EQUILIBRIUM
Quarterly Journal of Economics and Economic Policy
2014 VOLUME 9 ISSUE 4, December
p-ISSN 1689-765X, e-ISSN 2353-3293
www.economic-policy.pl

Pietrzak, M. B. (2014). The Modifiable Areal Unit Problem – Analysis of Correlation and Regression. *Equilibrium. Quarterly Journal of Economics and Economic Policy,* 9(3), pp. 113-131, DOI: http://dx.doi.org/10.12775/EQUIL.2014.028

Michał Bernard Pietrzak[*]
Nicolaus Copernicus University, Poland

The Modifiable Areal Unit Problem – Analysis of Correlation and Regression[**]

JEL Classification: *C01; C15; C21*

Keywords: *spatial econometrics; modifiable areal unit problem; scale problem*

Abstract: *The paper focuses on the issue of the modifiable areal unit problem, which means a possibility of obtaining various results for spatial economic analyses depending on the assumed composition of territorial units. The major research objective of the work is to examine the scale problem that constitutes one of the aspects of the modifiable areal unit problem. Analysis of the scale problem will be conducted for two research problems, namely, for the problem of the causal relationships between the level of investment outlays in enterprises per capita and the number of entities of the national economy per capita, and the issue of the dependence between the registered unemployment rate and the level of investment outlays per capita. The calculations based on the empirical values of those variables have showed that moving to a higher level of aggregation resulted in a change in the estimates of the parameters. The results obtained were the justification for undertaking the realisation of the objective. The scale problem was considered by means of a simulation analysis with a special emphasis laid on differentiating the variables expressed in absolute quantities and ones expressed in relative quantities.*

The study conducted allowed the identification of changes in basic properties as well as in correlation of the researched variables expressed in absolute and

Date of submission: December 19, 2013; date of acceptance: May 18, 2014

[*] Contact: pietrzak@umk.pl, Nicolaus Copernicus University, ul. Gagarina 13A, 87-100 Toruń, Poland

[**] The project was co-financed by Nicolaus Copernicus University in Toruń within a UMK research grant no. 1481-E.

relative quantities. Based on the findings, it was stated that a correlation analysis and a regression analysis may lead to different conclusions depending on the assumed level of aggregation. The realisation of the research objective set in the paper also showed the need to consider the adequate character of variables in both spatial economic analyses and during the examination of the scale problem.

Introduction

The content of the paper concerns the *modifiable areal unit problem* which was defined as the changeability of data properties under a change in a composition of territorial units. This issue is essential for conducting spatial analyses, since it indicates a possibility of obtaining different results under the impact of a change in the aggregation level, or under the impact of the assumption of another composition of territorial units within one aggregation level. The issue of the *modifiable areal unit problem* was already considered in the works of the following: Gehlke and Biehl (1934), Yule and Kendall (1950), Robinson (1950), Blalock (1964), Openshaw and Taylor (1979), Openshaw (1984), Anselin (1988), Reynolds (1988), Arbia (1989), Fotheringharn and Wong (1991), Holt *et al.* (1996), Tranmer and Steel (2001), Manley *et al.* (2006), Suchecki (2010), Flowerdew (2011), Pietrzak (2014a), Pietrzak (2014b).

Within the *modifiable areal unit problem* one of its aspects is going to be examined – the *scale problem*. The *scale problem* will be defined as a problem of changing spatial data properties and casual relationships under the impact of a change in the aggregation level. The *scale problem* ought to be analysed exclusively for compositions of territorial units forming a *quasi composition of regions*. Such an approach to the *scale problem* was described in the works of Pietrzak (2014a), Pietrzak (2014b). This approach necessitates determining a *quasi composition of regions* which means a set of particular composition of territorial units for subsequent aggregation levels. Particular composition of territorial units needs to be selected in a way that allows drawing correct conclusions within the research problem undertaken.

The major research objective of the present paper is to consider the *scale problem* applying a simulation analysis with a special emphasis laid on differentiating the variables expressed in absolute quantities and ones expressed in relative quantities. This results from the fact that spatial analyses conducted should be based predominantly on data expressed in relative quantities and referred to certain values describing an irregular region (size, population) with a view to providing data comparability and the correctness of the results obtained. Therefore, also simulation analyses conducted within the *scale problem* should be based on examining changes in the properties of the variables expressed in relative quantities. In the case of variables expressed in relative quantities, we face the problem of adequate aggregation which needs to be performed in an indirect way. On a lower aggregation level, every variable expressed in relative quantities should be split into two variables expressed in the absolute quantities and the aggregation should concern the variables expressed in the absolute quantities. Such an approach enables us to designate variables expressed in relative quantities at a higher aggregation level based on the aggregated variables expressed in absolute quantities.

