

# Exploring the variability and geographical patterns of population characteristics: Regional and spatial perspectives

Pavĺína Netrdov<sup>a\*</sup>, Vojtech Nosek<sup>a</sup>

## Abstract

*The variability and geographical patterns of population characteristics are key topics in Human Geography. There are many approaches to exploring and quantitatively measuring this issue. Besides standard aspatial statistical methods, there is no universal framework for incorporating regional and spatial aspects into the analysis of areal data. This is mainly because complications, such as the Modifiable Areal Unit Problem or the checkerboard problem, hinder analysis. In this paper, we use two approaches which uniquely combine regional and spatial perspectives of the analysis of variability. This combination brings new insights into the exploration of the variability and geographical patterns of population characteristics. The relationship between regional and spatial approaches is studied with models in a regular grid, using variability decomposition (Theil index) as an example of the regional approach, and spatial autocorrelation (Moran's I) as an example of the spatial approach. When applied to empirical data based on the Czech censuses between 1980 and 2011, the combination of these two approaches enables us to categorise the studied phenomena according to the regional and spatial nature of their variability. This is a useful advance, especially for assessing evolution over time or comparisons between different phenomena.*

**Keywords:** geographical patterns, regional variability, spatial autocorrelation, Theil index decomposition, Moran's I, Czech Republic

**Article history:** Received 18 April 2016; Accepted 3 January 2017; Published 30 June 2017

## 1. Introduction

Many scientific disciplines study the socio-economic characteristics of administrative regions or other spatial units and the differences between them. Although these disciplines often study the same phenomenon, the terminology is inconsistent and the applied methods differ. The reasons are manifold: the researchers' different professional backgrounds and routines in the field, varying research goals, issues concerning the availability of suitable data, etc. The result is that while some authors refer to variability, others use terms such as inequality, geographical concentration, spatial concentration, agglomeration, and polarisation. More importantly, while researchers with economic or regional science backgrounds usually prefer 'pure' quantification of regional differences by calculating some variability measures on a macro-scale level (such as NUTS 2 or NUTS 3 levels in the case of the EU<sup>1</sup>), geographers often try to look beyond the predefined

regions as a unit of analysis and focus on the micro-scale level (for example, municipalities). In this article, we aim to overcome the methodological divide between different approaches to measuring variability and geographical patterns by examining their relationship and joint use in empirical research.

The simplest way to measure the geographical variability of areal data is to apply standard statistical measures of variability to geographical data. These methods can be considered aspatial, however, as they do not work with the spatial information inherent in the data (Fotheringham et al., 2000). Therefore, they tend to not reveal much about the geographical organisation of phenomena in space and relationships to higher regional structures. Moreover, when studying geographical patterns, there are two main concerns associated with aspatial methods when applied to areal data – the Modifiable Areal Units Problem (MAUP) and the checkerboard problem. The MAUP deals with the fact

<sup>a</sup> Department of Social Geography and Regional Development, Faculty of Science, Charles University, Prague, Czech Republic (\*corresponding author: P. Netrdov, e-mail: [pavlina.netrdova@natur.cuni.cz](mailto:pavlina.netrdova@natur.cuni.cz))

<sup>1</sup> The term NUTS refers to "Nomenclature of Territorial Units for Statistics", a geocode standard for subdivisions of countries in the EU.

that areal data are sensitive to the definition of boundaries used for the construction of the units, which may not be meaningful for the analysed problem (Openshaw, 1984; Wong, 2009; Klapka et al., 2016). The checkerboard problem stems from the fact that the geographic positions of regions (and potential neighbourhood effects) is ignored even though they are based on spatial data (Guimaraes et al., 2011).

It is clear that aspatial methods are far from sufficient when assessing geographical patterns and variability, and this has been recognised by many authors (Arbia, 1989, 2001; Rey and Montouri, 1999). There are two fundamental approaches to addressing the MAUP and checkerboard problem and bringing spatial or regional perspectives into the analysis of variability. We call the approach to bringing spatial clustering into the analysis the “spatial approach”, and the approach to analysing variability at several regional levels and quantifying the importance of respective regional levels the “regional approach”. In this paper, the spatial approach is represented by spatial autocorrelation measures, which can quantify the level of spatial clustering in the whole study area and uncover local clusters, thus bypassing the checkerboard problem. We employ the example of Moran’s *I* as the most commonly used index of global spatial autocorrelation. The regional approach is represented by variability decomposition, which quantifies the importance of respective regional levels, thus addressing the MAUP problem. In this latter case, we use the example of the decomposition of the Theil index, a typical decomposable index used in inequality studies.

Sometimes aspatial methods are combined with spatial (Arbia, 2001; Lafourcade and Mion, 2007) or regional (Brühlhart and Traeger, 2005; Rey, 2004) approaches. It could be argued that differences of methodology account for the reason regional and spatial approaches are typically used separately and, as a consequence, have not previously been combined with aspatial methods by authors. This paper also responds partly to the challenges posed by Rey and Janikas (2005), with the goal of demonstrating the importance of studying the interrelationship of (regional) variability and spatial autocorrelation.

The main goal of this paper is to demonstrate how the conjoint use of regional and spatial approaches helps to uncover and understand the variability and geographical patterns of population characteristics that are unclear when only aspatial methods are used. We argue that using both approaches conjointly can offer more comprehensive and innovative results, as documented in some empirical studies (Blažek and Netrdová, 2012; Nosek and Netrdová, 2010). In order to fulfil this goal and to interpret results correctly, we need to understand the relationship between these two approaches, explore it on simulated data, and test it on empirical data. Moreover, we can categorise empirical data on the basis of their spatial and regional perspective of variability, which is useful especially when assessing evolution over time or comparisons across different characteristics.

The paper is organised as follows. In section 2, we describe and discuss the theoretical-methodological aspects of the approaches and highlight their potential complementarity. In section 3, we specify the methods and data used, where we also stress the importance of testing statistical significance and distinguishing between stochastic and spatially contingent components of measured values. The relationships between the methods are fully analysed in section 4. In section 5, we present empirical examples from the Czech Republic. Section 6 summarises and concludes the paper.

