Czech University of Life Sciences in Prague

Faculty of Environmental Sciences

Department of Applied Geoinformatics and Spatial Planning

Bachelor Thesis

Tools for spatial statistics in ArcGIS

Polina Moisseyenko

Supervisor

D.Sc. (Tech.) Olga Špatenková

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Environmental Sciences

BACHELOR THESIS ASSIGNMENT

Polina Moisseyenko

Applied Ecology

Thesis title

Tools for spatial statistics in ArcGIS

Objectives of thesis

Spatial dependency and autocorrelation are well-recognized issues that make statistical analysis different in the context of spatial data. The thesis will introduce the area of spatial statistics and evaluate the computational support for it in ArcGIS environment.

Methodology

The literature review will introduce the topic of spatial autocorrelation and describe core statistical methods for spatial data, such as local statistics, regression, kriging, or analysis of point patterns. The practical part of the thesis will illustrate how the concepts introduced are implemented in ArcGIS. The author will evaluate the findings and discuss the suitability of provided statistical tools.

The proposed extent of the thesis

30-40 pages

Keywords

spatial statistics, geostatistics, autocorrelation, ArcGIS

Recommended information sources

CRESSIE, N A C. *Statistics for spatial data.* New York ; Chichester ; Toronto ; Brisbane ; Singapore: John Wiley & Sons, Inc., 2015. ISBN 978-1-119-11461-1.

Esri, 2015. ArcGIS 10.4.Help. Esri, USA.

GETIS, A., 2008. A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective, Geographical Ananlysis, Volume40, Issue3,

CHUN, Y., GRIFFITH, D. 2012. Spatial Statistics and Geostatistics, Sage Publications Ltd.

OLIVER, M. – WEBSTER, R. *Geostatistics for environmental scientists.* Chichester: John Wiley & Sons, 2001. ISBN 0-471-96553-7.

Expected date of thesis defence

2018/19 SS – FES

The Bachelor Thesis Supervisor

D.Sc. (Tech.) Olga Špatenková

Supervising department

Department of Applied Geoinformatics and Spatial Planning

Electronic approval: 11. 3. 2019

doc. Ing. Petra Šímová, Ph.D.

Head of department

Electronic approval: 11. 3. 2019

prof. RNDr. Vladimír Bejček, CSc.

Dean

Prague on 11. 03. 2019

Declaration

I declare that I have worked on my bachelor thesis titled “Tools for spatial statistics in ArcGIS” by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break copyrights of any their person.

In Prague on _____

Acknowledgement

I would like to thank my supervisor, D.Sc. (Tech.) Olga Špatenková for guidance, valuable and wise advices, patience, and willingness to help. I would like to express my gratitude to my parents for giving me an opportunity to study abroad and for all the support and kindness they gave me.

Abstract

Spatial statistics is a field of study that combines methods for analyzing spatial distributions, patterns, and relationships among spatial data. The particular bachelor thesis evaluates the implementation of tools for spatial statistics in ArcGIS. The thesis is divided into a theoretical part as well as a practical part. The theoretical part defines important concepts of spatial statistics such as spatial autocorrelation, point pattern analysis, spatial regression models and spatial interpolation. What is more, the theoretical part defines the concept of GIS and introduces the ArcGIS environment. The practical part is targeted on the exploration of tools for spatial statistics in ArcGIS. The thesis provides outputs and comparison of explored tools.

It was observed that the ArcGIS provides with tools for spatial statistics that can assist in many issues of spatial analysis. The ArcGIS also has its advantages in interpreting the results. Still, not every described in the literature method of spatial statistics is implemented in the ArcGIS. On the other hand, not all those methods from the literature review are necessary for the ArcGIS.

Keywords: spatial statistics, geostatistics, autocorrelation, ArcGIS

Table of Content

List of figures

1 Introduction

Statistics is basically everywhere. Demography is statistics, the weather forecast is statistics, a branch of the economy – econometrics is statistics, population ecology is also statistics. Statistics can be the best assistant in such questions as decision making, finding trends, making predictions.

The origin of the word “statistics” is the Latin word *statisticus* – “state affairs”. The general usage of statistics as the collection and evaluation of data related to state affairs was designed by the German scientist Gottfried Achenwall (Gorroochurn, 2016). In present days statistics is defined as a branch of mathematics that deals with data collection, analysis, and interpretation of the results. However, in the context of spatially referenced data common statistical methods here are inappropriate. Spatial data can depend spatially and influence each other in time and space. In this case, methods of spatial statistics are more than appropriate. Spatial statistics combines all the spatial analysis methods. Spatial analysis is mainly recognizable in such techniques as spatial autocorrelation, point pattern analysis, and interpolation. All spatial analysis techniques are derived from mathematical equations and formulas. Nowadays there is no need to do those calculations by hand. All the calculations can be done within the console line in the form of commands. However, people perceive visual information better. Then software with a graphical user interface is more suitable. One of those programs is the ArcGIS. The ArcGIS provides with spatial statistics functions and visually interprets the results. This bachelor thesis will introduce spatial statistics techniques and evaluate the computational support for it in ArcGIS environment.

2 Objectives and methodology

2.1 Objectives

The main objectives of the thesis are to introduce spatial statistics and to evaluate spatial statistics tools implemented in the ArcGIS environment. The author will stick with the broader term of spatial statistics theory. The thesis will briefly introduce the philosophy of ESRI company. On the other side, not all the spatial statistics techniques are available in ArcGIS. And this is the most interesting part of the thesis.

The partial goals:

- Theoretical introduction to spatial statistics
- To introduce geographic information systems
- To introduce the ArcGIS environment
- To explore spatial statistics tools provided by ArcGIS
- To visualize and compare the outputs
- To evaluate the findings and discuss the suitability of provided statistical tools

2.2 Methodology

The thesis consists of the literature review and the practical part. The literature review includes an introduction to the spatial statistics techniques such as spatial autocorrelation, point patterns, GWR or kriging. The bachelor thesis is mainly focused on analyzing spatial patterns and associations in the data. The concepts of geostatistics are only briefly introduced. Furthermore, the literature review will define the term of geographic information system and the ArcGIS environment in general.

The practical part is based on the analysis of the tools for spatial statistics provided by ArcGIS. The author will explore the Spatial Statistics and Geostatistical Analyst toolboxes. There will be described such tools as High/Low Clustering tool, Geographically Weighted Regression tool, or Global Polynomial Interpolation tool. The map outputs will be used for the vivid comparison of the tools. The conclusion will be stated based on the theoretical review and practical outcomes.

3 Literature review

3.1 Basics of spatial statistics

3.1.1 Introduction to spatial statistics

Geography has a long history of developing mapping devices that enable insightful views of spatial data. The desire to make maps useful for analysis drove geographical writers to try to describe spatial distributions of data (Getis, 1999).

In fact, spatial statistics were developed from different fields of application. Those fields include mining engineering, agriculture, and forestry. An interest in spatial problems has grown during last 20 years. Besides that, computer technologies became more powerful and available. It made a positive impact on spatial statistics development. Such development has enabled a large collection of spatial datasets and encouraged usage of geographic information systems (Gelfand, et al., 2010). Today methods of spatial statistics are used in many quantitative disciplines (Cressie, 1991).

What exactly the spatial statistics are?

Spatial statistics embraces methods for analyzing spatial processes, spatial distributions, patterns, and relationships among geographic data. Spatial statistics methods include spatial relationships directly into their calculations (MIT, 2016). Spatial statistics uses spatial relationships of data to investigate the similarities among them (Hung, 2016).

Spatial data are defined by their location, attributes and topology (Krivoruchko, 2011). The attributes provide more information about features. The topology defines spatial relationships between features (de Smith, et al., 2007).

Spatial data are comprised of three major categories: continuous or geostatistical data, regional or lattice data, and spatial point patterns (Krivoruchko, 2011).

3.1.2 Types of spatial data

3.1.2.1 Geostatistical data

Geostatistics arose at the beginning of the 1980s in mining engineering, geology, and statistics. The original problem in geostatistics was to predict the amount of ore contained in the soil from observed samples (Cressie, 1991).

Geostatistics mainly deals with prediction at unobserved sites and reconstruction of spatial processes across the whole space. Here, a spatial set is a continuous subspace. Observation sites can be as regularly spaced as not (Gaetan & Guyon, 2010).

3.1.2.2 Lattice data

Lattice data are also known as aggregated data. These data are associated with areas and usually contain counts of an event within the polygon (Krivoruchko, 2011).

Here, a spatial set is a fixed discrete non-random set. These types of data are used for constructing and analyzing explicative models, prediction, and measuring spatial correlations (Gaetan & Guyon, 2010).

3.1.2.3 Point patterns

Point patterns types of data represent a collection of random events (Cressie, 1991). Point pattern data represent locations of events (Krivoruchko, 2011). Those types of data are used for analyzing the locations of the events (Cressie, 1991).

3.1.3 Descriptive spatial statistics

Describing data represents an important initial phase of any scientific method. The descriptive statistics are used for this purpose (Rogerson, 2015).

Descriptive statistics are very useful in providing summaries of spatial data. Descriptive spatial analysis is used to evaluate and understand such basic geographic concepts as central tendency and variability. In the context of spatial data, descriptive statistics are interested in distances between features and their mutual relationships which are closely related to those distances. The most common measure of those distances is Euclidean distance. In this chapter, the Mean Center and Euclidian Median will be defined

as an analogy for the central tendency and Standard Distance for describing the variability (Rogerson, 2015).

3.1.3.1 Mean Center

The mean center, the average location, is a measure of central tendency for a dataset. The coordinates of each point are required to determine the mean center. The mean center can be calculated by averaging the X and Y coordinates separately:

$$\bar{X}_c = \frac{\sum X_i}{n} \text{ and } \bar{Y}_c = \frac{\sum Y_i}{n} \quad (1)$$

where $\bar{X}_c$ is a mean center of X, $\bar{Y}_c$ is a mean center of Y, *i* represents a point, *n* is a total number of points (Sahoo, 2013).

3.1.3.2 Euclidean Median

The Euclidean median is considered as a more useful measure of center. The Euclidean median (X_e, Y_e) minimizes the sum of Euclidean distances from all the points to the central location. Mathematically this sum can be defined as:

$$\sum \sqrt{(X_i - X_e)^2 + (Y_i - Y_e)^2} \quad (2)$$

where X_i and Y_i represent coordinates of the point i, X_e and Y_e are coordinates of the Euclidean median (Sahoo, 2013).

3.1.3.3 Standard Distance

The standard distance is a spatial equivalent of the standard deviation. It measures the quantity of absolute dispersion in a point pattern. The formula for the standard distance is:

$$S_D = \sqrt{\frac{\sum(X_i - \bar{X}_c)^2 + \sum(Y_i - \bar{Y}_c)^2}{n}} \quad (3)$$

where S_D represents a standard distance, $\bar{X}_c$ and $\bar{Y}_c$ are mean centers of X and Y, X_i and Y_i are coordinates of the point i, n is a number of points in the distribution (Sahoo, 2013).

3.1.4 Global and Local statistics

Usually accepted statistical notion of pattern is a notion of complete spatial randomness. In this way pattern is equated with spatial homogeneity and heterogeneity resulting from deviations from complete spatial randomness. Many methods were elaborated to describe patterns in spatial data (Unwin, 1996).

Global statistics attempt to identify and measure the pattern of the entire study area. In global statistics, it is assumed that the pattern and the process are stable over space. But due to the fact, that areas are large such spatial homogeneity is extremely unlikely. Large areas of uninteresting spatial variation can swamp other of real interest (Unwin, 1996).

Local statistics help to identify variation across the study area. Methods of local statistics focus on individual features and their relationships to features in the neighborhood (MIT, 2016).

3.1.5 Spatial Autocorrelation

The Tobler's first law of geography states: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Main principles of spatial autocorrelation are based on this law (Unwin, 1996).

Spatial autocorrelation can be defined as the correlation of a variable with itself through space. Spatial autocorrelation measures the assumption of independence and the strength of autocorrelation. Because the correlation is within a single observation to differentiate this situation from conventional correlated samples situations, the prefix "auto-" was attached (Chun & Griffith, 2013). A variable is spatially autocorrelated if there are consistent patterns in its spatial distributions. That means that there might be dependencies within the data. Autocorrelation is positive when areas in neighborhood are alike. And vice versa autocorrelation is negative when nearby areas are unlike (Unwin, 1996).

3.1.5.1 Moran's I

Despite the existence of many ways to quantify spatial autocorrelation the most common index is Moran's I. This index is referred to global statistics which tell us whether an overall configuration is autocorrelated or not. To indicate where the interesting patterns are, we need local indicators (Pilz, 2009).

As was mentioned above spatial autocorrelation can be positive (nearby areas are alike), negative (neighbor areas are unlike), or neutral (the absence of spatial autocorrelation, areas are completely random) (Kirkegaard, 2015).

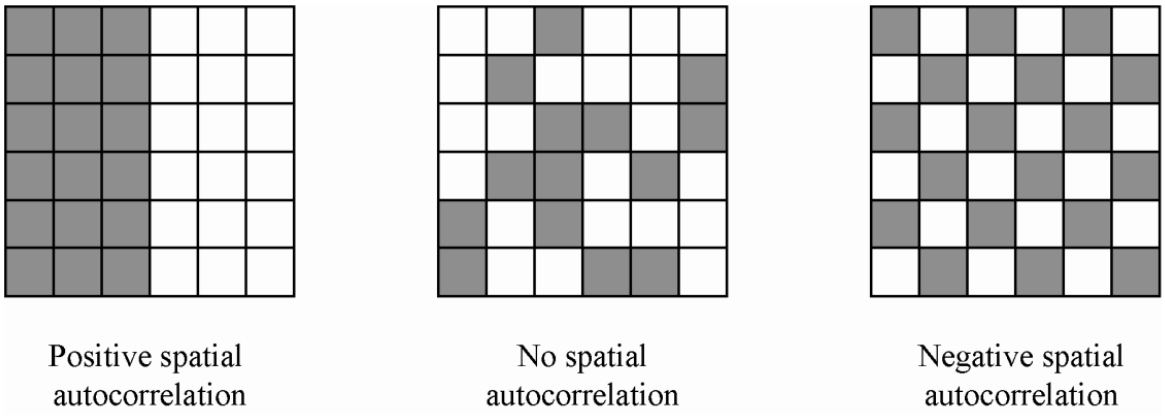

Figure 1: Examples of spatial autocorrelation. Source: (Kirkegaard, 2015)

Moran's I measures if the pattern of feature values is random, dispersed, or clustered. The main point of Moran's I is to compare the difference between the mean of the target feature and the mean for all features to the difference between the mean for each neighbor and the mean for all features (MIT, 2016).

The formula for Moran's I:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

where *i* and *j* represent different locations, so x_i and x_j are values of the variable in relevant locations *i* and *j*, $\bar{x}$ is the mean of the variable, w_{ij} is a spatial weight matrix, *n* represents a total number of observations. The value of Moran's I lies

in the interval from -1 to +1. Here the +1 value means positive autocorrelation, and -1 means negative autocorrelation (Xu & Kennedy, 2015).

There is also a local version of Moran's I called a Local Indicator of Spatial Associations (or LISA) by Anselin (Pilz, 2009). LISA measures the strength of patterns for each specific feature by comparing their values in a pair to the mean value for all the features in the interested area (MIT, 2016). It can be defined as

$$I_i = z_i \sum_j w_{ij} z_j \tag{5}$$

here w_{ij} represents a spatial weight matrix, z_i and z_j are the deviations from the mean (Xu & Kennedy, 2015).

A positive value indicates either a low value surrounded by low values and respectively a high value surrounded by high values. A negative value indicates either a low value surrounded by high values or a high value surrounded by low values (Xu & Kennedy, 2015).

3.1.5.2 Geary's ratio C

Another common spatial autocorrelation index is Geary's ratio C. It is also a measure of global clustering (Xu & Kennedy, 2015).

The formula of Geary's C is:

$$C = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

where n is a total number of observations, x_i and x_j are variables at locations i and j, $\bar{x}$ is the mean of the variable, w_{ij} represents a spatial weight matrix (Xu & Kennedy, 2015).

3.1.5.3 Getis–Ord General G

Like Moran's I, Getis-Ord General G is referred to global statistics. General G is used to indicate whether high or low values are clustered within the entire study area. General G excludes the value of the target feature in its calculation to see the effect of this feature on the surrounding area. Clusters

of high values called hot spots and vice versa clusters of low values called cold spots (MIT, 2016).

The General G can be defined as

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \quad (7)$$

where x_i and x_j are variables at locations i and j and w_{ij} represents a spatial weight matrix (ESRI, 2017).

There is a local version of General G – Getis-Ord G_i^*. It tells where the unusual pattern is. It is a fraction of total values in the neighborhood to the total values in the entire region. Originally defined with the actual point *i* excluded, but to avoid problems caused by a low number of neighboring zones, G* is often used including the own location value (Chun & Griffith, 2013).

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j - \bar{x}\sum_{j=1}^{n} w_{ij}(d)}{s_x\sqrt{\frac{(n-1)\sum_{j=1}^{n} {w_{ij}}^2(d) - \left[\sum_{j=1}^{n} w_{ij}(d)\right]^2}{n-2}}} \quad (8)$$

Here s_x represents the standard deviation of variable X for a selected distance *d,* w_{ij} represents a spatial weight matrix, x_j is a target feature (Chun & Griffith, 2013).