The simulation analysis conducted in the present paper allowed the identification of changes in basic statistics of the researched variables expressed in absolute and relative quantities, changes in the correlation between variables as well as in the regression dependence (causal relationships) between variables expressed in the relative quantities resulting from an aggregation process. Based on the results obtained, the varied reactions of variables expressed in absolute quantities and of variables expressed in relative quantities were confirmed. The realisation of the so formulated research objective in the paper allowed the identification of the need to consider, while researching the *scale problem*, an adequate character of variables.

Analysis of changes in the correlation of variables expressed in relative quantities with a common denominator

The research on the *scale problem* was started with raising the issue of research in the form of regression analysis for selected economic variables. The paper considers the causal relationship between the level of investment outlays in enterprises *per capita* (total capital expenditure for the period 2008–2011 relative to the population in 2011) and the number of entities of the national economy *per capita* (in 2011) in Poland. Apart from the complexity of the problem[1], the simplest equation is the positive effect of the number of entities on the size of the investment outlays. The increase in the number of entities should be translated by a better economic situation in the region to a higher level of investment outlays. In the case of spatial economic analysis, both variables should be expressed in *per capita* terms. This means that both variables are expressed in relative quantities and additionally have the same denominator (population). Table 1 contains designa-

[1] It is about the interaction of the two variables over time, with the omission of the impact of other variables, and the omission of the existing spatial dependence.

tions and description of the considered variables expressed in relative quantities and constituent variables expressed in absolute quantities.

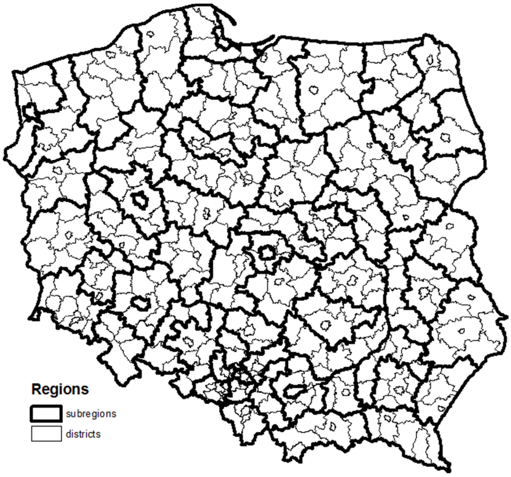
Table 1. Designation and description of selected economic variables

Variable	Description
y1	total investment outlays in enterprises for the period 2008-2011 relative to the population in 2011
y2	number of entities of the national economy *per capita* in 2011
x1	total investment outlays in enterprises for the period 2008-2011
x2	number of entities of the national economy in 2011
x3	population in 2011

Source: elaborated by the author.

The next step in researching the *scale problem* should consist in designating a *quasi composition of regions* for both variables in the research problem posed. The data on the size of investment outlays are available for the NUTS 4 (districts) – NUTS 0 (countries) compositions (aggregation levels), the number of entities for the NUTS 5 (municipalities) – NUTS 0 compositions and the population for NUTS 5 – NUTS 0 compositions. A *quasi composition of regions* must be created in such a way that analysis of the dependence between variables should ensure the correctness of the results obtained for each of its constituent particular composition of territorial units. In this case, out of the available systems ranging from NUTS 4 to NUTS 0 compositions, a *quasi composition of regions* should be established with NUTS 4 (districts) and the NUTS 3 (subregions) compositions (see Graph 1). For NUTS 2 (provinces), NUTS 1 (regions) and NUTS 0 (country – Poland) compositions, due to the large area, both variables are characterized by spatial heterogeneity (see Anselin, 1988). Consequently, the analysis of correlation and regression for the three levels of aggregation will lead to incorrect conclusions.

Graph 1. Division of Poland into districts (NUTS 4) and subregions (NUTS 3)



Source: elaborated by the author.

