Regional Disparities in Education Attainment Level in the European Union: A Spatial Approach

Michaela Chocholatá
Andrea Furková

Department of Operations Research and Econometrics,
Faculty of Economic Informatics,
University of Economics in Bratislava
Dolnozemská cesta 1,
Bratislava 852 35, Slovakia
E-mail: michaela.chocholata@euba.sk
E-mail: andrea.furkova@euba.sk

Abstract: *This article deals with the analysis of education attainment level across the 252 NUTS 2 regions of the European Union (EU) with consideration of the spatial aspect. Since the individual EU regions cannot be seen as isolated, the main aim of this article is to assess the impact of location on the education attainment level (percentage of population aged 25–64 with at least upper secondary education) during the period 2007–2015, as well as to investigate the impact of regional growth 2014/2007 on the education attainment level in 2015. The spatial analysis proved the existence of positive spatial autocorrelation and persistence of disparities in education attainment level across EU regions during the analysed period. The results of econometric analysis confirmed the expected positive impact of economic growth on education attainment level as well as the necessity to incorporate the spatial dimension into the model.*

Keywords: *education attainment level, EU regions, exploratory spatial data analysis, spatial econometrics*

doi: 10.1515/bjes-2017-0012 Baltic Journal of European Studies
Tallinn University of Technology (ISSN 2228-0588), Vol. 7, No. 2 (23)

1. Introduction

One of the priorities of the Europe 2020 Strategy is smart growth, which means developing an economy based on knowledge and innovation. The issue of education thus plays one of the key roles in the Europe 2020 Strategy in order to increase the share of people with higher education.

The five EU (European Union) targets for 2020 specified in this document (European Commission, 2010) concentrate on employment, research and innovation, climate change and energy, education, and combating poverty. These targets are specified in more detail in individual national strategies to reflect the current situation of each Member State with concentration on individual regions at different NUTS (Nomenclature of Units for Territorial Statistics) levels. However, it seems to be clear that all these targets are mutually interrelated. The level of attained education, for example, has an impact on the economic growth of individual regions, on people's employability, which in turn has a direct impact on the level of poverty and the economic well-being of individual people, regions and countries. The high level of economic growth, on the other hand, can act as a significant motivating factor for achieving of higher education.

With regard to regional modelling, the spatial aspect seems to be a significant factor of the analysis, since the regions of the individual EU Member States are largely interdependent and cannot be treated as isolated. However, in general, it is possible to identify quite huge regional disparities in individual EU regions regarding especially the socio-economic indicators, for example, GDP per capita, attained education level, migration and (un)employment rate. It is worth mentioning that in this context are very popular various analyses dealing with problems of convergence (see, e.g., Elias & Rey, 2011; Makrevska Disoska, 2016; Notermans, 2015; Paas & Schlitte, 2009).

Concerning the issue of education, Baumol *et al.* (1994), for example, stated that the levels of attained education are usually positively correlated with the economic growth of the analysed region. Tolley and Olson (1971) dealt with the interdependence between income and education expenditures and pointed out that the causal relationship between these two variables can be identified in both directions (see also, e.g., Elias & Rey, 2011).

Rodríguez-Pose and Tselios (2007), who analysed the European educational distribution in terms of educational attainment and inequality, stressed the problem of finding appropriate measurements of educational attainment across different regions and preferred a more complex view "focused on the educational attainment of individuals as measurement of human capital stock" (Rodríguez-Pose & Tselios, 2007, p. 5). However, in general, the level of attained education can be viewed as a very important measure and many research studies and papers dealing with this issue both across regions of developed and developing countries have been published (for a survey see, e.g., Rodríguez-Pose & Tselios, 2007; Elias & Rey, 2011; Sutton, 2012; Umar *et al.*, 2014; Chocholatá & Furková, 2016).

Although it is commonly known that there are certain patterns of geographical arrangement of highly developed and less developed regions in space, the spatial aspect of analysis has started to be taken into account relatively recently. Spatial analysis and spatial econometrics, reflecting the geographical location of the analysed region are the appropriate tools for such analysis. Although the term "spatial econometrics" was for the first time used in the 1970s and some of the techniques have been described already in the middle of the last century, the application of these techniques started at the turn of the new millennium (Arbia, 2006). Nowadays there exist various geographical information systems (GIS) and special software (GeoDa, R, SAS, MatLab, etc.) enabling to provide the spatial analysis and to estimate spatial econometric models.

Of the studies which have analysed the issue of attained education from the spatial econometric perspective, the following could be mentioned: Rodríguez-Pose and Tselios (2007), Elias and Rey (2011), Ahmed (2011), Wang (2012), Sutton (2012), Umar *et al.* (2014). While Rodríguez-Pose and Tselios (2007) studied the educational attainment and inequality within European regions, Sutton (2012) examined the spatial distribution of educational attainment in the US, and the remaining studies had to do with regions within the developing countries. Elias and Rey (2011) used both the exploratory spatial data analysis and the spatial econometrics in the analysis of educational convergence in Peru. They emphasised the use of social indicators since they "are a valuable complement to economic indicators when analysing spatial patterns in a given geographical region, and can often yield a more comprehensive view about regional socioeconomic behaviour" (Elias & Rey, 2011, p. 107). In her dissertation, Ahmed (2011) investigated the spatial aspects of income and education in Pakistan also based on spatial econometric techniques. Her analysis confirmed the importance of considering spatial effects in econometric modelling. Wang (2012) presented a spatial analysis of educational inequality in China's provinces. Umar *et al.* (2014) provided the spatial econometric analysis of educational distribution and regional income disparities in Nigeria. They came to the conclusion that "investing for equitable distribution of education will be a very effective policy strategy, both for improving regional economic performance and regional economic convergence in Nigeria" (Umar *et al*., 2014, p. 722).

The aim of this article is twofold: firstly it concentrates on the assessment of the persistence of regional disparities across EU regions in the area of education based on the spatial analysis of education attainment level (upper secondary, post-secondary non-tertiary and tertiary education of population aged 25–64, expressed in %) across 252 NUTS 2 regions of EU countries during the period 2007–2015, and, secondly, it investigates the impact of regional GDP growth 2014/2007 on the education attainment level in 2015 based on the estimation of corresponding non-spatial and spatial econometric models. The use of spatial analysis allows capturing and understanding the dynamics of development of the education attainment level across the EU regions during the analysed period 2007–2015 as well as assessing the persistence of regional disparities. The importance of spatial interaction and geographical proximity is also taken into account in the regression analysis testing the dependence between the attained education and regional economic growth and can be therefore considered as an interesting contribution to the discussion and to the empirical evidence in regional analyses of education attainment.

The rest of the article is organised as follows: Section 2 deals with the methodology and presents the analysed data, Section 3 illustrates the empirical results and Section 4 concludes the article.

