Spatial Analysis in Geomatics

3

Andreas Redecker, Jaroslav Burian, Nicolai Moos, and Karel Macků

Abstract

For processing geodata there are many different approaches of which all of them require their own specific input data and parameters to generate an outcome that suits the respective case of application. This chapter introduces the most common analyses that are conducted using a GIS. From basic tools like buffering certain vector geometries or merging operations of two different datasets to interpolating area wide raster datasets out of point data there is a huge variety of different toolsets that can be applied when using geodata. To understand why and how these toolsets are utilised, how they are parametrized and which other things are important to make proper use of all the different possibilities these toolsets are providing, this chapter sums up the analyses in reasoned groups and illustrates the many different approaches of spatial analyses through proper examples and depictions.

Keywords

Spatial analysis · Network analysis · GIS · Spatial statistics · Raster resolution · Geoprocessing · Data conversion

A. Redecker (✉) · N. Moos
Geography, Geomatics Group, Ruhr-University Bochum, Bochum, Germany
e-mail: andreas.redecker@rub.de; nicolai.moos@rub.de

J. Burian · K. Macků
Department of Geoinformatics, Palacký University Olomouc, Olomouc, Czech Republic
e-mail: jaroslav.burian@upol.cz; karel.macku@upol.cz

3.1 Simple Spatial Analysis (by Andreas Redecker)

This sub-chapter gives an overview of fundamental GIS methods for performing basic spatial analysis with feature data. Nevertheless, depending on the data involved and the workflow incorporating these methods, they can deliver highly valuable output. The process of manipulating geodata is called geoprocessing. To automate workflows all operators that are involved in an analysis can be combined with a geoprocessing model.

3.1.1 Selections

In many cases, not all features of a feature class are supposed to take part in an analysis. The selection of the desired objects can be performed based on the attributes of the features or incorporating their spatial characteristics. Depending on the GIS used, these two different methods can be applied successively or in one process.

3.1.1.1 Select by Attribute

This method is like selecting datasets in a database using a so-called WHERE-clause of the very common Structured Query Language (SQL). "The WHERE clause is used to extract only those records that fulfill a specified condition" (w3schools.com 2018) according to the feature's

V. Pászto et al. (eds.), *Spationomy*, https://doi.org/10.1007/978-3-030-26626-4_3

properties stored in the attribute table of a feature class (Fig. 3.1).

3.1.1.2 Linking Tabular Data

If the necessary properties for an attribute-based selection are not held in the attribute table of the feature class itself, they can be linked to it from external tabular data. For this, both tables need to contain a field (column) with matching entries. These must uniquely identify a feature in the attribute table as well as its corresponding data in the table to be linked.

3.1.1.3 Select by Location

The spatial approach for selecting features needs a second feature class whose features locations or extents determine which features of the original feature class will be selected. For this, the desired spatial relationship (e.g. intersect, contain, within a distance, etc.) and distance (optional) need to be specified (Fig. 3.2).

Expression:

Name LIKE 'fagus sylvatica'

AND

Age >= 50

ID	Name	Age
1	quercus robur	30
2	fagus sylvatica	79
3	acer pseudoplatanus	76
4	sorbus aucuparia	25
5	fagus sylvatica	37
6	quercus robur	88
7	fagus sylvatica	51

Fig. 3.1 Example of an attribute-based selection. (Source: Authors)

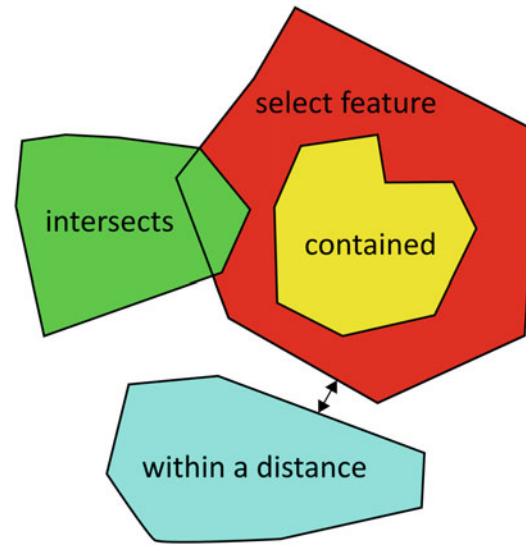

Fig. 3.2 Location-based selection methods. (Source: Authors)

3.1.2 Single Feature Class Operations

To prepare features for further analysis or to better visualise results, two major operations are available to change the structure of single feature classes.

3.1.2.1 Buffer

A Buffer is a proximity function, describing an equidistant line around a feature. Therefore, the resulting geometry type of a buffer operation always is a polygon – no matter if the input was a point, line or polygon-type feature class. The distance value for the construction of buffers around the features in a feature class is either defined by a single value or derived from an individual property in the attribute table for every single feature (Fig. 3.3).

3.1.2.2 Dissolve

The Dissolve operation consolidates the features in a feature class. Based on similar values in a specified attribute field it merges the geometries (if no attribute is specified, all features will be merged). With some dissolve operators at the same time, other attributes of the features merged get aggregated by previously specified statistical functions (mean, sum, min, max, count, etc.) (Fig. 3.4).

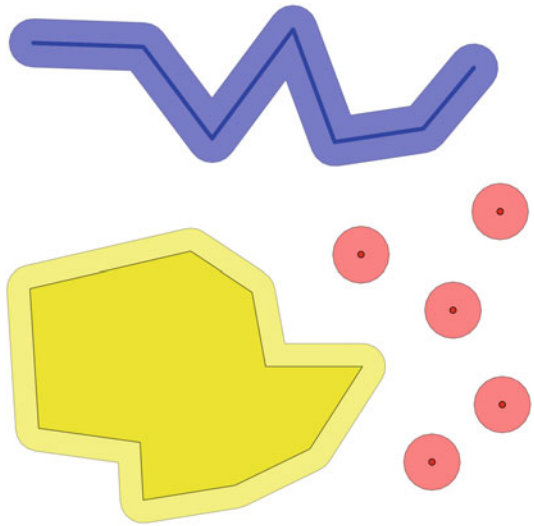

Fig. 3.3 Buffers with point-, line- and polygon-features. (Source: Authors)

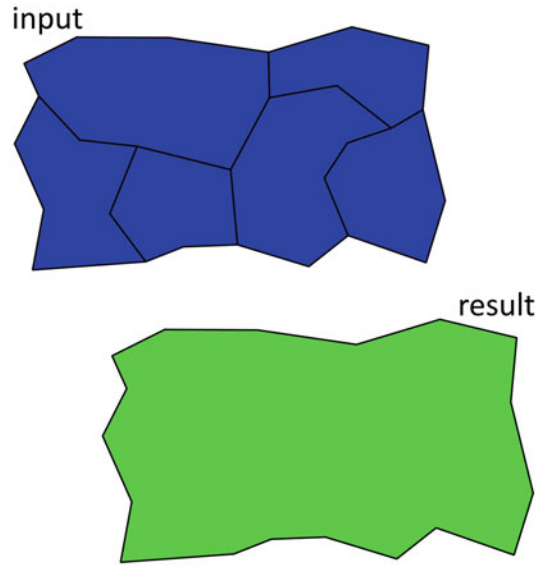

Fig. 3.4 Schematic example of a simple dissolve operation. (Source: Authors)

3.1.3 Overlay Operations

These operations combine two or more feature classes to gain new geo-datasets incorporating the extent of the features involved.

3.1.3.1 Clip

The clip-function creates a subset of features by cutting the features of one feature class by the polygon-features in another feature class. Only those parts of the features in the input layer that overlap with the polygons of the clipping layer will end up in the resulting feature class. It is often used to reduce the extent of a geo-dataset to that of the study area (area of interest, AOI) represented by a polygon feature. The attributes of the remaining features will not be changed (Fig. 3.5).

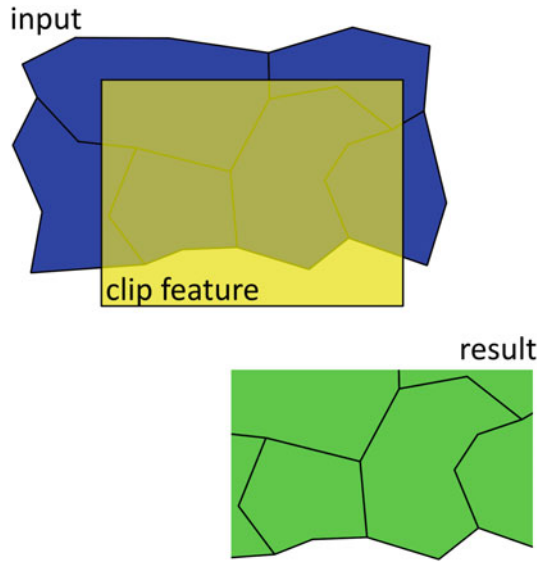

Fig. 3.5 Schematic example of a clip operation. (Source: Authors)

3.1.3.2 Difference

The difference-function also cuts the features of one feature class by the polygon features in another feature class. Only those parts of the features in the input layer that do not overlap with the polygons of the cut feature will end up in the resulting feature class. The attributes of the remaining features will not be changed (Fig. 3.6).

3.1.3.3 Union

This operator combines the polygon features of two or more feature classes. It does not create

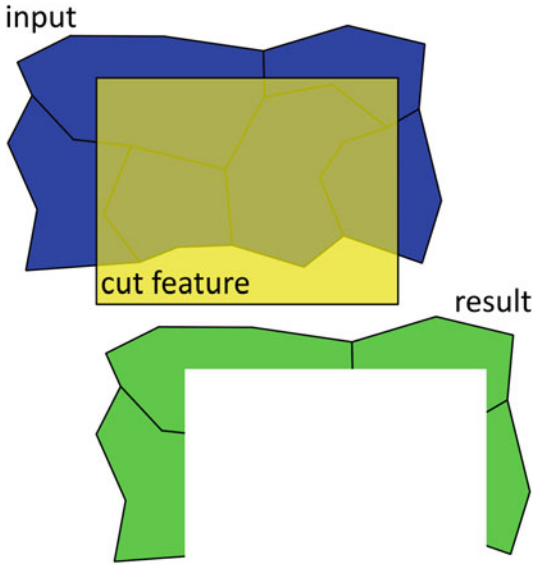

Fig. 3.6 Schematic example of a difference operation. (Source: Authors)

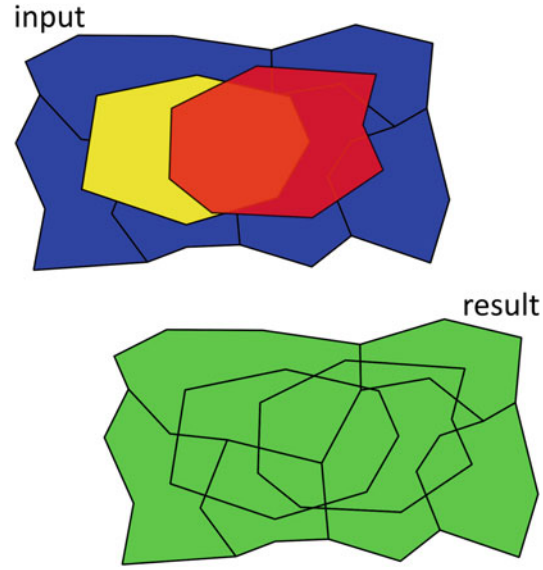

Fig. 3.7 Schematic example of a union operation. (Source: Authors)

overlapping features. Instead, it splits overlapping parts of features to subarea features and assigns the attributes of all involved objects to the new feature.