After determining the *quasi composition of regions*, the strength and the direction of the correlation were calculated for selected economic variables (see Table 1). In addition, the regression dependence was determined between investment outlays *per capita* (y1) and number of entities *per capita* (y2). The results of the calculations for both, the NUTS 4 level of aggregation, as well as for a higher level of aggregation NUTS 3, are shown in Table 2. The analysis of the results obtained leads to the conclusion that among all the variables there occurs a strong positive correlation. The regression parameter was statistically significant and indicates a positive dependence between the level of investment outlays and the number of entities. In addition, there are changes in correlation during the transition from NUTS 4 composition to a higher level of aggregation of NUTS 3 composition. There is an increase in the strength of correlation between investment outlays *per capita* (y1) and other variables (y2, x1, x2, x3). The same is true for the number of entities *per capita* (y2). However, close to one, the correlation between the variables expressed in the absolute quantities x1, x2, x3, decreased. The increase in the strength of the correlation for the pair of variables (y1, y2) had an impact on the estimate of the regression parameter, and contributed to a better fit of the model to empirical data.

Table 2. The correlation and regression dependence for the variables contained in Table 1

Level of aggregation (NUTS 4 composition)					
Correlation					
	y1	y2	x3	x1	x2
y1	1				
y2	0,35	1			
x3	0,31	0,45	1		
x1	0,45	0,39	0,89	1	
x2	0,32	0,51	0,97	0,96	1
Analysis of regression					
Regression coefficient	1,16	Standard error	0,16	Coefficient of determination	0,12
Level of aggregation (NUTS 3 composition)					
Correlation					
	y1	y2	x3	x1	x2
y1	1				
y2	0,69	1			
x3	0,57	0,47	1		
x1	0,85	0,63	0,81	1	
x2	0,75	0,76	0,88	0,91	1
Analysis of regression					
Regression coefficient	1,71	Standard error	0,22	Coefficient of determination	0,48

Source: elaborated by the author.

The following example shows that obviously there has been a change in the property of the variables during the transition to a higher level of aggregation. This is an important reason to consider the *scale problem* with simulation analysis. The analysis of this problem will consist in examining the changes in the properties of the simulated variables and dependence under the influence of the aggregation process. Aggregation of simulated variables must be performed within the taken *quasi composition of regions*.

Considering the *scale problem* will answer the question of whether there is a likelihood of changes in the results obtained under the influence of the aggregation process. This is important information, because the change in properties may result from sources other than the process of aggregation. A positive response also means that the information must be taken into account in the research as part of a research problem posed. A positive response, however, leaves open the question whether a change in the properties of variables results only from the aggregation process, or from other reasons (spatial trend, spatial autocorrelation).

Within the simulation analysis that was carried out in the paper, it was assumed that the simulated variables only have a linear correlation with each other. Therefore, it was examined whether, under such conditions, the aggregation process will change the direction and strength of the simulated correlation between variables. An additional assumption was made that the variables y1 and y2 are the variables expressed in relative quantities and have the same denominator. The variables y1 and y2 will be expressed by formulas.

As a result of the simulation analysis only variables y1, y2 and x3 were generated, and the values of the variables x1, x2 were calculated from formula 1. The assumed correlations between simulated variables y1, y2 and x3 are established in such a way (see Table 3) that they should be similar to the empirical correlation of the selected economic variables. The standard deviation values for the variables y1, y2 were chosen so that generated regression dependence was consistent with the value shown in Table 2 (1.16 value). The mean values for the variables y1, y2 have been assumed at such a level that the coefficient of variation showing the percentage of the standard deviation of the variable in the mean was equal to 10%[2]. The mean and the standard deviation for the variable x3 were assumed arbitrarily by the author. The variables y1, y2, x3 were generated as a vector of Gaussian random fields (see Arbia, 1989; Szulc, 2007). The assumed mean and standard deviation are presented in Table 4. In the case of the variables x1, x2, the mean, standard deviations and the correlation resulted from the calculations made.

Table 3. The assumed correlation during the simulation of the variables y1, y2, x3

Correlation			
	y1	y2	x3
y1	1		
y2	0,35	1	
x3	0,30	0,45	1

Source: elaborated by the author.

[2] The level was assumed arbitrarily by the author.

Table 4. The assumed mean and the standard deviation during the simulation of the variables y1, y2, x3

Properties			
Variables	y1	y2	x3
Mean	5	1,5	2500
Standard deviation	0,5	0,15	250

Source: elaborated by the author.

In the first step of the simulation analysis variables y1, y2 and x3 were generated at the NUTS 4 composition with assumed properties and correlation. Then, based on the realizations obtained, the values of the variables x1, x2 were designated according to the formula 1. It was decided to generate such variables because of the possibility of setting any regression dependence between variables y1 and y2. In this way, the resulting output is a set of variables x1, x2, x3, y1, y2 at the aggregation level NUTS 4 with specific properties[3].