## 2. Theoretical and methodological background

The variability of geographical phenomena is often studied with only aspatial methods, which are invariant to permutations across units and do not incorporate information about the absolute or relative position of the respective unit in the calculation (Fotheringham et al., 2000). Basic variability measures such as the variance and standard deviation are useful for quantifying absolute levels of variability, but they do not meet the independence of scale requirement and it is therefore difficult to use them for regional analyses. This can be solved by using the coefficient of variation. The drawback of the coefficient of variation is that it is calculated from the distribution’s mean, which is not resistant to the extreme values typical for asymmetrically distributed geographical phenomena (Imre et al., 2012; Korčák, 1938). A good way to assess the uneven distribution of phenomena in space is to construct a Lorenz curve, a graphical representation of the distribution of a studied variable (such as wealth) in a society or space. One statistic with a straightforward interpretation that can be easily derived from the Lorenz curve is the Gini coefficient, which, due to its relative independence of the mean is very popular, and probably the most widely used variability measure in the social sciences.

Each of these measures, including the Gini coefficient, satisfies the condition of anonymity, a property of being insensitive to any spatial permutation (Sen, 1972). This condition, however, is not always a desirable property of a variability measure and is more a pitfall (Arbia, 2001), especially from a geographical point of view. The total insensitivity to the geographical position of units under analysis leads to the same results when units with high concentration values are adjacent as when units are located in the opposite part of the study area. The problem of ignoring neighbourhood effects is known as the checkerboard problem (Guimaraes et al., 2011). Another well-known problem associated with the analysis of areal data is the Modifiable Areal Units Problem (Openshaw, 1984; Wong, 2009). This problem refers to the sensitivity of variability or other statistical measures to the exact delimitation of areal units. There are two components of this sensitivity: the zoning effect (the dependence of results on the changing zonal boundaries); and the scale effect (the dependence of results on the level of aggregation, e.g. from municipalities to a regional level of analysis) (Arbia, 1989). Marcon and Puech (2003) address this issue. They state that variability is measured at a single level (typically at a chosen administrative level). Since observations may differ at different geographical levels, however, it may be useful to measure concentration at different geographical levels simultaneously. When a less fragmented (i.e. more aggregated) regional structure is used, some local specifics may remain hidden in the regional means, and some interpretations may be biased.

Problems associated with aspatial methods can be partly overcome by analysing spatial variability through the concept of spatial autocorrelation. This approach enables us to incorporate the neighbourhood effects into measuring variability and to identify whether there is a significant spatial pattern. In this way, spatial autocorrelation addresses the issue connected with the checkerboard problem. There are many ways to measure spatial autocorrelation depending on the nature and properties of the data (Anselin, 1988; Cliff and Ord, 1973). In general, two forms of measuring spatial autocorrelation can be

identified – global and local. The indicators of global spatial autocorrelation measure the extent of spatial clustering of “similar” values. In local form, they identify exact spatial clusters and reveal their character. Examples of global spatial autocorrelation statistics include Geary’s *c*, Getis and Ord’s *G*, and Ripley’s *K*; local indicators include Getis and Ord’s *G<sub>i</sub>*, and Ord and Getis’ *O* (Getis, 2007). In recent studies, the most frequently used indicator of global spatial autocorrelation is Moran’s *I*, which is based on covariance and has many similarities with Pearson’s product-moment correlation coefficient. Anselin (1995) introduced local Moran’s *I*, the local indicator of spatial association (LISA) statistics. LISA cluster maps show statistically significant units in four types of spatial association.

The relationship between the aspatial concept of variability and the spatial approach to variability (quantified by spatial autocorrelation measures) has been studied by many authors (Arbia, 2001; Rey, 2004). Although one may expect an empirical relationship, as supported by empirical data (Rey, 2004), no theoretical or mathematical relationship exists. Consider the example of when all values in a studied area are modified in the same way. For instance, when values are increased or multiplied by a positive constant, the measure of spatial autocorrelation remains unchanged but the measure of aspatial variability changes quite markedly. This also applies conversely when we fix the values of variability and make random shuffles or specific arrangements of data in spatial units. As shown, there is no theoretical relationship between the overall variability and spatial variability. Therefore, observed spatial patterns of variability are only of an empirical nature. Rey’s (2004) finding of a strong positive relationship between measures of variability in state incomes and the degree of spatial autocorrelation in the US over the 1929–2000 period has no methodological justification and is therefore purely empirical. As shown, there is no theoretical relationship between the overall variability and spatial variability. Therefore, observed spatial patterns of variability are only of an empirical nature.

The utilisation of spatial approaches in assessing geographical variability helps to control the checkerboard problem, but invariably the MAUP effects remain. The reason is that both aspatial methods and spatial autocorrelation methods, representing a spatial approach, only work on one level of analysis. There are, however, typically more geographical levels that can be considered (for instance, a municipal level as a micro-scale and administrative regions, say NUTS3, as a macro-scale). Moreover, it is desirable not only to quantify variability at various, yet still single, geographical levels, but also to be able to quantify the relative importance of geographical levels compared with others. The regional approach in this way enables us to assess geographical variability at different scales and with distinct delimitation of regions, thus controlling the MAUP. First, in this approach, it is important to distinguish between overall variability, as measured between units at the most detailed sub-regional level, and regional variability, as measured between regional means. Second, there are two types of regional variability. Simple regional variability quantifies the differences between regional means, while relative regional variability quantifies the ratio between simple regional variability and overall variability. The latter enables us to quantify the importance of regional levels in overall variability and thus to assess the relative importance of a specific geographical level on the differentiation of particular phenomena.

For the regional approach, the aspatial methods are not sufficient. Unfortunately, the most commonly used coefficient to measure variability, the Gini coefficient, cannot be decomposed without a residuum into between-group and within-group components (for Gini coefficient decomposition, see Lambert and Aronson (1993); Mussard et al. (2003) necessary for quantifying relative regional variability, unless a spatial weight matrix is brought into the equation (see Rey and Smith, 2013). Gini decomposition without residual was proposed by Okamoto (2009), although the between-region variability of the Gini decomposition is null only if the distribution within each sub-region is identical to all the others. Decomposition enables us to quantify the share of a selected regional level and “scale down” the variability. This reveals the most important regional/local levels, and has many practical implications. For these purposes, it is possible to use indices from the generalised entropy class, which are decomposable without residuum. Of these the Theil index is the most widely used measure of regional variability (Cowell and Flachaire, 2007). The convenience of the various variability measures is illustrated by Shorrocks and Wan (2005) or Subramanian (2004), and Litchfield (1999) describes the axiomatic approach to properties of variability measures.