This spatial operation compares to the logical disjunction (OR) (Fig. 3.7).

3.1.3.4 Intersect

This operator combines the polygon features of two or more feature classes. Only those parts that are covered by a feature in every contributing feature class will be written to the result. The function does not create overlapping features. Instead, it clips the overlapping areas and assigns the attributes of all involved objects to the new feature. This spatial operation compares to the logical conjunction (AND) (Fig. 3.8).

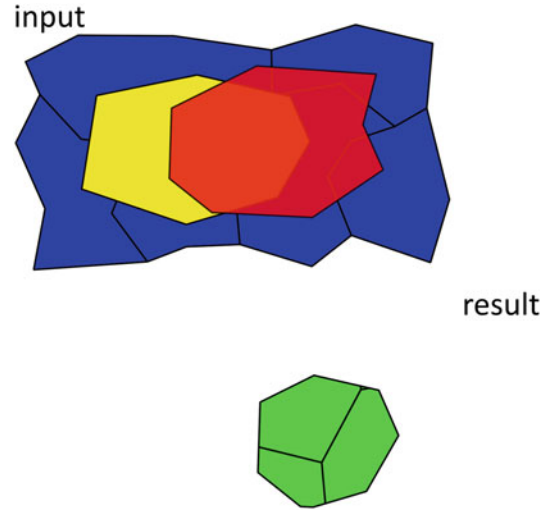

Fig. 3.8 Schematic example of an intersect operation. (Source: Authors)

3.1.3.5 Symmetrical Difference

The result of this function only contains those areas of the input features, that do not overlap. Hence it gives the same result as a union operation minus the result of an intersect.

This spatial operation compares to the logical exclusive disjunction (XOR) (Fig. 3.9).

3.2 Raster Analysis (by Jaroslav Burian)

Raster analysis (as part of spatial analysis) refers to the analytical operations with raster data. Map algebra (mathematical operations with rasters) is

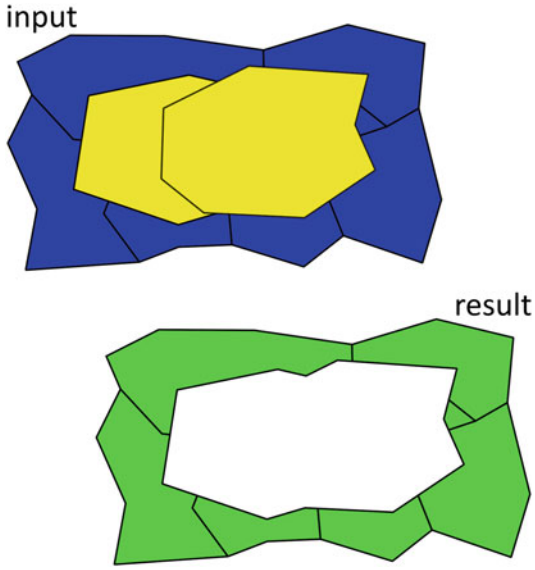

Fig. 3.9 Schematic example of a symmetrical difference operation. (Source: Authors)

used to processing this data. There exist many raster analysis options in GIS like hydrologic analysis, multi-criteria analysis, terrain analysis, surface modelling, surface interpolation, suitability modelling, statistical analysis, and image classification (processing of remote sensing data). Most of the application fields cover environmental issues (e.g. climatic change, weather forecasting, flood modelling) but there are also some focused on economic aspects (e.g. modelling of renewable energy potential, land suitability modelling, cost-distance analysis and many others).

3.2.1 Raster Data

As mentioned in Chap. 1, vector and raster data models are two main ways for geographic data representation. In a raster representation space is divided into an array of rectangular (usually square) cells (pixels). All geographic variation is then expressed by assigning properties or attributes to these pixels (Longley et al. 1999). The most significant characteristics of the raster is its spatial resolution that can be expressed as the length of a cell side as measured on the ground. As shown in Fig. 3.10, cell size can vary from centimetres (some aerial images) to kilometres (satellite images). The spatial resolution is a key characteristic that influences the quality and detail of any raster analysis. Higher spatial resolution leads to higher detail but also increases needed storage capacity and computational time.

3.2.2 Map Algebra

Mathematical operations that can be performed with rasters are referred to as raster or map algebra. Map algebra (also known as cartographic modelling) was defined by Dana Tomlin (Tomlin and Berry 1979; Tomlin 1994) as the informal computational language, that is the basis for raster data processing. Simply said, map algebra is the math applied to raster data. To formalise that, Tomlin defined raster operators and raster functions. Map algebra can be represented by arithmetic or simple analytical operations that are performed with one or more input raster layers (grids). In most software packages, the set of these features is referred to as a map or raster calculator (sometimes grid analysis) (Fig. 3.11).

3.2.3 Raster Operators

As part of the map algebra, operators and functions of mathematical language are used for data processing. Operators perform mathematical calculations with one or more raster layers. The basic type of operators are arithmetic operators (+, −, ∗, /). It is possible to add, subtract, multiply, divide, or perform the same single layer operations. In addition to arithmetic operators, there are Boolean operators (true, false), relational (greater than, smaller than or equal to), statistical (minimum, maximum, average and median), trigonometric (sine, cosine, tangent, arcsine), exponential and logarithmic.

3.2.4 Raster Functions

Tomlin (1994) classifies all GIS transformations of rasters into four basic classes, and it is used in several raster-centric GISs as the basis for their

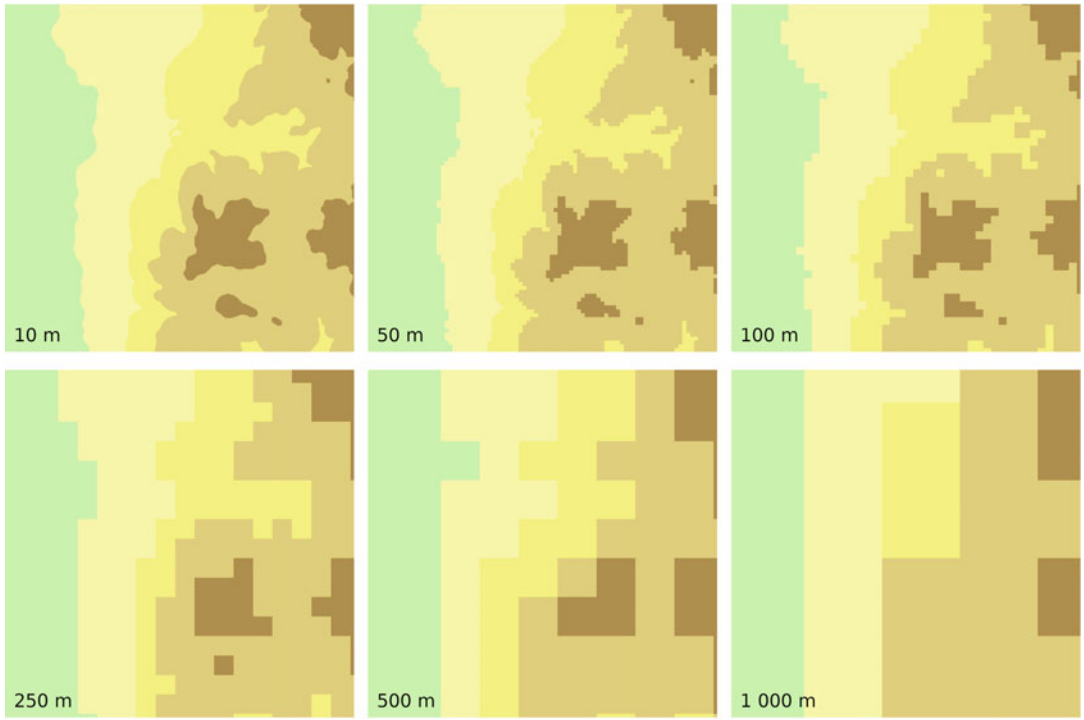

Fig. 3.10 Different cell size (10, 50, 100, 250, 500, 1000 m). (Source: Authors)

analysis languages. Depending on whether the functions work with only one raster cell or more, they are divided into local, focal, zonal, and global. Map algebra functions follow some rules (spatial resolution, the same coordinate system, mathematical operators) to combine all of its components.

3.2.4.1 Local

Local functions are always performed with one specific raster cell but in the entire grid. Using these functions, a new raster cell value is calculated from the values in one or more information layers. An example of a local function may be a simple combination of two raster layers (e.g. the combination of flood risk and earthquake risk) or multiplication of one raster layer by a specific value (e.g. prediction of the average temperature) (Fig. 3.12).

3.2.4.2 Focal

For focal functions, as with local functions, a new value is determined for each cell separately. However, it is calculated from the values in the defined area (neighbouring cells). The most common is the closest cell (3x3), but it can also be a larger area (square, triangle, circle, 4x4 matrix, etc.). On the principle of focal functions, the basic method of slope calculation works. For each cell in the defined area, the altitude difference is calculated from which the resulting gradient slope is calculated. The similar procedure is also used for aspect calculation and many hydrological modelling. Another example from the economic field is modelling of the city growth that uses cellular automata based on focal functions (Fig. 3.13).

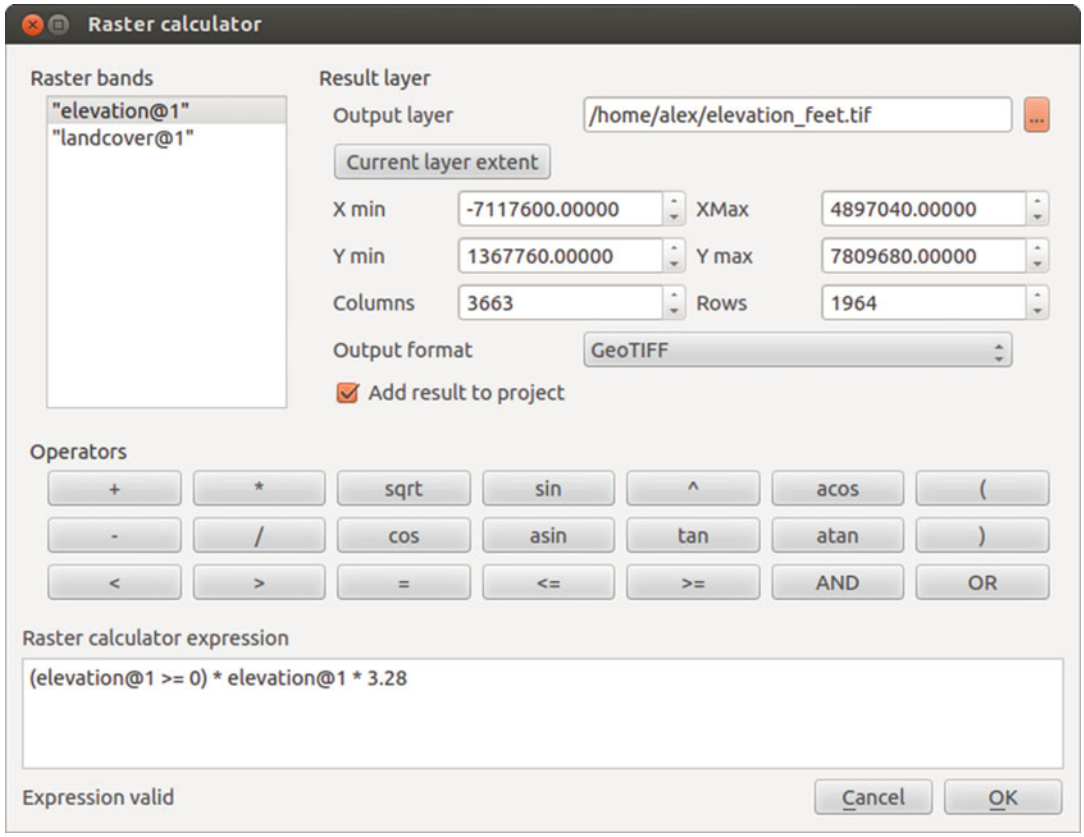

Fig. 3.11 Raster calculator in QGIS software. (Source: Authors)

3.2.4.3 Zonal

Calculation of zonal function is similar to the focal functions. The main difference is that the neighbourhood is defined by another raster (zonal) layer. The new value is calculated for each cell from the values that belong to the zone defined by another layer. The focal function is often applied to the calculation of a certain statistical indicator (average, median, etc.) for irregular areas defined in another layer (e.g. average altitude for individual forest areas or average highway accessibility for city districts) (Fig. 3.14).