3.1.6 Point pattern analysis

Point pattern analysis deals with the point patterns type of data. As was mentioned above, the pattern of the studied events is emerged by chance. Statistical approaches of point pattern analysis are usually based on a hypotheses of complete spatial randomness (CSR). That means that occurrence of the event does not depend on the presence of other events. The probability of the occurrence is the same across the whole study area (Getis, 1999). The goal of a point pattern analysis is to describe and explain the distribution of the events (Krivoruchko, 2011).

Methods of a point pattern analysis can be classified into two groups: distance based approach and density based approach. The density-based

methods describe so-called first-order effects, the overall intensity. The distance-based approaches describe the pattern in term of second-order effects, the interaction between points (Gimond, 2018).

3.1.6.1 Quadrat Density

One of the point pattern techniques is a quadrat analysis. It is a density based technique (Gimond, 2018). This technique is based on counting the number of events in subareas, called quadrats, which are located across the area of interest. It is the simplest method of measuring patterns (Hung, 2016). Quadrat shapes and quadrat numbers might influence the measure of density, so quadrats parameters must be chosen carefully. However, the quadrat method suffers from the modifiable areal unit problem (Gimond, 2018).

3.1.6.2 Kernel Density

The kernel density method is an extension of the quadrat analysis. The kernel approach also computes subareas but those subareas overlap one another. That provides a moving subarea window called kernel. This method generates a grid of density values, the cells of the grid are smaller than those of the kernel window. Then each cell is assigned the density value that is computed for the kernel window centered on that cell. The neighboring points contribute to the density calculation so that their influence decreases with the growing distance from their kernel center (Gimond, 2018).

3.1.6.3 Nearest Neighbor Index

This index was developed by Clark and Evans and is assumed as the most commonly used in spatial pattern analysis (Hung, 2016). The index is based on comparing the distances between nearest neighbors of observed points and random distances that would be expected on the basis of chance (Levine, 2010).

$$Nearest\ Neighbor\ Index = \frac{Observed}{Expected} \tag{9}$$

Observed and expected distance are defined as:

$$Observed = \frac{\sum_{i=1}^{n} d_i}{n} \tag{10}$$

$$Expected = \frac{0.5}{\sqrt{\frac{n}{A}}} \quad (11)$$

where d_i is the distance between *i* and its nearest neighbor, and *A* is the total study area, *n* is a total number of features (ESRI, 2017).

If the observed distance is smaller than expected, points are closer to each other, the index is less than 1. Then there is an evidence for clustering. Otherwise if the index is greater than 1, there is an evidence for dispersion (Levine, 2010).

3.1.6.4 K-function

Ripley's K-function is used to describe how events occur over a study area. K-function tests hypothesis about data randomness, clustering, or dispersion. The function calculates the average number of points within distance normalized on the average number of points per unit area. If patterns are clustered, Ripley's K-functions is larger than it would be for randomly distributed points (Krivoruchko, 2011).

$$\widehat{K}(h) = \frac{1}{\lambda n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{\omega(s_i, s_j)} I(h_{ij} \leq h) \quad (12)$$

Here *h* is a distance from an arbitrary point, *λ* is the intensity of the point pattern, h_{ij} is a distance between points *i* and *j*, *I* is an indicator function if $h_{ij} < h$ then *I* is 1, otherwise *I* is 0, $\omega(s_i, s_j)$ is the edge correction factor, *n* is a number of observations (Krivoruchko, 2011).

A normalized Ripley's K-function, the L-function, is used when the point pattern is close to randomly distributed points. The K-function transformation to L-function may reduce data variability, it makes data close to stationary (Krivoruchko, 2011).

The transformed K-function is given as:

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h \quad (13)$$

where *h* is a distance, $\hat{K}$ is the expected value of K-function (Krivoruchko, 2011).

The K-function allows to see how point patterns distribution can change within the scale. This is an important capability because most spatial processes characteristics can change across scales. For example, at far distances the points can be dispersed, while at near distances the points can cluster (Krivoruchko, 2011).

3.1.7 Spatial regression

Spatial regression can be distinguished from other spatial statistics techniques. Spatial regression does not answer the question WHERE did that happen, but WHY did that happen (MIT, 2016). Regression analysis attempts to establish an equation for predicting dependent variable from a set of covariates and therefore to explore possible relationships between variables (Chun & Griffith, 2013).

3.1.7.1 Ordinary Least Squares

Linear regression analyzes linear relationships among variables. Those relationships can be positive or negative. Ordinary least squares (OLS) is a common linear regression technique. Being a global model it establishes one equation to describe the entire dataset (MIT, 2016). This method assumes that the data are normally distributed, homogeneous in variance and are independent from each other (Burrough & McDonnell, 1998).

The goal of the ordinary least squares method is to minimize the sum of squared difference between the sampled variable Y_i and its predicted variable $\hat{Y}_i$:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \longrightarrow min \quad (14)$$

(Krivoruchko, 2011).

3.1.7.2 Geographically Weighted Regression

A geographically weighted regression (GWR) represents a collection of fitted classical linear regression models for each feature in the dataset. Unlike the OLS model, in GWR model coefficients can vary spatially. That means the coefficients at every location need to be estimated (Krivoruchko, 2011).

The GWR model can be expressed as:

$$y_i = \beta_{0i} + \beta_{1i}x_1 + \beta_{2i}x_2 + \cdots + \beta_{ni}x_n + \varepsilon \quad (15)$$

where β_{ni} represent coefficients, x_i is an explanatory variable at location i, ε is a random error (Xu & Kennedy, 2015).

The GWR is simply a collection of separately fitted classical linear regression models to every feature in the dataset (Krivoruchko, 2011).

3.1.7.3 Spatial Weight Matrix

Spatial weight matrix is used to specify interdependency among observations. Spatial weight matrix W can be defined as a n x n table of weights indexed by a list of neighbors. Wij > 0 indicates that there is a dependence between observation i and neighboring observation j. If observation j is not a neighbor of observation i, weights Wij will set to 0 (Gelfand, et al., 2010).

Spatial weights matrix in GWR can be defined as

$$W_i = \begin{pmatrix} w_{i1} & 0 & 0 & \dots & 0 \\ 0 & w_{i2} & 0 & \dots & 0 \\ 0 & 0 & w_{i3} & \cdots & 0 \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & w_{in} \end{pmatrix} \quad (16)$$

where w_{in} represents the relationship between i and n (Xu & Kennedy, 2015).

3.1.7.4 Spatial Lag Model

Spatial lag model is also known as the spatially autoregressive model. This model includes a spatially lagged dependent variable. That means that the means of the dependent variable in neighboring areas are assumed as an extra explanatory variable (Xu & Kennedy, 2015).

The model is expressed as

$$y_i = \beta_0 + \beta_1 x_i + \rho w_i y + \varepsilon_i \quad (17)$$

where y_i is the dependent variable, β_0 is the intercept, β_1 is a regression coefficient, w_i is a spatial weight matrix, x_i is a predictor, ρ is the lag parameter, ε_i is a random residual. The model shows that the value of dependent variable y_i at each location i is defined as by independent variable x_i as by the value of y for neighboring features (Xu & Kennedy, 2015).

3.1.7.5 Spatial Error Model

Unlike the spatial lag model, the spatial error model considers the error term as autoregressive (Xu & Kennedy, 2015).

$$y_i = \beta_0 + x_i\beta + \lambda w_i \xi_i + \varepsilon_i \quad (18)$$

Here λ represents the degree of spatial covariance between units, β_0 is the intercept, β is a regression coefficient, w_i is a spatial weight matrix, ξ_i is the part of the residual that is spatially correlated between units, ε_i is a random residual (Xu & Kennedy, 2015).

3.1.7.6 Generalized Linear Model

The models specified above assume that the distribution of the data is normal (Gaussian). But if the assumption of normal distribution is violated, the best approach is to use models that were specially designed for that distribution (Krivoruchko, 2011).

For this case generalized linear model (GLM) was designed. The GLM involves non-normal distribution models, such as those for Poisson and binomial random variables (Chun & Griffith, 2013).

3.1.8 Spatial Interpolation

Spatial interpolation is also known as a spatial prediction. It is a process to predict values of locations that were not sampled based on a set of sampled locations (Hung, 2016). Interpolation techniques are used with samples of a continuous field. Interpolation tools can be divided into two groups:

deterministic interpolation methods and statistical interpolation methods (Gimond, 2018).

3.1.8.1 Inverse Distance Weighted Interpolation

The inverse distance weighted interpolation (IDW) is one of the simplest deterministic interpolation methods. This technique uses values from nearby weighted locations to compute an average value for unsampled locations. Weights are inversely proportional to the power of distance between the unsampled location and the sampled locations (Mitas & Mitasova, 1999). The weights can be defined as:

$$\lambda_i = \frac{\frac{1}{d_i^p}}{\sum_{i=1}^{N} \frac{1}{d_i^p}} \tag{19}$$

where d_i is the distance between x_0 and x_i, p is a power parameter. The power parameter can affect the accuracy of IDW. Weights decrease as the distance increases, especially when the power parameter increases. So, nearby samples have heavier weights and have more influence on the estimation (Li & Heap, 2008).

3.1.8.2 Polynomial Interpolation

The polynomial interpolators represent values of the points in the form of coordinate polynomial. Where for the point *x* with coordinates *(x, y)*:

$$Z^*(x, y) = P_n(x, y) \tag{20}$$

P_n is a n-order polynomial (Demyanov & Savyelyeva, 2010).

Examples of polynomials:

A zero-order polynomial (constant):

$$\hat{Z}(x, y) = a_0 \tag{21}$$

A first-order polynomial (linear):

$$\hat{Z}(x, y) = a_0 + a_1 x + a_2 y \tag{22}$$

A second-order polynomial (quadratic):

$$\hat{Z}(x,y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 \quad (23)$$

(Krivoruchko, 2011)

The aim of the polynomial interpolation is to find the unknown coefficients a_i in such a way that polynomials were as good as possible fitted to the points. For this purpose, the coefficients are determined by minimizing the squared difference between predictions and the data:

$$E = \sum_{i=1}^{n} \left((\hat{Z}(x_i, y_i) - Z(x_i, y_i) \right)^2 \quad (24)$$

where the E is the squared difference, $\hat{Z}(x_i,y_i)$ are the predictions, $Z(x_i,y_i)$ are the data (Krivoruchko, 2011).

As the name suggests, the global polynomial interpolation uses the entire dataset while local polynomial uses the data from the selected area. Global polynomial interpolation can be used to track the large-scale trends (Demyanov & Savyelyeva, 2010). But the idea of local interpolation is to use larger weights for data separated by small distances from the assumed location, where the data closer to prediction location have larger weights (Krivoruchko, 2011).

3.1.9 Geostatistical methods

Geostatistics methods are based on the statistical interpolation methods. As was mentioned in the paragraph 3.1.2.1 geostatistical analysis deals with continuous data and is concerned with predictions. Geostatistical techniques are based on statistical models that consider spatial autocorrelation. Therefore, in addition to the actual prediction, they offer a measure of the accuracy of the prediction (Demyanov & Savyelyeva, 2010).

3.1.9.1 Variogram

The semivariance is a measure of the variance and depends on the distance between the samples. With the increasing distance increases the value of the semivariance. The semivariance increases until it will equal the

variance for the entire array of data. The variogram (or semivariogram) is the function that describes the semivariance and is usually exponential, spherical, or Gaussian (Getis, 1999).

3.1.9.2 Kriging

Kriging is a similar method to IDW. It also uses a linear combination of weights at sampled locations to predict data value of an unsampled location. Kriging uses variogram to express the spatial variation. The main advantage of kriging is that in addition to the prediction it also provides a measure of uncertainty of the prediction (Hung, 2016).

Kriging is the basic interpolation model of all geostatistical methods. There are many types of kriging (Demyanov & Savyelyeva, 2010).

Simple kriging requires a known stationary mean value as input to the model (Krivoruchko, 2011). The mean is known and stationary across the whole study area (Demyanov & Savyelyeva, 2010).

Simple kriging can be estimated as

$$Z^*(x) = m + \sum_{i=1}^{n(x)} \lambda_i(x) Y(x_i) \qquad (25)$$

where *Z(x)* is a second-order stationary random variable, *m* is a known stationary mean value, λ_i is a weight factor in point i, *Y* is a random function. The assumption of the stationary mean is the main disadvantage of simple kriging. Using the average of the sampled values make simple kriging less accurate (Demyanov & Savyelyeva, 2010).

The ordinary kriging is similar to simple kriging. The only difference is that ordinary kriging does not assume the known mean. Beyond that the ordinary kriging does not require the local mean to be stationary along the whole estimated area (Demyanov & Savyelyeva, 2010).

3.2 Geographic Information System

3.2.1 Definitions of geographic information system

Geographic Information System, or GIS, was developed by connection of several discrete technologies. That caused a synergistic effect. Beyond that, the maintaining of GIS was relied on innovations of many different disciplines. As a result, GIS is considered as a powerful technology (Fazal, 2008). In the most general sense, Geographic Information Systems are tools for managing geographic information. There is not only one accurate definition of GIS, a concept of GIS can change according to the intellectual, economic or political aims (DeMers, 2009).

According to Burrough & McDonnell (1998) GIS can be distinguished in three ways.

- GIS as a toolbox – a powerful set of tools that provides collecting and storing, transforming and displaying data which are spatially referenced to the real world.
- GIS as a database – a database system which consists of spatially indexed data, and a set of procedures operated upon the database used to handle spatial data with their attributes, location, and topology.
- GIS as an organization – an institute, which reflects an organizational structure integrating a database with technology, expertise and financial support.

Other definitions of GIS by the fields of use include:

A set of maps in a digital form – the general public.

A computer program that helps in solving geographic problems – decision making, planning.

A tool for detection what is otherwise hidden in geographic information – science, investigation (Longley, et al., 2016).

3.2.2 History of GIS

Development of geographical information systems is strictly connected with development of map-making and storing of spatial data. Until the nineteenth century geographical data were mainly used in fields of trade,

military and exploration operations. New needs such as planning roads or telegraph lines appeared with evolving infrastructure. The necessity of more detailed and specialized maps comes alongside. Those needs became main drivers of progress (Fazal, 2008).

In the early 1900s aerial photography emerged with the development of aviation. Aerial photography became an accelerator of the map-making progress. Photogrammetry, a technique of making measurements from photographs, was developed. Aerial photographs became an important source of quantitative information in estimating geological formation and vegetation (Fazal, 2008).

Aerial photography and satellite images made it possible to track landscapes changes over time. But satellite photos represented digital data, coded elements. New tools were needed to convert coded data into a familiar form of images (Burrough & McDonnell, 1998).

Evolution of GIS continued alongside with development of computer technologies. For the first time the term of GIS was used in the middle of 1960s in Canada when the first actual GIS was created. CGIS was designed as a measuring tool. With the development of microprocessors in the late 1980s computer capabilities became more advanced and available. Elaborating of the internet boosted GIS development. Data networks afforded new opportunities for search and distribution of geographical data (Longley, et al., 2016).

Nowadays it is fair to say that GIS is a very powerful technology. With all the progress that was made, GIS became more available and is now rich with data and techniques. Geographic information truly is a power and GIS helps us to use this information effectively (Heywood, et al., 2006).

3.2.3 Components of GIS

GIS consists of six main components without which GIS would not exist (Longley, et al., 2016).

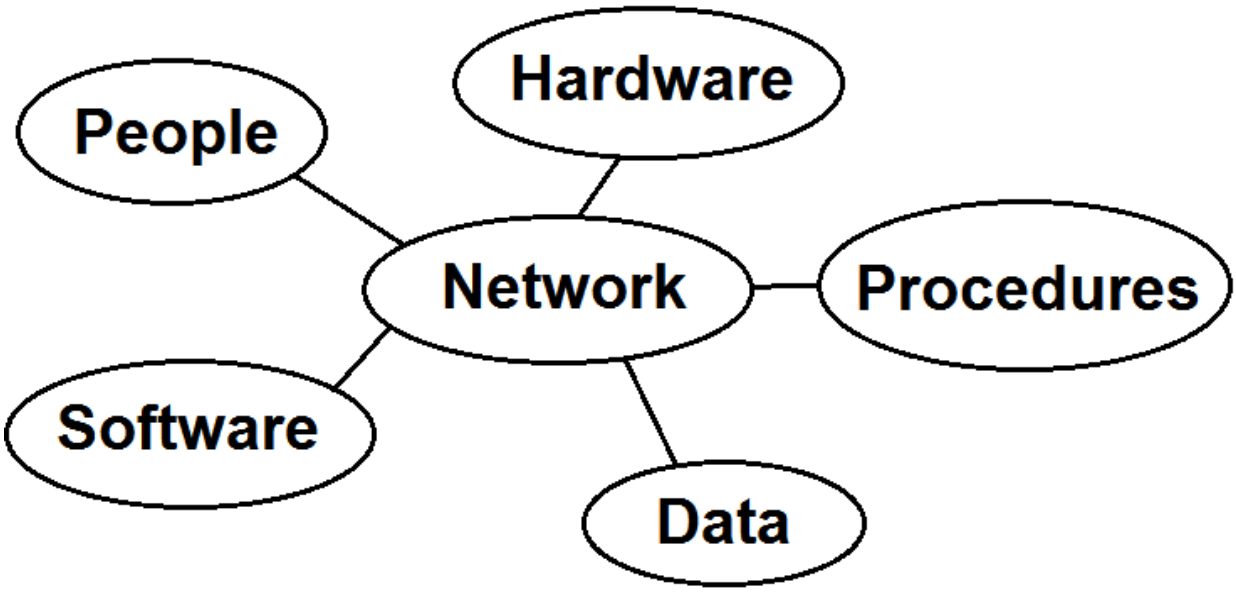

Figure 2: The six component parts of GIS. Source: author's own creation based on the scheme from Longley, et al., 2016.