The relationship between an aspatial concept of variability and a regional approach to variability (represented by simple regional variability) is similar to the spatial concept and is only of an empirical nature. As an example, consider when regional means remain the same, but the values within these regions change. Thus the same values for simple regional variability show different values for overall variability depending on the exact modifications of the data. This also applies conversely when we fix the values of overall variability and make random shuffle or specific arrangements of data in regions influencing the values of simple regional variability.

Methods of quantifying simple regional variability (aspatial) and relative regional variability (regional approach) use a predefined regional structure, and methods of measuring spatial autocorrelation (spatial approach) also require some predefinition. To determine spatial lag and measure spatial autocorrelation, one must define a spatial weight matrix that operationalises the concept of “near” spatial units that can influence each other (Cliff and Ord, 1973). By changing the regional structure, for example on different hierarchical levels, we can calculate the regional variability to assess the relevance of the specific regional delimitation and hierarchical level to a given spatial process. By changing the spatial weights matrix, for example by extending the distance of influence, we can assess the distance over which the spatial process operates. This idea, hidden in both approaches, highlights their similarity. The only difference is that one approach takes a discrete view of spatial processes embedded in regions when measuring regional variability, while the other takes a continuous view on spatial processes with distance-decay influence when measuring spatial autocorrelation. It can be argued that the method of repeating measures of regional variability for many regional systems “floating” in a given area and the method of using a discrete spatial weights matrix based on regions are interchangeable. But this contradicts the foundation of these approaches and we expect that combining the two (discrete and continuous) can help to better understand the regional and spatial consequences of a given process.



The difference as well as complementarity of regional and spatial approaches is summarised in Figure 1. There are four types of results when regional and spatial approaches are combined. If both relative regional variability and spatial autocorrelation are high, the characteristic under analysis may be considered spatially dependent and bounded in regions. The characteristic is concentrated, and within predefined regions. Sometimes the characteristic

concentrates across regional borders, meaning it is spatially dependent but with no relation to regions. If the characteristic does not form spatial concentrations nor concentrate in regions, it is both spatially and regionally independent. It is theoretically possible that the characteristic does not form any spatial concentrations yet it concentrates on a regional level. This characteristic would be spatially independent yet bounded in regions.

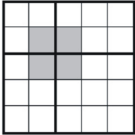
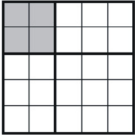
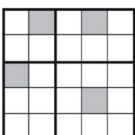
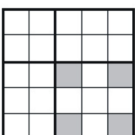
		Importance of regional perspective	
		NO / MINIMAL	YES / MAXIMAL
Importance of spatial perspective	YES / MAXIMAL	SPATIALLY dependent with weak relation to REGIONS <i>concentrations across regional borders</i> 	SPATIALLY dependent and bounded in REGIONS <i>concentrations in regions</i> 
	NO / MINIMAL	Both SPATIALLY and REGIONALLY independent <i>no concentrations</i> 	SPATIALLY independent yet bounded in REGIONS <i>no spatial concentrations</i> 

Fig. 1: Typology of areal data based on regional and spatial perspectives  
Source: authors' conceptualisation

### 3. Methods

The ways of measuring spatial autocorrelation and relative regional variability as tools for quantifying the spatial and regional approaches to geographical variability are manifold. In this paper, spatial (spatial autocorrelation) and regional (relative regional variability) approaches are represented by two specific measures: Moran's I and the Theil index (T) and its decomposition. The relationship between spatial and regional approaches is demonstrated in the examples of these two most widely-used methods and without loss of generality, some conclusions based on these two methods can be drawn.

The formulas for Moran's I (Equation 1) and Theil index T (Equation 2) can be written as:

$$(1) \quad I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} \sum_{i=1}^n (y_i - \bar{y})^2} ;$$

$$(2) \quad T = \left( \sum_{j=1}^k \frac{n_j}{n} \frac{y_j}{\bar{y}} \ln \frac{y_j}{\bar{y}} \right) + \left( \sum_{j=1}^k \frac{n_j}{n} \frac{y_j}{\bar{y}} \sum_{i=1}^{n_j} \frac{y_{ij}}{y_j} \ln \frac{y_{ij}}{y_j} \right) = T_B + T_W ;$$

where:

for I (Moran's I),  $n$  = number of units,  $i$  = index for individual units,  $j$  = index for regions,  $k$  = number of regions,  $\bar{y}$  = mean of the variable under analysis,  $w_{ij}$  = spatial weight matrix;

and for T (overall Theil index),  $T_B$  = between-region component of Theil index,  $T_W$  = within-region component of Theil index (see Anselin, 1988; Anselin, 1995; Elbers et al., 2008; Shorrocks and Wan, 2005).

While  $T_B$  is a measure of simple regional variability, the share of simple regional variability in overall variability:  $T_B/T$  measures relative regional variability. All computations regarding the Theil index and its decomposition were performed in MS Excel and EasyStat 1.0 (Novotný et al., 2014), and all computations regarding Moran's I were performed in GeoDa 1.4.0 (Anselin, 2003; Anselin et al., 2004).

As the formulas show, to measure spatial autocorrelation requires a spatial weight matrix ( $w_{ij}$ ) that operationalises the position and proximity of geographical units (Anselin, 1988; Cliff and Ord, 1973; Getis and Aldstadt, 2004). The selection of a particular spatial weight matrix is often considered crucial, and is said to have a significant effect on the resulting spatial autocorrelation values (Anselin and Rey, 1991). The selection of a spatial weight matrix is especially important when only one variable is studied and the results may vary significantly, or when systems with different regional structures are compared (Nosek and Netrdová, 2014). For studying general patterns and for interpretation (zero, high, or low Moran's I), however, the choice of a spatial weight matrix is not that important and the simplest spatial weights matrix, first-order contiguity, suffices (Stakhovych and Bijmolt, 2009). We have confirmed this by testing 18 different spatial weight matrices (this exercise is not included in this text). All analyses in this paper therefore use the 1<sup>st</sup> order queen spatial weights.

When using spatial autocorrelation, inference is commonly used and integrated in the software developed for this type of analysis. The inference is generally based on the comparison of random and empirical distributions of data in the studied area. To assess the significance of Moran's I against a null

hypothesis of no spatial autocorrelation, a permutation procedure is used, specifically the conditional permutation procedure embedded in GeoDa 1.4.0. A total of 9,999 permutations are used, which is sufficient to obtain stable results in most cases (Anselin, 2003). Due to a randomisation process, the results can differ slightly when replicated, and so it is better to speak of a pseudo-significance value (Anselin, 2003; Ord and Getis, 2012). There are ways to assess the sensitivity or 'stability' of the results, however, by increasing the number of permutations, repeating the permutation procedure several times, and changing the significance cut-off value (Anselin, Syabri, and Kho, 2004).

