Spatial Autocorrelation of a Demographic Phenomenon: a Case of One-Family Households and One-Person Households

Jaroslav Kraus[1] | *Charles University in Prague, Prague, Czech Republic*

Abstract

The paper aims to examine spatial distribution and to subsequently analyse a demographic phenomenon; share of one-family households and one-person households of all households by municipality in the Czech Republic (CR), because these two types of households are in a way contradictory. Although the Czech Republic is rather small and homogeneous with respect to demographic processes, it is questionable whether this also applies to the spatial distribution of households. Methods of local and global spatial autocorrelation represented by Moran's index were used; however, the challenge of normalisation of both variables using Box-Cox transformation had to be addressed before. For comparison classical measures of association on NUTS3 level were used with respect to the type of data being examined – e.g. the Pearson chi-square, the uncertainty coefficient, the lambda coefficient, as well as the Gini index (Gini ratio), which is a measure of statistical dispersion and the most commonly used measurement of inequality.

From the results a small statistically significant global autocorrelation was ascertained. Local results show that shares of both types of households are not a complementary phenomenon in terms of space. From the relationship of local and global Moran's index, it can be concluded that the global rate does not sufficiently depict detected local differences.

Keywords	***JEL code***
Population and housing census, households, spatial autocorrelation, Czech Republic	*C21*

INTRODUCTION – SOURCES AND REFERENCES

Spatial information related to demographic processes is an integral part thereof and becomes thus a subject of a demographic analysis. It can be presented clearly in population censuses – the largest demographic survey – in which specifically the analysis for small territorial units such as e.g. municipalities is one of the reasons why the censuses are carried out.

[1] Faculty of Science, Department of Demography and Geodemography, Albertov 6, 128 00 Prague, Czech Republic. E-mail: kraus@natur.cuni.cz. Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: jaroslav.kraus@czso.cz.

Households belong to the modern times' phenomena and are a long-term interest of demographers. The changing structure of households is one of the fundamental attributes of the second demographic transition, as mentioned in the standard contribution of demographic literature (van Kaa, 1987, p. 32). There, it is stated that the changes in the propensity to marry, divorce, separate, remarry, or cohabit, changes in fertility behavior and in the age at which children leave home, along with mortality trends and differentials, have had a marked impact on household patterns in Europe. Therefore, one-family households and one-person households were used as an example of possible spatial changes in this paper.

The second demographic transition, as one of the key elements of demographic development of contemporary societies, is closely related to spatial distribution of data. It is precisely described in (Howell et al., 2016, Chapter 6), where the local character of demographic data is mentioned and universal principles by means of a spatial analysis are being sought for distribution of the data. For example, retrospectively, a fertility decline in Europe is a classic example of spatial autocorrelation. In general, demographic transition – in all its components – is said to have an impact on all aspects of life in society.

Spatial analysis of (not only demographic) data can be defined as quantitative data analysis, in which the explanation is based on explicit spatial variables or prediction of the phenomenon observed based on spatial autocorrelation. Two elements are related to spatial distribution of data: spatial dependence and spatial heterogeneity. Spatial dependence is connected with Tobler's first law of geography – everything is related to everything else, but near things are more related than distant things (Tobler, 1970). Demographic data are governed by the same statistical principles as any other data of stochastic character. Spatial analysis of demographic data thus is the very essence of geodemography (unlike social geography) and is related to all components of demographic development (Howell et al., 2016, p. 102 and other) in their mutual relationships.

When abstracted from spatial autocorrelation, a variety of statistical methods can be applied. Variability rates are often used to quantify regional differences and to develop regional differentiation (NUTS3 level). Gini's concentration coefficient is used in geographic surveys, because it overcomes the deficiencies of the coefficient of variation depending on the average and is therefore more appropriate for affecting the variability of asymmetric distributions typical of socio-geographical phenomena (Netrdová, 2012, p. 270). In the case of measuring the intensity of statistical dependence, it is possible to use a classical chi-square test (non-parametric) for a nominal variable (NUTS3 level).

The aforementioned principles can be illustrated using examples of data for households, surveying and a subsequent analysis, which is an integral part of population censuses. Spatial analysis of households is a relatively new topic, especially a geostatistical approach, however, several interesting contributions with a focus on the Czech Republic have been published on this subject (see Bleha, 2019; or Netrdová, 2009). Changes occurred after 1981 when demographers started to be more aware of the issue of households. The paper is based on work with data on two types of households: one-family households and one-person households. Development and structure of families as one of the types of households, become a key element of demographic trend in many countries; the methodological definition is standardized and coordinated well – see, for example (UNECE, 2011, p. 10 and other). According to this, one-family households are comprised of one couple without children, a couple with one or more children, or a lone parent with one or more children, independently of their de jure marital status. A one-person household is made by a person who lives alone in a separate housing unit or who occupies, as a lodger, a separate room (or rooms) of a housing unit but does not join with any of the other occupants of the housing unit to form part of a multi-person household. In compliance with international recommendations, households are determined based on 'place of usual residence'.

One-person households belong to modern-day phenomena; they are constantly increasing in number (both in absolute and relative values) and their observation provides basic characteristics of population development in many countries including the Czech Republic. The subject of formation and dissolution of (multi-member) households is rather broad. This paper is devoted only to their current situation, i.e. regardless of their formation and possible dissolution due to the death of a household member or divorce of the married couple. Besides structure of households, it is also possible to pay attention to variables that are connected to households, e.g. income, housing prices, but also changes in gender roles (van Imhoff et al., 1995, p. 91 and other). Further, regarding these variables, it is possible, as with households, to expect spatial variability.

As stated previously, population censuses belong to basic sources of data on households (and characteristics of their members), as was the case in 2011. During the 2011 Census, both dwelling households and private households were surveyed in detailed structures. *The aim of the paper is to detect – on the distribution of the share of one-person households and of the share of households consisting of one complete family by municipality – whether there is a spatial autocorrelation and, further, whether the spatial autocorrelation is the same for both types of households or whether it is different (e.g. complementary).*

1 POPULATION AND HOUSING CENSUS IN THE CZECH REPUBLIC 2011

In the 2011 Population and Housing Census, over 4 375 thousand households were counted. The total amount of private households has been increasing over a long period of time. In the last ten years, it increased by 4% and since 1970 it has increased by over a million, expressed in an absolute value (see Table 1).