2. Methodology and data

To assess the impact of location on the attained education level, the article begins with the ESDA (Exploratory Spatial Data Analysis) which allows exploring the structure of the analysed data and detecting the presence of spatial dependence and patterns of spatial clusters. In general, it is useful to visualise the data prior to the analysis. As pointed out, for instance by Mitchell (2013, p. 5), the construction of various graphs (e.g., line graph, bar graph, box plot) enables assessing some of the key characteristics of the data, however, "in spatial data, traditional plots might obscure some of the spatial dependencies and further insights can be gained through visualisation via maps". Mapping of the data enables getting a visual impression of them and detection of clusters of similar or dissimilar values (Rodríguez-Pose & Tselios, 2007). There are various possibilities of mapping the data—it can be done either according to the unique values or by dividing the data in categories (e.g., percentile maps, quantile maps, box maps). While on the basis of some maps it is possible to identify some spatial patterns and extreme values, or outliers, they provide no information about statistical significance of clustering. The next step of ESDA is thus usually based on spatial autocorrelation analysis both on the global and the local level. The global Moran's *I* statistic and the local Moran's *I* statistic are the main instruments for this part of our analysis. The confirmation of the spatial autocorrelation implies the presence of spatial spill-over effects, which means that the data from one region can influence the data from some other region.

While the global Moran's *I* statistic provides us a measurement of the global spatial autocorrelation—that is, how strong the spatial association is across neighbouring regions (a single value for the whole data set)—its local version enables assessing the spatial autocorrelation for one particular spatial unit (region). The LISA (Local Indicators of Spatial Association) presented by Anselin (1995) can be used to determine the existence of local spatial clusters. The formulas for calculation of both these statistics can be found, for example, in Getis (2010). However, it is useful to mention that for the specification of the spatial interactions across analysed regions it is necessary to construct an appropriate spatial weight matrix **W** of dimension ($n \times n$), where *n* is the number of regions in the data set. The specification of **W** belongs to one of the most problematic issues heavily discussed in the literature since it can significantly influence the results (Anselin, 1988; Rodríguez-Pose & Tselios, 2007). The simplest and most commonly used is the contiguity weight matrix **W**, but one can also encounter the *k*-nearest neighbours weights, distance-based weights, combination of contiguity and distance, etc.

In this article we will consider three different specifications of the weight matrix—the queen contiguity matrix of the first order,[1] the 4 nearest neighbours weight matrix and the threshold distance matrix based on Euclidean distance metric.[2] First of all, it is necessary to define which regions are neighbours—that is, to decide which elements of matrix **W** will be non-zero. In case of the queen contiguity matrix, the regions are neighbours if they share any part of a common border, in the 4 nearest neighbours weight matrix each region has exactly 4 neighbouring regions and in case of Euclidean distance matrix is the specification of neighbouring regions based on distances between regions (the

[1] In general, the contiguous neighbours can be defined, e.g., analogously to the game of chess—the rook case, the bishop case, or the queen case. The "first order" means that only direct interaction between geographically neighbouring regions is taken into account (Fischer *et al*., 2010). More information about the first-order, second-order and higher-order neighbouring regions can be found, e.g., in LeSage and Pace (2009).

[2] Hereafter denoted as Euclidean distance matrix.

threshold value is usually chosen in order to ensure that each region has at least one neighbour). The diagonal elements of the matrix **W** are conventionally set to zero. To the creation of matrix **W** the appropriate software for spatial analysis can be used, such as, for example, free downloadable software GeoDa.[3] This software enables also visualisation of the spatial pattern and spatial clustering based on the Moran scatterplot capturing both the global and the local measures. In the Moran scatterplot, which is divided into four quadrants, it is possible to identify four different spatial associations: high-high (HH), low-high (LH), low-low (LL) and high-low (HL). The associations HH and LL indicate a positive spatial autocorrelation, while the associations LH and HL a negative autocorrelation. Confirmation of the positive spatial autocorrelation embodies the Tobler's First Law of Geography that near regions are more related than distant ones, that is, the regions with high (low) values tend to be located nearby other regions with high (low) values, whereas the negative spatial autocorrelation means the presence of spatial outliers, that is, regions with very different values from their neighbours (Elias & Rey, 2011; Rodríguez-Pose & Tselios, 2007). Further information is available from the LISA cluster maps which show only regions with statistically significant local Moran's *I* statistic, colour coded by type of spatial autocorrelation. In the software GeoDa the LISA maps share a common legend, which is as follows:

□ Not significant

🟥 High-High

🟦 Low-Low

🟪 Low-High

🟧 High-Low

The four categories (HH, LL, LH and HL) of the spatial association correspond to the four quadrants in the Moran scatterplot.[4]

Next we will concentrate on regression (econometric) analysis—estimation of the regression model reflecting the dependence between the attained education and regional economic growth. In general, there are two approaches to estimation: specific-to-general approach, also known as classical approach, and

[3] However, it is worth mentioning that the weight matrix is usually used in its row-standardised form.

[4] Software GeoDa enables choosing the significance level (p-value) of 0.0001–0.05 and randomisation approach involving 99–99999 permutations in order to test the robustness of the results.

general-to-specific approach, or Hendry's approach (for more information on both approaches in context of spatial econometrics see, e.g., Florax *et al*., 2003; Mur & Angulo, 2005; Mitchell, 2013). Below we will follow the specific-to-general approach.[5] As the first step it follows the estimation of the classic linear regression model by the ordinary least squares (OLS) method, i.e.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \{1\}$$

where $\mathbf{y}$ is a $(n \times 1)$ dimensional vector of a dependent variable, $\mathbf{X}$ is a matrix of independent variables of dimension ($n \times (k+1)$) and k is a number of independent variables, $\boldsymbol{\beta}$ is $((k+1) \times 1)$ dimensional vector of unknown parameters and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of independent identically distributed (i.i.d.) error terms. Since the presence of spatial effects can have significant implications on the quality of estimates, it proceeds to testing the presence of spatial dependence in order to choose the appropriate form of spatial model. In general, we can distinguish between two types of spatial dependence—spatial lag and spatial error, which also corresponds to the two forms of spatial models—SAR or Spatial Autoregressive Model and SEM or Spatial Error Model (see, e.g., Anselin, 1988). The analysis has usually been applied to the cross-sectional data or panel data. Since the cross-sectional data analysis deals with the data for individual regions in one defined time, the panel data analysis takes into account also the development of the cross-sectional data in time. However, in this article we will deal only with the cross-sectional data.

As mentioned by Anselin (1988), the use of the SAR model is appropriate in case the focus is on the assessment of the presence and strength of spatial interaction, and the SEM model in case of spatial dependence in the regression disturbance term. The SAR model can be formulated as follows:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad \{2\}$$

where ρ is the scalar spatial autoregressive parameter measuring the degree of dependence, $\mathbf{W}$ is a spatial weight matrix of dimension ($n \times n$), $\mathbf{Wy}$ is a $(n \times 1)$ dimensional vector of spatially lagged dependent variable, $\mathbf{u}$ is a $(n \times 1)$ dimensional vector of error terms and all other terms were previously defined above. Since the value $\rho \neq 0$ implies the existence of spatial effects across neighbouring regions (endogenous interaction effects[6]), a zero value indicates no spatial dependence between observations of the considered dependent variable.

[5] Support for the classical specific-to-general approach can be found, e.g., in Florax *et al*. (2003) who demonstrate that the classical approach outperforms the Hendry's strategy and provides for better inference.