When studying and measuring regional variability, the importance of statistical inference is often underestimated. In this case, one has to use non-parametric methods based on re-sampling – the confidence interval (or other desired characteristic) is constructed from the simulated values of the tested characteristics, which are calculated from data repeatedly generated from the original data set. Though not new (in a similar context see, for example, Longford et al., 2012), these methods are still underused in regional inequality research (Mills and Zandvakili, 1997; Stine, 1989).

In general, it is desirable to test whether the measured regional variability differs significantly from a situation where the data are randomly distributed in space (the null model). It is obvious that even in the null model some regional variability will be found. Therefore, regional variability can be understood as the sum of two components (for further explanation, see Novotný and Nosek, 2012): the stochastic component (regional variability of the null model) and the spatially contingent component (regional variability exceeding the null model, i.e. the measured regional variability minus the regional variability of the null model, referred to below as the adjusted relative regional variability). Isolating the spatially contingent component of regional variability helps when different systems are compared since each system has different stochastic variability embedded in the results (Novotný and Nosek, 2012).

#### 4. Regional and spatial approaches and their relationship

In order to utilise regional and spatial approaches conjointly and interpret the results properly, one must understand how their theoretical and methodological aspects are related. We model values of Theil index decomposition as a representation of a regional approach and Moran's I as a representation of a spatial approach in a regular grid by running series of simulations. Besides modelling the values we can study the relationship between these two approaches. The relationship between variability and spatial autocorrelation has not been studied in detail, with a few exceptions (Arbia, 2000, 2001; Rey, 2004; Rey and Janikas, 2005). Each of these authors, however, considered only a variability on a single level and did not take decomposition and relative regional variability into account.

The model consists of 10,000 log-normally distributed pseudo-random data. The log-normal distribution is often considered to represent socio-geographical data with the most accuracy (Novotný and Nosek, 2009). These data were distributed randomly and in several specific ways in a regular square tessellation with 100 rows and 100 columns. In the 100 by 100 grid, 100 'regional' units (10 by 10) are specified. In addition to overall variability measured by T (the

differences between 10,000 units), simple regional variability measured by  $T_B$  (the differences between 100 regional means) and relative regional variability measured by  $T_B/T$  (the share of simple regional variability in overall variability) can be calculated. The set of pseudo-random numbers helps to minimise the effect of differences in variability and the regular tessellations minimise the effects of regional delimitation. Different regular tessellations (triangles, squares, and hexagons) were compared in Boots and Tiefelsdorf (2000). Using different types of regular tessellations is far beyond the scope of this paper and not important for achieving its goals.

The relationship between relative regional variability ( $T_B/T$ ) and spatial autocorrelation (Moran's I) is expected to be rather complex. A strong positive relationship between the interregional inequality share and spatial clustering was found by Rey (2004), who pointed out the ease of change of this result through re-shuffling. Theoretically, when high spatial autocorrelation is observed, both high and low relative regional variability can be present. One might also assume that with very low spatial autocorrelation it is theoretically and mathematically impossible to have high relative regional variability. Figure 1 shows the theoretical possibilities which may occur. The main purpose of the simulation with the model data is to examine this relationship and to document the fact that certain values of spatial autocorrelation result in very different values of relative regional variability, and vice versa.

To simulate  $T_B/T$ , the value in each unit in the 100 by 100 regular square tessellation remains the same but the regional borders of 100 regions shift (assuming the grouping in regions is exhaustive, mutually exclusive and that there are neither enclaves nor exclaves). In this exercise,  $T_B/T$  can vary from 0 to 1 depending on the delimitation of regional borders and their correspondence with spatial clusters. The values in the regular square tessellation, which are generated to have fixed values of  $T_B/T$  (10% through 90%, with a maximum close to 100%), are rearranged within this tessellation with the aim of customising the levels of spatial autocorrelation. Several random shuffles and specific arrangements (ranging from arrangements with assumed maximal upper levels of positive spatial autocorrelation, such as by a concentration of high values in the same half of the tessellation, to chessboard arrangements with a supposed maximal negative spatial autocorrelation), were applied to test whether different relative regional variability can show different spatial concentrations when spatially distributed in specific ways. Another goal was to test the hypothesis of a statistically insignificant spatial autocorrelation when data with different levels of variability are randomly distributed. This relationship is depicted in Figure 2.

For fixed  $T_B/T$ , significant values of Moran's I are observed even when the data are randomly shuffled. This confirms the assumption that with some non-zero relative regional variability (even when  $T_B/T$  is only 10%), insignificant spatial autocorrelation cannot be observed. In fact, quite a clear pattern can be determined. With increasing  $T_B/T$ , Moran's I increases with roughly similar values. This relationship may therefore be considered purely stochastic in nature. The situation is slightly different when specific arrangements for calculating spatial autocorrelation are used (simulating a limited number of empirical relationships). The relationship for low values of  $T_B/T$  is rather loose (for  $T_B/T = 10\%$  it varies from + 0.5 to – 0.2). With increasing relative regional variability, the range of possible Moran's I

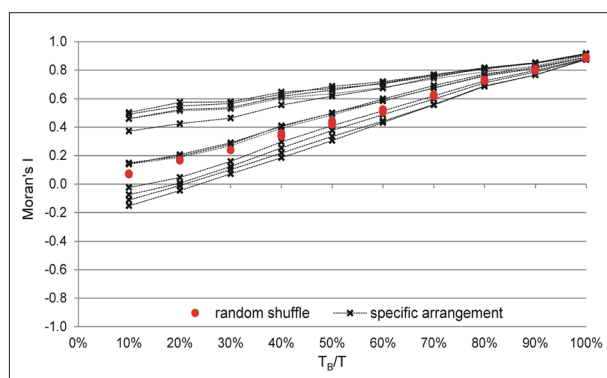


Fig. 2: The relationship between relative regional variability was measured by  $T_B/T$  and global spatial autocorrelation was measured by Moran's  $I$  (model data). Note: For the calculation of Moran's  $I$ , spatial weights based on queen contiguity (first order of contiguity) were used

Source: authors' simulation results

values narrows. With extreme values of relative regional variability, Moran's  $I$  is close to its possible maximum. From a stochastic perspective, there is a clear positive correlation. From the simulations, it is evident that only three basic combinations are possible: high  $T_B/T$  – high Moran's  $I$ , low  $T_B/T$  – high Moran's  $I$ , and low  $T_B/T$  – low Moran's  $I$ .