The first component, hardware, represents a physical equipment capable of accepting and storing data, and then to produce control outputs (Meulen, 2000). Hardware is the device that user interacts with. Usually the hardware is understood as desktop computers, but nowadays GIS functions can be delivered even through mobile phones (Longley, et al., 2016).

The second part of GIS is the software, the computer programs. The software should be based on the hardware functionality. The highly developed software requires the more advanced hardware (Harmon & Anderson, 2003). The GIS software consists of the applications and tools for creating maps and processing data (Harvey, 2008).

The third component of GIS is the database. Geographic data are a representation of the real world's aspects. Data can appear as a product of an aerial photography, so as a wholly digital product. Data serve as the main material for map-making and analysis (Decker, 2001).

The next important part of GIS is the network. Developing of the internet expanded GIS functions in data transmission. What is more, the internet provides a big count of free GIS data sources and freeware GIS applications (Heywood, et al., 2006).

Besides components that were described above GIS needs to be in control. An organization must establish procedures, hierarchical bindings and other mechanisms to ensure that GIS fulfills the needs of the organization (Longley, et al., 2016).

3.2.4 Analysis in GIS

Besides map making and data storing functions, GIS is also equipped with capabilities for data analysis (Harvey, 2008). Capabilities of spatial analysis distinguish GIS from other data processing systems. Those capabilities use spatial and non-spatial data for further data analysis and problem-solving (de By, 2001).

Spatial statistics analysis within the GIS emerges into a powerful set of tools. Not far ago statistical analysis was limited to exploratory data analysis and visualization. But at the present days, spatial statistics techniques provided by GIS allow to create models and make predictions (Krivoruchko & Gotay, 2003).

3.2.5 About ArcGIS

GIS software is based on computer capabilities such as data management, printing and display, and device drivers to provide a suitable environment for collection, storage, analysis and interpretation of geographic data. Nowadays there are many types of GIS software which can be used in many ways (Dawsen, 2011).

A popular commercial GIS software ArcGIS was developed by ESRI (Environmental Systems Research Institute) company. For the first time, it was released in 1999 and since then ESRI continues to upgrade ArcGIS with new tools for working with spatial data. These days ESRI is widely accepted as a leader in GIS applications field (Heywood, et al., 2006).

ArcGIS consists of:

- **ArcGIS Desktop** – Represents a suite of integrated GIS programs.
- **ArcGIS Server** – A platform where developers can build Web applications and Web services. The platform also supports building enterprise GIS applications.

- **ArcGIS Mobile** – A mobile version of GIS software for smartphones and tablets.
- **ArcGIS Engine** – Provides components for ArcGIS developers.
- **ArcGIS Online** – Cloud-based GIS applications and services that users are able to access via the internet.

(ESRI, 2012)

3.2.6 ArcGIS Desktop

In its origins, ArcGIS Desktop was designed for creating maps and spatial data. The analysis capability was added later. ArcGIS Desktop includes many components such as ArcMap, ArcCatalog and ArcGlobe applications. ArcMap application provides tools for viewing GIS data and tools for creating maps. ArcCatalog components are used to connect to geodatabases or to create one. ArcGlobe is a part of the 3D Analyst extension (Nasser, 2015).

ArcGIS Desktop provides three different levels of functionality:

The Basic license – Former ArcView, provides mainly basic features, such as mapping, data use, and simple editing (Nasser, 2015).

The Standard license – Former ArcEditor, includes the capability of viewing, creating and editing maps and spatial data (Nasser, 2015).

The Advanced license – Former ArcInfo, combines the functionality of both ArcView and ArcEditor along with advanced data analysis and modeling (Nasser, 2015).

Each license has a various set of analysis tools. The more advanced the license is the higher number of analysis tools it provides. Toolboxes are situated in the ArcToolbox application which is integrated into the ArcMap (ESRI, 2004).

The ArcToolbox is an integrated application which represents a suit of toolboxes where each set of tools provides different functions from simple data management tools to statistical analysis tools (ESRI, 2004).

4 Tools for spatial statistics in ArcGIS

4.1 Spatial statistics in ArcGIS

The ArcGIS Desktop is supported with functions for spatial statistics analysis. Those functions are situated in Spatial Statistics toolbox and Geostatistical Analyst toolbox (Griffith & Chun, 2018). Most of the tools were written in Python programming language and their source codes are fully accessible for ArcGIS users (Scott & Janikas, 2010).

The ArcGIS Desktop Advanced version 10.5.1 was used for evaluating the variety of spatial statistics tools.

The ESRI company has its own philosophy in grouping spatial statistics functions to certain toolboxes. The grouping depends on the method of application. For example, it is needed to perform an analysis in order to determine clustering or dispersing in the pattern in general. Then the most suitable tools for this purpose will be situated in a particular toolset. In this example, those tools are situated in the Analyzing Patterns toolset.

4.2 Used data

This chapter describes the data used to visualize how certain tools for spatial statistics work and provided outputs. The ArcGIS works with the shapefile format. A shapefile usually contains the geographic location and attributes of geographic features. The next paragraphs contain descriptions of three shapefiles used for the analysis.

4.2.1 Median housing dataset

The first shapefile is named *medianhousing.shp*. It is available for download from the open courseware of the Massachusetts Institute of Technology website. This shapefile represents the median housing values for each census tract in Middlesex Country provided by the American Community Survey. The attributes provide with information on the IDs, names, tract areas, and median housing values of each tract. The analysis will be conducted within the median housing values variable (Median_val).

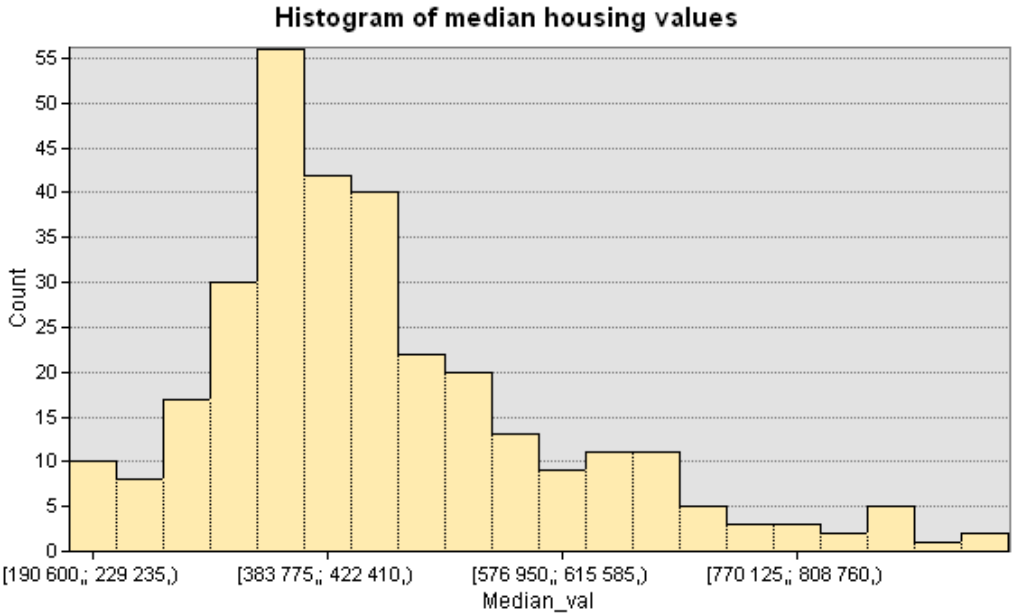

Figure 3: A histogram of a median housing values variable. Source: ArcMap.

Figure 3 represents a distribution of a median housing values variable. The overlook and the locations of the polygons within Massachusetts are illustrated in figure 4.

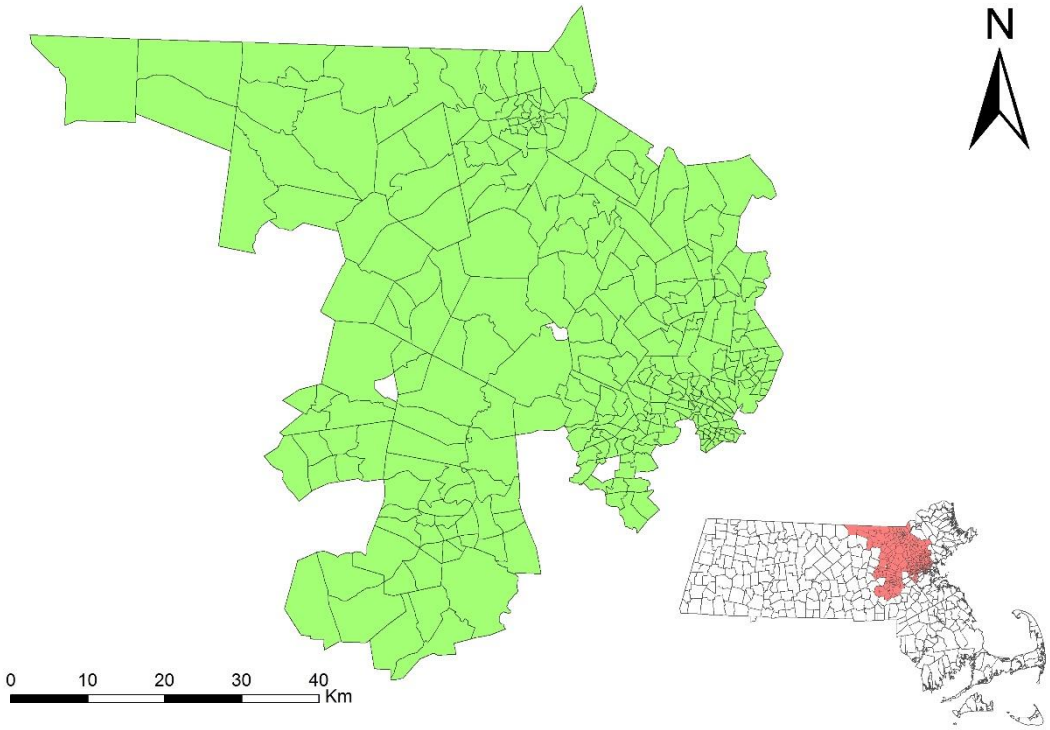

Figure 4:The medianhousing.shp shapefile. The small map represents the location of the Middlesex Country within the Massachusetts state. Source: own data processing.

4.2.2 South dataset

The second shapefile is also available from the MIT open courseware. The shapefile is named *south.shp*. It represents a national data set between the years 1960 and 1990. The shapefile includes such attributes as the names of regions and socio-economic characteristics. Variables used for the analysis are:

- Homicide Rate (HR) – is used as a dependent variable.

The explanatory variables are:

- Resource Deprivation (RD)
- Population Structure Component (PS)
- Unemployment rate (UE)
- Percent of divorced males (DV)
- Median age (MA)

A distribution of HR variable is illustrated in figure 5.

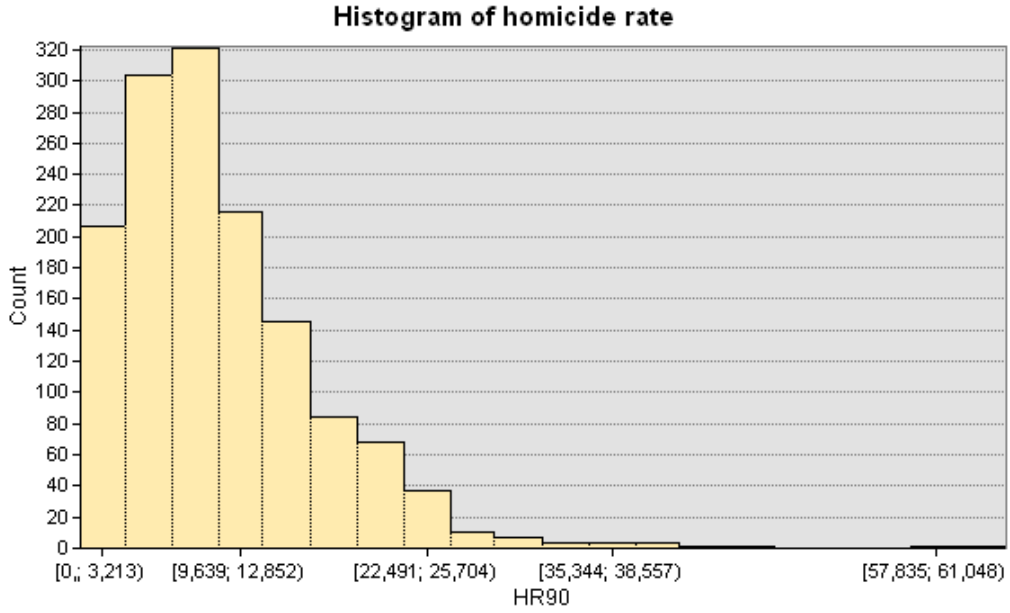

Figure 5: A histogram of a dependent variable. Source: ArcMap.

Figure 6 represents an overlook of the south shapefile and illustrates which regions are included in the dataset.

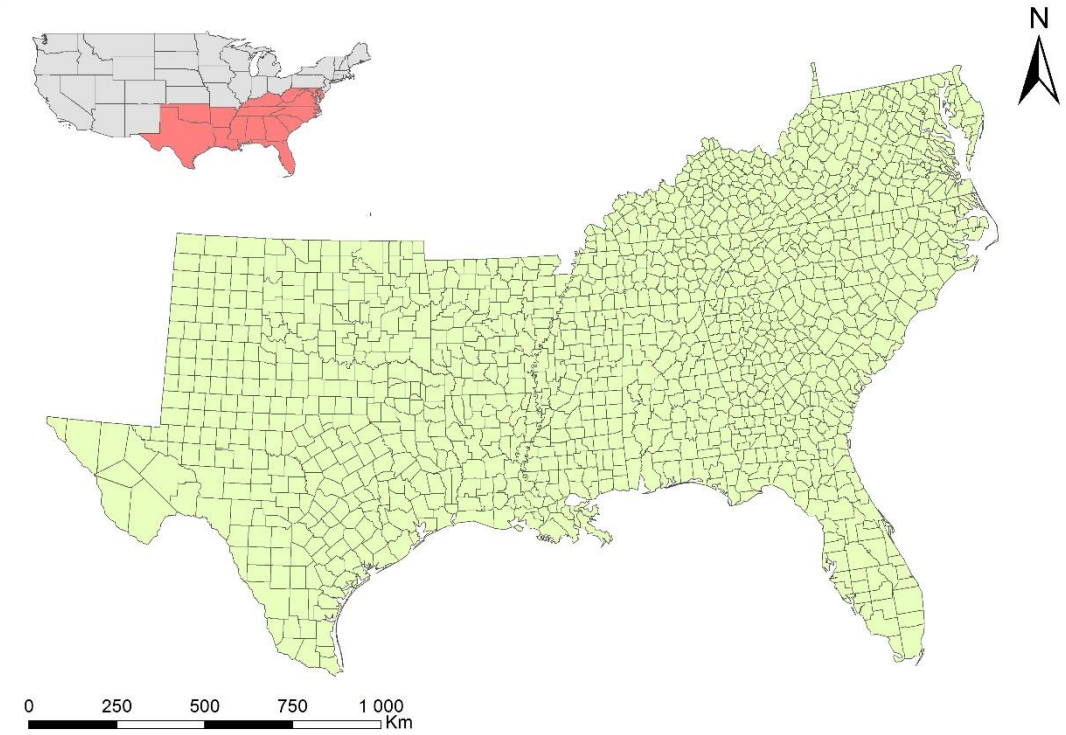

Figure 6: The south.shp. The small map represents which states are included into the dataset. Source: own data processing.

4.2.3 Heavy metals dataset

A *heavy_metals_tutorial.shp* shapefile was used to visualize the results of tools for prediction. This shapefile is a part of a large dataset used in the book “Spatial Statistical Data Analysis for GIS Users” by Krivoruchko. The shapefile represents sampling sites for measurement of heavy metals concentrations collected in Austria in 1995. The Measurements were done for such heavy metals as arsenic, cadmium, and mercury. The variable of mercury measurements is used for analysis in this thesis. Figure 7 represents a distribution of mercury measurements variable. The overlook of the shapefile is illustrated in figure 8.