3.2.4.4 Global

Global functions are performed from all grid cells. The result of global functions is usually several selected cells that meet the set conditions. They are mostly focused on distance analysis in the form of friction surfaces. An example of a global function may be to find the optimal route in a raster from A to B. For example, each cell in the entire raster represents the value of the friction (water, rock, forest – higher value, meadow, field – smaller value). The entire raster is then analysed to find the lowest cost path when moving from A to B (Fig. 3.15).

3.2.5 Selected Raster Analysis

Raster operators and raster functions can be applied to many different raster datasets to perform a wide range of raster analysis. For the purpose of this book, only a few selected analysis are described.

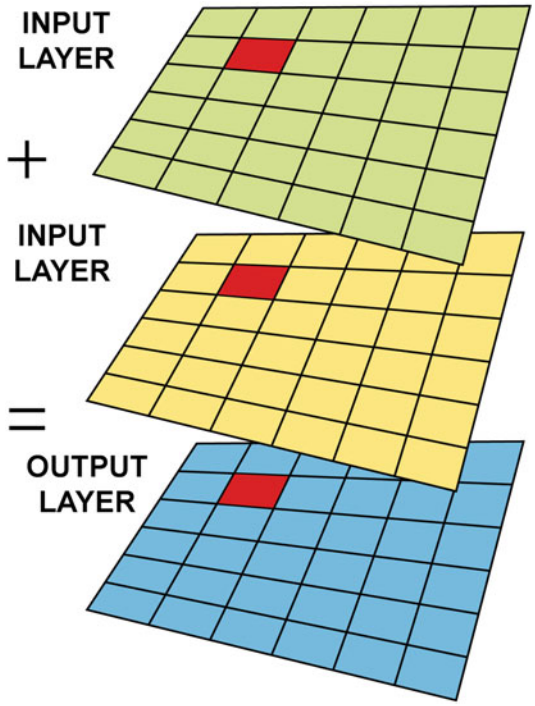

Fig. 3.12 Scheme of local function. (Source: Authors)

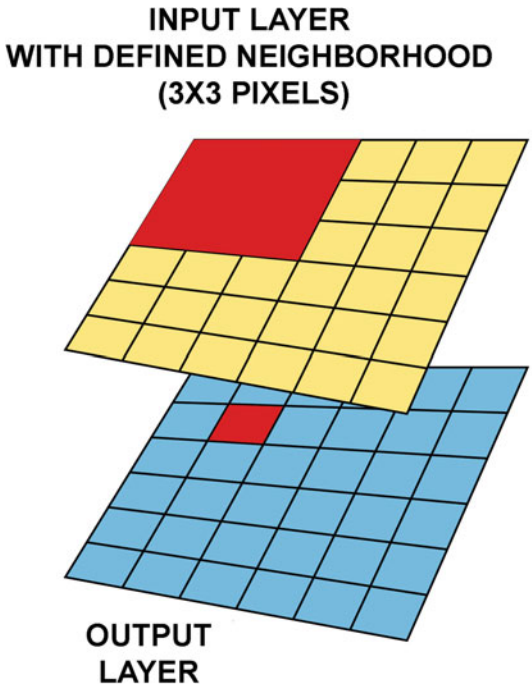

Fig. 3.13 Scheme of focal function. (Source: Authors)

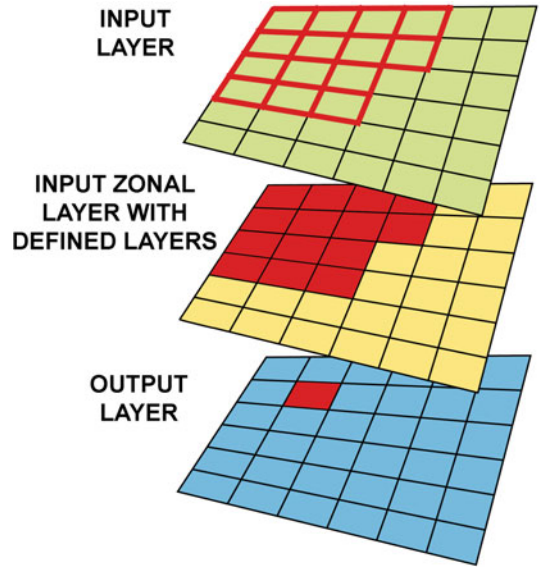

Fig. 3.14 Scheme of zonal function. (Source: Authors)

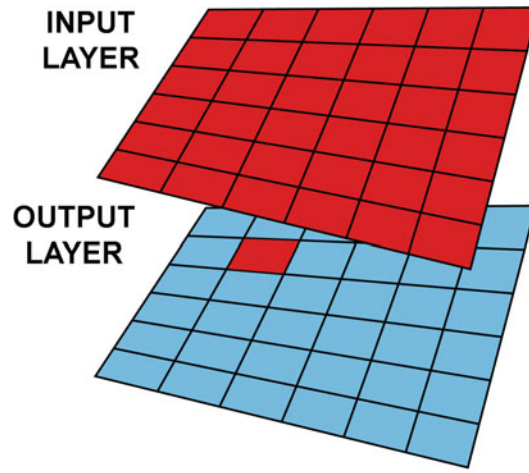

Fig. 3.15 Scheme of global function. (Source: Authors)

3.2.5.1 Resampling

To perform any raster analysis, input raster layers must have the same spatial resolution and coordinate system. Simply said, pixels (cell centres) have to match each other. To manage that several resampling methods are used. It means that one of the input rasters is resampled to the same resolution as another input layer. Original raster values are recalculated to the new ones based on nearest

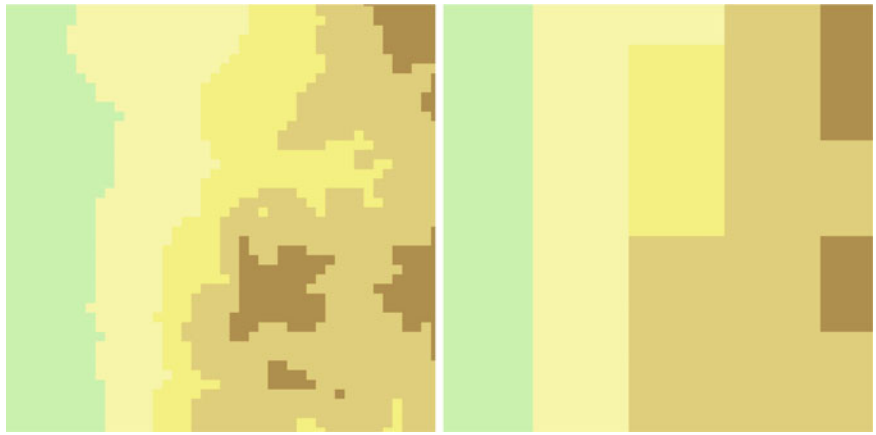

Fig. 3.16 Example of resampling from 100 to 1000 m. (Source: Authors)

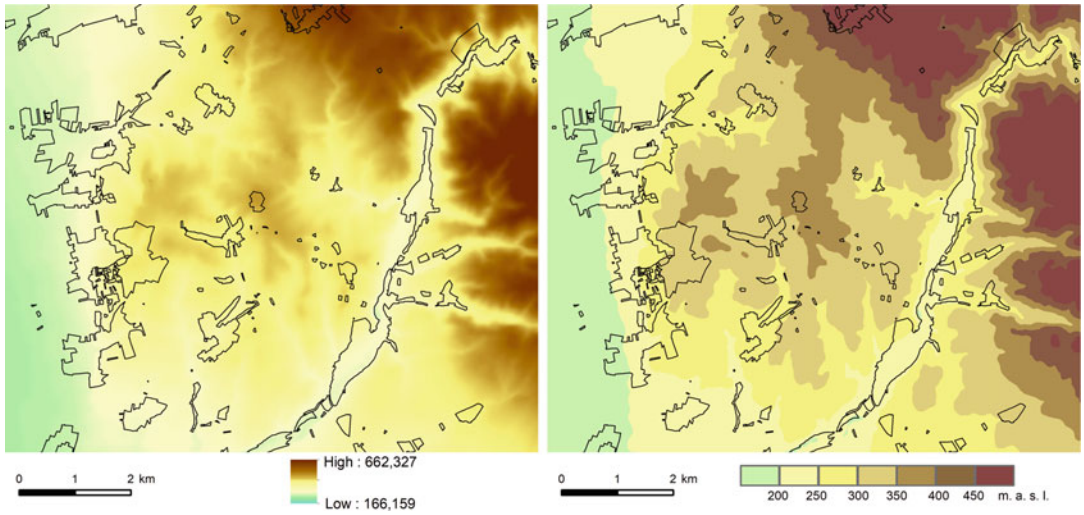

Fig. 3.17 Example of reclassification of elevation. (Source: Authors)

neighbourhood method (or other methods like bilinear, cubic convolution or majority) (Fig. 3.16).

3.2.5.2 Reclassification

One of the most common simple analysis is the reclassification. Reclassification is the process of reassigning a value, a range of values, or a list of values in a raster to new output values. In the case of continuous data (e.g. elevation, temperatures) reclassification creates a new raster with discrete values (several elevation zones – lowland, highland, etc.; or temperature zones). In the case of categorical data (e.g. 20 categories of land-use), reclassification creates a new raster with new discrete values (e.g. only 5 categories of land-use). Reclassification is a key process when you need to combine different data using a common value scale (Fig. 3.17).

3.2.5.3 Surface Analysis

Many raster analysis deal with the surfaces. Surfaces represent phenomena that have values at every point across their extent. Surfaces are derived from a limited set of sample values (e.g. elevation points, meteorological stations). A typical surface represents elevation, temperature, precipitation and many other continuous phenomena's. Surfaces can be represented by

contour lines, points or TINs (triangulated irregular networks); however most surface analysis in GIS is done on raster data.

Spatial Interpolation

There exist several ways to create surfaces. Spatial interpolation is the most common way to do it. Interpolation creates a continuous surface from discrete samples with measured values (point layer mostly). There exist several interpolation methods with a variety of parameters that influence the resulting surface. Each method is suitable for different data set (different phenomena with different spatial distribution). The most common interpolation methods are kriging, natural neighbours, spline and IDW (inverse distance weighting). Figure 3.18 shows different surfaces using the same input point elevation data.