[6] For more information and further explanation see, e.g., Anselin, 2003.

The SEM model is expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} = \lambda\mathbf{W}\boldsymbol{\varepsilon} + \mathbf{u} \qquad \{3\}$$

where λ is a spatial error parameter reflecting the intensity of spatial autocorrelation between regression residuals and $\mathbf{W}\boldsymbol{\varepsilon}$ is a $(n \times 1)$ dimensional vector of spatially lagged error terms. Both the SAR and the SEM model can be estimated by maximum likelihood (ML) method.

The appropriate form of spatial model, that is, SAR or SEM, can be chosen based on the Lagrange Multiplier (LM) test results or their robust modifications (Anselin & Florax, 1995). Both of these tests use the residuals from the OLS model in order to test the null hypothesis of no spatial dependence against the alternative hypothesis of spatial dependence. These tests are asymptotic and follow a $\chi2$ distribution with one degree of freedom.[7] In this context it is worth mentioning that some authors (e.g., LeSage & Fischer, 2008) criticise the choice of a spatial model specification based only on these diagnostic tests and highlight the use of the spatial mixed model (the SAR model augmented with the spatial lags of the explanatory variables, also known as the spatial Durbin model) especially in case of omitted variables (see also Ivanova, 2015; LeSage & Pace, 2009). The formulation of the spatial Durbin model (SDM) including the spatial lags of both dependent and explanatory variables is as follows:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\Theta} + \mathbf{u} \qquad \{4\}$$

where $\mathbf{WX}$ denotes the $(n \times k)$ dimensional matrix of spatially lagged explanatory variables[8], $\boldsymbol{\Theta}$ is a $(k \times 1)$ dimensional vector of parameters reflecting the exogenous interaction effects (Anselin, 2003) and all other symbols were previously explained above. Similarly, as SAR and SEM models, also SDM model can be estimated based on ML method. In general the SDM model plays an important role in the spatial econometric literature (see, e.g., LeSage & Pace, 2009; Mur & Angulo, 2005; Fischer *et al*., 2010), since it is possible to derive from it a number of other models as special cases, for instance, in case we cannot reject the null hypothesis $\boldsymbol{\Theta} = \mathbf{0}$, the SDM becomes a SAR model; if the null hypothesis $\boldsymbol{\Theta} = -\rho\boldsymbol{\beta}$ cannot be rejected, we will receive a SEM model from

[7] However, it can happen that the LM-tests do not indicate a clear-cut conclusion which of the two models is more appropriate (Paas & Schlitte, 2009). Some researches (see, e.g., Bivand, 2010) decide about the appropriate form of spatial model also based on lower value of the Akaike information criterion (AIC) or use the AIC, the Schwarz criterion (SC) and log likelihood in model selection (see, e.g., Sutton, 2012).

[8] Matrix X is here of dimension (n x k), i.e. without the first column of ones.

the SDM model and imposing the restrictions both $\rho = 0$ and $\boldsymbol{\Theta} = \mathbf{0}$ yields the classic linear regression model.

The analysis in this paper was done based on data retrieved from the Eurostat database General and Regional Statistics (Eurostat, n.d., b). The main focus was on the attained education characterised by the population aged 25–64 (in %) with upper secondary, post-secondary non-tertiary and tertiary education attainment[9] for 252 NUTS 2 regions of the EU countries (from the original data set of 272 regions were excluded 20 isolated regions of Cyprus, France, Finland, Greece, Italy, Malta, Portugal and Spain) during the period 2007–2015. In order to investigate the impact of region's growth on the attained education in 2015, the GDP per capita (defined at current market prices in Purchasing Power Standard, or PPS) growth rates from 2007 to 2014 were calculated (and expressed in natural logarithms) based on the data retrieved from the source mentioned above. The spatial analysis was carried out using the free downloadable software GeoDa[10] (n.d.). The shapefile (.shp) for the European regions was downloaded from the web page of Eurostat (n.d., a).

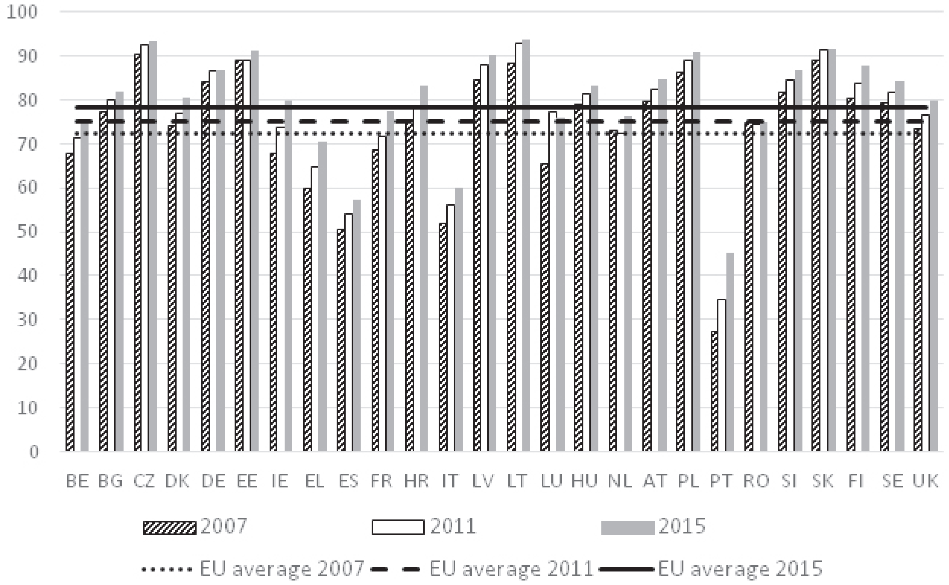
Although the whole analysis is based on data for 252 NUTS 2 EU regions, in order to evaluate country disparities in education attainment, Figure 1 depicts the country averages for attained education in 2007, 2011 and 2015. The horizontal straight lines illustrate the EU average values[11] of 72.40%, 75.21% and 78.25%, respectively. It is clearly visible that the share of people with at least upper secondary education show the rising tendency in almost all analysed countries. The lowest average values were identified for regions of Portugal (PT), Spain (ES), Italy (IT) and Greece (EL). However, in this context it is necessary to point out the fact that especially in Portugal the average values rose quite quickly—from 27.3% in 2007 to 45.1% in 2015. The highest average values (around 90%), on the other hand, are shown in Czech Republic (CZ), Lithuania (LT), Slovakia (SK), Estonia (EE), Poland (PL) and Latvia (LV).

[9] International Standard Classification of Education ISCED 2011: levels 3–8. For more information see, e.g., UNESCO, 2012.

[10] Box plots were graphically depicted using the software EViews.

[11] The average values were calculated on the basis of the 252 NUTS 2 regions values in individual years.