In the second step of the simulation analysis the aggregation of simulated variables for NUTS 3 composition was performed. The aggregation was carried out by the sum of the values of variables from the respective regions. It should be noted that the aggregation has been performed solely for the variables x1, x2, x3 expressed in absolute quantities. The values of variables y1, y2 at the NUTS 3 composition are obtained in accordance with formula 1. As a result of the simulation, 1000 realizations were obtained for each variable at each aggregation level. The possessed realizations of variables at both NUTS 4 and NUTS 3 compositions made it possible to check whether there was a change of their properties. A positive answer would mean that in the analysis of correlation and regression there are changes in the estimates of the parameters under the influence of the aggregation process.

In the third step, for each variable selected statistics were calculated at both levels of aggregation. Then, for each of the statistics, the average and standard deviation were determined based on a set of 1000 estimates. The results obtained are shown in Table 5 and Table 6[4].

[3] It should be noted that the simulation carried out has defects in the form of lack of control on some of correlation between variables. This is due to the fact that three variables are simulated with the properties contained in Table 3 and Table 4. The next two variables are determined indirectly, which does not provide for their properties consistent with those in Table 2.

[4] Based on the realisation of single variables a set of 1000 estimates for each of the statistics was received.

Table 5. The results of the simulation - basic statistics for simulated variables and results of regression analysis

Before aggregation (NUTS 4 composition)					
Variables	y1	y2	x3	x1	x2
Mean	5 (0,025)	1,5 (0,008)	2500 (13,320)	12524 (103,714)	3751 (34,738)
Standard deviation	0,5 (0,017)	0,15 (0,005)	250 (9,906)	2020 (71,491)	638 (9,489)
I statistics	-0,012 (0,032)	-0,006 (0,032)	0,001 (0,036)	-0,008 (0,035)	-0,011 (0,032)
P-value	0,584 (0,277)	0,532 (0,291)	0,474 (0,298)	0,535 (0,305)	0,552 (0,285)
Analysis of regression					
Regression coefficient	1,16 (0,179)	Standard error	0,16 (0,008)	Coefficient of determination	0,12 (0,032)
After aggregation (NUTS 3 composition)					
Variables	y1	y2	x3	x1	x2
Mean	5 (0,030)	1,5 (0,009)	14354 (77,351)	71994 (601,444)	12620 (200,117)
Standard deviation	0,25 (0,032)	0,075 (0,009)	5772 (81,275)	29210 (650,425)	8786 (208,626)
I statistics	-0,012 (0,032)	-0,006 (0,032)	0,001 (0,036)	-0,008 (0,035)	-0,011 (0,032)
P-value	0,584 (0,277)	0,532 (0,291)	0,474 (0,298)	0,535 (0,305)	0,552 (0,285)
Analysis of regression					
Regression coefficient	1,16 (0,557)	Standard error	0,38 (0,078)	Coefficient of determination	0,13 (0,101)

Source: elaborated by the author.

Table 6. The results of the simulation – correlation for simulated variables

Before aggregation (NUTS 4 composition)					
	y1	y2	x3	x1	x2
y1	1				
y2	0,35 (0,047)	1			
x3	0,30 (0,040)	0,45 (0,043)	1		
x1	0,80 (0,017)	0,49 (0,043)	0,80 (0,016)	1	
x2	0,38 (0,040)	0,85 (0,013)	0,85 (0,014)	0,76 (0,023)	1

Table 6 continued

After aggregation (NUTS 3 composition)					
	y1	y2	x3	x1	x2
y1	1				
y2	0,353 (0,153)	1			
x3	0,035 (0,170)	0,061 (0,169)	1		
x1	0,120 (0,171)	0,093 (0,179)	0,994 (0,001)	1	
x2	0,066 (0,171)	0,134 (0,178)	0,994 (0,001)	0,066 (0,171)	1

Source: elaborated by the author.

The fourth and the last step consisted in analysing the results and drawing conclusions. In the case of the variables y1 and y2 expressed in relative values, the received average for the mean remained unchanged, while the average for the standard deviation decreased as a result of the aggregation process. The average for the mean and the standard deviation for variables appointed x1, x2, x3 expressed in the absolute quantities increased as a result of the sum of the values of these variables in the process of aggregation.

The average obtained for the correlation between y1 and y2 has not changed. It should be noted, however, that in this case the standard deviation (error) of the received estimate of the Pearson coefficient increased significantly (more than three times). Although the average values are obtained at similar levels, the significant increase in the standard deviation means that as a result of the aggregation process the correlation coefficient value may change considerably.