The surface analysis involves several kinds of processing, including extracting new surfaces from existing surfaces, reclassifying surfaces, and combining surfaces (ESRI 2018a). The most common surface analysis (slope, aspect, hillshade, viewshed and watershed) are applied to the elevation data (terrain surfaces – digital elevation models).

3.2.5.4 Slope

The slope represents the rate of maximum change in z-value (elevation) from each cell. The slope is calculated as the maximum rate of change in values between each cell and its neighbours. The neighbourhood can be defined by 4 or 9 neighbouring cells, and there exist several methods for slope calculation. The most common method uses 3 × 3 cell neighbourhood. The slope

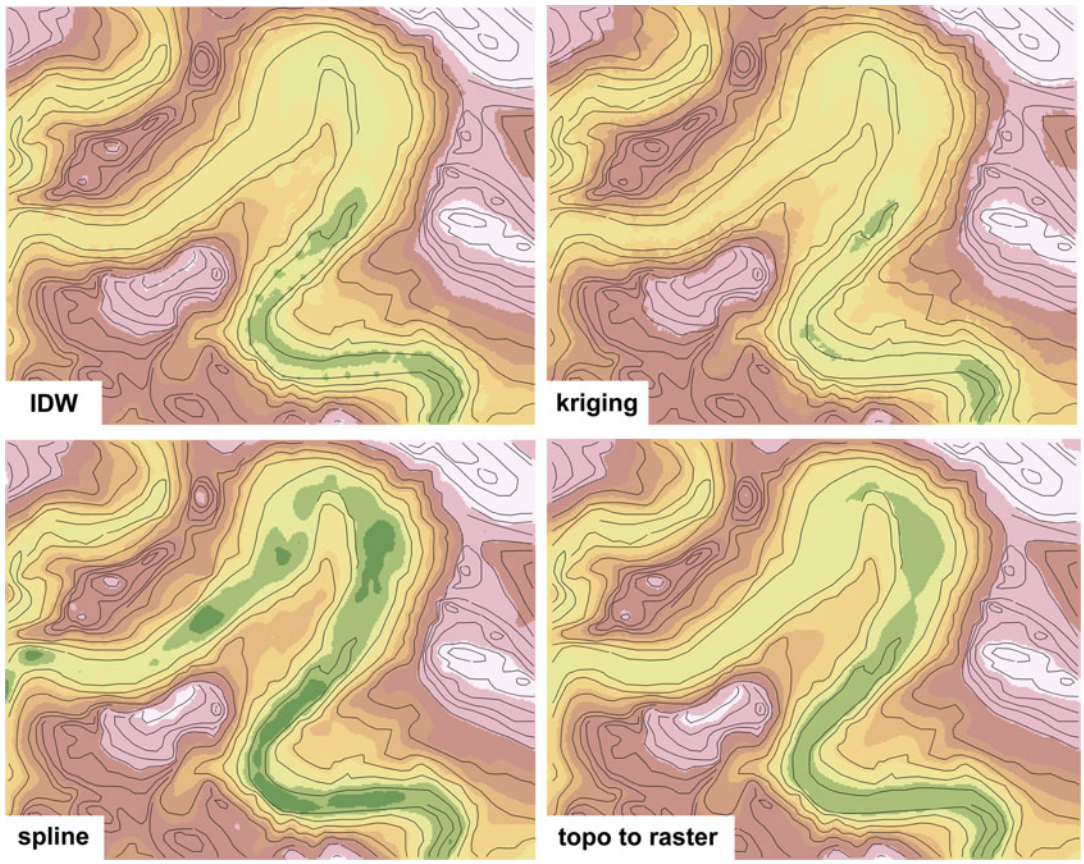

Fig. 3.18 Example of different interpolation methods. (Source: Authors)

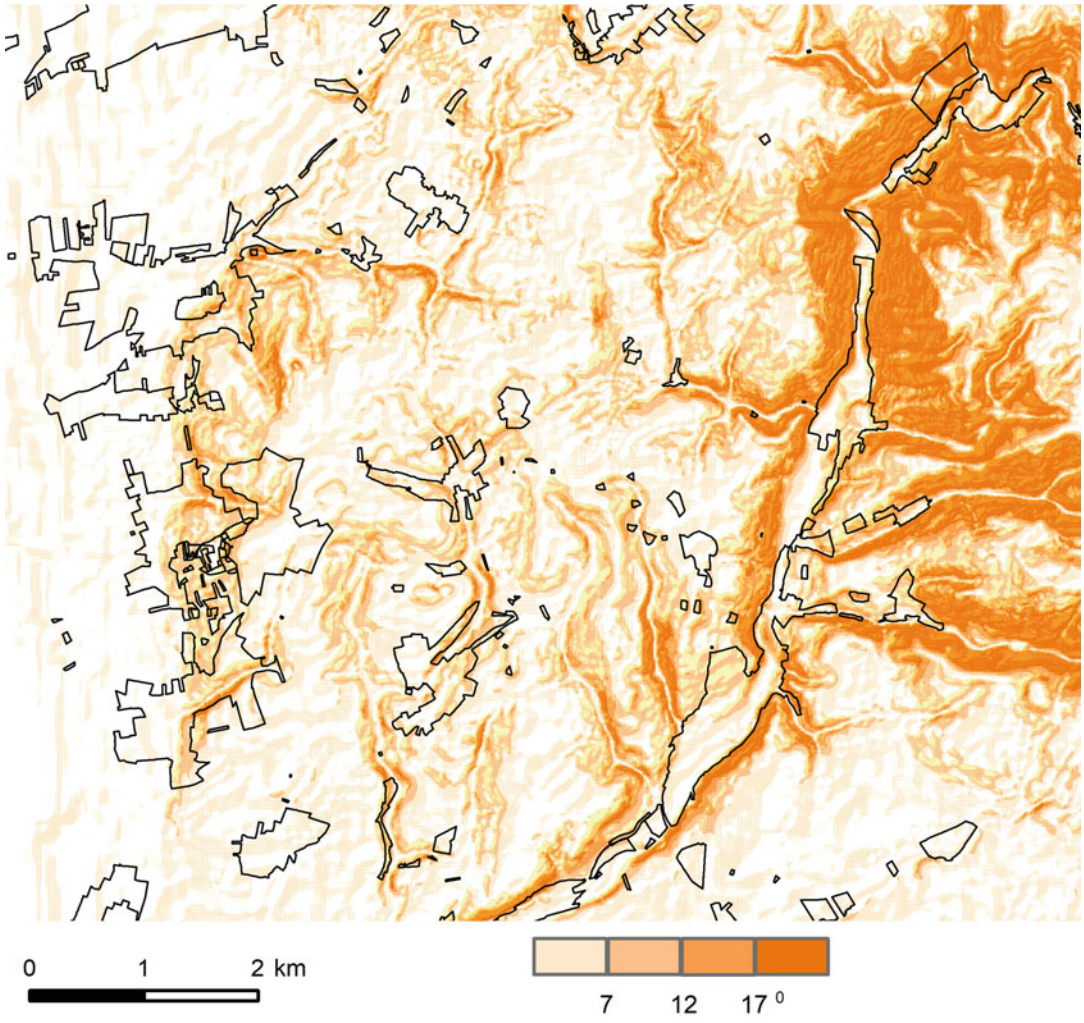

Fig. 3.19 Example of slope. (Source: Authors)

may be expressed as either degree (e.g., 45°) or percent (e.g., 50%). Information about slope can be used for location analysis in urban planning to find suitable places for new development (Fig. 3.19).

3.2.5.5 Aspect

The aspect identifies the orientation or direction of slope. Aspect is the down-slope direction of a cell to its neighbours. The cell values in an aspect grid are compass directions ranging from 0 to 360. North is 0, and in a clockwise direction, 90 is east, 180 is south, and 270 is west. Input grid cells that have 0 slope (flat areas) are assigned an aspect value of −1 (Albrecht 2005). Similarly, to slope analysis, aspect can be used for suitability and location analysis too (Fig. 3.20).

3.2.5.6 Hillshade (Illumination)

Hillshading is a technique used to create a realistic view (shades) of terrain by creating a three-dimensional surface from a two-dimensional display of it. Hillshading creates a hypothetical illumination of a surface by setting a position for a light source and calculating an illumination value for each cell based on the cell's relative orientation to the light or based on the slope and aspect of the cell (Albrecht 2005). Hillshades are often

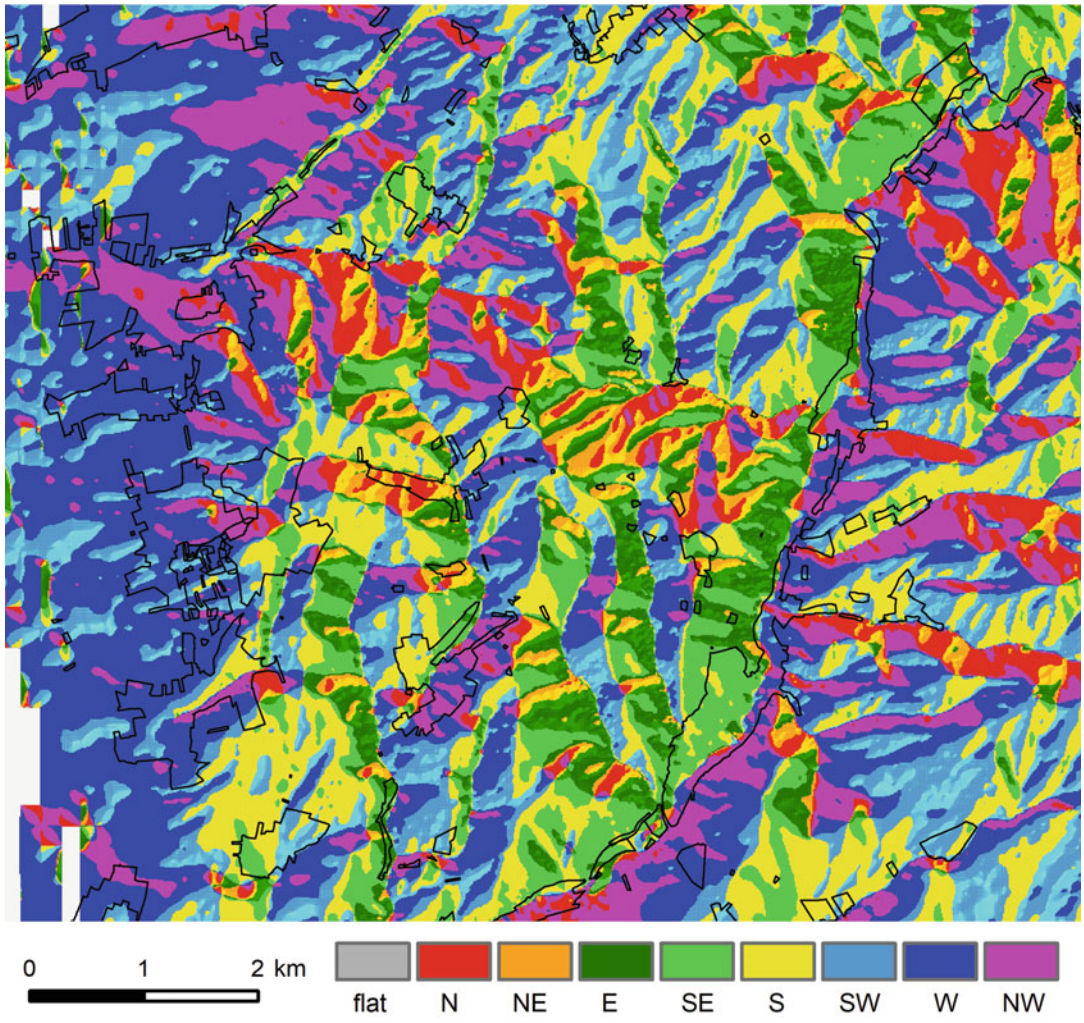

Fig. 3.20 Example of aspect. (Source: Authors)

used to increase the quality and readability of maps (Fig. 3.21).