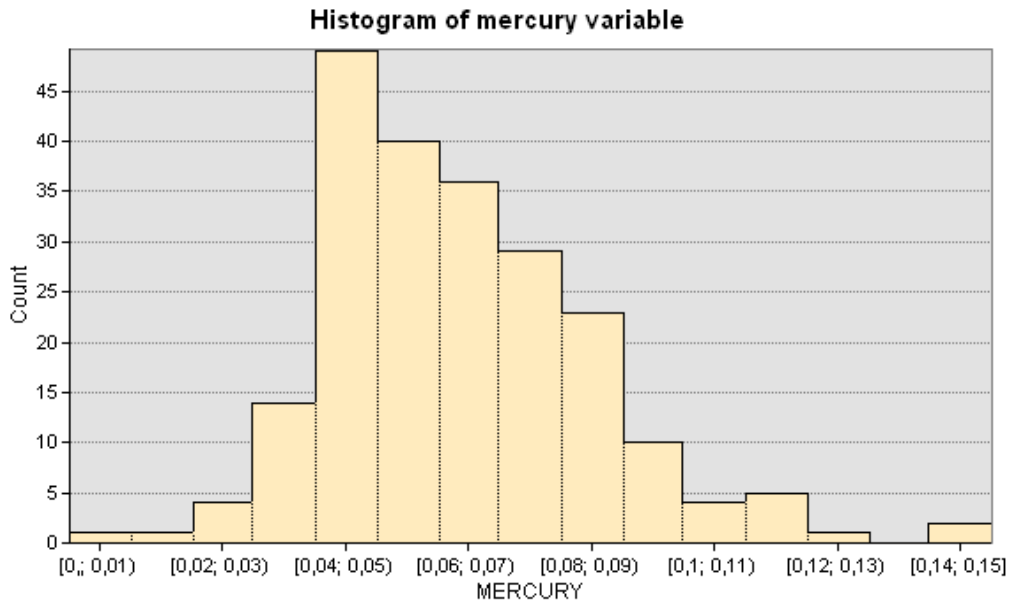

Figure 7: A histogram of a mercury variable. Source: ArcMap.

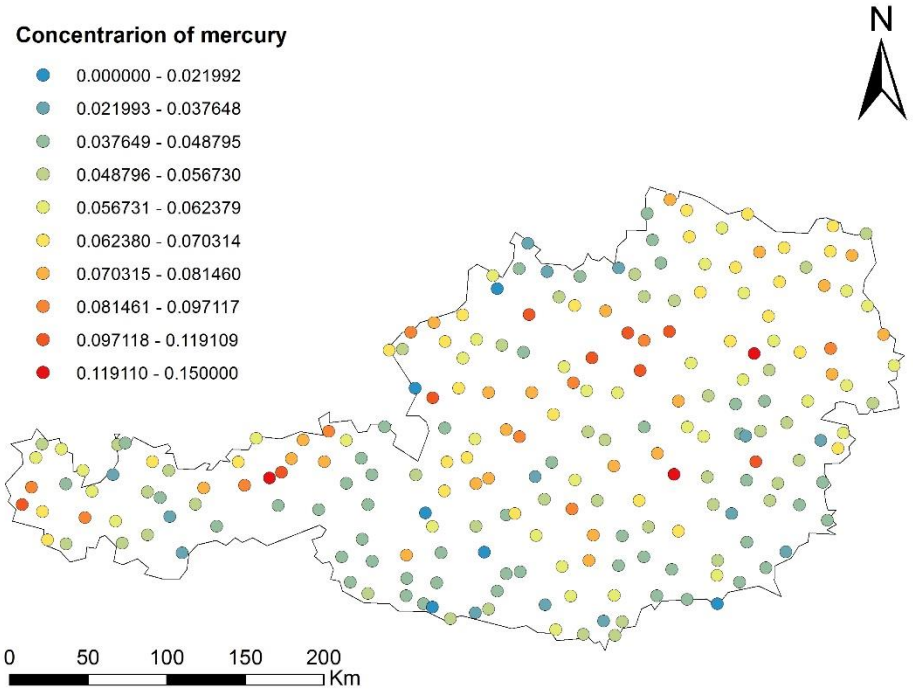

Figure 8: The heavy_metals_tutorial.shp -the concentration of mercury. Source: own data processing.

4.3 Spatial Statistics toolbox

The Spatial Statistics toolbox is divided into five toolsets:

- The first toolset is Analyzing Patterns. This toolset helps to identify spatial patterns. The tools are inferential statistics. They start with the null hypothesis that the pattern is the complete spatial randomness. Then the hypothesis is accepted or rejected based on a p-value. The

tools of the Analyzing Patterns toolset are: Average Nearest Neighbor, High Low Clustering, Incremental Spatial Autocorrelation, Multi-Distance Spatial Cluster Analysis, Spatial Autocorrelation.

- Another toolset Mapping Clusters is used for cluster analysis. These tools help to identify the location of statistically significant hot spots, cold spots, spatial outliers, and similar features. The tools of the Mapping Clusters toolset are: Cluster and Outlier Analysis, Hot Spot Analysis, Grouping Analysis, Optimized Hot Spot Analysis, Optimized Outlier Analysis, Similarity Search.
- The Measuring Geographic Distributions toolset provides with descriptive spatial statistics. This toolset allows to calculate a value that represents a characteristic of the distribution, such as the center or orientation of the data. The tools of the Measuring Geographic Distributions toolset are: Central Feature, Directional Distribution, Linear Directional Mean, Mean Center, Median Center, Standard Distance.
- The Modeling Spatial Relationships tools are used to construct spatial weights matrices or to model spatial relationships. Those tools use regression analysis. The tools of the Modeling Spatial Relationships toolset are: Exploratory Regression, Generate Network Spatial Weights, Generate Spatial Weights Matrix, Geographically Weighted Regression, Ordinary Least Squares.
- The last toolset is the Utility toolset. These tools were designed to be used in combination with other tools provided by the Spatial Statistics toolbox. The tools of the Utility toolset are: Calculate Distance Band from Neighbor Count, Collect Events, Convert Spatial Weights Matrix to Table, Export Features Attribute to ASCII.

The Spatial Statistics toolbox is situated in the ArcToolbox window. The ArcToolbox window can be found either on the Geoprocessing menu, on the Standard toolbar, or using the Search window.

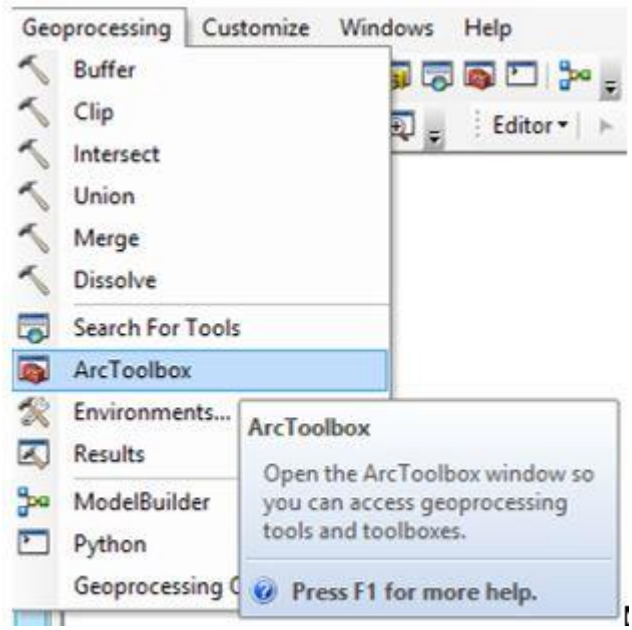

Figure 9: The Geoprocessing menu. Source: ArcMap.

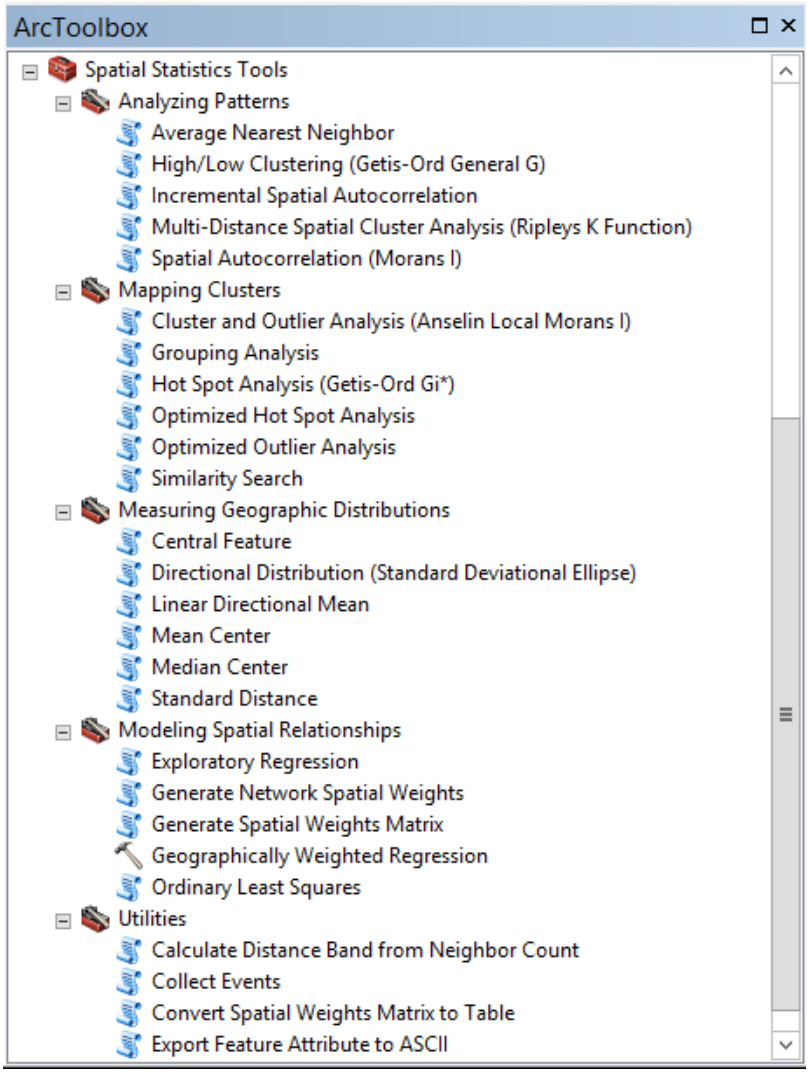

Figure 10: The ArcToolbox window. Source: ArcMap.

4.3.1 Is there a clustering in the pattern?

Going back to the example in chapter 4.1 it was described that the most suitable tools for the question are situated in the Analyzing Pattern toolset. The following paragraphs will examine every suitable tool and compare the results.

The medianhousing.shp shapefile will be used for the further analysis.

4.3.1.1 Average Nearest Neighbor tool

The first method to determine whether features are clustered or dispersed is to use the Average Nearest Neighbor tool. The tool calculates the nearest neighbor index which was described in the paragraph 3.1.6.3. Studying data are used as an input. There is an option to specify the Area value. If the Area value was not specified, so the area of minimum enclosing rectangle around studying features will be manually used. Then ArcMap optionally generates a report in HTML format.

The report can be found in the Results window. The output report provides with the information of the nearest neighbor index, a z-score, a p-value. Here, the nearest neighbor ratio is about 1.0. This is the evidence for the complete spatial randomness. The null hypothesis of complete spatial randomness is accepted based on the z-score and the p-value. The final answer is: the features are randomly distributed.

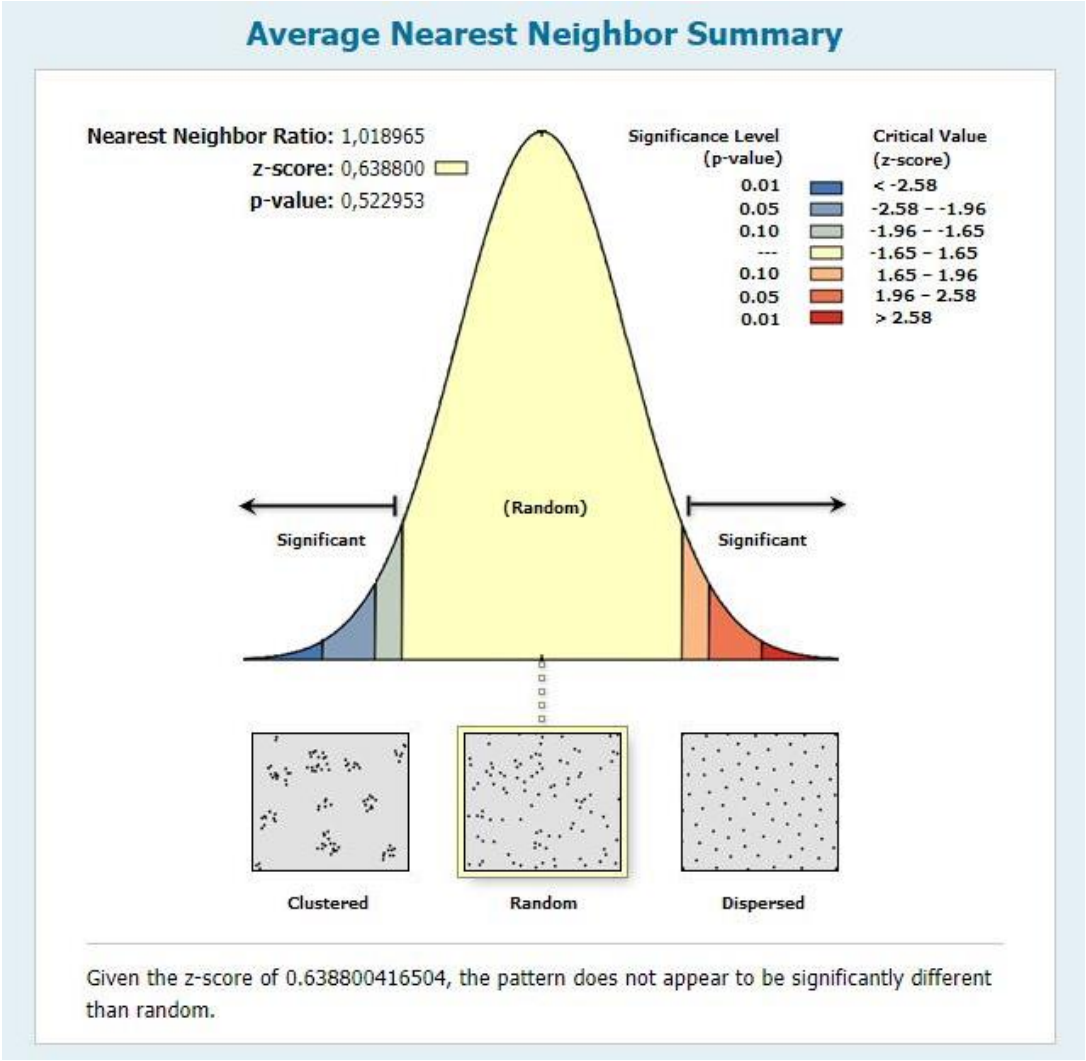

Figure 11: Average Nearest Neighbor with not specified Area value – the HTML report. Source: own data processing.

Figure 9 represents an output with the specified Area value. In the first example the generated Area value was around 6 100 km^2, while in the current example the Area value is specified as 10 000 km^2. As it can be seen the results differ. At the larger Area value the analysis shows that the pattern is clustered. This is a good example of how the Average Nearest Neighbor method is sensitive to the Area value.

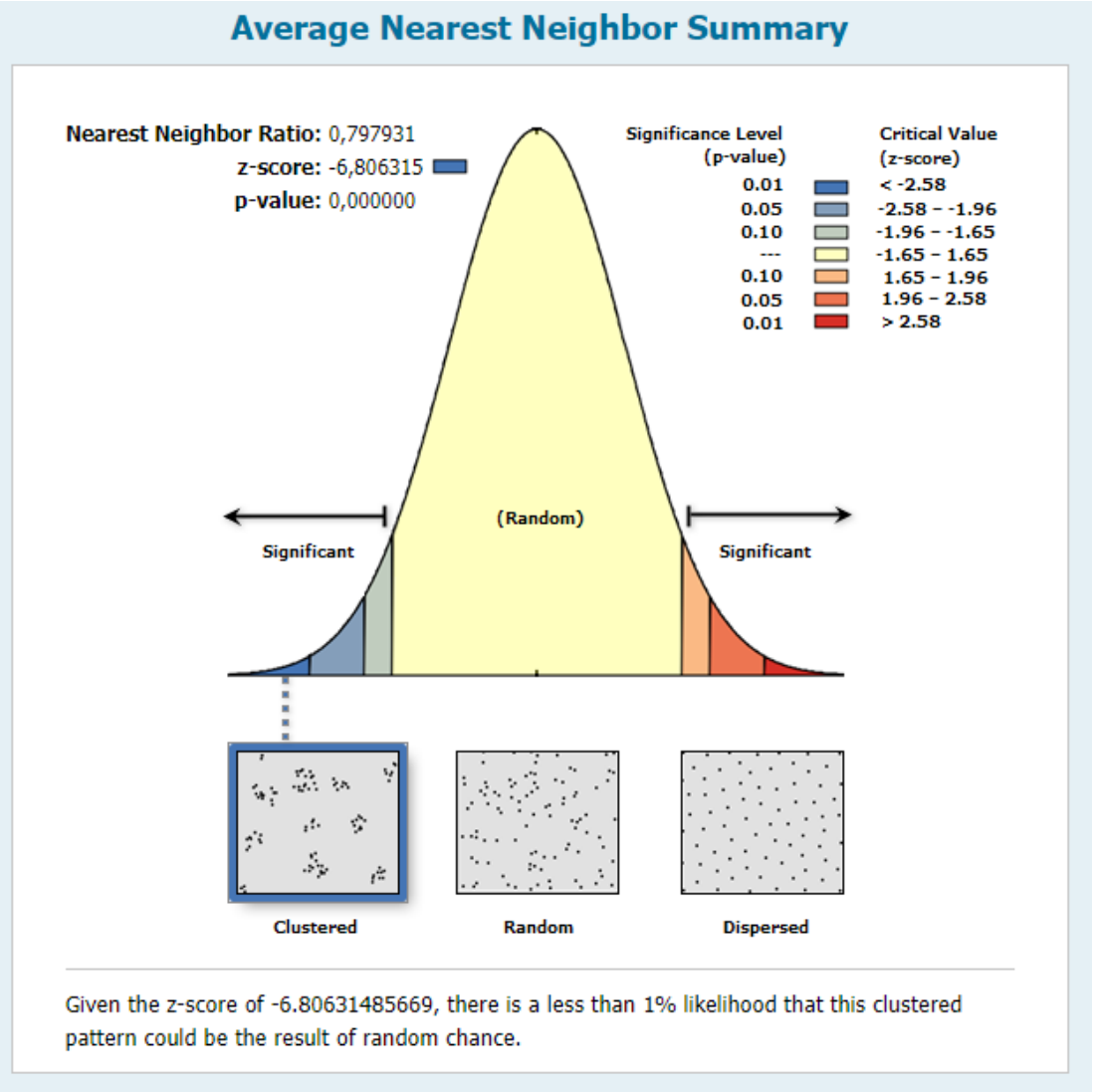

Figure 12: Average Nearest Neighbor with specified Area value (10 000 km^2) – the HTML report. Source: own data processing.