Table 1 Number of households in census years 1970–2011

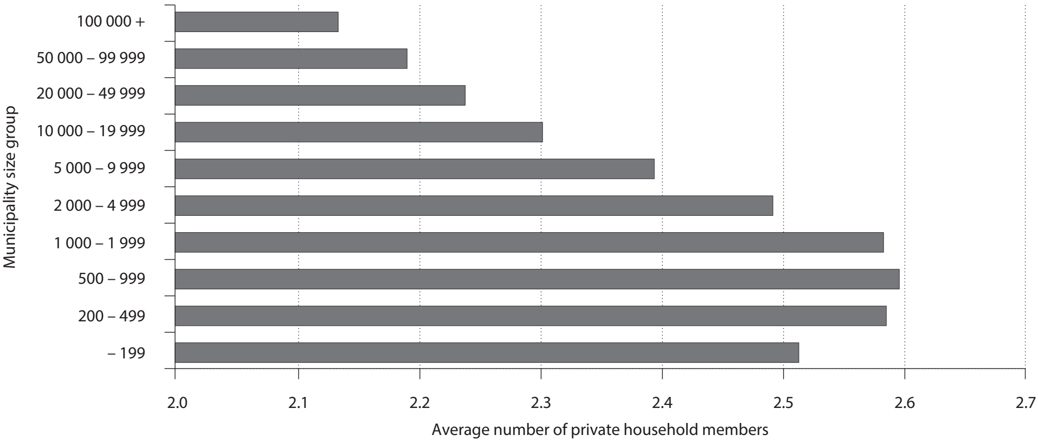
Households	1970	1981	1991	2001	2011	index 2011/1970	index 2011/2001
Total number of households	3 365 407	379 097	3 983 858	4 216 085	4 375 122	130	104
One-family households	2 526 778	2 760 247	2 856 608	2 803 340	2 667 867	106	95
One-person households	668 859	897 447	1 047 221	1 276 176	1 422 147	213	111

Source: 2011 Population and Housing Census (CZSO, 2013)

The basic characteristics of the development of households include the relative decline in the share of family households at the expense of uncompleted family households, and the absolute and relative growth of one-person households. While in 1970, private households consisting of one complete family made up two thirds of all private households, four decades later they made up hardly half. A decrease in their proportion was caused mainly by a marked absolute increase in one-person households. In 2011, one-person households comprised already a third of all households. The average size of a household has also been constantly decreasing for a long period of time. In 1970, on average 2.89 persons lived in a private household, whereas in 2011 it was only 2.34 persons. This, together with a change in the structure of private households, is a result of a long-term demographic trend, especially due to declines in fertility rate and a long-term high levels of divorce rate and an increasing availability of independent living (i.e. by frequent and simpler decomposition of complete families to singles and incomplete families) (CZSO, 2013, pp. 37–39).

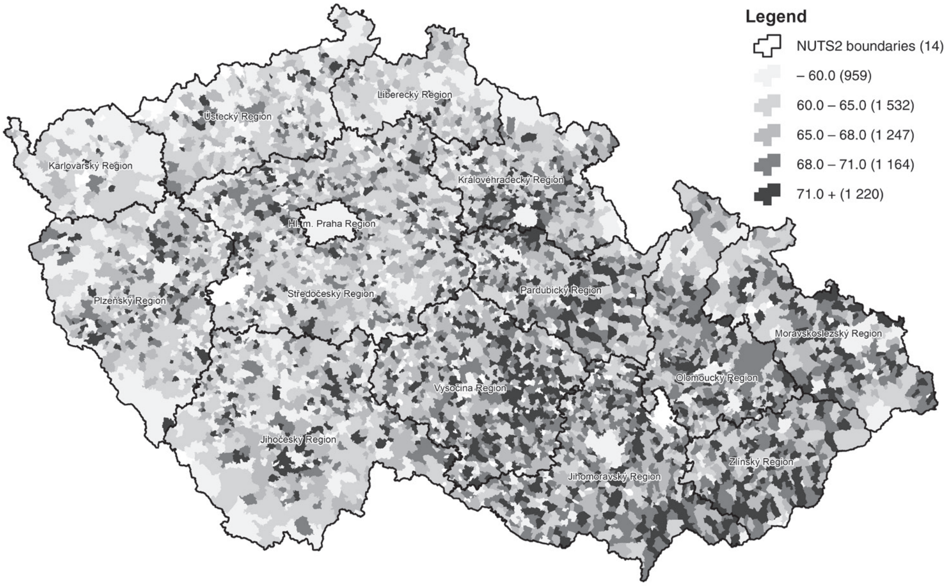
A household structure can be also viewed from the perspective of the number of households in a municipality (Figure 1). The highest average number of household members has been recorded in municipalities with a (usually resident) population of 200–1 999 (2.58–2.60 persons in a household); it is also valid that the bigger size of a municipality, the smaller the average size of a private household tends to be (CZSO, 2013, pp. 37–39).

Figure 1 The average number of private household members by municipality size group



Source: 2011 Population and Housing Census (CZSO, 2013)

Figure 2 Share of one-family households in all households by municipality in the CR

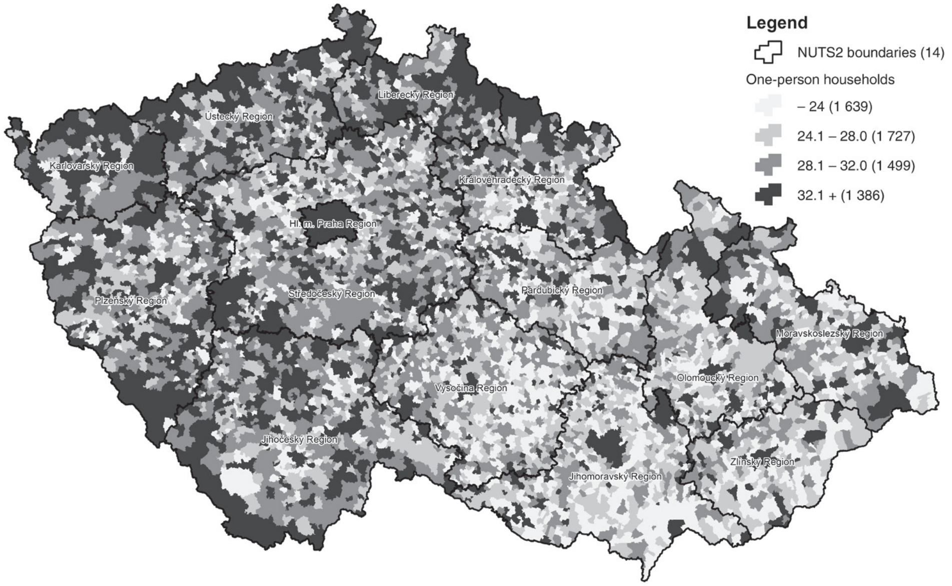


Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation based on data from the 2011 Population and Housing Census (CZSO, 2013)