3.2.5.7 Viewshed (Visibility Analysis)

The viewshed analysis identifies the cells in an input raster that can be seen from one or more observation points or lines. Each cell in the output raster receives a value that indicates how many observer points can see the location (Albrecht 2005). This raster analysis has a wide range of usage and applications. It can be used to determine the aesthetic impact of new city development (e.g. new houses), or for the placement of communications towers (if the direct visibility is needed) or optimal placement of a new lookout tower (Fig. 3.22).

3.2.5.8 Cost Distance Analysis (Least-Cost Path)

The cost distance analysis elaborates movement over continuous space, in which the cost of moving through any location is variable. Cost surface represents some factor or combination of factors that affect travel across an area (e.g. high values for steep terrain, low values for flat areas). In the second step, the least-cost path analysis

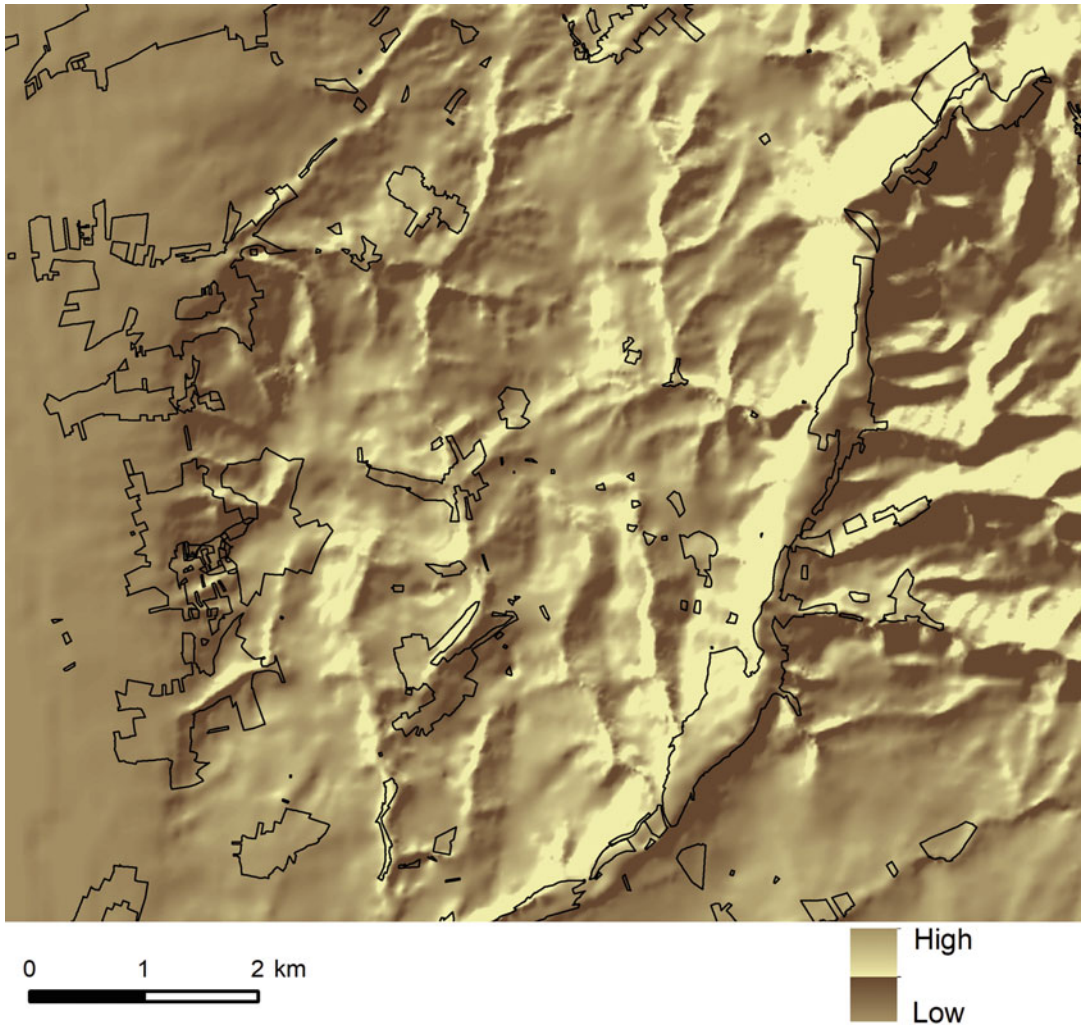

Fig. 3.21 Example of hillshade. (Source: Authors)

uses the cost-weighted distance and direction surfaces for an area to determine a cost-effective route between a source and a destination. This can be used e.g. for the planning of new highways to find the cheapest solution (Fig. 3.23).

3.2.5.9 Solar Radiation (Insolation) Analysis

The solar radiation analysis enables to calculate the amount of the solar energy over a geographic area for specific periods. It accounts for atmospheric effects, site latitude and elevation, steepness (slope) and compass direction (aspect), daily and seasonal shifts of the sun angle, and effects of shadows cast by surrounding topography (ESRI 2018a). Information about the amount of insolation is helpful for application in many fields, such as civil engineering, economy or agriculture research. It may be useful in localisation of a new site for a ski resort, wine yard or solar panels (Fig. 3.24).

3.2.5.10 Multi-Criteria Analysis

Multi-criteria analysis (MCA) is a method used to consider many different criteria when making a decision. In GIS, MCA is represented by overlay

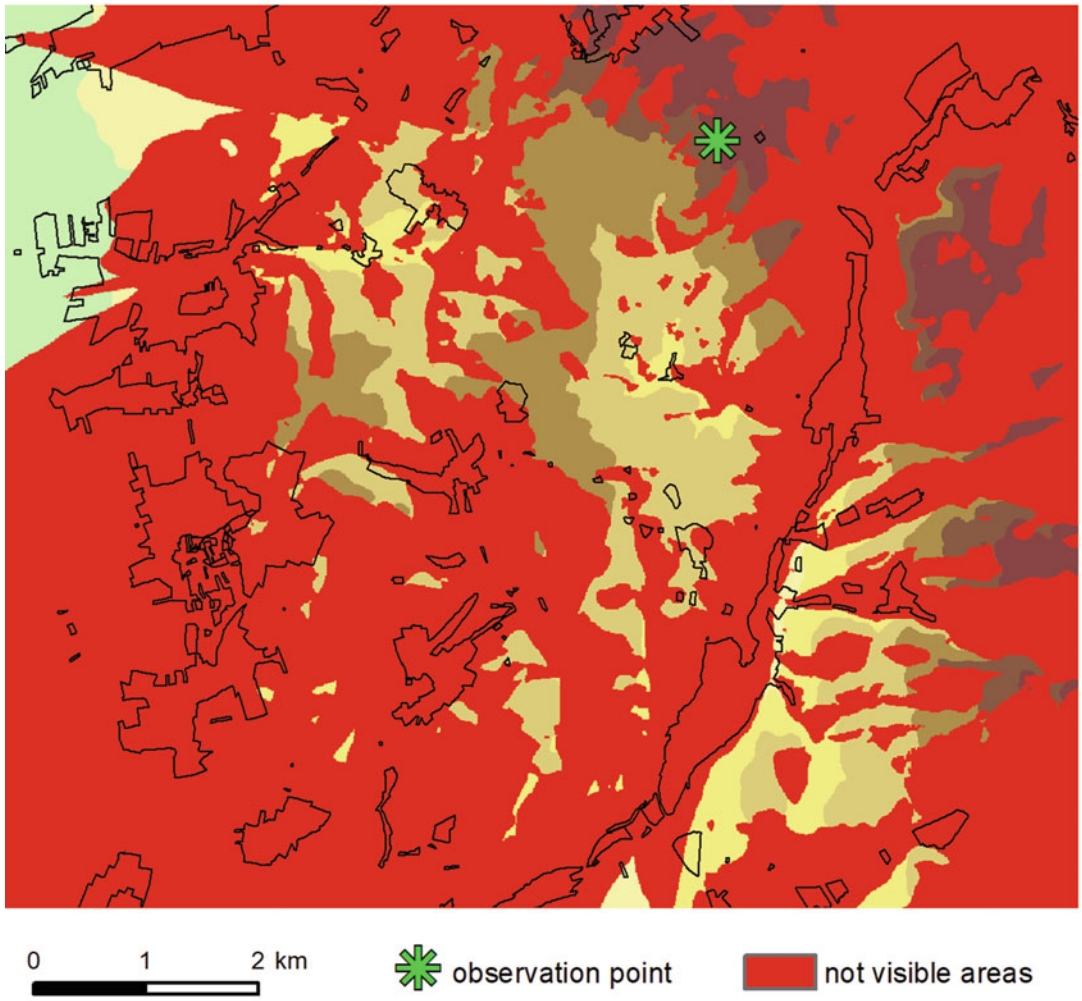

Fig. 3.22 Example of viewshed analysis. (Source: Authors)

analysis (weighted overlay) that overlays several rasters using a common measurement scale and weights each according to its importance. In this case, each criterion (or map layer) is brought to a common scale (reclassified) to simplify the process of combining the layers. Spatial MCA is used for decisions with a geographical factor (suitability analysis, location analysis), where multiple factors need to be considered (e.g. land-use, distances to public transportation, shops accessibility, park accessibility, etc.). In Fig. 3.25, you can see an example of multi-criteria analysis, that combines environmental (green), social (red), and economic suitability (blue) to obtain total land suitability for new housing development (dark green).