4.3.1.2 High/Low Clustering tool

Another method of determining the clustering is the High/Low Clustering tool. The tool is based on Getis-Ord general G function that was mentioned in paragraph 3.1.5.3. This tool differs from the previous tool that it determines whether the high or low values for the variable are clustered. The input is same as in the previous tool.

The output is also an HTML format report. As it can be seen in figure 10 the analysis shows clustering for high values for the pattern based on the z-score and p-value. The positive z-score indicates that the Observed General G value is greater than the Expected. That also indicates that the high-values for the variable are spatially clustered.

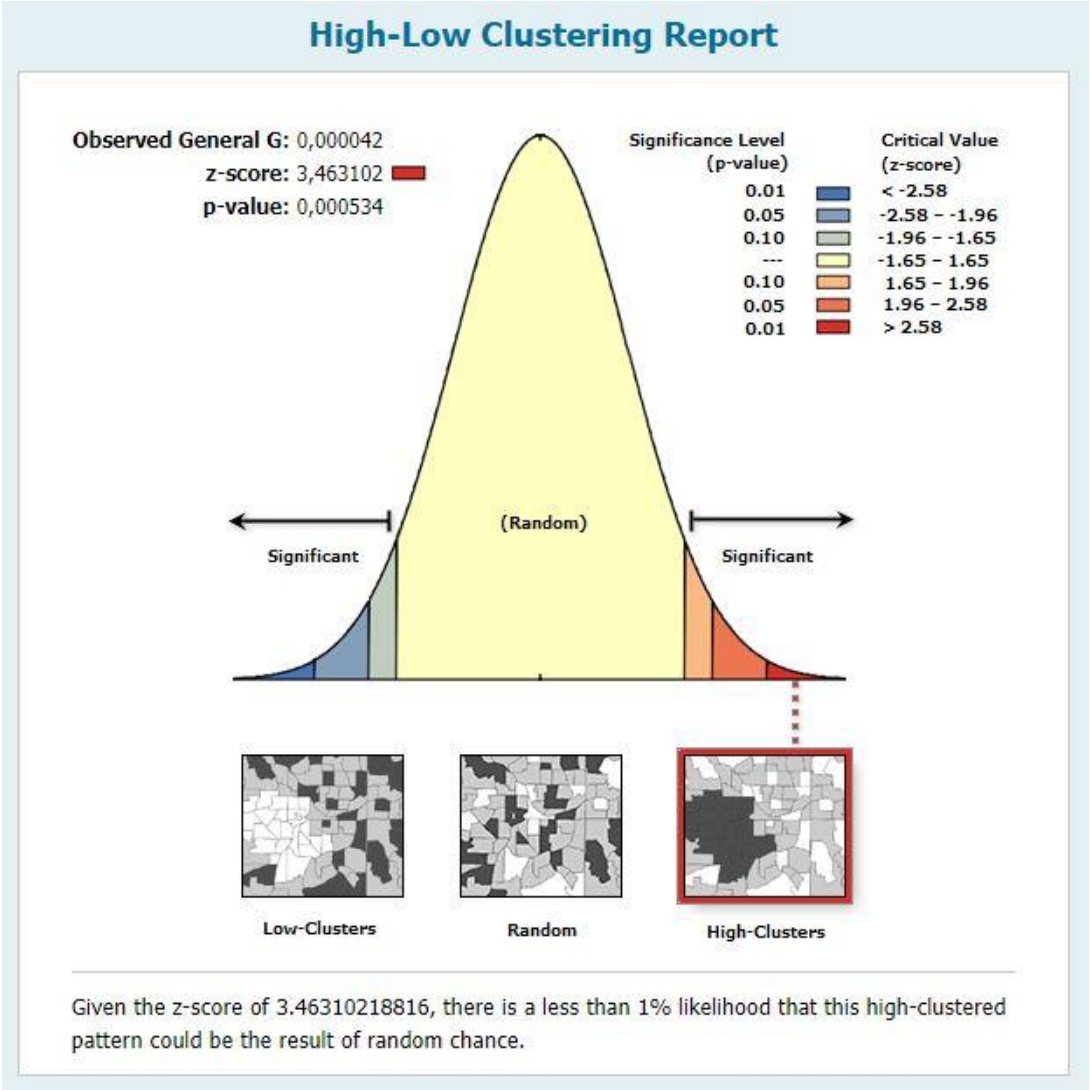

Figure 13:High/Low Clustering – the HTML report. Source: own data processing.

4.3.1.3 Multi-Distance Spatial Cluster Analysis tool

The next observed tool is the Multi-Distance Spatial Cluster Analysis tool. It is based on Ripley's K-function that was mentioned in the paragraph 3.1.6.4. According to ArcGIS Help guide, mathematically this tool uses the transformation of K-function, the L- function.

This tool calculates K values for input features. As the Average Nearest Neighbor, the K-function is also very sensitive to the scale. An output is dBASE table format. This table contains observed K values and expected K values. As an option, the tool calculates confidence envelopes.

medianhousing_MultiDistanceS2

OBJECTID *	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
1	2123,359151	4563,481872	2440,122721	1915,169136	2447,623121
2	4246,718302	8860,748019	4614,029717	4119,963633	4601,279931
3	6370,077453	12218,11867	5848,041217	6266,308025	6684,827388
4	8493,436604	15167,634429	6674,197826	8275,707607	8831,954974
5	10616,795755	17783,479566	7166,683812	10411,821116	11057,234165
6	12740,154906	20091,027948	7350,873042	12504,13002	13185,490955
7	14863,514057	22216,512312	7352,998256	14570,8169	15221,277866
8	16986,873208	24167,498684	7180,625477	16598,132516	17290,228802
9	19110,232358	25924,439498	6814,20714	18606,075973	19325,396801
10	21233,591509	27616,904227	6383,312717	20594,871635	21364,129247

Figure 14:Multi-Distance Spatial Cluster Analysis tool - an output table. Source: own data processing.

Here, on the output table (Figure 11) the observed K values are greater than the expected K values. That means that distribution at those distances is more clustered than dispersed. An optional line graph in figure 12 summarizes the results. The graph is built upon values from the output table.

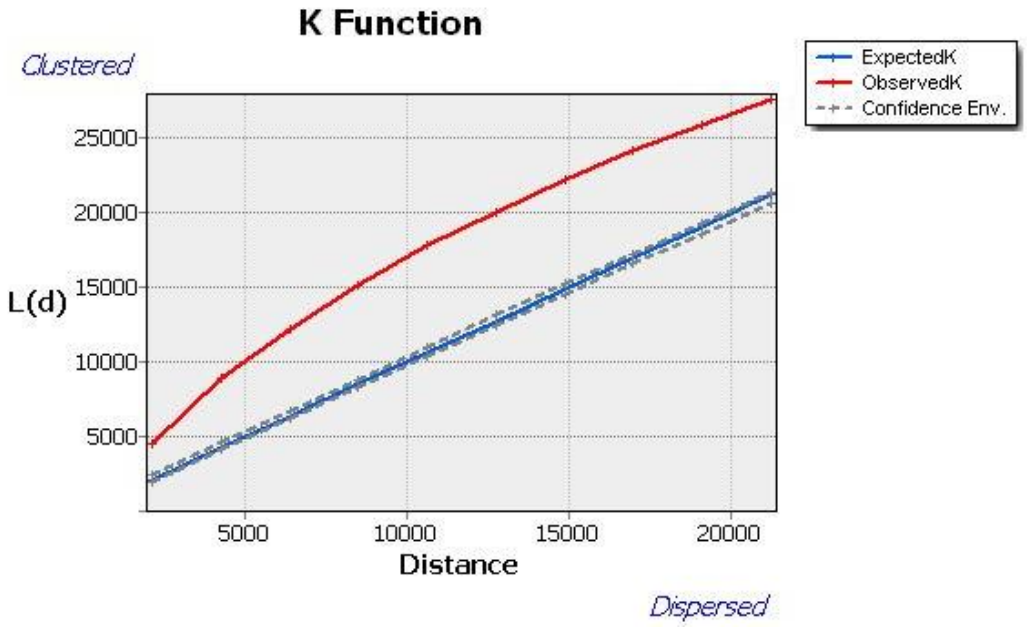

Figure 15:Multi-Distance Spatial Cluster Analysis tool - a Graphical output. Source: own data processing.

As it can be seen on graph when the red line of the observed K values is above the blue line of the expected K values, the distribution at those distances is more clustered. What is more, the observed K values are larger than the upper confidence envelope (HiConEnv) values. This indicates that the clustering for those distances is statistically significant. If the red line was below the blue line, that would indicate dispersing at the distances.

4.3.1.4 Spatial Autocorrelation tool

The last tool that can be used to determine clustering is the Spatial Autocorrelation tool. This tool measures spatial autocorrelation. The spatial autocorrelation is measured by global Moran's I spatial correlation index based on the selected feature and attribute value. As was mentioned in the paragraph 3.1.5.1 the value of spatial autocorrelation index lies in the interval from -1 to +1.

On the following example the Moran's I is equal to 0.356. The autocorrelation is positive. That means that the pattern is clustered. The null hypothesis is rejected based on p-value. The distribution of the median housing values variable is clustered.

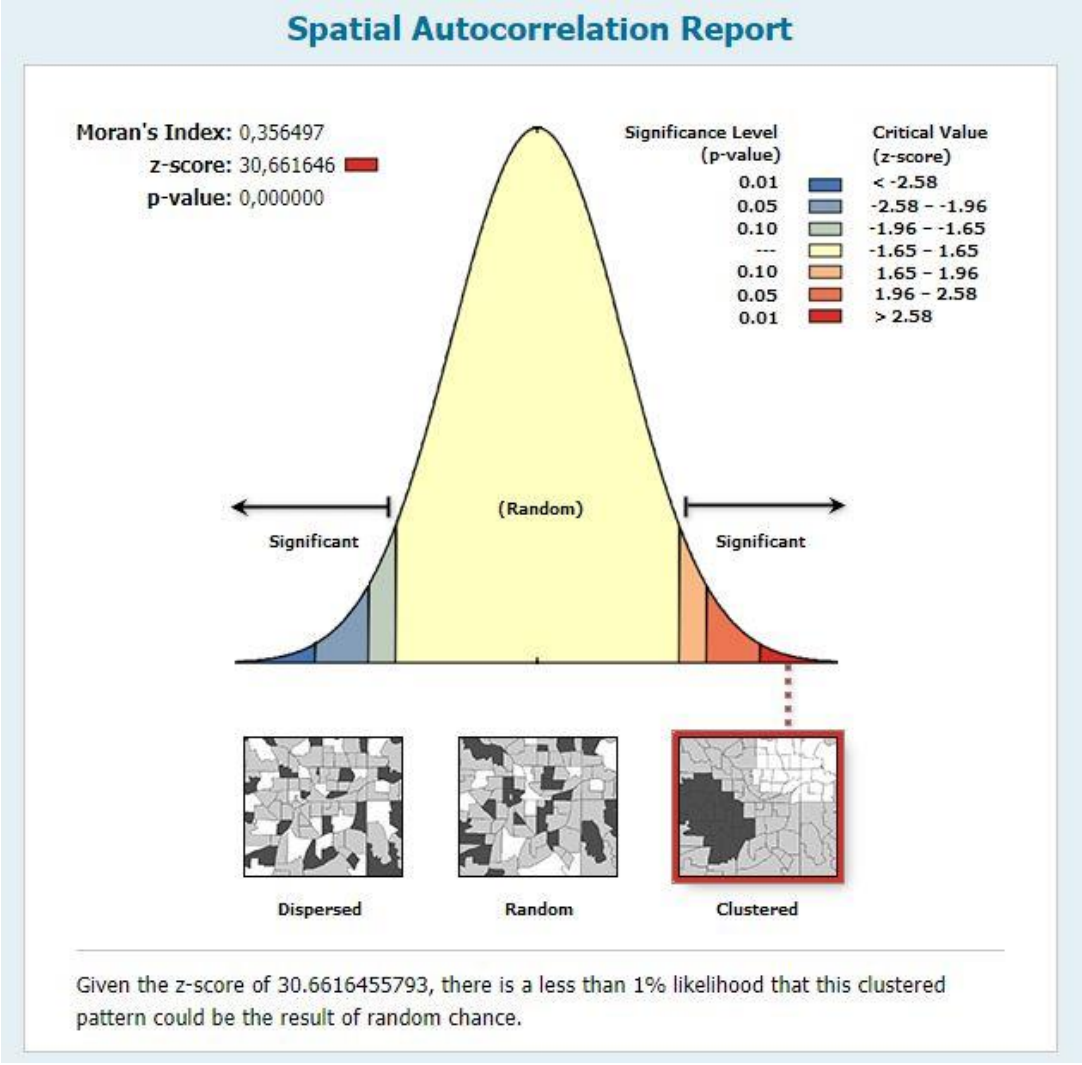

Figure 16: Spatial Autocorrelation – the HTML report. Source: own data processing.

The ArcGIS provides with the Incremental Spatial Autocorrelation tool. The tool is also based on Moran's I Spatial Autocorrelation Index. This tool measures spatial autocorrelation for a series of distances. Then the tool creates a line graph of those distances.

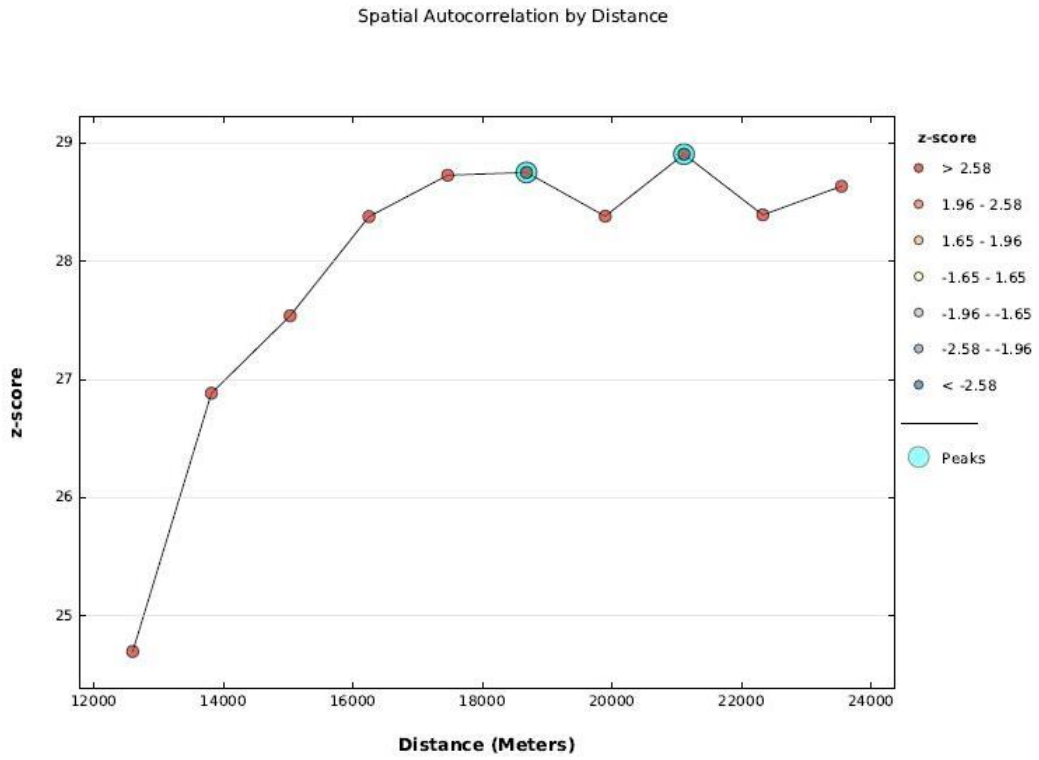

Figure 17: Incremental Spatial Autocorrelation - an output graph. Source: own data processing.

