straight line as regards the centroids served to determine the geographical distances between provinces. This model is based on province official statistics.

In contrast to the first model, the second and third models examine the role of the economic distance. But the second model employs province official statistics (ratio data) to determine the economic distance between provinces. In the third model, the economic distance between each pair of provinces is determined based on symbolic interval-valued data. Economic distances between provinces include their internal situation and economic disproportions between districts.

Table 6 presents the results of gravity models of population migration. All models use GDP *per capita* to explain pushing and pulling factors of flows but the models differ from the type of distance used and the method of its determination. The first model concerns the geographical distance. The other models consider the economic distance. One of them is based on official statistics recorded by provinces. The second one is based on symbolic data set.

Table 6. The estimates of gravity models of domestic population migration in Poland in 2011

Parameter		Model 1		Model 2		Model 3	
		estimate	p-value	estimate	p-value	estimate	p-value
Constant	α_1	-100,095	0,093***	-103,749	0,093*	109,936	0,047**
The impact of economic situation of an origin on population outflow from the origin	β_o	0,231	0,001***	0,005	0,000***	0,0034	0,002***
The impact of economic situation of a destination on population inflow to the destination	β_d	0,396	0,000***	0,005	0,000***	0,0033	0,003***
The impact of distance on population flows	γ	-102,008	0,001***	216,765	0,000***	583,379	0,000***
R^2 coefficient		0,46		0,54		0,51	

Significance level: *** 5%, ** 10%, * 15%

Source: own estimation in Gretl based on Local Data Bank of the Central Statistical Office of Poland.

The least squares method was used in estimation. All estimated values are statistically significant. Estimates of both origin-destination flows parameters (β_o and β_d) are positive. The better the economic situation in a province, the higher the population migration inflows, as well as population outflows.

The geographical distance estimate is negative, whereas the economic distance estimates are positive in both approaches. Therefore, the longer the geographical distance, the lower the migration flows. People migrate on relatively short distances in Poland. But the longer the economic distance, the higher the migration flows. Moreover, migration movement is big when the economic disparities occur in the country.

The estimates of the economic distance in both models are comparable due to using the same distance measure. The estimate in the model based on symbolic data is much higher than the estimate in the model based on official statistics. Therefore, the role of the economic distance in domestic migration flows is in fact more serious than official statistics show.

6. Conclusions

The presented study discussed the modifiable areal unit problem (MAUP) in spatial data analysis and proposed to use symbolic data analysis (SDA) to reduce the scale effect of MAUP in the gravity model of population migration.

Symbolic data was used to measure the economic disparities in Poland. This approach involves a special way of data aggregation. Its main advantage is to include details rather than "averaged" values. In fact, the scale of economic diversification in Poland is smaller whereas the role of the economic distance in domestic migration flows is more serious than official statistics show. The disadvantage of this approach is a highly complicated procedure, which includes preliminary data aggregation and the use of distance measure dedicated to symbolic data analysis.

An open issue is symbolic variables and construction of objects. In the following study, the second order symbolic objects (provinces) included district-level data due to statistical data availability. The ideal situation would be to dispose individual data for non-modifiable units (e.g. towns) as a starting point in data aggregation procedure. But some data is not presented due to statistical confidentiality (e.g. suicide data) and some statistical surveys are not carried out at the local level of territorial division (e.g. GDP *per capita*).

The second open issue is an appropriate construction of symbolic variables, which is an individual case and depends on the properties of a variable. In the study, aggregation of classical ratio data into symbolic interval data required removing outlier values. In this situation we lost some information. Moreover, if symbolic interval data is not a natural form of data, a symbolic interval-valued variable covers some additional values which were not recorded by any analysed units. In these situations we can try to use another type of symbolic variables, e.g. modal variables. This problem will be an objective of further research study.

REFERENCES

ANDERSON, J. E., (1979). A Theoretical Foundation for the Gravity Model, American Economic Review, 69 (1), pp. 106−116.

ANSELIN, L., (1988). Spatial Econometrics: Methods and Models, Kluwer Academic, Dordrecht.

ARBIA, G., (1989). Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems, Advanced Studies in Theoretical and Applied Econometrics, Vol. 14, Kluwer Academic Publishers, Dordrecht-Boston-London.

BEINE, M., BERTOLI, S., FERNÁNDEZ-HUERTAS MORAGA, J., (2015). A Practitioners' Guide to Gravity Models of International Migration, World Economy.

BENALI, H., ESCOFIER, B., (1990). Analyse factorielle lissée et analyse factorielle des différences locales, Revue de statistiques appliquées, XXXVIII (2), pp. 55−76.