These findings may help with the interpretation of empirical results in several ways. For example, the low values of relative regional variability do not necessarily mean that the studied phenomenon has a weak spatial pattern or none at all. This can be caused by the regional delimitation which may not be appropriate for the particular phenomenon, in which case, spatial autocorrelation statistics would provide significant added value.

## 5. Empirical evidence

In this section the theoretical assumptions and modelled data are tested empirically using the case of the Czech Republic. The data set contains empirical data from four

population censuses (1980, 1991, 2001, and 2011). All data were recomputed to the most current structure at a municipal level in 2011 and are thus directly comparable. We chose the Czech Republic as an empirical example because comparable data were available at a very detailed level in a very fragmented regional structure (6,251 municipalities). Besides the municipal level, the data were studied at the Czech regional structure – 13 NUTS 3 units (administrative macro-regions). At this regional level, the Prague region has been merged with its surrounding region (Central-Bohemian Region) in order to represent geographical processes more accurately (see Hampl, 1999). Due to their consistent statistical nature, characteristics with a minimum structural variable of 0 and a maximum structural variable of 1 were chosen for the study. Further, the characteristics were chosen based on their supposed behaviour in the geographical context of regional variability and spatial autocorrelation (following the results documented in Netrdová and Nosek, 2009). Based on these two requirements, the following characteristics were chosen:

- Unemployment (economic) – the unemployed population, normalised by the economically active population;
- Agriculture (economic) – the population employed in agriculture, normalised by the economically active population;
- Education (social) – the university-educated population, normalised by the population over the age of 15 years; and
- Age (demographic) – the population 65 years and older, normalised by the overall population.

Figure 3 captures the empirical results: the spatial approach to geographical variability (Moran's  $I$ ) on the vertical axis; and the regional (Theil index decomposition) on the horizontal. The setup is based on the theoretical assumptions presented in Figure 1. There are three combinations (types) represented in this empirical case. The unemployment rate (red) proved to be both spatially dependent and bounded in regions. This result was expected due to the fact that regional delimitation matches quite well the labour market delimitation (based on work flows).

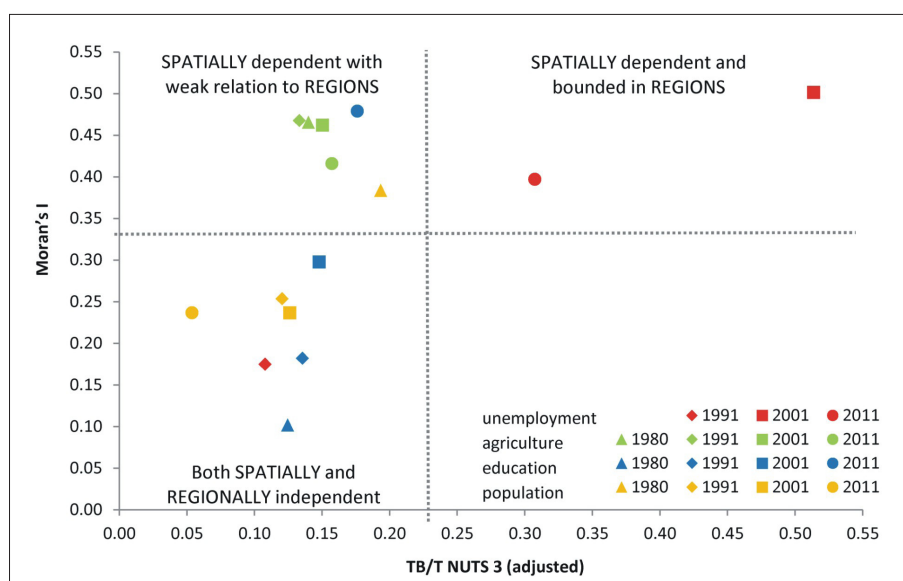


Fig. 3: Empirical results of regional and spatial approaches in the Czech Republic between 1980 and 2011

Note: For the Moran's  $I$  calculation, spatial weights based on queen contiguity (first order of contiguity) were used. The Theil index was weighted by population

Source: Czech Statistical Office (population census 1980, 1991, 2001, 2011); authors' calculations



The only exception is the year 1991, when the regional and spatial patterns were not yet developed. The employment in agriculture (green) is a typical representative of a spatially dependent characteristic with a weak relation to regions. In this case, the administrative regions do not fit with relatively highly agricultural areas determined to a large extent by physical geography. The share of university educated (blue) has always been regionally less dependent with clusters around larger cities. The clustering (concentration) has been steadily increasing since 1980. Demographic characteristics, such as the share of the 65 years and older population, are typical representatives of both spatially and regionally independent characteristics. The slightly higher values of both Moran's  $I$  and  $T_B/T$  in 1980 is caused mainly by a different age structure in Sudetenland.

For a proper interpretation and understanding of the empirical results, it is critical to also consider the local level. Although both aforementioned methods (regional and spatial) are suitable for studying the geographical variability of the studied phenomena, several important questions remain unanswered:

- What role does “the spatial” play in the distribution?;
- What is the nature of spatial clustering – can we identify development axes or nodes, areas of peripheries, and so on?; and
- In what localities does statistically significant clustering occur?

Answering these questions is common in most geographical research (see for example, Sun and Jones, 2013) and crucial for a geographical contextual understanding of the studied processes, which is naturally more important than mere quantifications of the differences.

Local statistics of spatial autocorrelation present the most suitable way to support simple graphical visualisation by identifying and testing spatial clusters. Local statistics have many advantages over simple visualisation, as well as when compared with global statistics showing the average for the entire studied area. They eliminate the problems of analysing spatial aggregated data, help to discover deviations from global statistics and thus help to better map spatial processes (Fotheringham, 1997; Unwin and Unwin, 1998). The advantages of LISA cluster maps are documented using empirical examples with Czech data.