The correlation between the variables y1, y2 and the variables x1, x2, x3 at a higher level of aggregation was reduced almost to zero with a significant increase in the error of received estimates. The correlation between the variable x3 and the variables x1, x2 increased. However, the correlation between the variable x1 and x2 decreased to almost zero, with a significant increase in error (the standard deviation value).

In the case of the linear regression between the variables y1 and y2 we received similar average of the regression parameter and the coefficient of determination for both levels of aggregation. It should be stressed, however, that both the regression parameter and the coefficient of determination obtained a much higher standard deviation at the NUTS 3 composition (aggregation level). In addition, as a result of aggregation, the value of the standard error of the estimate of the regression parameter increased significantly. This means that as a result of the aggregation process the output regression dependence is not maintained and often changes substantially, with the possible lack of statistical significance of the regression parameter.

In addition, using Moran's I test, the existence of spatial autocorrelation for all variables was examined. For none of the variables, there was no spatial autocorrelation at a higher NUTS 3 composition. This means that the aggregation of variables did not affect the appearance of spatial dependence.

Analysis of changes in the correlation of variables expressed in relative quantities with a different denominator

Research on the *scale problem* was expanded by a further research problem in the form of regression analysis between the registered unemployment rate and investment outlays in enterprises *per capita* in Poland, both variables for the 2011. A negative regression dependence should occur between these variables. Regions with a higher level of investment outlays *per capita* are usually characterized by a lower unemployment rate compared to regions with lower levels of investment outlays. The registered unemployment rate is expressed as a percentage of the unemployed to the economically active population. Investment outlays *per capita* are expressed in relation to the population. Both variables are expressed in relative quantities, and each has a different denominator. Notation and a description of the variables considered are shown in Table 7.

Table 7. Designation and description of selected economic variables

Variable	Description
y1	registered unemployment rate in 2011
y2	total investment outlays in enterprises for the period 2008-2011 relative to the population in 2011
x3	number of the unemployed in 2011
x1	number of the economically active population in 2011
x2	total investment outlays in enterprises for the period 2008-2011
x4	population in 2011

Source: elaborated by the author.

Also in this case it is necessary to determine *quasi composition of regions*. Both the data for the registered unemployment rate and investment outlays *per capita* are available for NUTS 4 – NUTS 0 compositions. The NUTS 2 – NUTS 0 compositions seem to be a very high level of aggregation due to the spatial heterogeneity of the variables considered. As in the case of investments outlays and number of entities, a *quasi composition of regions* will consist of only two levels of aggregation: NUTS 4 and NUTS 3 compositions.

After determining the *quasi composition of regions*, the correlations for all pair of the variables and the regression dependence between the unemployment rate (y1) and the size of investment outlays (y2) were designated. The results obtained are shown in Table 8. The analysis of the results leads to the conclusion that between the majority of variables there is a strong correlation, both positive and negative. Between the registered unemployment rate (y1) and the variables y2, x1, x2, x3 and between investment outlays (y2) and the variables x1, x2, x3 there was an increase in the strength of the correlation. In the case of pairs of the variables (x2, x4), (x2, x3), (x3, x4) expressed in the absolute quantities, the strength of correlation dependence remained at a similar level, for the pairs of the variables (x1, x2), (x1, x4) the correlation strength significantly decreased, and for a pair of the variables (x1, x3) it is close to zero.

The regression parameter proved to be statistically significant and indicates a negative regression dependence between the unemployment rate and the level of investment outlays *per capita*. Also in this case, the increase in the strength of the correlation between variables y1 and y2 influenced the change in the estimate of regression parameter and the coefficient of determination.

The analysis of the results contained in Table 8 shows that also in this case there has been a change in the properties of variables in the transition to a higher level of aggregation. This justifies reconsidering the *scale problem* using a simulation analysis with the assumption that all variables are characterized by a mutual linear correlation only. It was also assumed that the variables y1 and y2 are the variables expressed in relative quantities and have different denominators, which can be expressed by means of formulas:

$$Y_1 = \frac{X_1}{X_3},\ Y_2 = \frac{X_2}{X_4}. \qquad (2)$$

Table 8. The correlation and regression dependence for the variables contained in Table 7