3.3 Network Analysis (by Nicolai Moos)

3.3.1 Introduction

Most people are familiar with using a navigation system, which means that they have at least once processed a basic network analysis by looking for

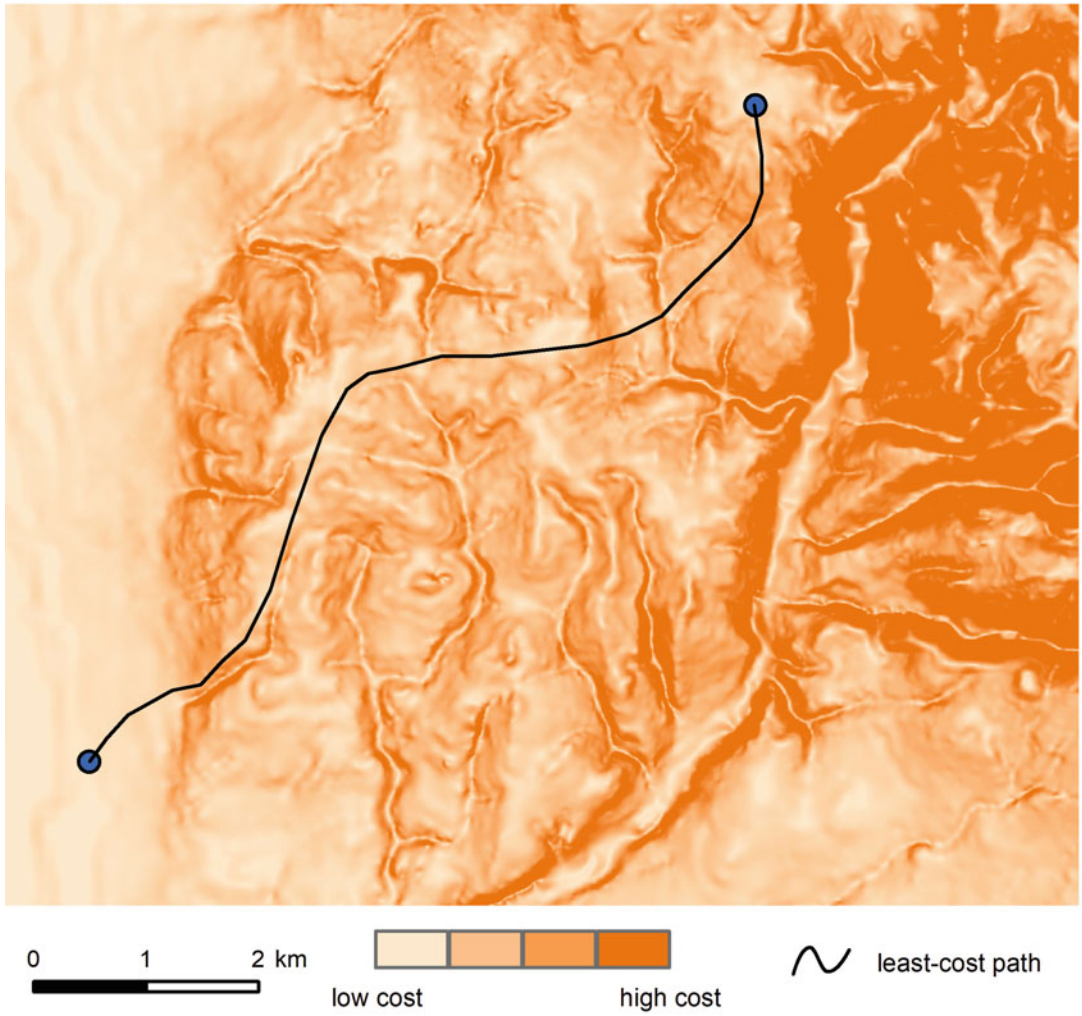

Fig. 3.23 Example of cost distance analysis. (Source: Authors)

the shortest path or fastest route to a different location from their own. Since GIS-Software is a lot more sophisticated than a common navigation system and consequently offers many more possibilities, this chapter will characterize the fundamental functions and approaches of network analyses in GIS.

When dealing with network analysis tools and functions it is necessary to prepare a suitable spatial network in form of a network dataset which is able to perform all functions that are included in a network analysis (DeMers 2008). Network datasets typically consist of line features that stand for the routes of motion in the network, enhanced with further features and premises to ensure proper usage (ESRI 2018b). Regular line features are generally not related to each other and have no or only a few connectivity rules. This means for instance if two different lines are intersecting each other, none of them is aware of it, what makes the dataset restrictive as you cannot turn from one to another. To make sure that the network recognizes these crossings as such, it is necessary to transfer the network into a new one that has nodes which allow a turn from one edge to another, except for over- or underpass lines (e.g. tunnels, bridges, etc.), where this intentionally should not be possible. Basically, it is necessary to decide whether the network dataset should have an end point or any vertex connectivity (see Fig. 3.26).

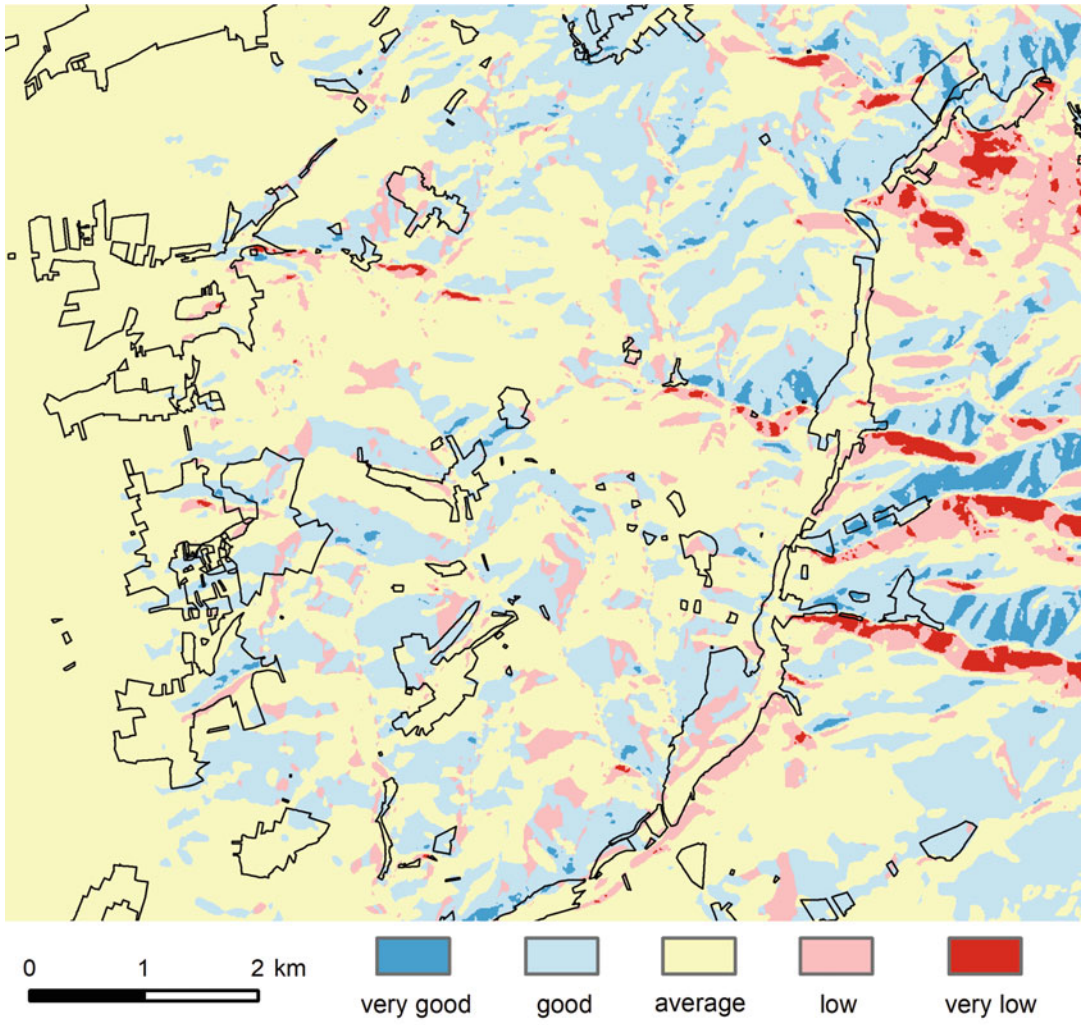

Fig. 3.24 Example of solar radiation intensity. (Source: Authors)

Furthermore, a working network dataset needs information on directions linked to each street segment as there are one-way streets as well as streets that can be driven on in both directions. If an outcome of the analysis should deal with a time or street capacity component these are also figures that have to be included as impedances constructing a network dataset (Chang 2010).

Once the network dataset is built up, there are several different opportunities of calculations in a network analysis.

3.3.2 Optimal Routes

The most basic function is the calculation of an optimal route from point A to point B and any number of intermediate stops, while the order of these stops is determined by the user and not the tool. This route can be either focusing on the shortest distance or the fastest route, depending on the needs of the respective user (Fig. 3.27). Relevant factors for this calculation can be the existence of one-way streets, barriers like construction sites

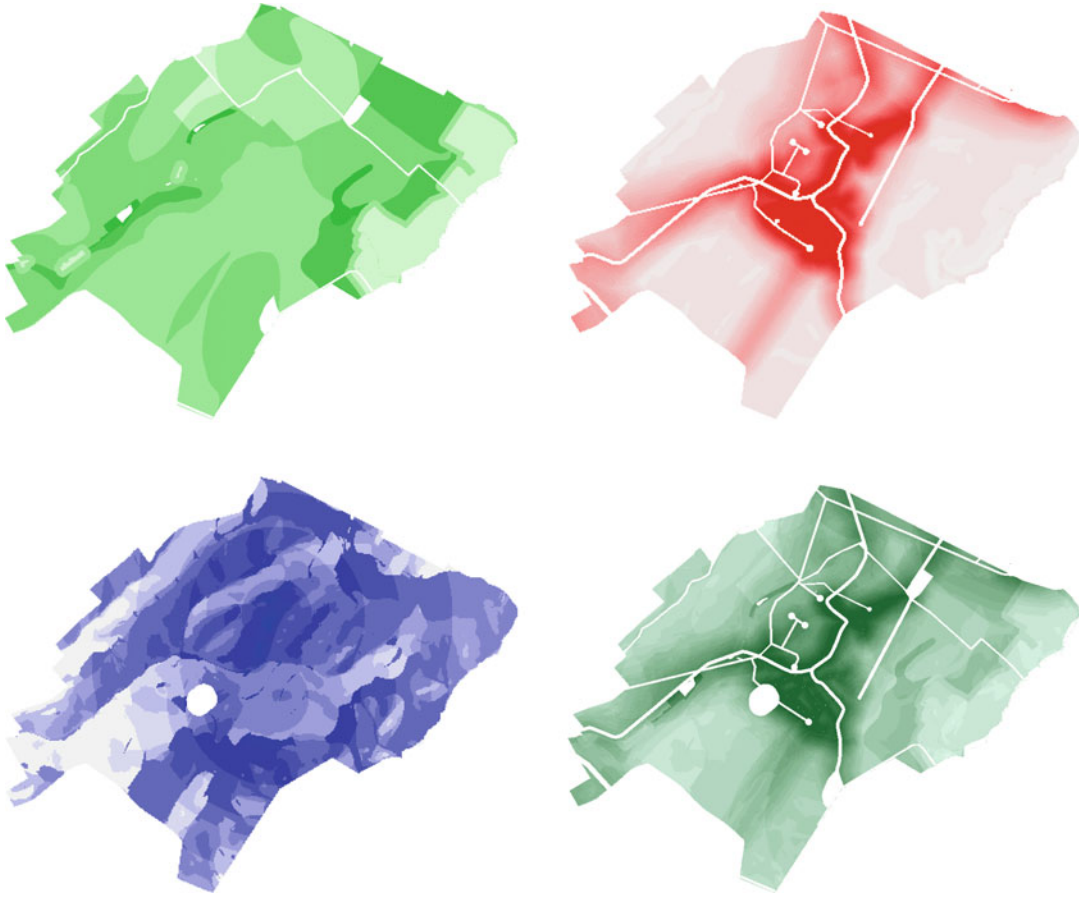

Fig. 3.25 Example of multi-criteria analysis. (Source: Authors)

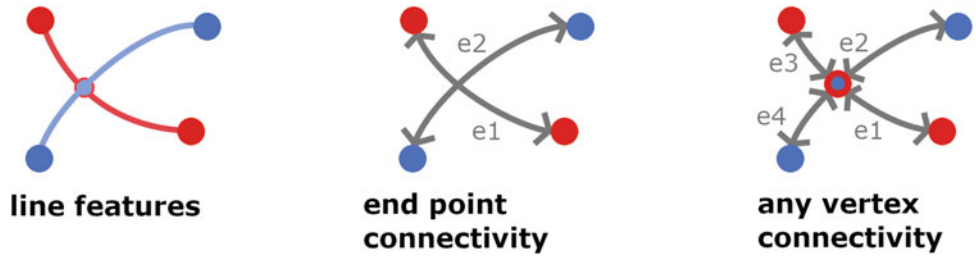

Fig. 3.26 Line Feature input with nodes and edges and different output connectivities for a network dataset. (Source: Authors)

or other obstacles, prohibited turns or other restrictions that influence the way the route will be computed (ESRI 2018b). All the different factors can be parametrized before the calculation and adjusted afterwards, as the result is only a virtual and therefore temporary layer in the map that can be permanently exported if necessary. Additionally to the line feature of the route, the user can generate driving directions containing specific information about the route, e.g. how long to stay on a certain street and when turn to another.