The graph on the figure 14 is used to identify peaks. Here the higher peaks represent distances where the spatial autocorrelation is significantly stronger.

Even though it is not quite regular spatial statistics method but it can be practically used in analyzing patterns to determine at which distance spatial autocorrelation is more significant. The result of the analysis can be interpreted as: the median housing values clustering is more significant on distances of 21 000 m and 19 000 m.

4.3.1.5 Results

All tools described above, tell whether there is clustering or not, but do not determine the location of clusters. Even though the tools managed to perform pattern analysis, the results are interpreted differently.

The Average Nearest Neighbor and the Multi-Distance Clustering Analysis tools techniques, as was mentioned in chapter 3.1.6, are point pattern analyses and based on the distances between points. The ArcGIS did not show any problems with using a polygon type data but a point type data would be more appropriate. Even though the results still differ. The Average Nearest Neighbor calculates distances between the neighboring features while the Multi-Distance Spatial Cluster Analysis tool calculates distances between every feature. Therefore, the Average Nearest Neighbor tool identifies clusters within neighboring features, while the Multi-Distance Spatial Cluster Analysis tool identifies clustering within distances. The Average Nearest Neighbor tool comes in handy, for example, for the analysis of predominance of invasive species. The Multi-Distance Spatial Cluster Analysis tool can be practically used if it is needed to analyze at which distances the distribution is more clustered.

As was mentioned in chapter 3.1.5 The Spatial Autocorrelation and the High/Low Clustering tools techniques refer to global statistics. Those tools determine clustering within values of a selected attribute of the variable. The main difference between those tools is that the High/Low Clustering tool estimates clustering and what is more it determines if high or low values are clustered, while the Spatial Autocorrelation tool measures clustering for all values. The High/Low Clustering tool can be used for the concentration of high values analysis. The Spatial Autocorrelation tool will be appropriate, for example, to determine a concentration or spreading of trends.

4.3.2 Where are the clusters?

It is noticeable that the previous tools could only determine the clusters, while the following tools allow to identify the locations of those clusters.

4.3.2.1 Cluster and Outlier Analysis tool

The first tool that can answer the question is the Cluster and Outlier Analysis tool. It calculates Anselin's local Morans' I (LISA). This technique was described in the paragraph 3.1.5.1.

As an output, there will be created a new shapefile. This shapefile identifies statistically significant cold spots (Low-Low), hot spots (High-High)

and outliers (either Low-High or High-Low). The High-High values represent high-value clusters. The High-Low values represent outliers. In that outliers, the high value is surrounded by low values. The created shapefile also includes information on the z-scores and the p-values. The results of cluster analysis are based on those values. The results of the Cluster and Outlier Analysis tool can be seen on the figure 15.

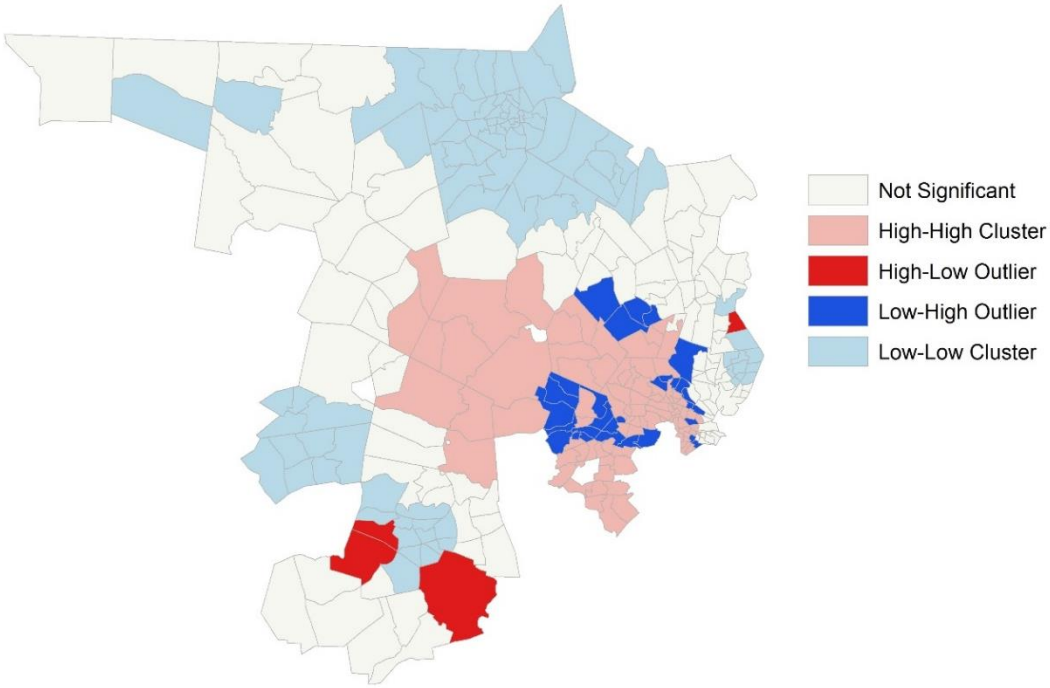

Figure 18: Cluster and Outlier Analysis – the output. Source: own data processing.

4.3.2.2 Hot Spot Analysis tool

The Hot Spot Analysis tool is another tool that can locate clusters. This tool identifies the clusters of high values and the clusters of low values. The tool uses Getis-Ord G* technique which was mentioned in the paragraph 3.1.5.3.

As a tool above, the hot spot analysis tool also gives an output in the form of shapefile which identifies statistically significant cold spots (the clustering of low values) and hot spots (the clustering of high values). Attributes include p-values, z-scores and the number of neighbors. The results can be interpreted based on the z-scores: the larger positive z-score is, the more intense the clustering of high values. The same is with the negative z-scores and low values.

Figure 16 represents the result of the hot spot analysis.

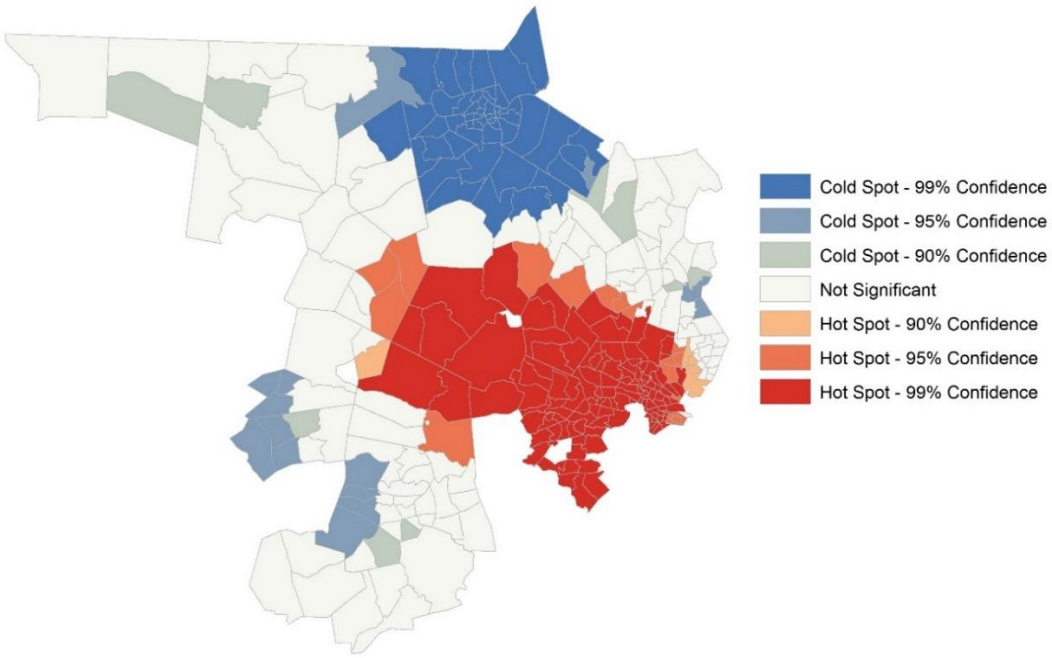

Figure 19: Hot Spot Analysis – the output. Source: own data processing.

4.3.2.3 Results

As was already mentioned the techniques those tools are based on, refer to local statistics. Both tools locate clusters. However, there are differences in results and interpreting.

Both the Cluster and Outlier Analysis tool and the Hot Spot Analysis tool measure clustering of high and low values. The difference is that the Cluster and Outlier Analysis tool also measures outliers. This is very useful when it is needed to find anomalies in patterns. As for the hot spot analysis tool, it will assist in the examination of the intensities of clustering.

4.3.3 Characteristics of the distribution

The Spatial Statistics toolbox also includes a toolset that consists of descriptive statistics tools.

The Measuring Geographic Distributions toolset includes tools that are used to identify the median center, the mean center, the standard distance, and the central feature. The calculations of geographic distributions are described in the paragraph 3.1.3.

The descriptive statistics of the studied data are depicted in figure 17.

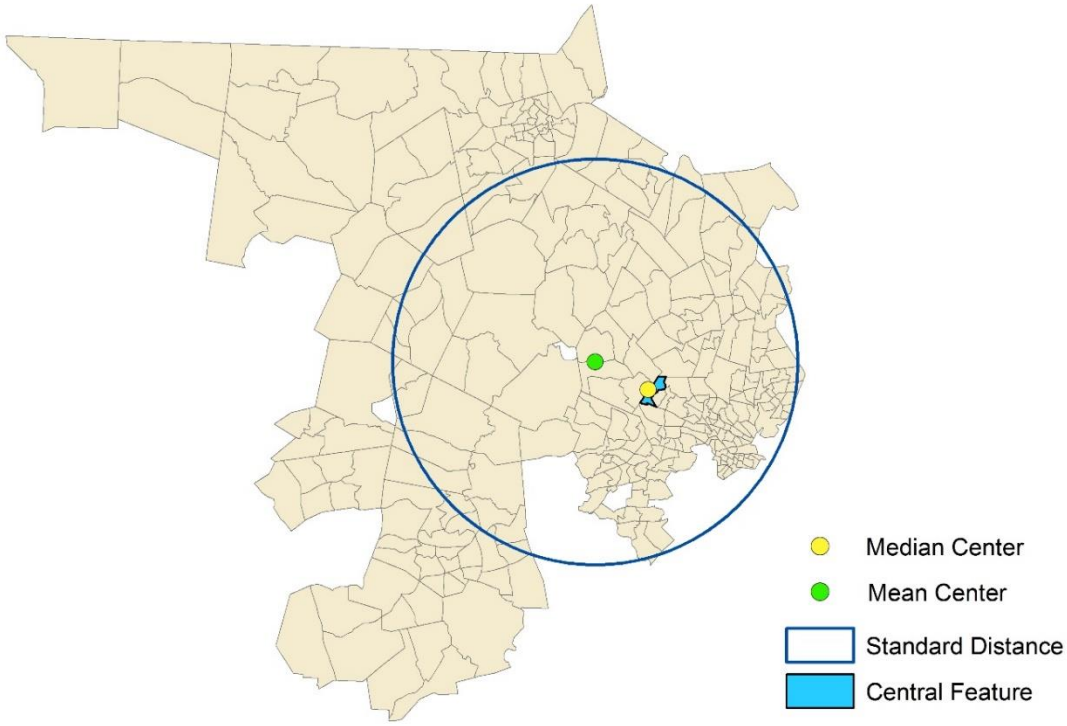

Figure 20: Descriptive statistics of the studied data. Source: own data processing.

The tools for descriptive statistics are used to summarize studying data. Such basic characteristics as central tendency and variability could give a basic picture of the distribution of data before conducting any analysis.

4.3.4 Regression analysis

As was mentioned in chapter 3.1.7 spatial regression answers the question WHY did something happened. ArcGIS provides with spatial regression analysis tools. The following paragraphs will examine the OLS tool and GWR tool and compares those tools.

4.3.4.1 Ordinary Least Squares tool

The Ordinary Least Squares tool provides the linear regression that is used to generate prediction or to model relationships among variables. This tool is situated in the Modeling Spatial Relationships toolset. The OLS technique was described in the paragraph 3.1.7.1.

The tool output consists of the report in pdf format and the map of residuals. The report consists from the summary table, the diagnostics table of

the model, plots of the relationships between the dependent variable and each explanatory variable, and the histogram of standardized residuals.

Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	8,962537	1,781336	5,031357	0,000001*	1,709932	5,241457	0,000000*	--------
RD90	4,587789	0,214570	21,381309	0,000000*	0,291864	15,718950	0,000000*	2,090407
PS90	1,955899	0,205401	9,522349	0,000000*	0,330308	5,921435	0,000000*	1,226016
UE90	-0,524402	0,070028	-7,488488	0,000000*	0,082467	-6,358941	0,000000*	1,892543
DV90	0,461590	0,115173	4,007811	0,000072*	0,115710	3,989207	0,000078*	1,101253
MA90	-0,049482	0,048901	-1,011869	0,311763	0,049898	-0,991655	0,321524	1,263381

Figure 21: The summary table. Source: own data processing.

OLS Diagnostics

Input Features:	south	Dependent Variable:	HR90
Number of Observations:	1412	Akaike's Information Criterion (AICc) [d]:	9008,825986
Multiple R-Squared [d]:	0,309158	Adjusted R-Squared [d]:	0,306701
Joint F-Statistic [e]:	125,839368	Prob(>F), (5,1406) degrees of freedom:	0,000000*
Joint Wald Statistic [e]:	330,817288	Prob(>chi-squared), (5) degrees of freedom:	0,000000*
Koenker (BP) Statistic [f]:	64,758576	Prob(>chi-squared), (5) degrees of freedom:	0,000000*
Jarque-Bera Statistic [g]:	2833,424057	Prob(>chi-squared), (2) degrees of freedom:	0,000000*

Figure 22: The diagnostics of the model. Source: own data processing.

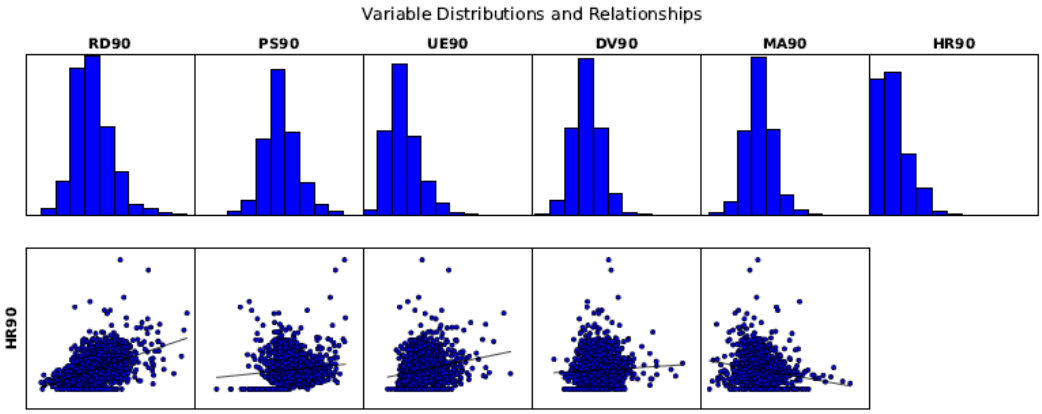

Figure 23: The relationships between dependent variable and each explanatory variable. Source: own data processing.

Figure 18 represents the summary table for each explanatory variable. The column Probability [b] represents a p-value. On the basis of p-value all explanatory variables, except for a Median age variable (MA90), influence on the dependent variable of Homicide Rate.

Figure 19 contains the diagnostic output table. Multiple R-Squared and the Akaike ‘s Information Criterion are measures of model fit. In this example,

Multiple R-Squared is 0.3091. That means the model explains 30.9 % of the variability.

Figure 20 represents histograms showing the distribution of each variable in the OLS model. The scatterplots represent the relationships between dependent variable and each explanatory variable.

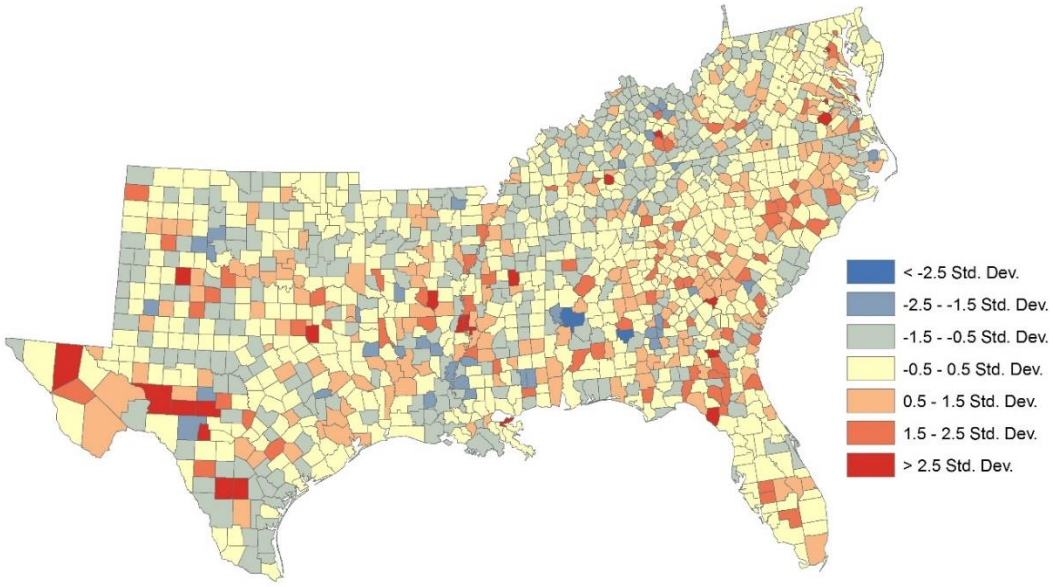

Figure 24: The map of residuals – OLS model. Source: own data processing.