Level of aggregation (NUTS 4 composition)						
Correlation						
	y1	y2	x2	x4	x1	x3
y1	1					
y2	-0,36	1				
x2	-0,30	0,33	1			
x4	-0,32	0,31	0,98	1		
x1	0,03	0,16	0,78	0,84	1	
x3	-0,31	0,31	0,99	0,99	0,82	1
Analysis of regression						
Regression coefficient	-0,24	Standard error	0,03	Coefficient of determination		0,13
Level of aggregation (NUTS 3 composition)						
Correlation						
	y1	y2	x2	x4	x1	x3
y1	1					
y2	-0,63	1				
x2	-0,46	0,69	1			
x4	-0,38	0,57	0,94	1		
x1	0,57	0,19	0,27	0,43	1	
x3	-0,47	0,85	0,92	0,81	0,07	1,00
Analysis of regression						
Regression coefficient	-0,44	Standard error	0,068	Coefficient of determination		0,39

Source: elaborated by the author.

Within the simulation analysis only variables y1, y2, x3, x4 were generated and the values of the variables x1, x2 are calculated from the formula 2. The correlation between the generated variables y1, y2, x3, x4 have been established again at a level similar to the empirical correlation based on selected economic variables (see Table 9). The standard deviation values for the variables y1, y2 were selected in a way to make the generated regression dependence consistent with the value shown in Table 8 (value -0.24). The values of the mean for the variables y1, y2 are determined in such a way that the coefficient of variation was 10%. The values of the mean and of the standard deviation for the variables x2, x4 were determined arbitrarily by the author. The assumed mean and standard deviation are shown in Table 10. In the case of the variables x1, x2, the mean, standard deviations and correlation resulted from the calculations made.

Table 9. The assumed correlation during the simulation of the variables y1, y2, x2, x4

Correlation				
	y1	y2	x2	x4
y1	1			
y2	-0,35	1		
x2	-0,30	0,30	1	
x4	-0,30	0,30	0,95	1

Source: elaborated by the author.

Table 10. The assumed mean and the standard deviation during the simulation of the variables y1, y2, x2, x4

Properties				
Variables	y1	y2	x2	x4
Mean	0,4	0,6	400	1000
Standard deviation	0,04	0,06	40	100

Source: elaborated by the author.

Simulation analysis has been initiated to generate variables y1, y2, x3, x4 at NUTS 4 composition. Then, based on formula 2 the values of the variables x1, x2 were determined. As a result of the simulations, the output set of variables x1, x2, x3, x4, y1, y2 was obtained at the aggregation level – NUTS 4 composition. Within the simulation performed 1000 realisations were obtained for each of the variables.

In the next step, basing on the possessed realisations, aggregation of simulated variables was performed for the NUTS 3 composition. In the first step aggregation was performed for the variables x1, x2, x3, x4 expressed in absolute quantities. Then, the obtained realizations of variables expressed in the absolute quantities at the aggregation level - NUTS 3 composition allowed us to determine the values of the variables expressed in the relative quantities y1, y2 using formula 2.

The obtained realisations of the variables at both levels of aggregation allowed the calculation of selected statistics for each of the variables. Again, basing on a set of calculated estimates, the average and standard deviation were calculated for each of the statistics. The results of the calculation are presented in Table 11 and Table 12.

Table 11. The results of the simulation - basic statistics for simulated variables and results of regression analysis

Before aggregation (NUTS4 composition)						
Variables	y1	y2	x2	x4	x1	x3
Mean	0,4 (0,002)	0,6 (0,003)	400 (2,079)	1000 (5,094)	159 (0,911)	601 (5,811)
Standard deviation	0,04 (0,001)	0,06 (0,002)	40 (1,250)	100 (3,817)	19 (0,661)	96 (3,077)
I statistics	-0,012 (0,032)	-0,006 (0,032)	-0,008 (0,035)	-0,011 (0,032)	-0,004 (0,030)	-0,009 (0,037)
P-value	0,584 (0,277)	0,532 (0,291)	0,535 (0,305)	0,552 (0,285)	0,522 (0,277)	0,564 (0,290)
Analysis of regression						
Regression coefficient	-0,24 (0,031)	Standard error	0,03 (0,001)	Coefficient of determination		0,12 (0,031)
After aggregation (NUTS 3 composition)						
Variables	y1	y2	x2	x4	x1	x3
Mean	0,4 (0,002)	0,6 (0,003)	2297 (11,943)	5746 (29,117)	915 (5,363)	1396 (26,245)
Standard deviation	0,02 (0,002)	0,03 (0,003)	922 (11,664)	2307 (27,853)	369 (5,437)	58379 (1183)
I statistics	-0,012 (0,032)	-0,006 (0,032)	-0,008 (0,035)	-0,011 (0,032)	-0,004 (0,025)	-0,010 (0,035)
P-value	0,584 (0,277)	0,532 (0,291)	0,535 (0,305)	0,552 (0,285)	0,522 (0,294)	0,552 (0,288)
Analysis of regression						
Regression coefficient	-0,23 (0,114)	Standard error	0,077 (0,012)	Coefficient of determination		0,12 (0,092)

Source: elaborated by the author.