From Figures 2 and 3 it is obvious that the distribution of the two aforementioned household types is not homogeneous in the territory of the Czech Republic.[2] It appears that in general, one-person

[2] Processing thereof is based on results for municipalities. As at the Census date, 6 251 municipalities were in the Czech Republic; they are grouped together (with a hierarchy) to higher territorial units (NUTS).

Figure 3 Share of one-person households in all households by municipality in the CR



Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation based on data from the 2011 Population and Housing Census (CZSO, 2013)

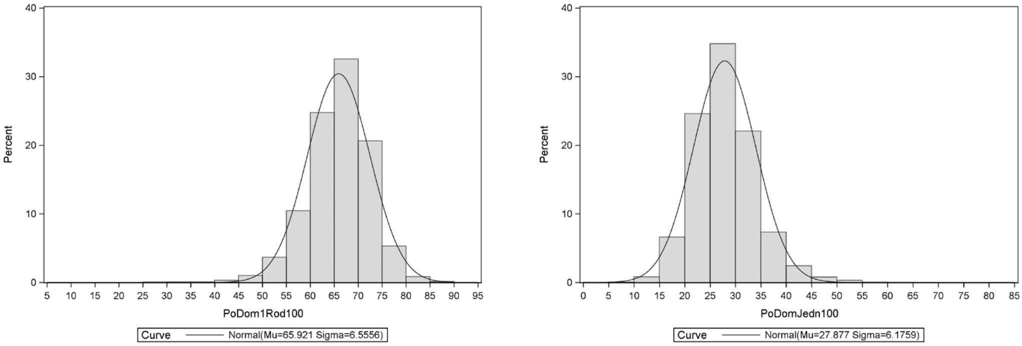
households are more frequent in the western part of the Czech Republic, while households consisting of one complete family are more common in the eastern part, while households consisting of one complete family are more common in the eastern part, relative to households of a given type. However, there are frequent exceptions from that distribution. The mentioned results raise the question *to what extent shares of one-person households and of one-family, households are influenced by their spatial distribution, i.e. whether they are spatially autocorrelated.* From a methodological point of view, variables can be seen as a continuous variable; a figure for a given territorial unit (e.g. a municipality) is a result of settlement of the entire municipality and, similarly, the settlement and structure of the entire CR results from data for all municipalities.

2 METHODS AND METHODOLOGY

Spatial autocorrelation may be a result of unobserved or hard-to-quantify processes, combined in various places, consequently causing spatial structuring of a given phenomenon. In the context of specification of econometric models (i.e. searching for explanatory variables), measuring of spatial autocorrelation can be considered to be a diagnostic tool. If there is a spatial autocorrelation, it is determined by examining whether the variable value for a given (e.g. geolocalized) observation is associated with values of the same variable for neighboring observations (INSEE, 2018, p. 67). Spatial autocorrelation may be positive, negative, or there is no spatial autocorrelation among given data. Spatial autocorrelation can be measured globally or locally; both ways assess the same, i.e. whether there is a spatial correlation of a given phenomenon – however, it is not the same. There are different ways of measuring spatial autocorrelation; Moran‘s I is often used.

The use of Moran's I requires data normality and stationarity (that is, the same data mean and data variance at any location). Moran's I, however, is rarely used in geostatistics in which data stationarity is the main assumption and data normality is a desirable feature. (Krivoruchko, 2011, p. 61). Although it is not entirely clear from Figure 4 (especially after addition of a bell-shaped curve), this condition has not been met; the results of statistical tests that are used for tests for normality have not confirmed the hypothesis that the distribution of both the variables meets the condition. Many tests were elaborated to test the normality; in this case Kolmogorov-Smirnov test has been chosen followed by Cramér-von Mises test and Anderson-Darling test, respectively, for verification (SAS Base).

Figure 4 Histogram of frequency distribution of the share of one-family households (PoDom1Rod100) and the share of one-person households (PoDomJedn100)



Source: Own calculation

Based on a detailed data analysis the primary cause of why the condition of normality has not been met, was determined to be due to extreme values for some observations. However, because spatial autocorrelation is based on the concept of continuity of a variable, it was necessary to minimize the amount of excluded (i.e. extreme) observations so that continuity of data is disrupted as little as possible. The Czech Republic comprises 6 251 municipalities; after exclusion of extremely low and extremely high values of one-person households and one-family households we got 6 122 municipalities with the population of 10 413 thousand, equating to 99.8% of the total number of the usually resident population (10 437 thousand). However, that was still not enough to ensure normality; it was necessary to transform data in order to achieve the required normality. A commonly applied method of Box-Cox transformation (SAS Stat) was used, which generally is in the following form:

$$\begin{array}{ll} (y^{\lambda} - 1) / \lambda, & \lambda \neq 0, \\ \log (y), & \lambda \neq 0, \end{array} \qquad (1)$$

where we work with one lambda parameter. When lambda is approaching zero, it is basically a logarithmic transformation. An advantage of this type of transformation is that it is selected from a solution set based on individual values of the lambda parameter from a fixed interval so that the plausibility function (its logarithm) is maximized. For the computation itself we chose the software environment of the product (SAS Stat), namely the TRANSREG procedure, which enables data transformation without using an explicative variable or, to put it more precisely, with using a fictitious (constant) variable. The result of data transformation is satisfactory and the result of testing is that the transformed variable meets the condition of normality – see Table 2.