BERTOLI, S., FERNÁNDEZ-HUERTAS MORAGA, J., (2013). Multilateral resistance to migration, Journal of Development Economics, 102, May, pp. 79−100.

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

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

BUNEA, D., (2012). Modern Gravity Models of Internal Migration. The Case of Romania, Theoretical and Applied Economics, Vol. XIX, No. 4(569), pp. 127−144.

CHOJNICKI, Z., (1966). Application of gravity and potential models in spatio-economic research (in polish), PWN, Warsaw.

CHOJNICKI, Z., CZYŻ, T., RATAJCZAK, W., (2011). Potential model. Theoretical basis and applications in spatio-economic and regional research (in polish), Bogucki Wydawnictwo Naukowe, Poznan.

CLIFF, A. D., ORD, J. K., (1981). Spatial processes: models and applications, Pion, London.

CONLEY, T. G., TOPA, G., (2002). Socio-economic distance and spatial patterns in unemployment, Journal of Applied Econometrics, Vol. 17/4.

DARK, S. J., BRAM, D., (2007). The modifiable areal unit problem (MAUP) in physical geography, “Progress in Physical Geography”, Vol. 31, No. 5.

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

DUDEK, A., PEŁKA, M., WILK, J., (2013). symbolicDA package of R-CRAN, http://cran.r-project.org/web/packages/symbolicDA/index.html.

FISCHER, M. M., WANG, J., (2011). Spatial data analysis: models, methods and techniques, Springer Briefs in Regional Science, Springer, Berlin.

FOTHERINGHAM, A. S., CHARLTON, M.E., BRUNSDON, C. F., (2001). Spatial Variations in School Performance: a Local Analysis Using Geographically Weighted Regression, Geographical & Environmental Modelling, Vol. 5, pp. 43−66.

FOTHERINGHAM, A. S., O’KELLY, M. E., (1989). Spatial interaction models: formulations and applications, Kluwer, Dordrecht,

GATNAR, E., WALESIAK, M. (Ed.), (2011). Qualitative and symbolic data analysis using R program (in polish), C.H. Beck, Warsaw.

GEHLKE, C. E., BIEHL, K., (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material, Journal of the American Statistical Association, No. 29.

GHATAK, S., MULHERN, A., WATSON, J., (2008). Inter-regional migration in transition economies. The case of Poland, Review of Development Economics, No. 12(1), Oxford, pp. 209−222.

GOTWAY CRAWFORD, C. A., YOUNG, L. J., (2004). A spatial view of the ecological inference problem, In: G. King, O. Rosen, M. Tanner (Eds.), Ecological Inference: New Methodological Strategies, Cambrige University Press, pp. 233−244.

GRABIŃSKI, T., MALINA, A., WYDYMUS, S., ZELIAŚ, A., (1988). International statistics methods (in polish), PWE, Warsaw.

GREENWOOD, M. J., (1997). Internal migration in developed countries, In: Rosenzweig, M. R. and Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1B, Elsevier, Amsterdam, pp. 648−720.

GRIFFITH, D. A., FISCHER, M., (2013). Constrained variants of the gravity model and spatial dependence: Model specification and estimation issues, Journal of Geographical Systems, 15(3), pp. 291−317.

HOLZER, J. Z., (2003). Demography (in polish), PWE, Warsaw.

HORNING, A., DZIADEK S., (1987). Outline of land transport geography (in polish), PWN.

ICHINO, M., YAGUCHI, H., (1994). Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 24, No. 4.

ISARD, W., (1960). Methods of regional analysis, MIT Press, Cambridge.

JELINSKI, D. E., WU, J., (1996). The modifiable areal unit problem and implications for landscape ecology, Landscape Ecology, Vol. 11, No. 3, pp. 129−140

KING, G., (1997). A solution to the ecological inference problem: reconstructing individual behaviour from aggregate data, Princeton University Press, Princeton.

KING, G., TANNER M. A., ROSEN O. (Eds.), (2004). Ecological Inference: New Methodological Strategies, Cambridge University Press, New York.

KAPISZEWSKI, M., DURHAM, H., REES, P., (1999). Internal Migration and Regional Population Dynamics In Europe: Poland Case Study, In: P. Rees, M. Kupiszewski (Eds.), Internal Migration and Regional Population Dynamics in Europe: A Synthesis, Collection Demography, Council of Europe, Strasbourg.

LESAGE, J. P., PACE, R. K., (2008). Spatial econometric modeling of origin-destination flows, Journal of Regional Science, Vol. 48, No. 5, pp. 941−968.

Local Data Bank of the Central Statistical Office of Poland, http://www.stat.gov.pl/bdl.

LUCAS, R., (1997). Internal Migration in Developing Countries, In: M.R. Rosenzweig, O. Stark (Eds.), Handbook of Population and Family Economics, Elsevier Science B.V, Amsterdam, pp. 721−798.