The analysis of the results leads to the following conclusions: in the case of the variables y1 and y2 expressed in relative quantities, the obtained average for the mean remained unchanged, and the average for the standard deviation decreased as a result of the aggregation process. As a result of the aggregation of the variables x1, x2, x3, x4 expressed in the absolute quantities, the average for the mean and standard deviation increased.

The average for the correlation of a pair of variables (y1, y2) remained unchanged at more than three times the standard deviation of growth. This means that the level of the correlation between these variables may change significantly under the impact of the aggregation process.

Table 12. The results of the simulation – correlation for simulated variables

Before aggregation (NUTS 4 composition)						
	y1	y2	x2	x4	x1	x3
y1	1					
y2	-0,35 (0,043)	1				
x2	-0,30 (0,046)	0,30 (0,050)	1			
x4	-0,30 (0,042)	0,30 (0,043)	0,95 (0,002)	1		
x1	0,59 (0,033)	-0,04 (0,047)	0,59 (0,039)	0,55 (0,037)	1	
x3	-0,40 (0,043)	0,80 (0,018)	0,77 (0,020)	0,80 (0,017)	0,31 (0,045)	1
After aggregation (NUTS 3 composition)						
	y1	y2	x2	x4	x1	x3
y1	1					
y2	-0,35 (0,147)	1				
x2	-0,01 (0,162)	0,026 (0,167)	1			
x4	-0,01 (0,161)	0,025 (0,166)	0,99 (0,001)	1		
x1	0,07 (0,163)	-0,01 (0,166)	0,99 (0,001)	0,99 (0,001)	1	
x3	-0,04 (0,159)	0,11 (0,165)	0,99 (0,001)	0,99 (0,001)	0,98 (0,002)	1

Source: elaborated by the author.

The correlation dependence between the variables y1 and y2 and the variables x1, x2, x3 at a higher level of aggregation was reduced almost to zero with a significant increase in the error (the value of standard deviation) of received estimates. The correlation for pairs of variables (x1, x3) and (x2, x3) increased. However, the correlation for a couple of variables (x1, x2) decreased to almost zero, with a significant increase in error.

As a result of the regression analysis at both levels of aggregation similar average values of the estimates of regression parameter and the coefficient of determination were obtained. However, in both cases it was connected with an increase in the standard deviation.

Also in this case, there was an increase in the average and standard deviation of the standard error of the regression parameter. This means that the process of aggregation can lead to significant changes in the estimate of the regression parameter obtained with large variations in the level of statistical significance of the regression parameter.

The existence of the spatial autocorrelation for all variables was examined, too. The values of Moran's I statistics did not indicate the occurrence of spatial dependence resulting from the aggregation process.

Conclusions

The paper considered the *scale problem* which is one of the aspects of the *modifiable areal unit problem*. This problem is very important because its presence can lead to changes in the results of spatial economic analysis. In the research conducted on the *scale problem* emphasis was laid on differentiating variables expressed in the absolute and relative quantities. This is due to the fact that spatial economic analyses are usually performed on the basis of variables expressed in relative quantities.

The paper considered two research problems concerning selected economic variables. The calculations based on the empirical values of these variables showed that during the transition to a higher level of aggregation a change in the estimates of the parameters occurred. The results obtained were the reason for considering the *scale problem* by means of a simulation analysis.

Based on the simulation analysis the following conclusions were drawn: in the case of variables expressed in relative quantities the mean does not change under the influence of the aggregation process and, in addition, the standard deviation decreases. For variables expressed in absolute quantities the mean and standard deviation are rising, which results from the summing of the values of variables. The results are an important argument for the use of data expressed in relative quantities while dealing with spatial economic analysis.

Analysis of the results for the correlation between variables expressed in relative quantities showed that the average of the correlation did not change. It should be noted, however, that the standard deviation of the correlation increased significantly. This means that as a result of the aggregation process values of the Pearson correlation coefficient can be changed substantially. There has been no systematic regularity for the other correlation between variables expressed in absolute quantities and ones expressed in relative quantities.