In the category of variables “spatially dependent and bounded in regions” (see Fig. 1), one cannot assess the character of clustering and its relation to regional organisation. There may be differences, however, in the type of clustering. Areal clusters, axes, centres, or other specific patterns may form. From the studied characteristics, unemployment proved to have the highest relative regional variability, as well as global spatial autocorrelation. As shown in Figure 4, the unemployment rate forms spatial clusters with low-low types of clusters organised in axes connecting Prague with other regional centres in Bohemia

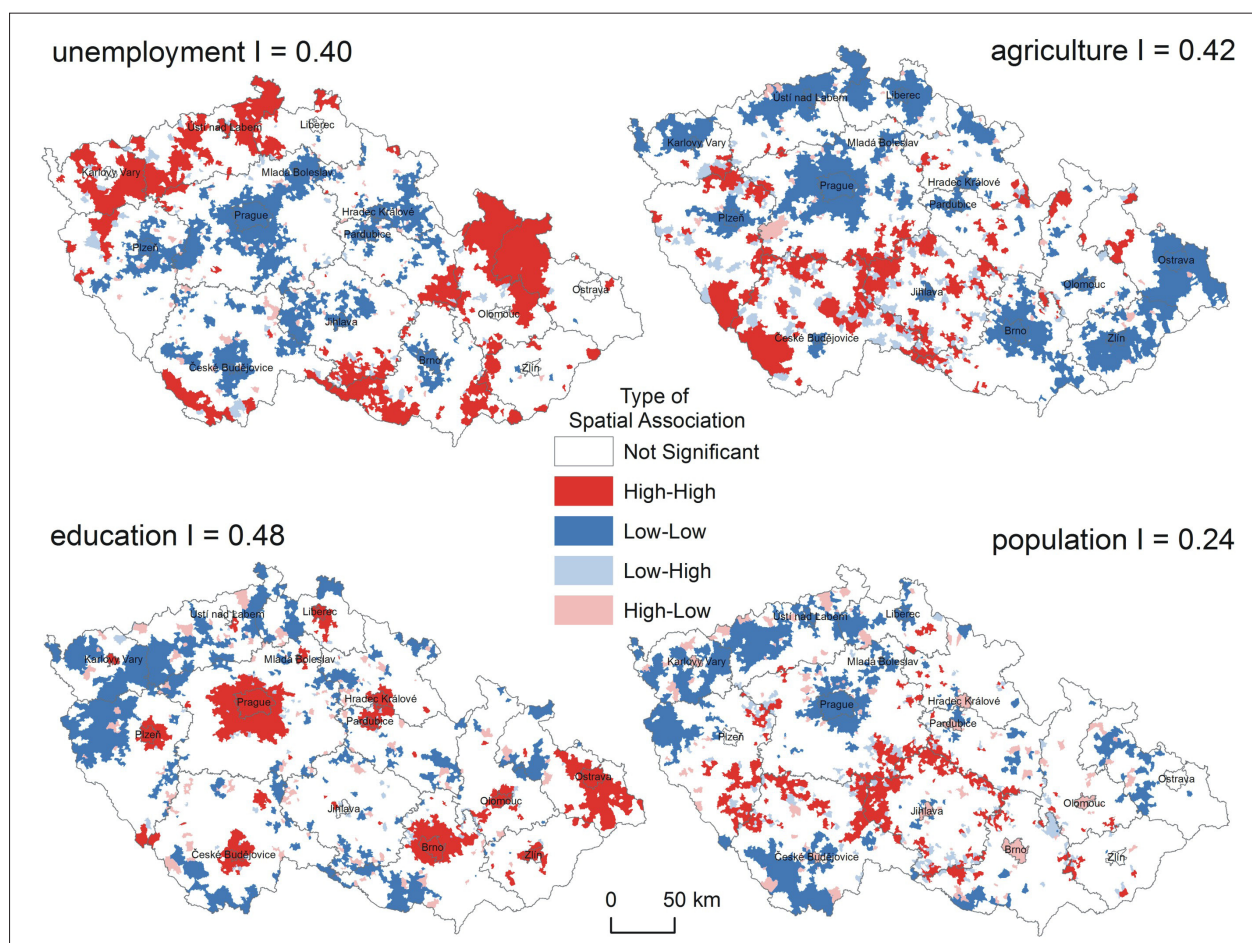


Fig. 4: LISA cluster maps for empirical Czech data, weighting scheme queen contiguity 1<sup>st</sup> order

Note: The High-Low type of spatial association indicates that a municipality with a value above the mean is surrounded by municipalities with values below the mean, and so on. The significance cut-off value of 0.05 is used (after carrying out 9,999 permutations). The permutation procedure was performed using GeoDa 1.4.0.

Source: authors' calculations based on Czech Statistical Office (population census 2011)

(Plzeň, Liberec, České Budějovice). The high-high types of unemployment rate clusters are located in structurally affected regions that were formerly oriented on heavy industry (northern Bohemia, the Ostrava region) and peripheral regions (such as southern parts of Moravia). The regional distribution of clusters corresponds with the evidence of high regional variability in the unemployment rate – in the majority of NUTS 3 regions, low-low or high-high clusters dominate.

The question of the different character of spatial clusters is also relevant when variables belong to the second category “spatially dependent with weak relation to regions” (Fig. 1). We can only assume that spatial clusters are formed across regional borders, i.e. they do not comply with the regional structure. Different spatial patterns are documented by employment in agriculture and the share of the population that is university-educated. Employment in agriculture is regionally more bounded than expected; however, the  $T_B / T$  is lower at the higher geographical level (NUTS 3), indicating the possibility that the concentrations run across these regions. This characteristic can, to a large extent, be determined by physical-geographic attributes (visible in Fig. 4). Spatial clusters of the low-low type are located especially in border regions (regions with higher altitude) and in metropolitan areas (around Prague, Brno, Ostrava, and so on) that are organised in “areal clusters”. The share of the university-educated population shows a slightly different pattern.  $T_B / T$  is relatively low, but Moran’s  $I$  is rather high because the university-educated population is concentrated in bigger cities around which high-high clusters can also be found (especially university cities such as Prague, Brno, Ostrava, Olomouc, Hradec Králové, and so on). There are not many low-low types of clusters for this variable.

Local analysis is also important for the category of “both spatially and regionally independent variables” (see Fig. 1). For these variables, no statistically significant spatial autocorrelation was observed. As these results were obtained for the whole area under study (in our case 6,251 municipalities in the Czech Republic) as average values, we cannot be sure whether the variable does not cluster in the whole area or whether there are some local clusters. Although the share of the population 65 and over did not appear to have levels of spatial autocorrelation as high as other variables, the LISA cluster map in Figure 4 shows significant spatial clusters similar to that of the agriculturally-employed population and the share of the university-educated population. The identification of spatial clusters of the low-low type is determined historically and is related to the displacement of ethnic Germans from the Sudetenland after World War II. To summarise, the typology suggested in Figure 1 was supported by empirical findings.