Figure 1. Country averages in upper secondary, post-secondary non-tertiary and tertiary education attainment in 2007, 2011 and 2015, %

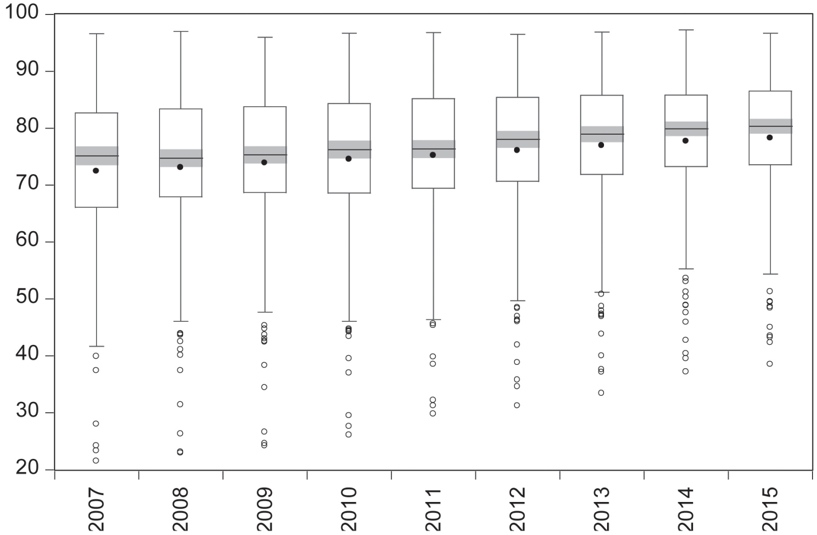


Source: Authors' illustration

Figure 2 illustrates the box plot for the education attainment during the period 2007–2015. While the median values during the first three analysed years remained almost constant, in the following years they rose slowly (from 76.4 in 2010 to 80.5 in 2015). The mean values have had a gradually rising tendency from 72.4 in 2007 to 78.25 in 2015. The interquartile range representing the middle 50% of the data, that is, the difference between the first and third quartiles, did not show a clear down- or upward tendency. It slightly dropped from 16.6 in 2007 to 15.1 in 2009, followed by a slow upward movement in the next two years (from 15.7 in 2010 to 15.75 in 2011). During the next three years it declined to the value of 12.55 in 2014 and finally it rose to 12.95 in 2015. Whiskers and staples show the values that are within the inner fences[12]—the whiskers in the lower part are much longer than in the upper part. Furthermore, it is necessary to mention that there are several lower outliers in all analysed years and no upper outliers.

[12] The first and the third quartile can also be termed "hinges". The inner fences are defined as hinges +/- 1.5*interquartile range. Values outside the inner fences are denoted as outliers (see, e.g., EViews, 2014).

Figure 2. Box plot for education attainment in 2007–2015, %



Source: Authors' illustration

3. Empirical results

After the brief characteristics of the data presented in the previous section we start the empirical part of our analysis with mapping of the data via box maps for 2007 and 2015 (see Fig. 3) in order to illustrate the unequal distribution of attained education over space, that is, also the existence of some differences across regions inside the analysed countries.[13] Box map is a special form of a quartile map and consists of six categories. Besides the four categories corresponding to the four quartiles, two extra categories are specified for upper and lower outliers,[14] respectively. Therefore, as Figure 3 and its legend reveal, the first and last quartile no longer correspond to exactly one fourth of the observations, since the lower and upper outliers, respectively, are depicted as extra categories (Anselin *et al*., 2010). Six lower outliers were detected in 2007 (all the analysed Portuguese regions and the Spanish region Extremadura) and

[13] The regional disparities in education attainment are in general considered to be higher across regions from different countries than across regions within countries, since the education systems are in general determined at the national level (see, e.g., Rodríguez-Pose & Tselios, 2007).

[14] Outliers are defined in the same way as in the box plot.

10 lower outliers in 2015 (4 out of 5 analysed Portuguese regions[15]; 4 Spanish regions: Castilla-la Mancha, Extremadura, Andalucía and Región de Murcia; 2 Italian regions: Campania and Puglia), respectively. Since no upper outliers were detected, the regions with the highest values (around 96%) were Chemnitz and Dresden in Germany and the Praha region in the Czech Republic.

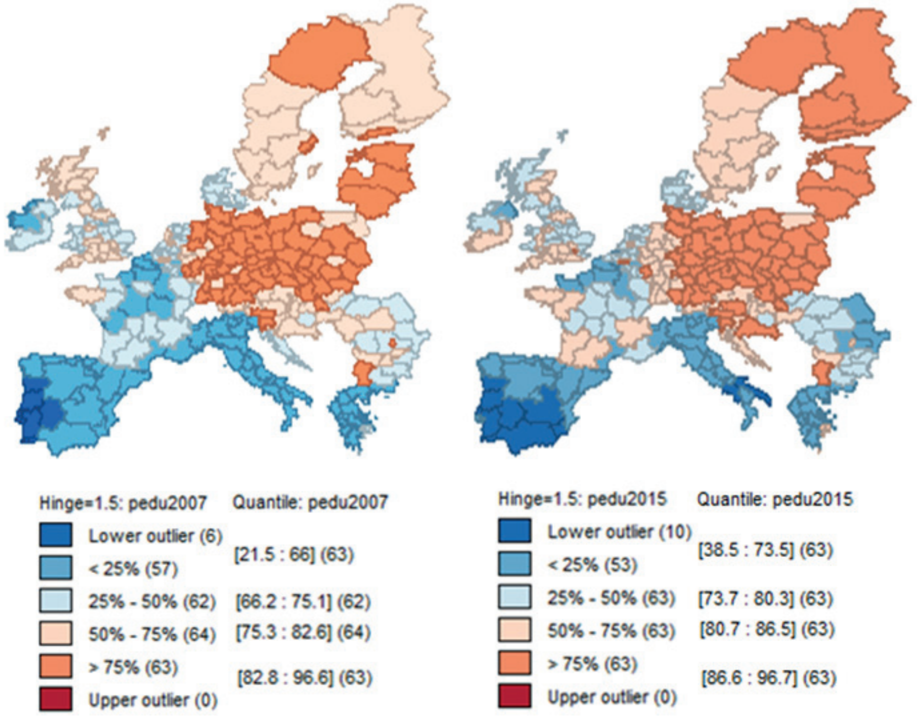
Figure 3 reveals that, in general, the regions with higher (lower) percentage of population with considered education attainment level tend to be gathered together, but on the other hand, it is also possible to identify some regions with a higher percentage of "educated" population in comparison to the neighbouring regions (in 2007, e.g., Bucuresti-Ilfov in Romania, Helsinki-Uusimaa in Finland, Stockholm region in Sweden, Yugozapaden in Bulgaria, or Közép-Magyarország in Hungary and in 2015, e.g., Trier in Germany, Prov. Brabant Wallon in Belgium, Kontinentalna Hrvatska in Croatia, Outer London in the UK, Yugozapaden in Bulgaria or Közép-Magyarország in Hungary). This is mostly the case of capital city regions or regions with universities and/or high concentration of R&D (Research and Development) activities, since these regions attract people with higher qualification and enable them better opportunities for further education, better career prospects and higher salaries.

Although the visualisation using box maps enables analysing whether education attainment is randomly distributed across the analysed regions or whether there exists spatial clustering of regions with similar percentage of people with at least upper secondary education, this approach does not provide any information about the statistical significance or insignificance of the clustering (Rodríguez-Pose & Tselios, 2007; Mitchell, 2013).