Regression analysis was performed only for variables expressed in relative quantities. The resulting average of the regression parameter and the coefficient of determination were at a similar level at both levels of aggregation. Both in the case of the regression parameter and the coefficient of determination, the values of the standard deviation increased within the process of aggregation. It also significantly increased the standard error of the regression parameter.

Based on the outcome of the research, no occurrence of spatial dependence in the data resulting from the aggregation process was found.

To sum up the results, it can be concluded that the analysis of correlation and regression can lead to different results, depending on the assumed level of aggregation. The increase in the average of standard error also points to the possibility of the emergence of statistical insignificance of the regression parameter. The obtained results, however, should be approached cautiously as they can, in part, be derived from the imperfectly conducted simulation. In spite of this the results indicate a need to consider the character of the variables. This means that the character of the variables is a key element in the consideration of the *scale problem*. The next step in the research should be a consideration of the *scale problem* with the assumption of additional properties for variables such as spatial autocorrelation or a spatial trend. It is also necessary to consider the issue of improving a simulation procedure for generating variables expressed in absolute quantities and for variables expressed in relative quantities with the established correlation.

References

Tate, N. & Atkinson P. M (Eds.) (2001). *Modelling scale in geographical information science*. Chichester: John Wiley & Sons.

Anselin, L. (1988). *Spatial Econometrics: Method and Models.* Netherlands: Kluwer Academic Publishers.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economics and Related Problems*. Dordrecht: Kluwer Academic Publisher.

Arbia, G. (2006). *Spatial Econometrics, Statistical Foundations and Applications to Regional Convergence*, Berlin-Heidelberg: Springer-Verlag.

Blalock, H. (1964). *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press.

Flowerdew, R. (2011). How serious is the Modifiable Areal Unit Problem for analysis of English census data?. *Population Trends*, 145.

Holt, D., Steel, D. G., & Tranmer, M. (1996). Area homogeneity and the modifiable areal unit problem. *Geographical Systems*, 3.

Fotheringharn, A. S., & Wong, D. W. S. (1991). The modifiable area unit problem in multivariate analysis. *Environment und Planning A*, 23.

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*, 29.

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying spatial patterns of British census variables Computers. *Environment and Urban Systems*, 30.

Marble, D. F. (2000). Some thoughts on the integration of spatial analysis and Geographic Information Systems. *Journal of Geographical Systems*, 2.

Openshaw, S. (1977a). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers*. *New Series*, 2.

Openshaw, S. (1977b). Algorithm 3: a procedure to generate pseudo-random aggregationsof N zones into M zones, where M is less than N. *Environment and Planning A*, 9.

Openshaw, S. (1977c). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, 9.

Openshaw, S., & Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem In: N. Wrigley (Ed.). *Statistical methods in the spatial sciences*. London: Pion.

Openshaw, S. (1984a). *The Modifiable Areal Unit Problem*. Norwich: GeoBooks, CATMOG 38.

Openshaw S. (1984b). Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16.

Openshaw, S. & Rao, L. (1995). Algorithms for re-engineering 1991 census geography. *Environment and Planning A*, 27.

Pietrzak, M. B. (2014a). Redefining The Modifiable Areal Unit Problem Within Spatial Econometrics. The Case Of The Scale Problem. *Equilibrium*, 9(2).

Pietrzak, M. B. (2014b). Redefining the Modifiable Areal Unit Problem within spatial econometrics, the case of the aggregation problem. *Equilibrium*, 9(3).

Reynolds, H. D. (1998). The Modifiable Area Unit Problem: Empirical Analysis by Statistical Simulation. Doctoral thesis, Graduate Deparment of Geography, University of Toronto.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3).

Suchecki, B. (2010). *Ekonometria Przestrzenna*. Warszawa: Wydawnictwo C.H.Beck.

Suchecki, B. (2012). *Ekonometria Przestrzenna II*. Warszawa: Wydawnictwo C.H.Beck.

Szulc, E. (2007). *Ekonometryczna analiza wielowymiarowych procesów*. Toruń: Wyd. UMK.

Tranmer, M. & Steel, D. (2001). Using Local Census Data to Investigate Scale Effects In: N. Tate, P. Atkinson (Ed.). *Modelling scale in geographical information science*. Chichester: John Wiley & Sons.

Yule, G. U. & Kendall, M. G. (1950). *An introduction to the theoryof statistics*. London: Griffin.

Zeliaś, A. (1991). *Ekonometria Przestrzenna*. Warszawa: PWE.