## 6. Conclusions

Two approaches for assessing the variability and geographical patterns of population characteristics were introduced and their relationships to spatial measures were discussed on (simulation) model and empirical examples. We use spatial autocorrelation measures for the spatial approach, and relative regional variability for the regional approach.

These two approaches are widely used in geography and regional science, but only one method is normally used, depending on the researchers’ methodological backgrounds,

preferred research field, or main research goals. By assessing the relationship between Moran’s  $I$  and the Theil index decomposition, we documented a complex relationship between spatial autocorrelation and relative regional variability. This relationship was studied both theoretically (on simulated data) and empirically (with the example of empirical data for the Czech regional structure).

Using theoretical simulations with modelled data, we demonstrated that relative regional variability highly correlates with values of spatial autocorrelation. This correlation is predominantly caused by methodological similarities of both statistics, while the correlation of regional variability and spatial autocorrelation is purely empirical. With three possible types, however, the relationship between relative regional variability and spatial autocorrelation is slightly more complex. This typology helps in assessing how the spatial concentration of the respective variables corresponds with regional delimitations.

In summary, the regional variability and spatial autocorrelation approaches are strongest when used conjointly rather than separately (Rey, 2004). They produce important complementary findings about spatial aspects of variability. In the relative regional variability approach, differences are attributed to geographical levels, while global spatial autocorrelation and its local form help to uncover local specifics that are unrelated to regional structure.

It is also important to mention the restrictions on how the methods proposed in this paper should be used. First, very detailed data are required for measuring spatial autocorrelation. This is often a problem in the social sciences. Analyses are therefore often limited to a few variables and frequently to data from population censuses. Second, the dependence of spatial autocorrelation statistics on the subjective choice of a spatial weighting scheme may be considered important. Our tests, however, suggest that the choice of spatial weight matrix does not influence the final interpretation. Finally, although the combination of both methods helps mitigate the checkerboard problem and MAUP, they should still be taken into account. For example, the results of the regional variability analyses depend on the chosen regional structure and thus can directly face MAUP. On the other hand, comparing simple and relative regional variability for different regional structures can uncover the effects of that structure on interpretations of the final results.

The combination of spatial and regional viewpoints can have interesting implications for public policy as well. It is clear that attributing processes to different regional levels (and thanks to local analyses, also to specific localities) can have strong practical implications. In the geographical context of the European Union, it is highly relevant to study the role of international borders and/or other regional borders (such as NUTS 2, the basic regional units for the EU’s convergence policy). The study of geographical patterns and variability should be extended in research that deals with local or micro-regional data and when comparing the regional structure of socio-demographic and socio-economic indicators (such as in Kládvo et al., 2012). Not only can this help to study cross-border cooperation (or segregation), but it can also help to understand the EU’s integration process better.

Despite the interesting findings of this study, there are still many avenues for future research. The generalisations presented here should be tested repeatedly with other



variability and spatial autocorrelation statistics. Importantly, these methods should be employed in empirical research. By capitalising on the advantages and complementarity of both approaches, interesting and innovative outcomes can be uncovered to reach a better understanding of both regional differences and dependencies, as well as the spatial effects of variability.

## Acknowledgement

*This work was supported by the Czech Science Foundation (GACR) under Grant No. 15-10493S – “Evolutionary dynamics of spatial differentiation of socioeconomic phenomena and the role of regions in Czechia spatial and multilevel approach”.*

## References:

- ANSELIN, L. (1988): Spatial Econometrics: Methods and Models. Dordrecht, Kluwer.
- ANSELIN, L. (1995): Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2): 93–115.
- ANSELIN, L. (2003): An introduction to spatial autocorrelation analysis with Geoda. Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, USA. [online]. [cit. 05.04.2013]. Available at: [http://www.utdallas.edu/~briggs/poec6382/geoda\\_spauto.pdf](http://www.utdallas.edu/~briggs/poec6382/geoda_spauto.pdf)
- ANSELIN, L., REY, S. J. (1991): Properties of tests for spatial dependence in linear-regression models. *Geographical Analysis*, 23(2): 112–131.
- ANSELIN, L., SYABRI, I. KHO, Y. (2004): GeoDa: An introduction to spatial data analysis. Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, USA. [online]. [cit. 05.04.2013]. Available at: <http://www.csiss.org/events/workshops/geodaGA.pdf>
- ARBIA, G. (1989): Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems. Boston, Kluwer.
- ARBIA, G. (2000): Some critique to statistical measures of spatial concentration and convergence. *International Advances in Economic Research*, 6(3): 590.
- ARBIA, G. (2001): The role of spatial effects in the empirical analysis of regional concentration. *Journal of Geographical Systems*, 3(3): 271–281.
- BLAŽEK, J., NETRDOVÁ, P. (2012): Aktuální tendence lokální diferenciace vybraných socioekonomických jevů v Česku: směřuje vývoj k větší mozaikovitosti prostorového uspořádání? *Geografie*, 117(3): 266–288.
- BOOTS, B., TIEFELSDORF, M. (2000): Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, 2(4): 319–348.
- BRÜLHART, M., TRAEGER, R. (2005): An account of geographic concentration patterns in Europe. *Regional Science and Urban Economics*, 35(6): 597–624.
- CLIFF, A. D., ORD, J. K. (1973): Spatial Autocorrelation. London, Pion.
- COWELL, F. A., FLACHAIRE, E. (2007): Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, 141(2): 1044–1072.
- Czech Statistical Office (2014): Results of population censuses 1980, 1991, 2001, and 2011 for municipalities in a regional structure in 2011. Prague, Czech Statistical Office.
- ELBERS, C., LAJOUW, P. F., MISTIAEN, J. A., ÖZLER, B. (2008): Reinterpreting between-group inequality. *The Journal of Economic Inequality*, 6(3): 231–245.
- FOTHERINGHAM, A. S. (1997): Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21(1): 88–96.
- FOTHERINGHAM, A. S., BRUNSDON, C., CHARLTON, M. (2000): Quantitative geography – Perspectives on spatial data analysis. London, SAGE.
- GETIS, A. (2007): Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, 37(4): 491–496.
- GETIS, A., ALDSTADT, J. (2004): Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2): 90–104.
- GUIMARAES, P., FIGUEIREDO, O., WOODWARD, D. (2011): Accounting for neighboring effects in measures of spatial concentration. *Journal of Regional Science*, 51(4): 678–693.
- HAMPL, M. [eds.] (1999): Geography of Societal Transformation in the Czech Republic. Prague, Department of Social Geography and Regional Development, Faculty of Science, Charles University in Prague.
- IMRE, A. R., NOVOTNÝ, J., ROCCHINI, D. (2012): The Korcak-exponent: a non-fractal descriptor for landscape patchiness. *Ecological Complexity*, 12: 70–74.
- KLADIVO, P., PTÁČEK, P., ROUBÍNEK, P., ZIENER, K. (2012): The Czech-Polish and Austrian-Slovenian borderlands – similarities and differences in the development and typology of regions. *Moravian Geographical Reports*, 20(3): 22–37.
- KLAPKA, P., HALÁS, M., NETRDOVÁ, P., NOSEK, V. (2016): The efficiency of areal units in spatial analysis: Assessing the performance of functional and administrative regions. *Moravian Geographical Reports*, 24(2): 47–59.
- KORČÁK, J. (1938): Deux types fondamentaux de distribution statistique. *Bulletin de l'Institut International de Statistique*, 30: 295–299.
- LAFOURCADE, M., MION, G. (2007): Concentration, agglomeration and the size of plants. *Regional Science and Urban Economics*, 37(1): 46–68.
- LAMBERT, P. J., ARONSON, J. R. (1993): Inequality decomposition analysis and the Gini coefficient revisited. *The Economic Journal*, 103(420): 1221–1227.
- LITCHFIELD, J. A. (1999): Inequality: Methods and Tools. World's Bank Web Site on Inequality, Poverty, and Socio-economic Performance. [online]. [cit. 30.05.2013]. Available at: <http://siteresources.worldbank.org/INTPGI/Resources/Inequality/litchfie.pdf>
- LONGFORD, N. T., PITTAU, M. G., ZELLI, R., MASSARI, R. (2012): Poverty and inequality in European regions. *Journal of Applied Statistics*, 39(7): 1557–1576.
- MARCON, E., PUECH, F. (2003): Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography* 3: 409–428.