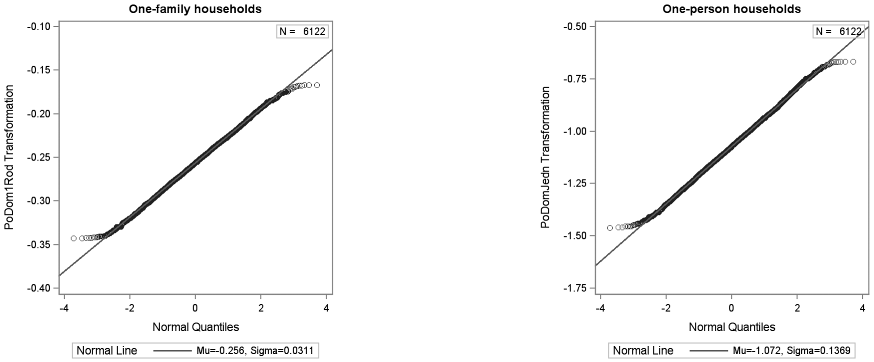
Table 2 Fitted Normal Distribution for transformed variables

One-family Household					Single Household			
Test	**Statistic**		**p Value**		**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.011	Pr > D	0.067	D	0.008	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.058	Pr > W-Sq	>0.250	W-Sq	0.033	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.465	Pr > A-Sq	>0.250	A-Sq	0.354	Pr > A-Sq	>0.250

Note: If p Value > 0.05, the hypothesis on data normality is not declined.
Source: Own calculation

Similarly, for the result of a test, a shift of both variables to normality can be documented in a histogram of frequency distribution of both variables or by other graphic procedures such as a QQ plot – see Figure 5.

Figure 5 Quantile distribution of empirical data of a given type of household compared to quantiles of normal distribution



Source: Own calculation

Now we can start to investigate a spatial autocorrelation for both types of households on a newly defined set of 6 122 municipalities. Basic information can be found e.g. in Shekhar and Xiong (2008, p. 360, 644). In this paper we used Moran's I. The principle of computation is that it takes into account the difference between the value of the variable (i.e. the share of households) and the average of values of that variable for a given area (neighborhood). Moran's index is used preferably (in comparison to other ones), because it is more stable against extreme values, further it can be used in two ways (see below). The index can be written in several ways, it is frequently written as follows:

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \,. \tag{2}$$

Null hypothesis H_0, states that there is no spatial correlation in the given territory. Vice versa, if $I_w > 0$, then there is a positive autocorrelation, which means that high values are neighboring high ones and low values are neighboring low ones. In the case of a negative autocorrelation it would be the contrary.

Depending on the distribution of a spatial variable, the calculation of a median value:

$$E(I_W) = E(c_W) = -\frac{1}{n-1}\,, \tag{3}$$

and testing statistics:

$$\frac{I_W - E(I_W)}{\sqrt{\mathrm{Var}(I_W)}} \sim \frac{c_W - E(c_W)}{\sqrt{\mathrm{Var}(c_W)}} \sim \mathrm{N}\,(0{,}1) \tag{4}$$

can be defined.

A key element for calculation of indices of spatial autocorrelation is to determine the neighborhood, i.e. to select spatial entities that are neighbors *in definition*. The defining of the neighborhood is a rather complex issue, which should always be based on knowledge of the examined issue (i.e. determination of a working hypothesis on why given spatial elements are selected to be neighbors) and that has a major influence on the result of calculation of a spatial autocorrelation. Based on a definition of neighborhood, it is then possible to start calculation of the so-called spatially weighted matrix; in the I_w calculation formula, the elements of the matrix are denominated as w_{ij}. For more details about the issue see INSEE (2018, p. 57). Modelling of spatial correlations is also described clearly in (ArcGISPro), in which computations of spatial indices for this paper were also made.

Two working hypotheses of spatial autocorrelation were stated: the *influence of immediate neighborhood* and the *influence of a local center.* In the ArcGIS environment it meant that, the Moran's I index was calculated by the contiguity-edges-corners method and by the fixed-distance method. The first hypothesis results from an assumption that if two municipalities are neighboring (in terms of topology by their edge or a corner), then there is the biggest interaction between them. The second hypothesis is calculated as follows: neighbors are those whose centroids were less than 25.8 km apart from each other. This distance has been determined in such a way that each municipality has at least one municipality with more than 1000 inhabitants as a neighbor – thus it was defined as a local center.

Transformed variables of the shares (Box-Cox transformation, see above) of both one-family households and one-person households were worked with.

Table 3 Spatial autocorrelation (Moran's I) for the share of one-family households and the share of one-person households by chosen neighbourhood method

Test	One-family households		One-person households	
	Contiguity edges corners	Fixed distance	Contiguity edges corners	Fixed distance
Moran's Index	0.193**	0.117**	0.204**	0.112**
Expected Index	0.000	0.000	0.000	0.000
Variance	0.000	0.000	0.000	0.000

Note: ** p Value<0.05.
Source: Own calculation

The result (see Table 3) can be interpreted as follows: It is clear that in the data (i.e. in the share of one-family households or one-person households) there is a positive autocorrelation, which is not very high given the interval in which the theoretical values of Moran I are located. However, especially in the case of the fixed-distance method and which in both cases is statistically significant. The value of the Expected Index indicator means that the index is calculated from randomly generated data streaming from a normal distribution, which is the case with this household data.

The Moran's index is a global statistic, which provides no information about the extent of local variation in spatial variability. For that there are tools that enable us to assess the local level of spatial autocorrelation (LISA) and to measure the intensity and importance of autocorrelation between the value of the variable in a spatial unit and the value of the same variable in neighboring spatial units. These indicators examine the following two features:

- for each observation they show the intensity of clustering of similar/opposite values around that observation;
- the sum of local indices at all observations is proportional to the corresponding global index, e.g. to global Moran's I.

In the case of Moran's I, its local value can be written as follows:

$$I_i = (y_i - \bar{y}) \sum_j w_{ij} (y_j - \bar{y}) \,, \tag{5}$$

and the value of the global index is as follows:

$$I_w = \text{konst} \cdot \sum_i I_i \,, \tag{6}$$

where:

$I_i > 0$ indicate clustering of similar values (higher or lower than the average for a given neighborhood),

$I_i < 0$ indicate clustering of different values.