MOELLERING, H., TOBLER, W. R., (1972). Geographical variances, Geographical Analysis, Vol. 4, pp. 34−50.

OPENSHAW, S., TAYLOR, P. J., (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem, In: N. Wringley (Ed.), Statistical Applications in the spatial sciences, Pion, London, pp. 127−144.

OPENSHAW, S., (1984). The Modifiable Areal Unit Problem, GeoBooks, CATMOG 38, Norwich.

PARIKH, A., VAN LEUVENSTEIJN, M., (2002). Internal migration in regions of Germany: A panel data analysis, Working Paper, No. 12, European Network of Economic Policy Research Institutes.

PIETRZAK, M., ŻUREK, M., MATUSIK, S., WILK, J., (2012). Application of Structural Equation Modeling for analysing internal migration phenomena in Poland, Przegląd Statystyczny (Statistical Review), No 4, Vol. LIX, pp. 487−503.

PIETRZAK, M. B., DRZEWOSZEWSKA, N., WILK, J., (2012). The analysis of interregional migrations in Poland in the period of 2004-2010 using panel gravity model, Dynamic Econometric Models, Vol. 12(2012), pp. 111−122.

PIETRZAK, M., WILK, J., (2014). The economic distance in spatial phenomena modelling with the use of gravity model (in polish), In: K. Jajuga, M. Walesiak (red.), Taksonomia 22. Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 37, pp. 177−185.

Reshaping Economic Geography, (2009). World Bank, Washington.

ROY, J. R., (2004). Spatial Interaction Modeling: A Regional Science Context, Springer-Verlag, Berlin.

SANTOS SILVA, J., TENREYRO, S., (2006). The log of gravity, The Review of Economics and Statistics, 88, pp. 641−58.

SEN, A., SMITH, T. E., (1995). Gravity models of spatial interaction behavior, Springer, Berlin-Heilderberg-New York.

SHEPHERD, B., (2013). The Gravity Model of International Trade: A User Guide, United Nations ESCAP & ARTNET.

SUCHECKA, J. (Ed.), (2014). Spatial statistics. Spatial structures analysis methods (in polish), C.H. Beck., Warsaw.

SUCHECKI, B. (Ed.), (2010). Spatial econometrics. Spatial data analysis methods and models (in polish), C.H. Beck, Warsaw.

TATE, N. J., ATKINSON, P. M. (Eds.), (2001). Modelling scale in geographical information sciences, Wiley & Sons, London.

TOBLER, W., (1979). Smooth pycnophylactic interpolation for geographical regions, Journal of the American Statistical Association, Vol. 74, pp. 519−536.

TOBLER, W., (1989). Frame independent spatial analysis, In: Accuracy of Spatial Databases, M. Goodchild S. Gopal (Eds.), CRC Press, pp.115−122.

TODARO, M., (1980). Internal Migration in Developing Countries. A survey, In: R. A. Easterlin, Population and Economic Change in Developing Countries, University of Chicago Press, Chicago, pp. 361−402.

VAN DER GAAG, N. (Ed.), (2003). Study of past and future interregional migration trends and patterns within European Union countries: in search of a generally applicable explanatory model, Report on behalf of Eurostat.

WHITE, M. J., LINDSTROM, D. P., (2006). Internal Migration, In: D.L. Poston, M. Micklin (Eds.), Handbook of Population, Springer, Berlin-Heilderberg.

WILK, J., (2012). Symbolic approach in regional analyses, Statistics in Transition – new series, Vol. 13, No 3, December 2012, pp. 581−600, http://stat.gov.pl/cps/rde/xbcr/pts/SIT_13_3_December_2012.pdf.

WILK, J., (2011). Cluster analysis based on symbolic data (In Polish), In: E. Gatnar, M. Walesiak (Eds.), Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R, C.H. Beck, Warsaw, pp. 262−279.

WILK, J., (2006). Problems of symbolic objects classification. Symbolic distance measures (In Polish), In: J. Garczarczyk (Ed.), Ilościowe i jakościowe metody badania rynku. Pomiar i jego skuteczność, Zeszyty Naukowe Akademii Ekonomicznej w Poznaniu, 71, pp. 69−83.

WONG, D., The modifiable areal unit problem (MAUP), (2009). In: Fotheringham A.S., Rogerson P.A. (Eds.), The SAGE Handbook of Spatial Analysis, SAGE Publications Ltd., pp. 105−123.

YULE, U., KENDALL M. S., (1950). An introduction to the theory of statistics, Charles Griffin, London.

ZELIAŚ, A. (Ed.), (1991). Spatial econometrics (In Polish), PWE, Warsaw.