- MILLS, J. A., ZANDVAKILI, S. (1997): Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics*, 12(2): 133–150.
- MUSSARD, S., SEYTE, F., TERRAZA, M. (2003): Decomposition of Gini and the generalized entropy inequality measures. *Economics Bulletin*, 4(7): 1–6.
- NETRDOVÁ, P., NOSEK, V. (2009): Approaches to measuring the significance of geographical dimension of societal inequality. *Geografie*, 114(1): 52–65.
- NOSEK, V., NETRDOVÁ, P. (2010): Regional and spatial concentration of socio-economic phenomena: empirical evidence from the Czech Republic. *Ekonomický časopis/ Journal of Economics*, 58(4): 344–359.
- NOSEK, V., NETRDOVÁ, P. (2014): Measuring Spatial Aspects of Variability. Comparing Spatial Autocorrelation with Regional Decomposition in International Unemployment Research. *Historical Social Research*, 39(2): 292–314.
- NOVOTNÝ, J., NOSEK, V. (2009): Nomothetic geography revisited: statistical distributions, basic generative mechanisms, and inequality measures. *Geografie*, 114(4): 282–297.
- NOVOTNÝ, J., NOSEK, V. (2012): Comparison of regional inequality in unemployment among four Central European countries: an inferential approach. *Letters in Spatial and Resource Sciences*, 5(2): 95–101.
- NOVOTNÝ, J., NOSEK, V., JELÍNEK, K. (2014): EasyStat 1.0. Prague, Faculty of Science, Charles University. [cit. 19.05.2015]. Available at: <https://web.natur.cuni.cz/~pepino/EasyStat.zip>
- OKAMOTO, M. (2009): Decomposition of gini and multivariate gini indices. *The Journal of Economic Inequality*, 7(2): 153–177.
- OPENSHAW, S. (1984): Concepts and Techniques in Modern Geography. Volume 37: The Modifiable Areal Unit Problem. Norwich, Geo Books.
- ORD, J. K., GETIS, A. (2012): Local spatial heteroscedasticity (LOSH). *The Annals of Regional Science*, 48(2): 529–539.
- REY, S. J. (2004): Spatial analysis of regional income inequality. In: Goodchild, M., Janelle, D. [eds.]: *Spatially Integrated Social Science: Examples in Best Practice 1* (pp. 280–299). Oxford, Oxford University Press.
- REY, S. J., JANIKAS, M. V. (2005): Regional convergence, inequality, and space. *Journal of Economic Geography*, 5(2): 155–176.
- REY, S. J., MONTOURI, B. D. (1999): US regional income convergence: a spatial econometric perspective. *Regional Studies*, 33(2): 143–156.
- REY, S. J., SMITH, R. J. (2013): A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences*, 6(2): 55–70.
- SEN, A. (1972): *On economic inequality*. Oxford, Clarendon Press.
- SHORROCKS, A., WAN, G. (2005): Spatial decomposition of inequality. *Journal of Economic Geography*, 5(1): 59–81.
- STINE, R. (1989): An introduction to bootstrap methods: examples and ideas. *Sociological Methods Research*, 18(2–3): 243–291.
- STAKHOVYCH, S., BIJMOLT, T. H. A. (2009): Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science*, 88(2): 389–408.
- SUBRAMANIAN, S. (2004): Indicators of inequality and poverty. Research Paper No. 2004/25, World Institute for Development Economics Research. [online]. [cit. 30.05.2013]. Available at: <https://www.wider.unu.edu/sites/default/files/rp2004-025.pdf>
- SUN, W., JONES, B. (2013): Using multi-scale spatial and statistical analysis to assess the effects of brownfield redevelopment on surrounding residential property values in Milwaukee County, USA. *Moravian Geographical Reports*, 21(2): 56–64.
- UNWIN, A., UNWIN, D. (1998): Exploratory spatial data analysis with local statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3): 415–421.
- WONG, D. (2009): The Modifiable Areal Unit Problem (MAUP). In: Fotheringham, A. S., Rogerson, P. A. [eds.]: *The SAGE Handbook of Spatial Analysis* (pp. 105–125). London, SAGE.

**Please cite this article as:**

NETRDOVÁ, P., NOSEK, V. (2017): Exploring the variability and geographical patterns of population characteristics: Regional and spatial perspectives. *Moravian Geographical Reports*, 25(2): 85–94. Doi: 10.1515/mgr-2017-0008.