Spatial clustering of similar or different values is observed *as follows: as High-High values (HH), Low-Low values (LL), High-Low values (HL), or Low-High (LH) values*. If we mean high value surrounded by another high values, resp. low value surrounded by another low value then they are referred to as hot spots, resp. cold spots. If we mean a high value surrounded by low values or a low value surrounded by high values, then these are spatial outliers (Anselin, 1995).

The significance of each local indicator is based on spatial distribution of data and statistics that is asymptomatically approaching the normal distribution:

$$z(I_i) = \frac{I - E(I_i)}{\sqrt{\text{Var}(I_i)}} \sim \text{N}\,(0{,}1)\,. \tag{7}$$

Since the global rate of spatial autocorrelation (Moran's I) proved to be distinctively higher in the case of usage of the neighboring municipalities method, local rates of Moran's I were further computed only for this method of neighborhood determination. Further, numbers of households that live in each of them were determined.

Another possibility of exploring spatial variability is to use the classical measures of association with respect to the type of data being examined. One test statistic for the hypothesis of no general association is the Pearson chi-square. This statistic is defined for *i* is from 1 to *s* and the summation for *j* is from 1 to *r*:

$$Q_P = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \,, \text{ where:} \tag{8}$$

$$m_{ij} = E\,\{n_{ij} \mid H_0\} = \frac{n_{i+} n_{+j}}{n} \tag{9}$$

is the expected value of the frequencies in the *i* th row and *j* th column.

Measures of association when one or both variables are nominally scaled are more difficult to define, since you cannot think of association in these circumstances as negative or positive in any sense. However, indices of association in the nominal case have been constructed, and most are based on mimicking R-squared in some fashion. One such measure is the uncertainty coefficient, and another is the lambda coefficient (Stokes, 2012, p. 129).

Asymmetric lambda λ (*Colums* | *Rows*), is interpreted as the probable improvement in predicting the column variable *Y* given knowledge of the row variable *X*. The range of asymmetric lambda is $0 \leq \lambda(C \mid R) \leq 1$. Asymmetric lambda (C|R) is computed as:

$$\lambda\,(C|\,R) = \frac{\sum_i r_i - r}{n - r}\,, \tag{10}$$

and its asymptotic variance is:

$$\mathrm{Var}(\lambda\,(C|\,R)) = \frac{n - \sum_i r_i}{(n-r)^3}\left(\sum_i r_i + r - 2\sum_i\,(r_i \mid l_i = l)\right). \tag{11}$$

The nondirectional lambda (symmetric) is the average of the two asymmetric lambdas, ($\lambda(C \mid R)$ and ($\lambda(R \mid C)$. Its range is $0 \leq \lambda \leq 1$. Lambda symmetric is computed as:

$$\lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c} = \frac{w - v}{w}\,, \tag{12}$$

and its asymptotic variance is computed as:

$$\mathrm{Var}(\lambda) = \frac{1}{w^4}\left(wvy - 2w^2\left(n - \sum_i\sum_j\,(n_{ij} \mid j = l_i, i = k_j)\right) - 2v^2\,(n - n_{kl})\right). \tag{13}$$

The uncertainty coefficient U is the symmetric version of the two asymmetric uncertainty coefficients. Its range is $0 \leq U \leq 1$. The uncertainty coefficient is computed as:

$$U = 2(H(X) + H(Y) - H(XY))\,/\,(H(X) + H(Y))\,, \tag{14}$$

and its asymptotic variance is:

$$\mathrm{Var}(U) = 4\sum_i\sum_i \frac{n_{ij}\left(H\,(XY)\ln\left(\frac{n_i n_j}{n^2}\right) - (H\,(X) + (H\,(Y))\ln\left(\frac{n_{ij}}{n}\right)\right)^2}{n^2\,(H\,(X) + (H\,(Y))^4}\,, \tag{15}$$

where $H(X)$, $H(Y)$, and $H(XY)$ are defined in the previous section. See (SAS Stat) for completed description.

Gini index (or Gini ratio), is a measure of statistical dispersion and it is the most commonly used measurement of inequality preferably used in economics. It measures the inequality among values of a frequency distribution. Index of zero expresses perfect equality, where all values are the same, index of 1 (or 100%) expresses maximal inequality among values. The sample Gini coefficient was calculated using the formula:

$$G = \frac{1}{2\overline{X}n(n-1)}\sum_{i=1}^{n}(2i-n-1)X_i\,, \tag{16}$$

where X_i are the sizes sorted from smallest to largest, $X_1 \le X_2 \le X_n$ (Dixon, 1987).

3 RESULTS

It is clear from the Table 4 that the highest number and the highest share of both one-family households and one-person households is made by households in municipalities, in which the value of the local index of spatial autocorrelation is not statistically significant (the *Not Significant* line). In the case of complete households consisting of one family it is 4 805 municipalities, in the case of one-person households it is 4 766 municipalities from the total number of 6 121 municipalities, which entered the computation after data transformation.

Table 4 Absolute and relative frequencies of individual types of households by municipality of the CR

Statistic	One-family households		One-person households	
	Frequency	Percent	Frequency	Percent
Not Significant	1 747 947	65.7	715 930	50.5
HH	118 119	4.4	117 824	8.3
HL	22 167	0.8	540 553	38.1
LH	542 743	20.4	9 078	0.6
LL	231 240	8.7	36 017	2.5
Total	2 662 216	100.0	1 419 402	100.0