In order to confirm that location affects the share of people with analysed education attainment level, that is, to test for spatial autocorrelation, it is necessary to define spatial neighbourhoods and spatial weights. The spatial weight matrix **W** was specified in three different ways in order to show the sensitivity of its specification to the results. Firstly, a contiguity weight matrix of queen case definition of neighbours was specified (two regions are considered as neighbours if they share any part of a common border); secondly, we used the weight matrix of 4 nearest neighbours; and, finally, **W** was based on Euclidean distance metric. Since in case of queen contiguity weight matrix the regions had 1 to 11 neighbours (with the highest frequency of 5 neighbours), based on Euclidean distance weights the number of neighbours was substantially higher spanning from 1 to 70 neighbours (with the highest frequency of 56 neighbours).

[15] The exception was the region of the capital city of Portugal—Área Metropolitana de Lisboa.

Figure 3. Box maps for considered education attainment level in 2007 and 2015

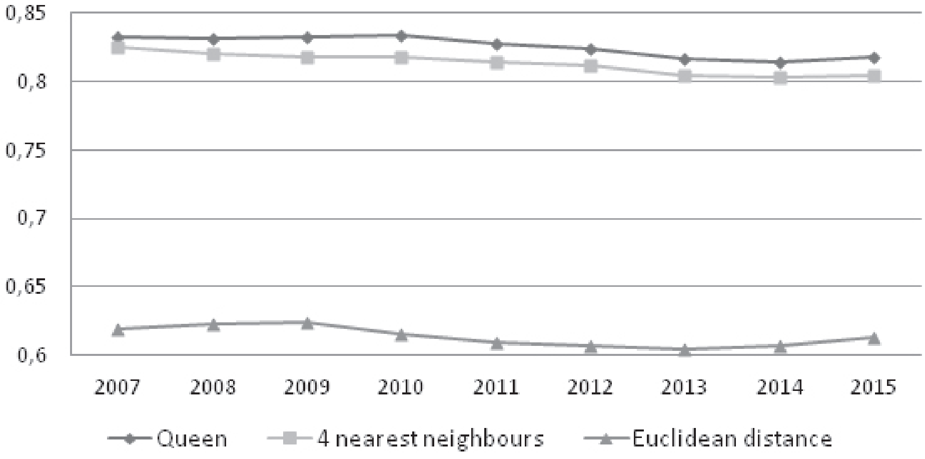


Source: Authors' illustration

Concerning the different specifications of the **W** matrix, the global Moran's *I* statistics were calculated. The dynamics of Moran's *I* during the period 2007–2015 for the three above-mentioned specifications of weight matrices is captured in Figure 4.

Since the global Moran's *I* values based on queen case definition and 4 nearest neighbours definition of weight matrix were similar—varying between 0.8 and 0.85, the global Moran's *I* values based on Euclidean distance weight matrix were substantially lower, between 0.6 and 0.65. In this context, it is important to mention that these weight matrices are used in a row-standardised form—that is, the elements of each row sum to one and each neighbour of the concrete region is given equal weight. As pointed out, for example, by Mitchell (2013, p. 16) "the row-standardised weights increase the influence of likely spill-overs where a region has few neighbours relative to regions where spill-overs will occur between many neighbours". As mentioned above, the number of neighbours was extremely different in weight matrices used and can therefore serve as one of the possible explanations of the different global Moran's *I*

Figure 4. Dynamics of global Moran's I during 2007–2015 for different specification of weight matrices



Source: Authors' illustration

values. The values of global Moran's *I* statistics are larger than the expected value $E(I) = -1/(n-1) = -0.00398$ which indicates the positive spatial autocorrelation, which means that it is much more likely that regions with high (low) percentage of population with considered education attainment level will have neighbours with also high (low) share of population with at least upper secondary education than in case of pure randomness.

On the one hand, the global Moran's *I* only provides a measurement of global spatial association and ignores the region-specific details. On the other hand, the LISA indicators allow for the decomposition of global statistics and enable identification of the contribution of each individual observation (Anselin, 1995; Ahmed, 2011). The spatial pattern and spatial clustering based on LISA cluster maps concerning the different weights (queen, 4 nearest neighbours, Euclidean distance) both for 2007 and 2015 are visualised in Figure 5.

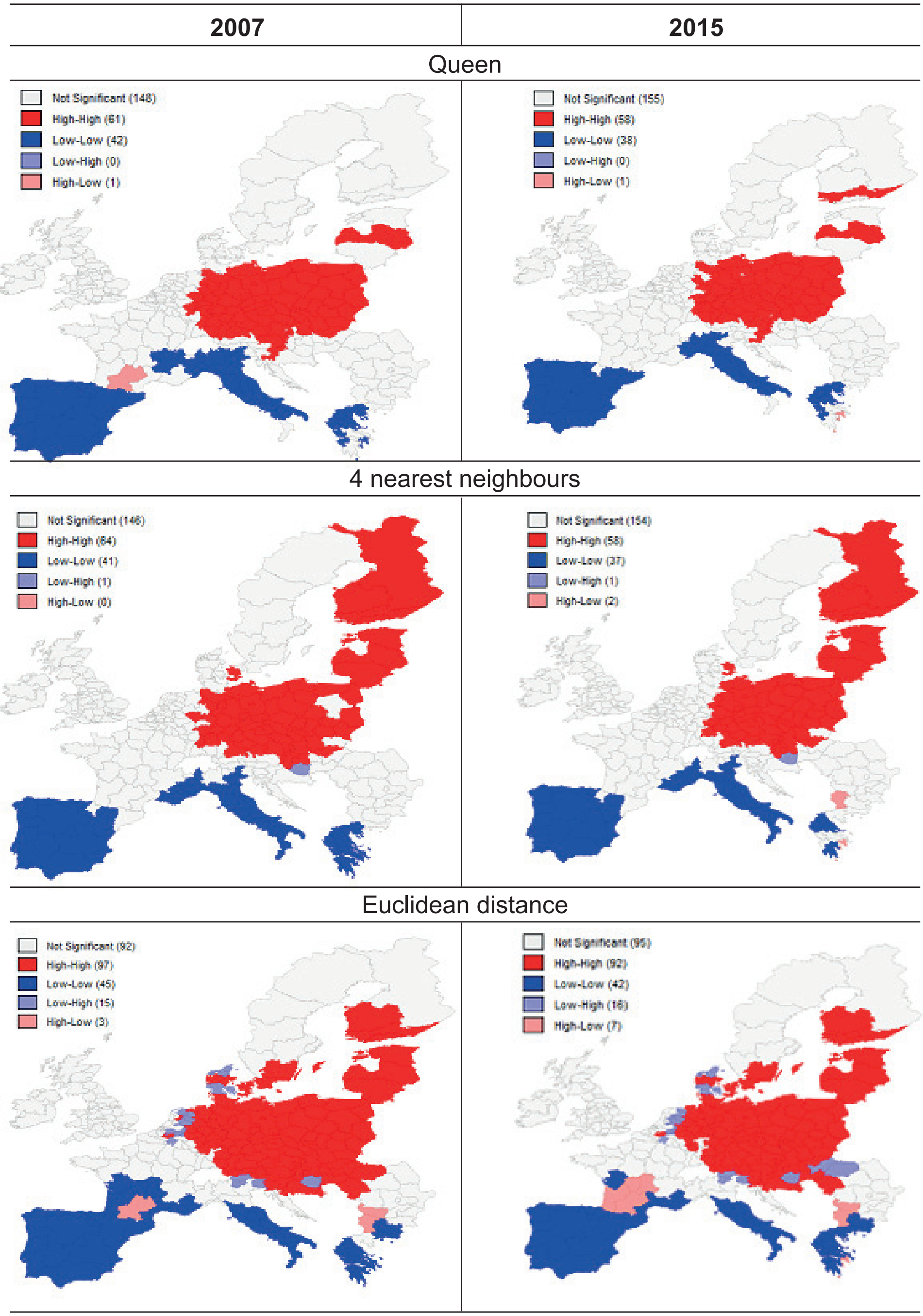
According to Figure 5, there are many regions representing positive spatial autocorrelation and only few regions with significant negative spatial autocorrelation. However, the results for individual years differ due to the used weight matrix. In 2007, based on queen contiguity weight matrix and 4 nearest neighbours, 103 and 105 regions with statistically significant positive spatial autocorrelation were identified, respectively. On the basis of Euclidean distance weights, more expanded clusters are depicted—significant positive spatial autocorrelation was proved for 142 regions. The results for 2015 were only slightly different with 96 and 95 statistically significant regions with positive autocorrelation based on queen and 4 nearest neighbours weight matrix, respectively. The Euclidean distance weight matrix identified 134 regions to be statistically significant with positive autocorrelation. The significant positive autocorrelation of HH was detected for regions located mainly in the central and eastern EU countries and some regions of northern Europe, while the significant positive autocorrelation of LL was identified for regions situated mainly in Portugal, Spain, Italy, Greece and France. The location of the regions did not change significantly in comparison with the year 2007. The same is true for the spatial outliers—that is, regions with statistically significant negative spatial autocorrelation of LH and HL, respectively. In case of queen contiguity weight matrix only one region of this type (HL) was identified both in 2007 (Midi-Pyrénées in France) and 2015 (Attiki in Greece). Using the 4 nearest neighbours specification of weight matrix, we specified one LH region (Dél-Dunántúl in Hungary) both in 2007 and 2015 and two HL regions (Attiki in Greece, Yugozapaden in Bulgaria) in 2015. The highest number of regions with statistically significant negative spatial autocorrelation was specified for the Euclidean distance weight matrix with 18 regions in 2007 and 23 regions in 2015 belonging to the different EU countries (Fig. 5).