The map of residuals can be then analyzed in order to find clusters. If the analysis shows clustering, then that will be an evidence of a missing of at least one explanatory variable. The Exploratory Regression tool can be used for constructing of more appropriated OLS model.

4.3.4.2 Generate Spatial Weights Matrix tool

This tool is used to construct a spatial weights matrix which was described in the paragraph 3.1.7.2. The spatial weights matrix will be generated in swm format. Then the output matrix can be used in other tools such as Hot Spot Analysis tool.

The swm file format then can be transformed to the table. This process can be done by using the Convert Spatial Weights Matrix to Table tool that can be found in the Utilities toolset.

4.3.4.3 Geographically Weighted Regression tool

This tool is used to construct the GWR model which was described in the paragraph 3.1.7.3. The tool is also situated in the Modeling Spatial Relationships toolset. The GWR tool outputs are similar to the OLS tool outputs, but the report is generated in the form of a dBASE table.

To begin with, the GWR tool returned the error message while building a model. The error was caused by the multicollinearity, the redundancy among explanatory variables. This problem can be solved by removing non-significant variables. Going back to the summary of the OLS model a single non-significant variable was the MA90 variable. Therefore, this variable can be removed.

GWR_supp

OID	VARNAME	VARIABLE	DEFINITION
0	Neighbors	145	
1	ResidualSquares	37466,172664	
2	EffectiveNumber	157,713666	
3	Sigma	5,465392	
4	AICc	8896,719563	
5	R2	0,46407	
6	R2Adjusted	0,39711	
7	Dependent Field	0	HR90
8	Explanatory Field	1	RD90
9	Explanatory Field	2	PS90
10	Explanatory Field	3	UE90
11	Explanatory Field	4	DV90

Figure 25: The summary of the GWR model. Source: own data processing.

The GWR model has explained 46.4 % of the variability based on Multiple R-squared (Figure 22). As is with the OLS model the map of residuals also can be analyzed for clustering for the possibility of missing variables.

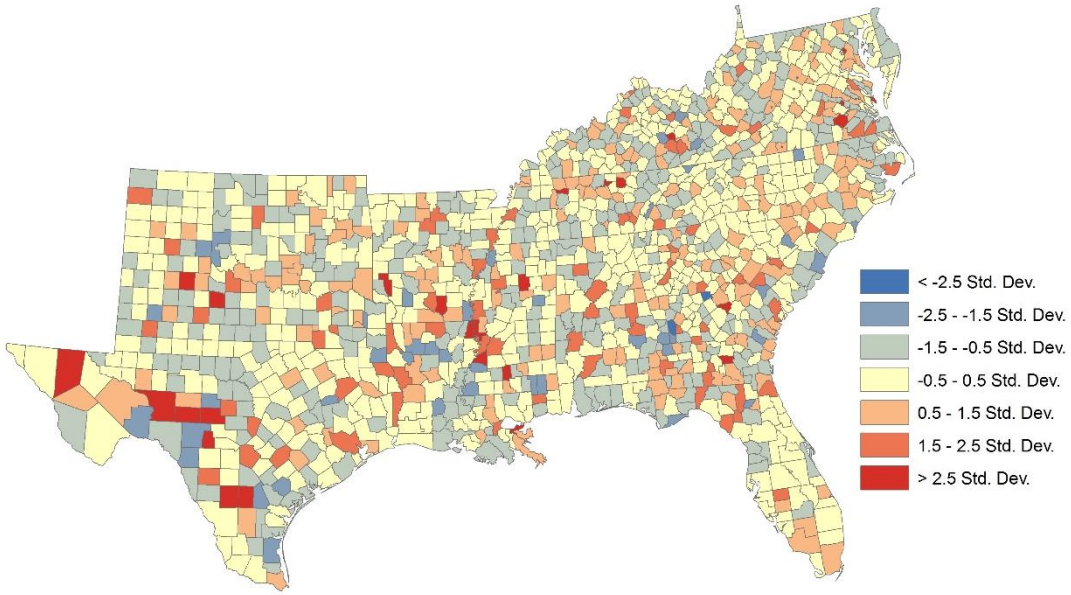

Figure 26: The map of residuals – GWR model. Source: own data processing.

Unlike the OLS tool, the GWR tool does not provide with information about how each explanatory variable influence the dependent variable. Evaluating influences of each explanatory variable on the dependent variable requires some extra actions. For example, it is required to evaluate the influence of unemployment rate on the homicide rate. In that case, the variable of unemployment should be specified as a quantity value on the Symbology menu.

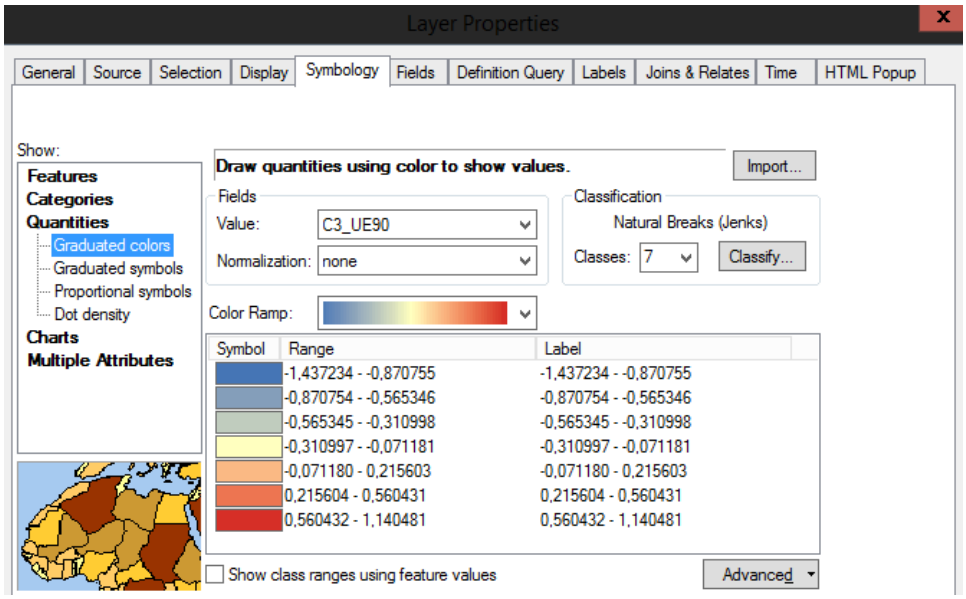

Figure 27: The Symbology menu. Source: own data processing.

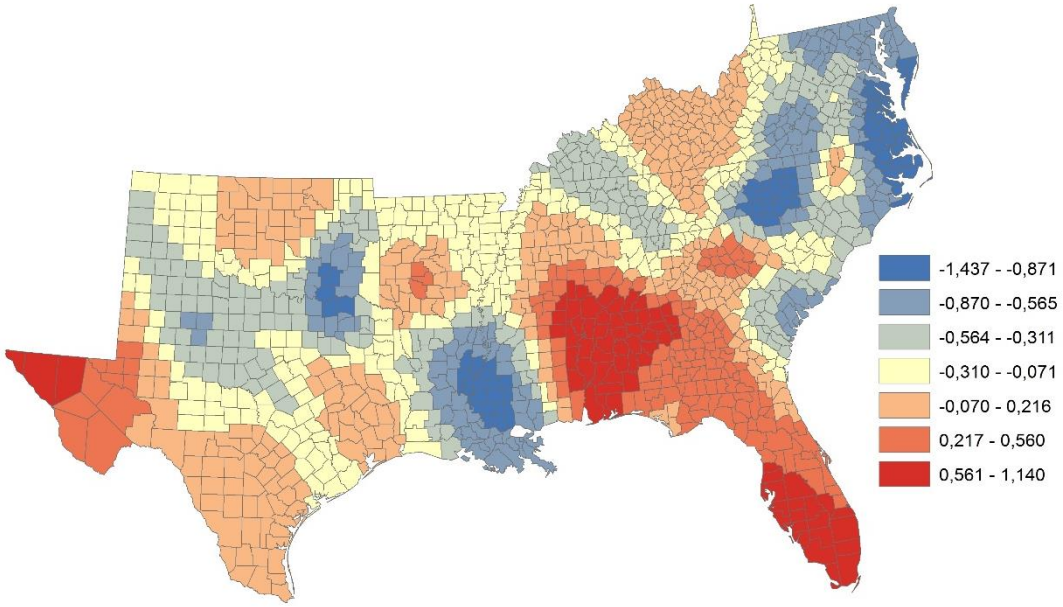

Figure 28:The influence of the unemployment rate on the homicide rate. Source: own data processing.

Figure 25 shows how the relationship between the unemployment rate and the homicide rate changes over the study area. Warm colors represent a strong influence while cold colors represent a weak influence.

4.3.4.4 Results

Both the OLS and the GWR methods explain why does the dependent variable behave this way. Moreover, the regression analysis tools can also be used to predict the behavior of values in the regions of influence of the explanatory variables. The difference between models is in their equations: the OLS is a global model so it establishes one equation for all features (paragraph 3.1.7.1), while the GWR is a local model that creates equations for every feature (paragraph 3.1.7.2).

The OLS tool could be used as an experimental model. For example, if the OLS model does not fit to the data good enough, the best decision will be to move to GWR.

In comparison with the OLS model, the GWR model exceeds in R-square and Akaike's Information Criterion parameters. The OLS model explained 30.9 % of the variability, while the GWR model explained 46.4 % of the variability. Doubtless, the GWR model fits into the data better.

4.4 Geostatistical Analyst Toolbox

Geostatistical Analyst toolbox is another toolbox equipped with spatial statistics functions. The toolbox is also divided into different toolsets. The Geostatistical Analyst toolbox includes the following five toolsets.

- The Interpolation toolset provides with functions which can predict values on unsampled locations. The tools of the Interpolation toolset are: Diffusion Interpolation With Barriers, Empirical Bayesian Kriging, Global Polynomial Interpolation, IDW, Kernel Interpolation With Barriers, Local Polynomial Interpolation, Moving Window Kriging, Radial Basis Functions.
- The next toolset is the Sampling Network Design toolset. Those tools can assist in the placement of new sampling sites. The tools of the Sampling Network Design toolset are: Create Spatially Balanced Points, Densify Sampling Network.
- The Simulation toolset includes tools that are used to perform geostatistical simulations. The tools can assist in the analysis of the results of those simulations afterwards. The tools of the Simulation toolset are: Extract Values To Table, Gaussian Geostatistical Simulations.
- The tools provided by the Utilities toolset assist with pre- and postprocessing when developing interpolation models. The tools of the Utilities toolset are: Cross Validation, Neighborhood Selection, Semivariogram Sensitivity, Subset Features.
- The Working with Geostatistical Layers toolset mainly contains tools used to manipulate with the properties of geostatistical layers. The tools of the Working with Geostatistical Layers toolset are: Areal Interpolation Layer To Polygons, Calculate Z-value, Create Geostatistical Layer, GA Layer To Contour, GA Layer To Grid, GA Layer To Points, Get Model Parameter, Set Model Parameter.

As the Spatial Statistics toolbox, the Geostatistical Analyst toolbox is also can be found in the ArcToolbox window.

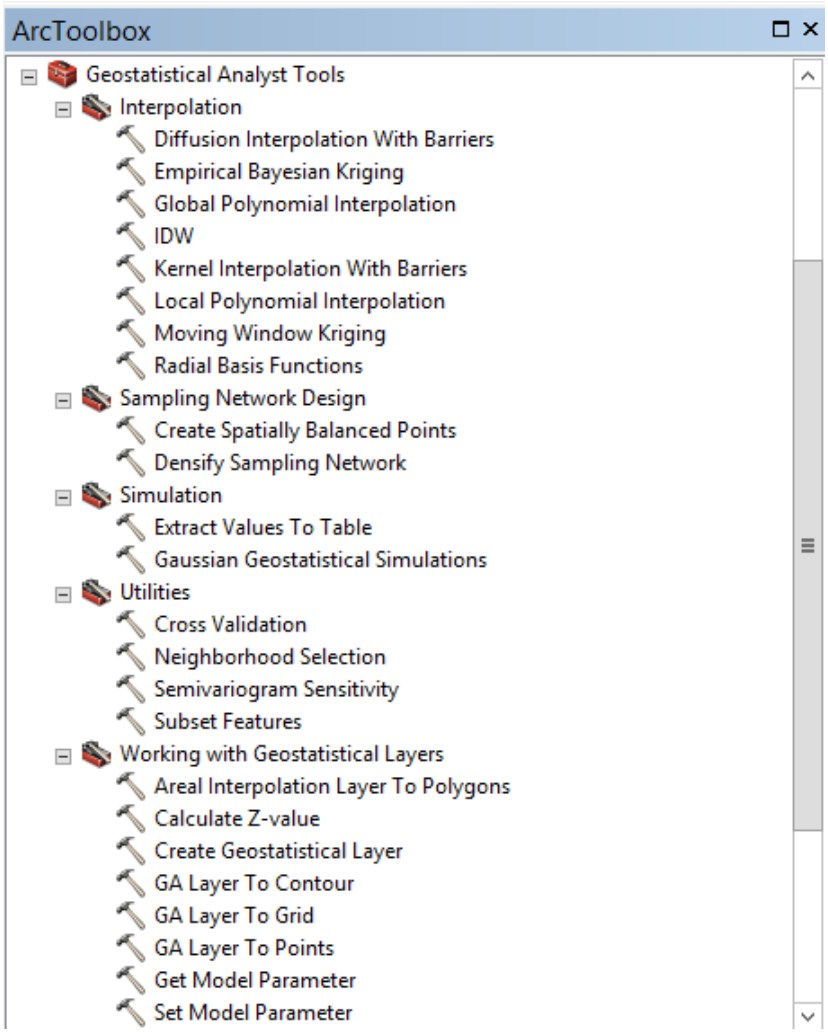

Figure 29: Location of the Geostatistical Analyst toolbox in the ArcToolbox window. Source: ArcMap.

4.4.1 Making predictions

4.4.1.2 IDW tool

The IDW tool uses the Inverse Distance Weighted Interpolation technique to predict values for unsampled locations. The IDW interpolation technique was described in the paragraph 3.1.8.1. As an input was used the heavy_metal_tutorial shapefile. The tool requests the z value which holds values of each point. For the analysis was used a mercury variable. As an output, the new raster file will be created (Figure 27).

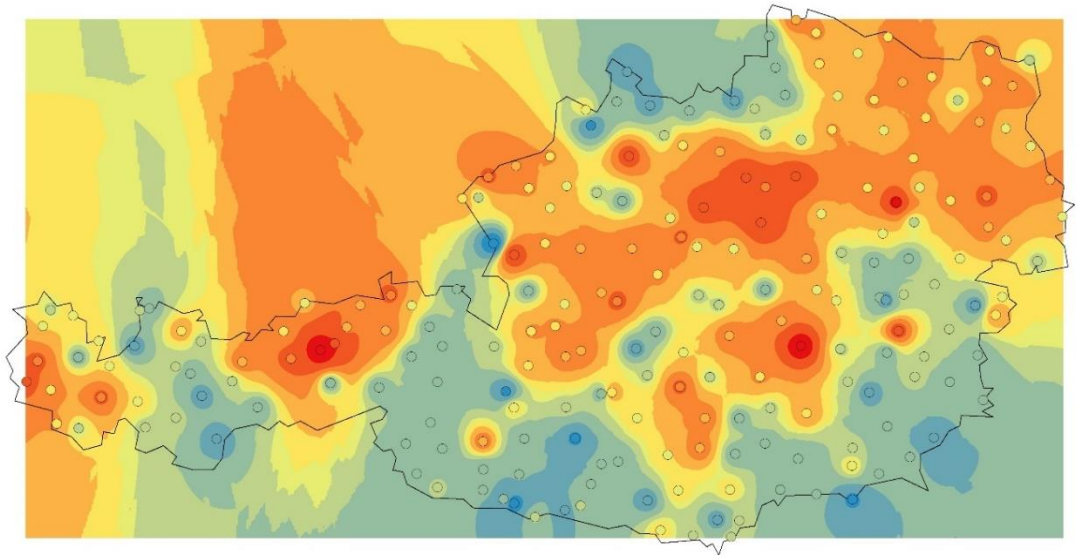

Figure 30: The IDW prediction. Source: own data processing.

The surface contains predictions for unsampled locations. The principles of the IDW tool work are to take the known points and then to create a surface by estimating unknown points. As can be seen, the colors of predicted location coincident with the most colors of the sampled points on the location.