Source: Own calculation

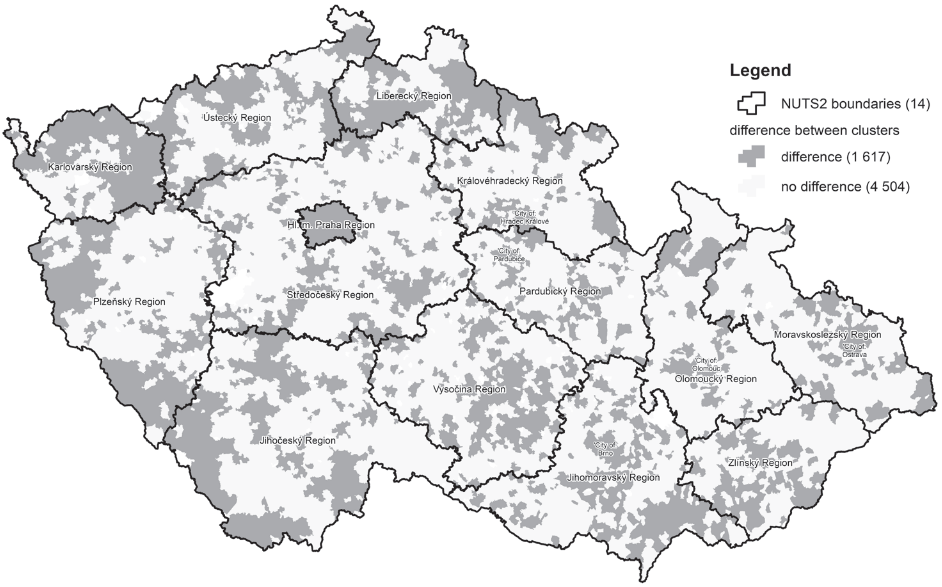
From map outputs (see the online version of Statistika journal No. 4/2019) it is possible to find that for both variables there are territories in which the spatial autocorrelation is higher than in the remaining territory of the Czech Republic. In the case one-family households it applies to many municipalities in the *Karlovarský* Region, the *Ústecký* Region, the *Liberecký* Region, and partially also the *Plzeňský* Region, and the *Jihočeský* Region, where there are mainly Low-Low clusters. In the case of one-person households, the situation in those municipalities is – quite logically – the opposite, i.e. High-High clusters. An interesting situation is observed in the eastern part of the Czech Republic. For the share of the one-family households variable, High-High clusters (hot spots) are quite frequent (a high share of complete households consisting of one family), meaning that this share is significant in many municipalities. But also Low-High cases (spatial outliers), i.e. the low values of this share accompanied by a high share in neighbouring municipalities are significant. Similarly, for one-person households in the eastern part of the Czech Republic, there are Low-Low clusters (cold spots), meaning there is a lower share of one-person households spread in a significant way. At the same time a High-Low type (spatial outliers), i.e. the high value of the share of one-person households in the given municipality is observed in the neighbourhood of municipalities with the low share. This is the case for the cities of *Brno*, *Ostrava*, *Pardubice*, *Hradec Králové*, but also for many smaller towns in the eastern part of the country. The same situation is seen for the share of one-person households in the capital city of Prague, while in the neighbouring municipalities of the *Středočeský* Region this phenomenon is less frequent. Interpretation of hot spots and cold spots for both variables (one-person households, one-family households) is interesting and will be addressed in the upcoming article on spatial regression analysis.

It is interesting to compare agreement of both types of clusters, i.e. one-family households and one-person households by municipality (see Figure 6). It turns out that the agreement or disagreement is not the same in all regions and that it differs even within the regions. For example, virtually the entire *Středočeský* Region contains the not significant result for the clusters of both types of households and therefore the agreement is high there (light colour is prevailing). In contrast, the *Capital City of Prague* is, in the case of one-person households, the High-Low type, i.e. a high share of one-person households in Prague surrounded by a low share of one-person households in neighbouring municipalities (on average). In the case of one-family households, it is the *Not Significant* type and therefore the value for Prague is marked with a dark colour on the map.

As in the case of Prague and other big cities, also in the remaining (i.e. smaller) municipalities of the Czech Republic, the situation is differentiated and *it cannot be said that shares of one-family households and one-person households are a complementary phenomenon: i.e. where there is a high share of one-person households the opposite is true for one-family households*.

Frequencies of significant clusters are different for both types of households. As established, the High-Low type refers to the high value on the given territory (municipality) surrounded by low values in the neighbourhood (on average) and a Low-High type means that the low value on the given territory is surrounded by high values in its neighbourhood (on average). For the case of one-person households it means that the High-Low types occur in big cities (e.g. *Brno*, *Ostrava*, *Olomouc*, *Hradec Králové*, *Pardubice*),

Figure 6 Cluster and outlier analysis according to the agreement of individual clusters of municipalities in the Czech Republic (comparison between one-family households and one-person households)

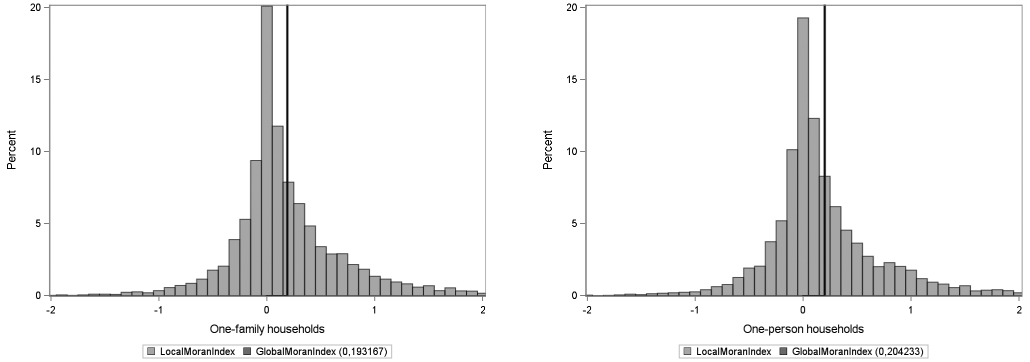


Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation

while households consisting of one family (of Low-High type) are more frequently observed in smaller municipalities and are therefore more likely to occur in the western part of the country.

Local indicators of spatial autocorrelation enable to identify areas in which similar values are clustered in a statistically significant way. In general, if the global spatial autocorrelation is strong or at least observable (as it is in the case of households) then local indicators indicate those areas that have a special impact on the global process (local autocorrelation is higher than the total autocorrelation) or, vice versa, where an obvious autocorrelation exists although the global autocorrelation is not significant.