The results of the spatial analysis revealed the persistence of disparities in considered education attainment level across EU regions during the analysed period (similarly to Rodríguez-Pose & Tselios, 2007). Moreover, as we can see in Figure 3, in 2015 even a greater number of lower outliers were identified. The statistical significance of regions with HH and LL spatial associations did not change considerably during the analysed period (Fig. 5), it can be therefore a useful information and a great challenge for the EU authorities to distribute the limited resources more effectively to create more attractive job opportunities and thus to encourage the population to achieve higher education, better career prospects and higher standard of life.

Furthermore, we will present the results of regression (econometric) analysis—estimation of the regression model reflecting the dependence between the attained education and regional economic growth. The estimation was done on the basis of "classical approach", that is, beginning with the OLS estimation which is (with regard to the diagnostic check) followed by the estimation of spatial models. The results are presented separately for queen case weights (Table 1) and Euclidean distance weights (Table 2).[16]

[16] The 4 nearest neighbours weight matrix was not used for estimation due to non-symmetric weights.

Figure 5. LISA Cluster Maps for the attained education in 2007 and 2015, %



Source: Authors' illustration

In the first step the classic linear regression model (1) was estimated based on OLS. Estimation results are gathered in Tables 1 and 2 (column: Linear model). Both parameters β_0 and β_1 were statistically significant, the expected positive relationship between the attained education and regional economic growth was confirmed. However, the R^2 was very low and the diagnostic statistics—Moran's *I* applied on regression residuals and the Lagrange Multiplier tests—indicated that we can clearly reject the null hypothesis of non-spatial dependence (see Tables 1 & 2), which means that the spatial aspect should be taken into account. Since both tests, and their robust versions, are highly significant, it is difficult to provide a clear-cut conclusion about which of the two models, either SAR (2) or SEM (3), is more suitable. We therefore decided for estimation of both of them.

The estimation results by ML for the SAR and the SEM model are given in Tables 1 and 2 (columns: SAR model and SEM model). All the estimated parameters were statistically significant, the positive value of regression parameter β_1 confirms the positive impact of growth on the attained education. The statistical significance of spatial parameters ρ and λ confirms the strong positive and significant spatial autocorrelation, that is, presence of spatial effects across neighbouring regions. The higher percentage of "educated" population in a specific region will tend to push up the rate in the neighbouring regions. The statistical adequacy of the SAR and SEM models was proved by the low values of Moran's *I* statistic applied on corresponding spatial residuals indicating no further evidence of spatial autocorrelation. The inclusion of the spatial component thus seems to be inevitable in order to overcome the problem of possibly biased results and hence misleading conclusions from estimation of non-spatial models.

Also, the SDM model was estimated (see columns: SDM model). As we can see both in Table 1 and Table 2, its parameters β_0, β_1 and ρ were statistically significant at 1% significance level. A different situation occurs in the case of parameter θ which was in the queen case definition of the weight matrix (Table 1) statistically significant at 5% significance level, but under Euclidean weights (Table 2) no statistical significance was proved. The negative sign of this parameter indicates that higher growth in neighbouring regions connected *inter alia* with better career opportunities and better economic well-being will attract certain share of "better educated" population to move to such regions and thus the percentage of population with considered education attainment level in analysed region will go down.

Table 1. Estimation results of linear regression model, SAR model, SEM model and SDM model, weights: queen

	Linear model	SAR model	SEM model	SDM model
Estimation	OLS	ML	ML	ML
β_0	74.962***	16.2201***	75.8947***	15.0507***
β_1	51.5028***	15.6248***	30.3175***	26.1037***
λ	-	-	0.8112***	-
ρ	-	0.7789***	-	0.7966***
θ	-	-	-	-13.8315**
R^2	0.2866	0.7880	0.7932	0.7942
AIC	1863.63	1613.05	1611.65	1611.27
Diagnostic tests				
Moran's I (error)	16.5693***	-	-	-
LM (lag)	261.0349***	-	-	-
Robust LM (lag)	6.8265***	-	-	-
LM (error)	262.7246***	-	-	-
Robust LM (error)	8.5162***	-	-	-
Moran's I (spatial residual)	-	-0.0594	-0.0851	-0.0788

Note: Symbols *** and ** indicate statistical significance at 1% and 5% level of significance, respectively.