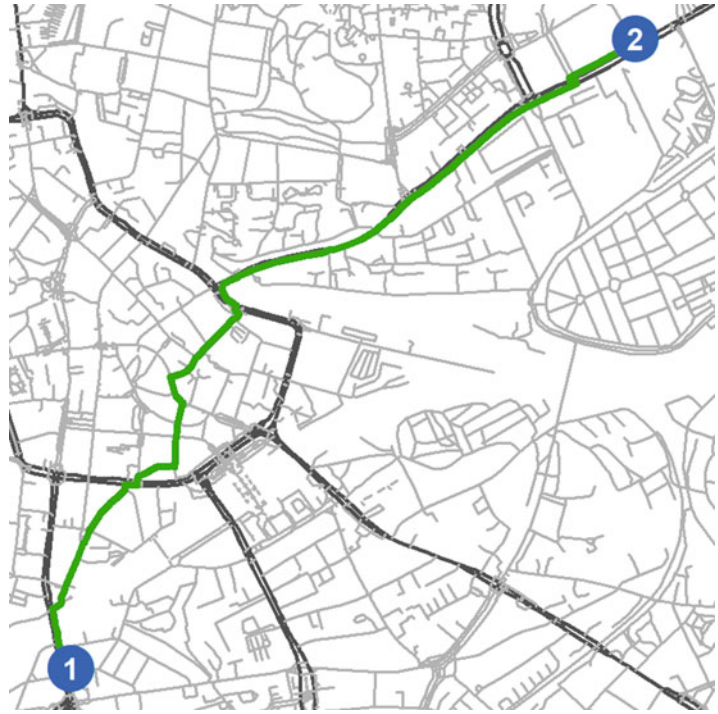

Fig. 3.27 Optimal Route from location 1 to location 2. (Source: Authors)

3.3.3 Traveling Salesman Problem

The traveling salesman problem is an issue that is not only present in classical spatial analyses, but also in various other fields, like e.g. logistics or designing products that contain several different spots which need to be connected in a specific chronological order. The most relevant aspect in this network analysis is the efficiency of a route no matter if it is a real person that is travelling or any other subject that is moving between several different locations (Curtin 2007).

The starting situation is a given amount of different locations in a network and a person or subject that has to visit each of these locations in a certain order. That order has to be the most efficient one regarding distance, required time or costs (see Fig. 3.28). Depending on how these different factors are weighted, the routes can vary significantly what makes choosing the right parameters essential for getting a proper result.

3.3.4 Service Areas

If there is a location in a network dataset and it is desirable to know for instance how far one can get with a car in a determined time range or how long it will last to reach a certain distance from that starting point, then a service area analysis is a reasonable approach. All that has at least to be parametrized are the different breaks in terms of certain time ranges or distances that define the borders between the different areas (see Fig. 3.29). These analyses could for example help to find a suitable location for a new hospital

Fig. 3.28 Fastest Route to stop once at every location, order calculated by the tool. (Source: Authors)

Fig. 3.29 Locations of hotels in blue, service areas for accessibility from five minutes (darkest blue) to 15 minutes (lightest blue). (Source: Authors)

as their output is the information about the size of an area (and the number of inhabitants) that is covered by e.g. an ambulance within two, four or 10 min during a specific time of the day.

3.3.5 Location-Allocation Analysis

How can we save money for transportation? Where should we place a new facility? How big is the potential area that is covered by the store? Are stores reachable for all customers in a certain amount of time? The location-allocation analysis has several approaches like minimizing impedances, number of facilities or maximize area coverage, accessibility or market shares (Chang 2010). It therefore combines the different methods of a network analysis. Each of these tasks implies the preparation of an analysis layer

that can calculate the optimal location for the particular case of application.

Necessary inputs for this layer are a network dataset as well as facility locations and demand points. The facilities are split up into candidate facilities that represent the potential location of a new facility, competitor facilities that mark the existing sites of present competitors and required facilities that represent existing sites of say one's own organization. Demand points are locations that represent the different factors that determine the grade of suitability for a new candidate facility. These can be centroids of districts or other administrative units as well as different kinds of demand profiles like accumulation of students, families or workers of a certain business. The demand points contain information like income, age, social status, etc.

As there are too many possible cases of application, the maximize attendance approach as the most frequently used one is presented in this chapter. In this example the target is to detect the locations that would generate the most efficient business (maximized attendance) for a retail chain, assuming that the customers rather frequent stores that are close to dense population centers and don't want to travel for more than 5 min. Once everything is set up, the generated output layer shows the detected site(s) connected with lines to the most valuable demand points (number of population) that determine the chosen target classes (see Fig. 3.30).

3.3.6 Origin-Destination Matrices

For the creation of an origin-destination (OD) matrix it is essential to set a certain amount of starting point features as well as a certain amount of target point features that are all located within the network dataset. The analysis settings can vary between different impedances, barriers in the network, a certain point of time and other parameters that influence the result concerning the properties of the network dataset (Curtin 2007).

The result layer shows the shortest routes and directions from all starting point features to all target point features that are within a determined range of distance or time (see Fig. 3.31). This can be used e. g. for creating a new model of pedestrian movements or checking the suitability of new sites within a network dataset.

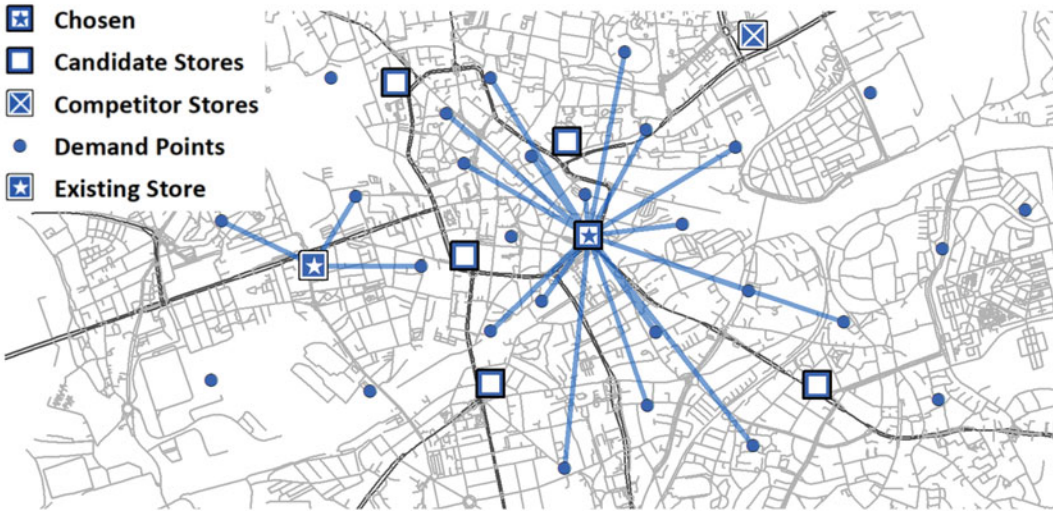

Fig. 3.30 Location-Allocation Analysis result with lines showing the connection to valuable and affecting demand points. (Source: Authors)

Fig. 3.31 OD Matrix for accessibility from certain locations (blue circle) to hotels (blue square) with needed amount of travel time (in minutes). (Source: Authors)

3.4 Spatial Statistics (by Karel Macků)

In the context of the Tobler's first law of geography saying "Everything is related to everything else, but near things are more related than distant things", spatial statistics is a set of exploratory techniques for describing and modelling spatial distributions, patterns, processes and relationships. This group of analyses is necessary for a deeper understanding of spatial data, which is provided with the use of statistical methods. In this chapter, the most frequently used methods of spatial statistics are briefly introduced.

Spatial statistics is a subcategory of spatial data analysis which is closely linked to mathematical statistics. Spatial statistics is a set of exploratory techniques for describing and modelling spatial distributions, patterns, processes and relationships (Bennett et al. 2017). According to Haining (2003) some of the spatial analyses include mathematical modelling where model outcomes are dependent on the spatial interaction between objects in the model, or spatial relationships of the geographical positioning of objects within the model. This statement represents the difference between simple spatial analyses and more advanced methods that approach the tasks using mathematical and statistical apparatus. Question is why any events happen on their location and not elsewhere? Is there any association with the environment? Are the events spread or clustered in any area? With proper data, these types of questions can be answered with spatial statistics.

Spatial statistics methods are based on the assumption that elements that are close to each other are also more closely related. A direct link to Tobler's first law of geography can be observed here: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970, p. 236). Spatial statistics can also be viewed as a complementary tool to spatial data analysis – it offers a mathematical apparatus and methods for evaluating spatial information, on the other hand, stands geography or other spatial science, which formulates a hypothesis or identifies the key parameters of these spatial data (Getis 2005). In the search for a high degree of certainty, the statistical approach is always recommended.

There should be no confusion between the terms spatial statistics and geostatistics – geostatistics is one of the spatial statistics sub-disciplines and has emerged as a tool for a probability prediction of the distribution of ore

deposits in the mining industry (Longley et al. 2010).

Spatial statistics include methods based on stochastic (i.e. random) nature and pattern of phenomena. These tasks can be divided into descriptive (producing essential information about a set of elements) and interference – analysis of patterns and behaviour of spatial data. This type of analyses is the subject of this chapter.

3.4.1 Pattern Analysis

One of the frequently asked questions in the advanced spatial analysis regarding the spatial distribution is the question whether elements are in a random structure or there is a pattern in their behaviour. In particular, it is essential whether the presence of one value causes an increase or decrease in the probability of occurrence of another value in its vicinity (Longley et al. 2010). By their deployment in space, the elements can create one of the following structure:

- Random structure
- Cluster structure
- Regular/scattered structure

At the beginning, the term ‚cluster'should be clarified. Clustering is a global property of the spatial pattern in a dataset, measured by a single statistics (Anselin 2005). Then cluster is a group of features, whose value and/or its locations are closer together than they would be by random. The purpose of pattern analyses is to determine whether the spatial behaviour of the geographic elements follows one of the above-mentioned options and if this behaviour is somehow statistically demonstrable. Actual spatial distribution is therefore tested against one of these options. Confirming the existence of significant clusters of similar values/clusters of points near one another is one of the most common tasks. Such a task could be based only on a visual analysis of spatially visualised data; however, the use of spatial statistics underlies this estimate by numerical tests and makes it more reliable. The resulting finding helps to understand the behaviour of the observed phenomenon and to support the hypotheses that explain this behaviour. The following lines will describe selected spatial pattern analysis.

3.4.2 Point Patterns

3.4.2.1 Ripley's K Function

The *K function* is one of the methods for assessing the randomness of the distribution of the set of point data. It allows seeing if the elements appear to be dispersed, clustered, or randomly distributed throughout the area of interest. The basis of this method is to monitor the occurrence frequency in a defined space – for example, the area in the distance d from each point. The K function is defined as the ratio of the number of occurrence points in the defined area (grid or defined distance d) and the expected density of points per area unit, how would it be within the random distribution of the elements (most often represented by the homogeneous Poisson process, also known as complete spatial randomness). This principle allows identification of deviations from spatially evenly distributed data (Dixon et al. 2002). If the number of observed points within a given area is higher than for a random distribution, the distribution is clustered. If the number is smaller, the distribution is dispersed (Gillan and Gonzalez).