Figure 7 Indexes of global and local rates of spatial autocorrelation (Moran's I)

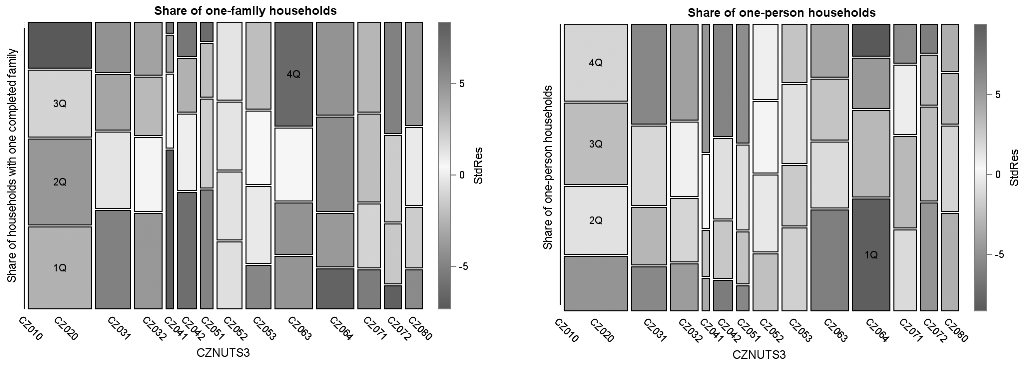


Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: Own calculation

The relationship can be observed in the histogram of frequencies showing local and global rates of autocorrelation of one-family households or one-person households as computed by nearest neighbour method.

From mutual comparison of the global rate and local rates (see Figure 7) it is apparent in both cases (i.e. in the case of one-family households as well as in the case of one-person households) the global rate in the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

Figure 8 Frequency plot by type of households and NUTS3



Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: Own calculation

The previous results can be compared with the results of spatial variability measurement, i.e. by measures of association at the NUTS3 level.

The mosaic plot (Figure 8) is a graphical depiction of the frequency table. It shows the distribution of the weight categories by dividing the x axis into 14 intervals (NUT3 level). The length of each interval is proportional to the percentage of the share of households, which are divided into four quartiles by type. Within each quartile category, the share of households is further subdivided by NUTS3 level. In order to make the residuals comparable across cells, the standardized residuals were added on the right side of the both graphs. The width of the column indicates the frequency of the phenomenon monitored.

The results show that the distribution is neither identical nor complementary. Both types of households create separate spatial patterns and show a relatively large variability of the observed phenomenon.

Table 5 Statistics for share of households by NUTS3

Statistic	Share of one-family households			Share of one-person households		
	DF	Value	Prob	DF	Value	Prob
Chi-Square	39	634.432	<0.0001	39	580.641	<0.0001
Likelihood Ratio Chi-Square	39	655.759	<0.0001	39	594.397	<0.0001
MH Chi-Square (Rank Scores)	1	17.470	<0.0001	1	322.076	<0.0001
Phi Coefficient		0.322			0.308	
Contingency Coefficient		0.306			0.294	
Cramer's V		0.186			0.178	
Gini index		0.050			0.113	

Source: Own calculation

Output in Table 5 displays the Chi-Square statistics *QP = 634.4328* with *39* df and *p<0.0001* for variable share of one-family households and *QP = 580.641* with *39* df and *p<0.0001* for variable share of one-person households. Both results are statistically significant on *0.05* level. Other statistics, such as the Mantel-Haenszel Chi-Square statistic with the result that is also statistically significant at *0.05* and that also shows a statistically significant dependence of both variables, was calculated too.

Interesting is the comparison with the result of the GINI index calculation, which is a measure of statistical dispersion and the most commonly used measurement of inequality. Multiple approaches can be used to estimate the Gini coefficient. One of the frequently used estimates is the so-called Somers' d statistics, but in this case the GINI index was calculated directly according to the procedure described in (Dixon, 1987). The results show that the value of the Gini index (especially in the case of share of one-family households) is approaching zero, therefore *there is no significant diversity in the data at NUTS3 level.*

Previous results of the Gini index confirm the association rates contained in Table 6. The entry was the share of one-family households resp. share of one-person households by NUTS3 (nominal variable) and the output by Lambda Statistics and Uncertainty Coefficients. Again, the results contained in Table 6 show that the association rates do not deviate significantly from zero and thus confirm the lack of diversity of the shares of both types of households at the NUTS3 level.

Table 6 Measures of association for share of households by NUTS3

Statistic	Share of one-family households		Share of one-person households	
	Value	ASE	Value	ASE
Lambda Asymmetric C|R	0.012	0.004	0.009	0.004
Lambda Asymmetric R|C	0.128	0.010	0.123	0.011
Lambda Symmetric	0.068	0.006	0.064	0.007
Uncertainty Coefficient C|R	0.022	0.002	0.020	0.002
Uncertainty Coefficient R|C	0.039	0.003	0.035	0.003
Uncertainty Coefficient Symmetric	0.028	0.002	0.025	0.002

Source: Own calculation

The relationship can be observed in the histogram of frequencies showing local and global rates of autocorrelation of one-family households or one-person households as computed by nearest neighbour method.

From mutual comparison of the global rate and local rates (see Figure 7) it is apparent in both cases (i.e. in the case of one-family households as well as in the case of one-person households) the global rate in the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

CONCLUSION – DISCUSSION

The aim of the paper was to study the issue of population trend, represented by households in a slightly different way, preferably by means of spatial autocorrelation. It is indisputable that the Czech Republic is a rather homogeneous territory and that changes resulting from the development of the population take place over time (population ageing, change in the structure of households). However, this does not imply that differences among individual parts of the country cannot be observed and that subsequent impacts of these changes cannot be investigated on the level of education or economic characteristics.

The highest number and highest share of both one-family households and one-person households are made up of households in municipalities, in which the value of the local index of spatial autocorrelation is not statistically significant. From map outputs it is, however, possible to conclude that for both variables there are territories in which the spatial autocorrelation is higher than in the remaining territory of the CR.

Frequencies of significant clusters are different for both types of households. In the case of one-person households it means that the High-Low types occurs in big cities, while households consisting of one family are more frequently cases of smaller municipalities and are more likely to occur in the western part of the country.

From mutual comparison of the global rate and local rates it is obvious in both cases (i.e. in the case of one-family households as well as one-person households) that the global rate on the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

Comparison with statistics calculated on NUTS3 level (again) show, that the distribution is neither identical nor complementary. Both types of households create separate spatial patterns and show a relatively large variability of the observed phenomenon. The value of the Gini index (especially in the case of share of one-family households) is approaching zero and therefore there is no significant diversity in the data at NUTS3 level.

The basic fact that has been ascertained is that shares of one-family households and one-person households are not a complementary phenomenon in the territory of the Czech Republic; i.e. in municipalities (that are defined as neighbouring), in which there is a higher share of one-person households there is, in general, a lower share of one-family households and vice-versa. In a given territory, other factors also exert an influence (e.g. size group of the municipality of the place of residence); they are modelling the situation and deserve further attention.