Source: Authors' calculations

Table 2. Estimation results of linear regression model, SAR model, SEM model and SDM model, weights: Euclidean distance

	Linear model	**SAR model**	**SEM model**	**SDM model**
Estimation	**OLS**	**ML**	**ML**	**ML**
β_0	74.962***	5.6813**	68.5007***	5.0002**
β_1	51.5028***	16.481***	28.4534***	23.8815***
λ	-	-	0.9480***	-
ρ	-	0.9083***	-	0.9187***
θ	-	-	-	-9.5553
R^2	0.2866	0.7250	0.7329	0.7275
AIC	1863.63	1644.48	1639.14	1645.07
Diagnostic tests				
Moran's I (error)	30.7675***	-	-	-
LM (lag)	608.8735***	-	-	-
Robust LM (lag)	15.4382***	-	-	-
LM (error)	803.8638***	-	-	-
Robust LM (error)	210.4285***	-	-	-
Moran's I (spatial residual)	-	0.0467	0.0287	0.0408

Note: Symbols *** and ** indicate statistical significance at 1% and 5% level of significance, respectively.

Source: Authors' calculations

Regarding the statistical significance of parameters and other model characteristics (especially AIC values) we will prefer the SDM model for queen case weights and SEM model for Euclidean distance weights. We can also conclude that although the results are sensitive to the choice of the weight matrix, it was proved that in both cases the consideration of the space in econometric modelling is an inevitable part of the estimation procedure in order to receive econometrically correct results.

4. Conclusion

In this paper we carried out the spatial analysis of education attainment level (percentage of population aged 25–64 with at least upper secondary education) across 252 NUTS 2 regions of EU countries during the period 2007–2015 in order to assess the impact of location on the percentage of population with at least upper secondary education as well as to investigate the impact of regional growth on the share of people with specified education attainment level.

The main contributions of this article can be summarised as follows. Firstly, the studies dealing with the analysis of educational attainment across EU regions are quite scarce. Secondly, the complexity of the paper, since besides the spatial analysis of the attained education also the dependence between the attained education and regional economic growth has been investigated based on the estimation of both non-spatial and spatial econometric models.

The ESDA tools based on the graphic visualisation and mapping of the data accompanied by the calculation of the global and local Moran's *I* statistics enabled us to assess the existence of clusters and the presence/significance of spatial dependence. The article presents *inter alia* the box plot for considered education attainment level in individual years (Fig. 2) indicating slightly rising mean value of percentage of population with at least upper secondary education and simultaneously the persistence of disparities during the period under consideration. Although visualisation via box maps (Fig. 3) for the first and the last analysed year enabled us to identify the clusters of regions with similar values, the information about statistical significance or insignificance of the clusters was provided by the calculation of the global and local Moran's *I* statistics. The neighbourhood of regions was characterised by three different types of spatial weight matrices—the queen contiguity matrix of the first order, the 4 nearest neighbours weight matrix, and the Euclidean distance weight matrix. The results proved strong positive spatial autocorrelation in the case of all three specifications of the weight matrix (Fig. 4), that is, the regions with a high (low) percentage of population with upper secondary education or higher tend to be located nearby to and clustered with other regions that have also high (low) shares of population with considered education attainment level (similar results were presented also, e.g., by Sutton, 2012; Rodríguez-Pose & Tselios, 2007). Based on the local Moran's *I* statistics and LISA cluster maps, the significant positive autocorrelation of HH was detected for regions located mainly in the central and eastern EU countries and some regions of northern Europe, while the significant positive autocorrelation of LL was identified for regions situated mainly in Portugal, Spain, Italy, Greece and France. It was also proved that the regions with similar percentage of population with considered education attainment level tend to be more spatially clustered than could be expected from pure chance. Furthermore, it should be pointed out that the results did not change significantly during the considered period (Fig. 5); the huge disparities across regions thus have a tendency to persist over time (which is in line with the results presented by Rodríguez-Pose and Tselios, 2007). An analysis of this type can serve as a useful tool both for the EU authorities and the authorities of individual countries to make proper decisions (in accordance with the Europe 2020 Strategy) in order to motivate people to achieve higher education attainment levels by supporting of the appropriate regions—not only the less developed regions, but also regions which can serve through the spill-over effects to the rising share of "better educated" population in nearby regions.

In the second part of the paper, the regression model reflecting the dependence between the attained education and regional economic growth was estimated using both non-spatial and spatial approaches. Since the residuals from the classic (non-spatial) regression model were spatially autocorrelated, after diagnostic checking the spatial models (SAR, SEM and SDM) based on different weight matrices (queen, Euclidean distance) were estimated (Tables 1 & 2). With regard to statistical significance of parameters as well as AIC values, we preferred the SDM specification for queen case weights and SEM specification for Euclidean distance weights. The positive relationship between education attainment level and growth rate was clearly confirmed. Although the results are sensitive to the choice of the weight matrix, in both cases statistical significance of spatial parameter confirms the presence of spatial effects across neighbouring regions indicating that the higher percentage of "educated" population in a concrete region will tend to push up the rate in neighbouring regions. The SDM specification furthermore pointed out the fact of possible outflow of some people with considered education attainment level to the neighbouring regions with higher rates of economic growth.

The results presented in this paper clearly proved that the impact of location on the share of population with at least upper secondary education does matter and therefore the regression analysis should be enriched by inclusion of the spatial dimension in order to avoid possibly biased results and, hence, misleading conclusions.

The inclusion of further factors into analysis reflecting, for example, the migration of workforce with considered education attainment level, funds invested into R&D, foreign direct investments, unemployment rate and the poverty level could serve as interesting challenges for further research.

Acknowledgements

This work was supported by the Grant Agency of Slovak Republic VEGA Grant No. 1/0248/17 "Analysis of Regional Disparities in the EU based on Spatial Econometric Approaches".

Michaela Chocholatá works as associate professor at the Department of Operations Research and Econometrics at the Faculty of Economic Informatics, University of Economics in Bratislava. She received her PhD degree (in Econometrics and Operations Research) in 2006, and after successfully completed habilitation in 2014 she became associate professor (docent, in Slovakia). In her research activities she deals mainly with modelling of financial time series, analysis of stock returns, exchange rates as well as with issues concerning spatial econometrics. She has published her scientific work in various journals, conference proceedings, and has been a member of various research teams. She took part in Erasmus teaching mobility at the University of Akureyri, Iceland, and ISCTE Business School Lisbon, Portugal.

Andrea Furková works as assistant professor at the Department of Operations Research and Econometrics at the Faculty of Economic Informatics, University of Economics in Bratislava since 2000. She obtained her PhD degree (in Econometrics and Operations Research) in 2007. Her main research activities include multicriteria decision-making methods, exploratory spatial data analysis and spatial econometrics. Her recent research covers exploratory spatial data analysis and spatial econometrics within the area of regional income convergence modelling and innovative activity in the European Union regions. Her bibliography includes scientific works in international scientific journals, publications in international scientific conferences and textbooks. Her interests include, e.g., microeconomics, multicriteria decision-making models and methods. She took part in Erasmus teaching mobility at University of Akureyri, Iceland, and ISCTE Business School Lisbon, Portugal.