4.4.1.3 Global Polynomial Interpolation tool

This tool is based on the polynomial interpolation technique, the global polynomial interpolation to be precise. This technique was described in the paragraph 3.1.8.2. As an output will be created a raster map. The first-order polynomial was used as an example.

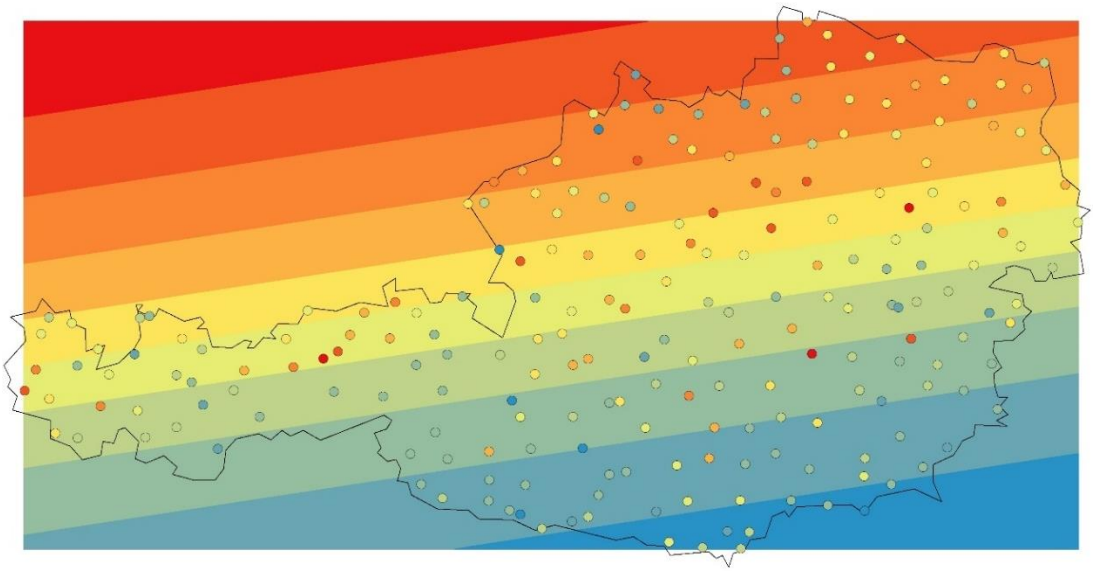

Figure 31: The Global Polynomial Interpolation output. Source: own data processing.

This tool work with the entire data set to identify global trends. It fits a first-order polynomial into the data. It is noticeable that it is hard, even impossible to trace the trends from the output depicted in figure 28.

4.4.1.4 Local Polynomial Interpolation tool

The Local Polynomial Interpolation tool refers to the local polynomial technique which was mentioned in the paragraph 3.1.8.2. For the example, the first-order polynomial was used. The local polynomial interpolation tool applies the same approach as the global method but locally in the smaller neighborhood.

Figure 29 represents the predicted locations.

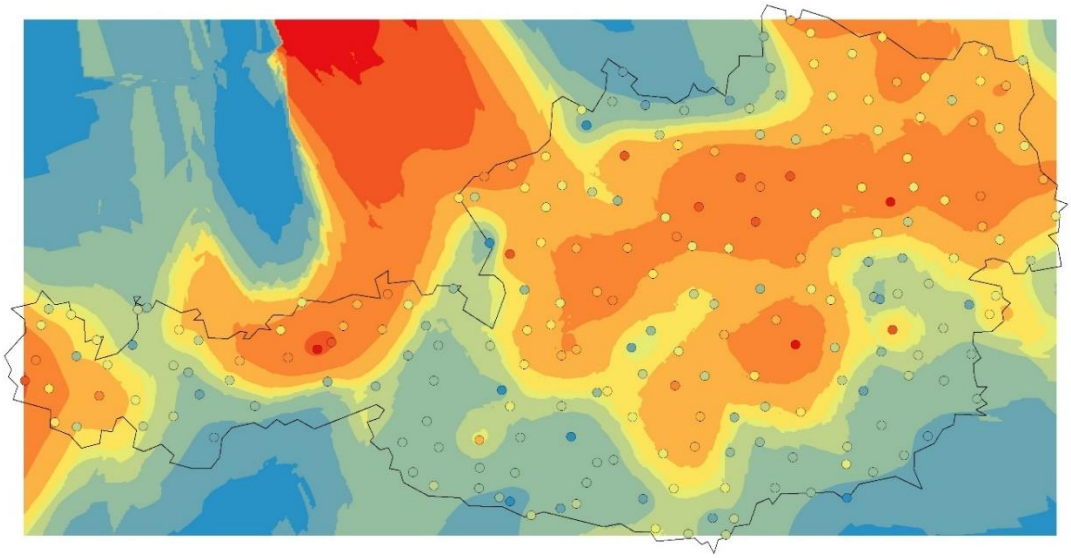

Figure 32: Local Polynomial Interpolation output. Source: own data processing.

The local polynomial interpolation tool works with shorter distances. The idea of local interpolation is that the data closer to the prediction location have larger weights. As can be seen in figure 29 there are some points of high values surrounded by points of low values. For this reason, the nearest predicted location is neither high or low values. Obviously, it models the studied phenomenon better than the global method.

4.4.1.5 Kriging

As was mentioned in the paragraph 3.1.9.2 there are a lot of types of kriging. Here will be examined the Empirical Bayesian Kriging tool.

Kriging is a robust estimator. It considers the distance and the degree of variation between the data, but its effectiveness depends on the correct specification of the semivariogram model. The optimal solution can be reached after selecting appropriate parameters. As a result, the raster map of predictions will be created.

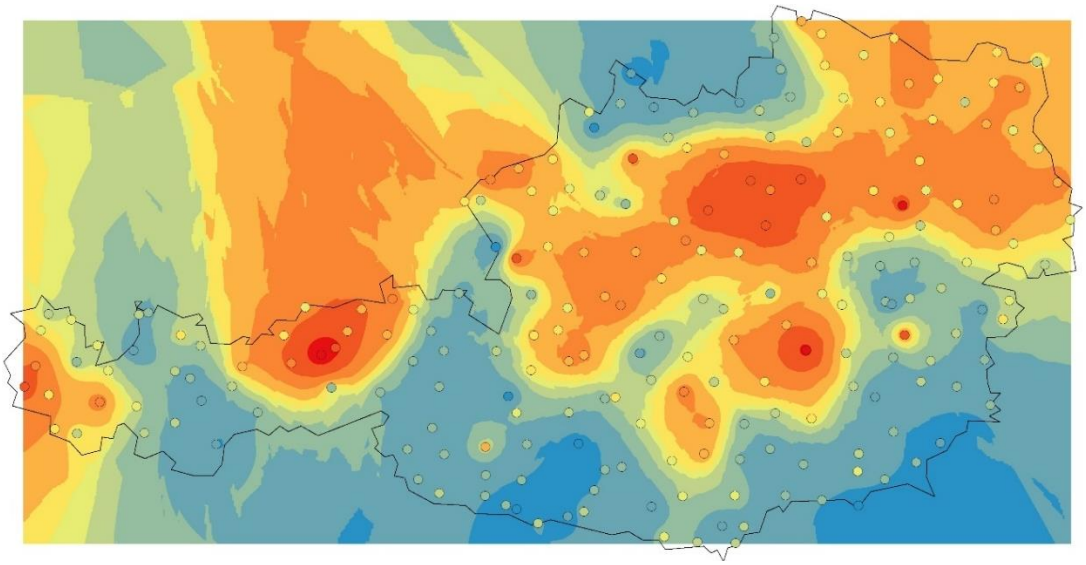

Figure 33: Empirical Bayesian Kriging tool output. Source: own data processing.

As well as the IDW method, kriging also uses measured values at sampled locations to predict values of the unsampled locations. Weights do not depend only on distances, but on the overall spatial distribution of the phenomenon, which needs to be studied and spatial dependency quantified first.

4.4.1.6 Results

In comparison with all examined tools kriging is assumed to be the best prediction method. The IDW tool has a disadvantage in the uncertainty of the prediction. This method does not provide prediction standard errors information. The global polynomial showed the worst prediction between the tools because it is meant to be used on bigger datasets.

The Geostatistcal Analyst toolbox also provides with such geostatistical tools as semivariogram, which was mentioned in the paragraph 3.1.9.1. Most of kriging tools are available only with the Geostatistical Analyst license. Another two basic types of kriging require a Spatial Analyst extension license.

5 Discussion

It is clear from the observations that tools for spatial statistics implemented in the ArcGIS can practically assist in many issues of spatial analysis. The grouping of tools for spatial statistics is user-friendly. It is more convenient to find the most appropriate tool from the prepared set of tools than to scroll through the entire list of tools to find one. Doubtless, the user interface is a strong side of the ArcGIS. It is noticeable that the ArcGIS has its advantages in interpreting the results. In this bachelor thesis three types of outputs were explored: maps, html reports and dBASE tables. Every output provides with essential information that is easy to understand.

Although, not all methods of spatial statistics, which are described in the literature review, are implemented in the ArcGIS. For example, such regression analysis as a spatial error model or a generalized linear model that were mentioned in chapter 3.1.7. Those techniques could assist in building a model that would be a better fit for data. For example, the spatial error regression model is available in the open-source GIS named Geoda (MIT, 2016).

On the other hand, not all those methods must be definitively implemented into the ArcGIS environment. For example, ArcGIS already provides with two measuring spatial autocorrelation methods in its environment, though the difference between those two techniques was described in chapter 4.3.1, so implementing the Geary's C method in the ArcGIS seems unnecessary. It is noticeable, that the ESRI company targets on meeting wider user needs and makes an effort to implement as many spatial statistics techniques in the ArcGIS as possible.

Moreover, in the last two years it became possible to combine the ArcGIS and R. R is an open-source programming language that provides with thousands of functions include spatial statistics. The R-ArcGIS Bridge extended the potential of the ArcGIS in spatial analysis. This extension allows to create custom tools that are powered by R language (ESRI Canada, 2018). Besides, at the beginning of 2019 ESRI released the ArcGIS Pro 2.3. This new software provides with more spatial statistics functions which are not

implemented in ArcMap 10.5.1. ESRI will eventually replace ArcMap with ArcGIS Pro (ESRI, 2018).

6 Conclusion

Doubtless, the path that spatial statistics has made from numbers and equations on paper to powerful computer calculations is fascinating. Manipulation with spatial data became easier by means of integration of spatial statistics techniques into the GIS environment. ESRI made a great effort to build such GIS software within advanced yet user-friendly functions.

This bachelor thesis shows that the ArcGIS is capable to solve a wide spectrum of spatial statistics problems. For sure, ESRI has its huge potential for future development of spatial statistics functions in the ArcGIS environment.

7 Bibliography

BURROUGH, P. A. & MCDONNELL, R. A., 1998. Principles of Geographical Information Systems. Oxford: Oxford University Press.

CHUN, Y. & GRIFFITH, D. A., 2013. Spatial Statistics & Geostatistics: Theory and Applications for Geographic Information Science and Technology. Los Angeles: SAGE.

CRESSIE, N., 1991. Statistics for Spatial Data. New York: John Wiley & Sons.

DAWSEN, C. J., 2011. Geographic Information Systems. New York: Nova Science Publishers.

DE BY, R. A., 2001. Principles of Geographic Information. Enschede: ITC.

DE SMITH, M. J., GOODCHILD, M. F., LONGLEY, P., 2007. Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools. Leicester: Troubador Publishing Ltd.

DECKER, D., 2001 GIS Data Sources. New York: John Wiley & Sons.

DEMERS, M. N., 2009. Fundamentals of Geographic Information Systems. New York: J. Wiley.

DEMYANOV, V.V., SAVYELYEVA, E. A., 2010. Geostatistics: Theory and Practice. Moscow: Nauka.

ESRI, ©2004. ArcGIS 9. What is ArcGIS? [Online]. Redlands: ESRI Press. [cit. 2018-10-12]. Available at: <http://downloads.esri.com/support/documentation/ao_/698What_is_ArcGis.pdf>.

ESRI, ©2017. ArcGIS Desktop: ArcGIS 10.5.1 Help. Redlands, CA: Environmental Systems Research Institute.

ESRI Canada, ©2018. Tutorial for ArcGIS Pro 2.2 / ArcGIS Desktop 10.6.1: Getting Started with the R-ArcGIS Bridge [Online]. [cit. 2019-3-1]. Available at: <https://esricanada-ce.github.io/r-arcgis-tutorials/1-Getting-Started.pdf>

ESRI, ©2018. ArcGIS Pro tool reference—ArcGIS Pro | ArcGIS Desktop [Online]. [cit. 2019-3-1]. Available at: <https://pro.arcgis.com/en/pro-app/tool-reference/main/arcgis-pro-tool-reference.htm>

ESRI, ©2012: Welcome to the ArcGIS Help Library [Online]. [cit. 2018-10-15]. Available at: <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html>.

FAZAL, S., 2008. GIS Basics. New Delhi: New Age International.

GAETAN, C. & GUYON, X., 2010. Spatial Statistics and Modeling. New York: Springer New York.

GELFAND, A. E., DIGGLE, P. J., MONTSERRAT, F. & GUTTORP, P., 2010. Handbook of Spatial Statistics. Boca Raton: CRC Press.

GETIS, A., 1999. Spatial Statistics. Geographical Information Systems, Volume 1, pp. 239-251.

GIMOND, M., ©2018: Intro to GIS and Spatial Analysis [Online]. [cit. 16.12.2018]. Available at: <https://mgimond.github.io/Spatial/index.html>.

GORROOCHURN, P., 2016. Classic Topics on the History of Modern Mathematical Statistics: From Laplace to More Recent Times. Hoboken: John Wiley & Sons.

GRIFFITH, D. A. & CHUN, Y., 2018. GIS and Spatial Statistics/Econometrics: An Overview. In Comprehensive Geographic Information Systems. Amsterdam: Elsevier.

HARMON, J. E. & ANDERSON, S. J., 2003. The Design and Implementation of Geographic Information Systems. Hoboken: John Wiley & Sons.

HARVEY, F., 2018. A Primer of GIS: Fundamentals of Geographic and Cartographic Concepts. New York: Guilford Press.

HEYWOOD, I., CORNELIUS, S. & CARVER, S., 2006. An Introduction to Geographical Information Systems. Harlow: Pearson Prentice Hall.

HUNG, M. -C., 2016. Applications of Spatial Statistics. Rijeka: InTech.

KIRKEGAARD, E. O. W., 2015. Some Methods for Measuring and Correcting for Spatial Autocorrelation [Online]. [cit. 15.12.2018]. Available at: <https://thewinnower.com/papers/2847-some-methods-for-measuring-and-correcting-for-spatial-autocorrelation>.

KRIVORUCHKO, K., 2011. Spatial Statistical Data Analysis for GIS Users. Redlands: ESRI Press.

KRIVORUCHKO, K. & GOTAY, C. A., 2003. Using Spatial Statistics in GIS. International Congress on Modelling and Simulation, pp. 713-736. ESRI.

LEVINE, N., 2010. CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incidents Locations (v 3.3) – manual. Houston: Ned Levine & Associates. [cit. 17.12.2018]. Available at: <https://www.icpsr.umich.edu/CrimeStat/download.html>.

LI, J. & HEAP, A. D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Canberra: Geoscience Australia. [cit. 6.01.2019]. Available at: <https://pdfs.semanticscholar.org/686c/29a81eab59d7f6b7e2c4b060b1184323a122.pdf>.

LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. & RHIND, D. W., 2016. Geografické Informace: Systémy a Věda. Olomouc: Univerzita Palackého v Olomouci.

MEULEN, M. van der., 2000. Definitions for Hardware and Software Safety Engineers. London: Springer-Verlag.

MITAS, L. & MITASOVA, H., 1999. Spatial Interpolation. Geographical Information Systems: Principles, Techniques, Management and Applications, Volume 1, pp. 481-492.

MIT, ©2016: Introduction to Spatial Statistics [Online]. [cit. 14.12.2018]. Available at: <https://ocw.mit.edu/resources/res-str-001-geographic-information-system-gis-tutorial-january-iap-2016/spatial-statistics/MITRES_STR_001IAP16_spati.pdf>.

NASSER, H., 2015. ArcGIS By Example. Birmingham: Packt Publishing Ltd.

PILZ, J., 2009. Interfacing Geostatistics and GIS. Berlin: Springer.

ROGERSON, P. A., 2015. Statistical Methods for Geography: A Student's Guide. Los Angeles: SAGE Publications.

SAHOO, P. M., 2013. Statistical Techniques for Spatial Data Analysis. New Delhi: IASRI.

SCOTT, L. M. & JANIKAS, M. V., 2010. Spatial Statistics in ArcGIS. Berlin: Springer.

TOBLER, W. A., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, Volume 46 (sup1).

UNWIN, D., 1996. GIS, Spatial Analysis and Spatial Statistics. Progress in Human Geography, Volume 20, pp. 540-551.

XU, Y. & KENNEDY, E., 2015. An Introduction to Spatial Analysis in Social Science Research. The Quantitative Methods for Psychology, Volume 11, pp. 22-31.