Comparing the agreement of both types of clusters, i.e. one-family households and one-person households by municipality, shows that the agreement or disagreement is not the same in all regions and that it differs even within regions.

Obviously, there is a problem in the interpretation of the results with respect to the above calculations. The Gini coefficient implies low variability between NUTS3 regions, which is in part contradictory to the results of the spatial autocorrelation. Does it therefore make sense to focus on regional differentiation of the share of different household types?

There could be two explanations. The first computes the so-called MAUP, which is a source of statistical bias that can significantly impact the results of statistical hypothesis tests (Openshaw, 2000). The results of the Gini coefficient, as well as the association rates were calculated at the NUTS3 level and only then interpreted, while the results of spatial autocorrelation were calculated directly at the municipal level.

The second explanation is essentially related to data. If we respect the spatial autocorrelation in the data, then it is necessary to choose such methods that allow spatial autocorrelation – *by definition*, such as Moran's Iindex. This explanation is more realistic according to the author of the paper.

Searching for causes of existence or non-existence of spatial autocorrelation on the level of municipalities is a different challenge; it deserves more intensive attention and an explanation based on age, education or economic variables may be present a first possible option.

ACKNOWLEDGMENT

This paper is supported by grant GAČR, 2018–2020 No. 18-12166S.

References

ANSELIN, L. *Local Spatial Autocorrelation Clusters.* The Center for Spatial Data Science, 2016.

ANSELIN, L. Local Indicators of Spatial Association – LISA. Geographical Analysis, 1995, 27, pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

ARCGIS PRO [online]. ESRI. [cit. 3.11.2019] <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm>.

BIVAND, R., PEBESMA, E., GÓMEZ-RUBIO, V. Applied Spatial Data Analysis with R. 2nd Ed. *Modelling Areal Data*, Springer, Science + Business Media, 2013.

BLEHA, B. AND ĎURČEK, P. An interpretation of the changes in demographic behaviour at a sub-national level using spatial measures in post-socialist countries: A case study of the Czech Republic and Slovakia. *Papers in Regional Science*, February 2019, Vol. 98, No. 1, pp. 331–352.

CZSO. *Atlas sčítání 2011* (in Czech). Prague: Czech Statistical Office, 2013.

CRESSIE, N. *Statistics for spatial data.* Rev. Ed. New York: Wiley, 1993.

DE SMITH, M. J., GOODCHILD, M., LONGLEY, P. A. *Geospatial Analysis.* 6th Ed. Global spatial autocorrelation, 2018.

DIXON P. et al. Bootstrapping the Gini Coefficient of Inequality. *Ecology*, Oct. 1987, Vol. 68, No. 5, pp. 1548–1551.

FISCHER, M. AND GETIS, A. eds. *Handbook of Applied Spatial Analysis, Software Tools, Methods and Applications.* Springer, 2010.

GRIFFITH, D. A., CHUN, Y., DEAN, D. J. *Advances in Geocomputation, Geocomputation 2015.* The 13th International Conference, Springer.

HOWELL, F. M. et al. Recapturing Space New Middle-Range Theory in Spatial Demography. *Demography Is an Inherently Spatial Science*, Spatial Demography, Springer, 2016.

CHING-LAN, CH., YI-CHI, CH., TZU-MING, L., YEA-HUEI, K. Y. *Using spatial analysis to demonstrate the heterogeneity of the cardiovascular drug-prescribing pattern in Taiwan.* BMC Public Health, 2011.

INSEE-EFGS-EUROSTAT. *Handbook of Spatial Analysis.* 2018.

KRIVORUCHKO, K. Spatial Statistical Data Analysis for GIS Users. *Statistical Approach to GIS Data Analysis, Principles of Modeling Regional Data.* ESRI Press, 2011.

NETRDOVÁ, P. AND BLAŽEK, J. Aktuální tendence lokální diferenciace vybraných socioekonomických subjektů v Česku: směřuje vývoj k větší mozaikovitosti prostorového uspořádání? (in Czech), *Geografie*, 2012, 3, pp. 266-288

NETRDOVÁ, P. AND NOSEK, V. Vývojové pravidelnosti a specifika geografické diferenciace obyvatelstva a jeho struktury na úrovni obcí v Česku v transformačním období (in Czech). *Geografie*, 2018, 123, 2, pp. 225–251.

OPENSHAW, S. AND ABRAHART, R. J. eds. *Geocomputation. The modifiable areal unit problem.* Taylor & Francis, 2000, pp. 36–38. ISBN 0-203-30580-9.

VAN IMHOFF, E. et al. *Household Demography. Theories of Household Formation: Progress and Challenges.* Springer, Science + Business Media, 1995.

VAN KAA, D. *Europe's Second Demographic Transition, Population Bulletin.* March 1987, Vol. 42, No. 1.

SAS. *Base SAS(R) 9.4 Procedures Guide: Statistical Procedures* [online]. 3rd Ed. [cit. 3.11.2019] <http://support.sas.com/documentation/cdl/en/procstat/67528/HTML/default/viewer.htm#procstat_univariate_details52.htm>.

SAS. *Stat SAS(R) 9.2 User's Guide* [online]. 2nd Ed. [cit. 3.11.2019] <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm>.

SHEKHAR, X. *Encyclopedia of GIS.* Springer Science + Business Media, 2008.

STOKES, E. et al. *Categorical Data Analysis.* 3rd Ed. SAS Institute Inc., 2012.

TOBLER, W. A Computer Movie Simulating Urban Growth in the Detroit Region [online]. *Economic Geography. Supplement: Proceedings,* International Geographical Union, Clark University, 1970, Vol. 46, pp. 234–240. [cit. 3.11.2019] <http://www.jstor.org/stable/143141>.

UNECE. *Measurement of emerging forms of families and households.* NY and Geneva, 2011.

ANNEX

Figure A1 Cluster and outlier analysis of share of one-family households by municipality in the Czech Republic

Note: See the online version of *Statistika journal* No. 4/2019.
Source: Own calculation

Figure A2 Cluster and outlier analysis of share of one-person households by municipality in the Czech Republic

Note: See the online version of *Statistika journal* No. 4/2019.
Source: Own calculation