References

Ahmed, S. (2011), *Essays on Spatial Inequalities in Income and Education: Econometric Evidence from Pakistan*, PhD dissertation, University of Trento. Retrieved from http://eprints-phd.biblio.unitn.it/621/1/PHD_SOFIA_AHMED__oct_2611__.pdf [accessed 15 Feb 2015]

Anselin, L. (1995), 'Local indicators of spatial association – LISA,' *Geographical Analysis*, vol. 27, no. 2, pp. 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

—— (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer. https://doi.org/10.1007/978-94-015-7799-1

—— (2003), 'Spatial externalities, spatial multipliers and spatial econometrics,' *International Regional Science Review*, vol. 26, pp. 153–166. https://doi.org/10.1177/0160017602250972

Anselin, L. & Florax, R. (1995), *New Directions in Spatial Econometrics*, Berlin, Heidelberg & New York: Springer. https://doi.org/10.1007/978-3-642-79877-1

Anselin, L.; Kim, Y. W. & Syabri, I. (2010), 'Web-based analytical tools for the exploration of spatial data,' in M. M. Fischer & A. Getis (eds.) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Berlin & Heidelberg: Springer-Verlag, pp. 151–173. https://doi.org/10.1007/978-3-642-03647-7_10

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

Baumol, W. J.; Nelson, R. R. & Wolff, E. N. (1994), *Convergence of Productivity: Cross-National Studies and Historical Evidence*, New York: Oxford University Press.

Bivand, R. S. (2010), 'Spatial econometric functions in R,' in M. M. Fischer & A. Getis (eds.) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Berlin & Heidelberg: Springer-Verlag, pp. 53–71. https://doi.org/10.1007/978-3-642-03647-7_4

Chocholatá, M. & Furková, A. (2016), 'Spatial econometric analysis of attained education across NUTS 2 regions of European Union,' in *Proceedings of the 8th International Conference 'Economic Challenges in Enlarged Europe'*, Tallinn: Tallinn University of Technology, pp. 1–9.

Elias, M. & Rey, S. J. (2011), 'Educational performance and spatial convergence in Peru,' *Région et Dévelopment*, no. 33, pp. 107–135.

European Commission (2010), Communication from the Commission Europe 2020: A strategy for smart, sustainable and inclusive growth, COM (2010) 2020 final. 3.3.2010. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF [accessed 10 Feb 2016]

Eurostat (n.d., a) Administrative units, statistical units, Eurostat. Retrieved from http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units [accessed 15 Feb 2015]

—— (n.d., b), General and Regional Statistics, Eurostat. Retrieved from http://ec.europa.eu/eurostat/ [accessed 10 Feb 2016]

EViews (2014), 'Users Guide.' Retrieved from http://www.eviews.com/EViews8/EViews8/EViews%208%20Users%20Guide%20I.pdf [accessed 25 Jun 2016]

Fischer, M. M.; Bartkowska, M.; Riedl, A.; Sardadvar, S. & Kunnert, A. (2010), 'The impact of human capital on regional labor productivity in Europe,' in M. M. Fischer & A. Getis (eds.) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Berlin & Heidelberg: Springer-Verlag, pp. 585–597. https://doi.org/10.1007/978-3-642-03647-7_28

Florax, R. J. G. M.; Folmer, H. & Rey, S. J. (2003), 'Specification searches in spatial econometrics: the relevance of Hendry's methodology,' *Regional Science and Urban Economics*, vol. 33, pp. 557–579.
https://doi.org/10.1016/S0166-0462(03)00002-4

Getis, A. (2010), 'Spatial autocorrelation,' in M. M. Fischer & A. Getis (eds.) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Berlin & Heidelberg: Springer-Verlag, pp. 255–278.
https://doi.org/10.1007/978-3-642-03647-7_14

GeoDa (n.d.), [Home page]. Retrieved from https://geodacenter.asu.edu/software/downloads [accessed 15 Feb 2015]

Ivanova, V. (2015), 'How space channels wage convergence: the case of Russian cities,' National Research University Higher School of Economics (HSE), Basic Research Program Working Papers, WP BRP 120/EC/2015.
https://doi.org/10.2139/ssrn.2717492

LeSage, J. P. & Fischer, M. M. (2008), 'Spatial growth regressions: model specification, estimation and interpretation,' *Spatial Economic Analysis*, vol. 3, no. 3, pp. 275–304. http://dx.doi.org/10.1080/17421770802353758

LeSage, J. P. & Pace, R. K. (2009), *Introduction to Spatial Econometrics*, Boca Raton, London & New York: Chapman & Hall/CRC.
https://doi.org/10.1201/9781420064254

Makrevska Disoska, E. (2016), 'Re-shaping the model of economic growth of the CEE countries,' *Baltic Journal of European Studies*, vol. 6, no. 2(21), pp. 137–159.
https://doi.org/10.1515/bjes-2016-0016

Mitchell, W. (2013), 'Introduction to spatial econometric modelling,' Centre of Full Employment and Equity Working Paper, no. 1–13.

Mur, J. & Angulo, A. (2005), 'A closer look at the Spatial Durbin Model,' 45th Congress of European Regional Science Association, Amsterdam.

Notermans, T. (2015), "The EU's convergence dilemma," *Baltic Journal of European Studies*, vol. 5, no. 1(18), pp. 36–55. https://doi.org/10.1515/bjes-2015-0004

Paas, T. & Schlitte, F. (2009), *Spatial Effects of Regional Income Disparities and Growth in the EU Countries and Regions*. Retrieved from http://ec.europa.eu/eurostat/documents/1001617/ 4398377/S3P2-SPATIAL-EFFECTS-TIIU-PAAS-FRISO-SCHLITTE.pdf [accessed 5 Feb 2015]

Rodríguez-Pose, A. & Tselios, V. (2007), 'Analysis of Educational Distribution in Europe: Educational Attainment and Inequality within Regions,' in Papers DYNREG08, Economic and Social Research Institute (ESRI), No. 8/2007.

Sutton, F. (2012), *The Nexus of Place and Finance in the Analysis of Educational Attainment: A Spatial Econometric Approach*. Retrieved from http://arizona.openrepository.com/arizona/bitstream/10150/265348/1/azu_etd_12498_sip1_m.pdf [accessed 11 Jul 2016]

Tolley, G. S. & Olson, E. (1971), 'The interdependence between income and education,' *Journal of Political Economy*, vol. 79, no. 3, pp. 460–480. https://doi.org/10.1086/259763

Umar, H. M.; Ismail, R. & Eam, L. H. (2014), 'A spatial econometrics analysis of educational distribution and regional income disparities in Nigeria,' in *Proceedings Book of ICETSR 2014, Malaysia Handbook on the Emerging Trends in Scientific Research*, pp. 722–731.

UNESCO (2012), *International Standard Classification of Education ISCED 2011*, Institute for Statistics. Retrieved from http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf [accessed 15 Feb 2015]

Wang, Z. (2012) *A Spatial Analysis of Educational Inequality in Mainland China*, Stats 499 Paper, Statistics Department, University of Michigan. Retrieved from http://deepblue.lib.umich.edu/bitstream/handle/2027.42/91791/wangzh.pdf?sequence=1 [accessed 15 Feb 2015]