For an example, data of position of small and medium enterprises in Olomouc region has been analyses with K function. In such a data, it is expected that companies are located in the sites that means they will be clustered within the city.

The result of point pattern analysis can be presented as a graph – see Fig. 3.32. The vertical axis is the K-function value; the horizontal axis is the searching distance d. The blue line represents the k value of the random distribution of the points, and the red line represents the K function value – the real observed distribution of the points (the position of companies). If the observed value is above the random, it means that points are clustered. If the observed were under the random, the data would be dispersed. In this case, result points to strongly clustered data. This supports the original hypothesis about the location of enterprises.

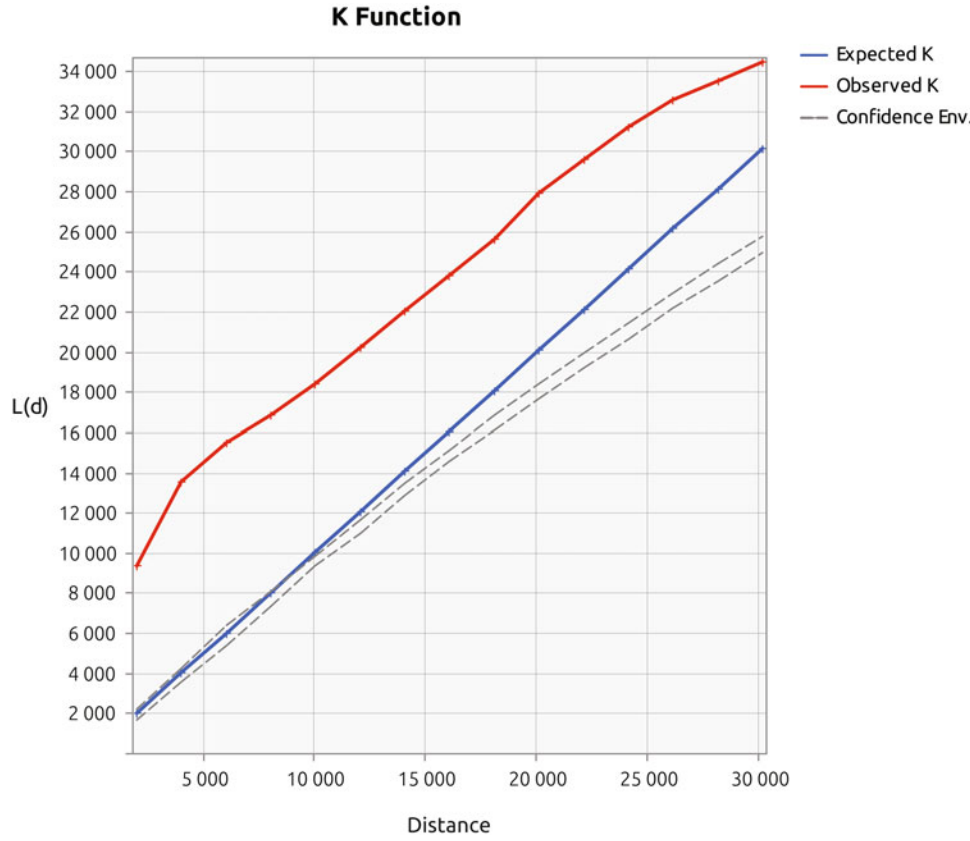

Fig. 3.32 A graphical output of K function. (Source: Authors)

3.4.2.2 Kernel Density

Kernel smoothing methods are used to transform data from a discrete representation (geolocated points) into a continuous array. This process is particularly useful for better interpretation of spatial distinction of variables behaviour. The kernel density estimate works with localised data, which are used for the expression of the spatially smoothed estimate of the local intensity of the occurrence of objects/events. This local smoothed intensity can also be understood as the surface of the risk of occurrence of these objects/events (INSEE Eurostat 2018). The application on spatial data is based on density estimation, a function of estimating the values occurrence based on observed data (Silverman 1986).

Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point (ESRI 2018c). The final surface is created by estimating the intensity at any point using the appropriate probability density function (K – kernel function). It is necessary to determine the area in which the algorithm will assess the density of the phenomenon. This sphere – so-called bandwidth, might be calculated on all input points and median distance between its centre and all input points. The bandwidth parameter essentially determines the degree of smoothing of the resulting surface. The different kernel functions can be used to make the result of density estimation different. The application on the spatial data implemented in ArcMap software uses the quartic function, which approximates to the normal distribution.

The resulting surface is represented in the form of a raster, which can be conveniently visualised for the purpose of overview of the phenomenon

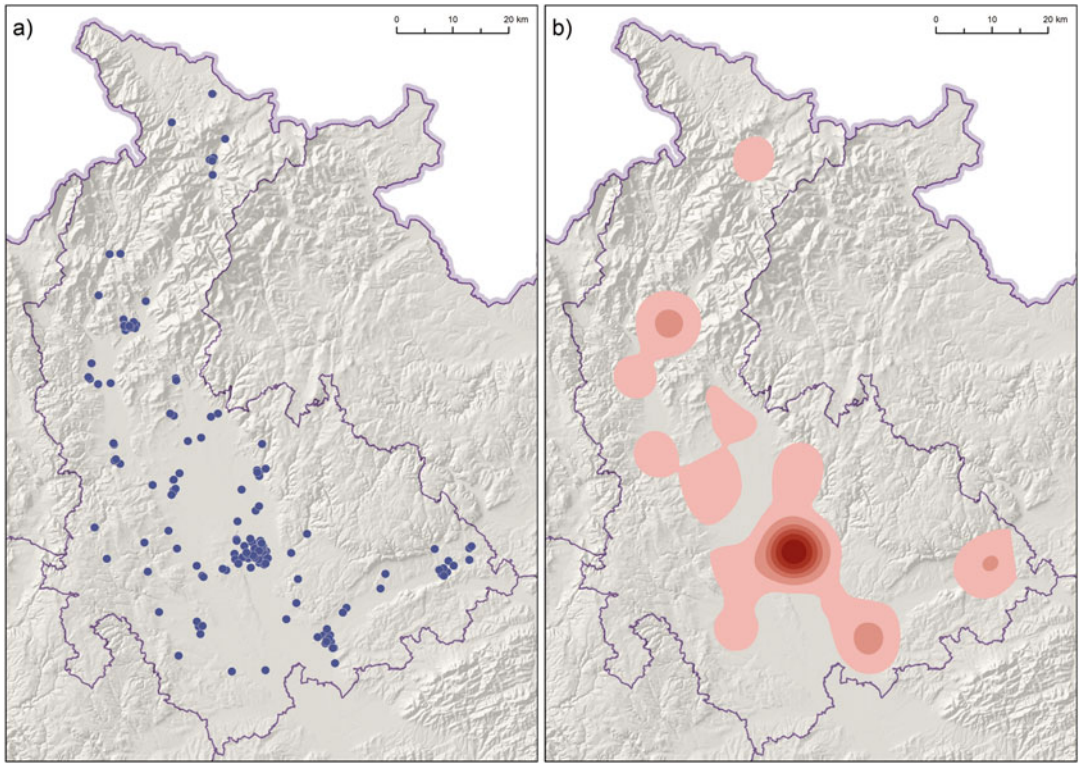

Fig. 3.33 (**a**) input point data, (**b**) result of Kernel density function. (Source: Authors)

and revealing point patterns. The same data describing location of enterprises in Olomouc region (as for the K function in previous chapter) was used for demonstration of Kernel density function. Figure 3.33 shows the visualisation of input points and output surface. The bandwidth was set to five kilometres, with the aim to produce more smoothen output. The output presents the probability of enterprise occurrence, this type of visualisation brings a generalized overview of the spatial distribution of points in the area of interest. The aim is not to estimate the correct probability of occurrence, but rather to get overall impression about spatial distribution of points. For that reason, a legend for interpretation of result is not included.

3.4.3 Spatial Autocorrelation

The previous chapter has described how clusters of point phenomena can be identified based on their location. A following task can be to identify clusters based on the location combined with the value of the observed phenomenon at the same time. Such an analysis makes possible to evaluate whether there are spatially closer elements that have similar values of the observed phenomenon and form together high or low-value clusters, or whether the elements in the space are located at random. In the natural world, we expect some influence of environment on the monitored phenomenon. For example, analysing a strong economic region concerning the GDP per capita, we naturally assume that the regions in its immediate neighbourhood will be similar, as the whole area is characterised by similar conditions. Similarly, we expect these regions to differ from other, more remote areas. To support such a claim, an analysis of spatial autocorrelation can be used.

Spatial autocorrelation is a correlation between the values of one variable, and it allows to evaluate the degree of similarity of one object with objects in its neighbourhood and comparison

with more remote objects (Cliff and Ord 1973). First, it is necessary to define relations of the object with its surrounding objects, which is provided by the matrices of spatial weights. Here, the distance of objects enters as a weight for defining spatial relationships – the autocorrelation of neighbouring objects will have more importance than the autocorrelation of distant objects. If positive autocorrelation occurs, we conclude that objects with similar values are spatially located near each other, forming spatial clusters of similar values. Negative autocorrelation indicates the proximity of different values, autocorrelation around zero indicate randomity in the spatial distribution of values.

Autocorrelation can be measured by several measures – an example of them is the **Moran's I** or **Geary's** criterion. Positive index value indicates a positive autocorrelation, and negative values represent negative autocorrelation. These indicators, however, measure autocorrelation only at the *global* level, that is the whole area of interest. If the result of these tests come out positively, it makes sense to ask how the autocorrelation varies in the space. A *local* test – **LISA** (Local Indicators of Spatial Association) serves for this task. Since the method of identifying spatial autocorrelation is based on traditional statistical methods, the calculation is complemented by the statistical significance, represented by p-value. This makes it possible to assess whether the result obtained is statistically significant or not.

The initial analysis of autocorrelation reveals spatial dependence, so it is known that clusters of high and low values occur in the area of interest. Local Moran's I can be visualised to identify these areas. However, it is still unknown whether the high value of autocorrelation means clustering of high or low values. For a deeper understanding of the phenomenon, it is possible to visualise the observed variable depending on the average value in its surroundings – this is presented by Moran's plot (Anselin 1996).

Using LISA and Moran's plot as supporting tools, all objects can be classified into four groups corresponding to the quadrants in Moran's plot. Spatial clusters showing above-average or below-average values of a variable in a particular unit consistent with its surroundings are found in the graph in the top right (hot spots, high-high) and left-low (cold spots, low-low) quadrants. This is evidence of high autocorrelation. On the contrary, the areas identified in the left upper (LH) or right lower (HL) quadrants are characterized by the existence of a low value surrounded by high and vice versa (Anselin 1995) (Fig. 3.34).

Similar output as provided by LISA is available also with **Getis – Ord G∗.** The main difference is that for LISA, the value of the feature being analysed is not included in that analysis, only neighbouring values are. Alternatively, when the local analysis is being done with Getis-Ord Gi∗, the value of each feature is included in its analysis (Getis and Ord 2010). The local sum for a feature and its neighbours is compared proportionally to the sum of all features; when the local sum is very different from the expected local sum, and when that difference is too large to be the result of random chance, a statistically significant z-score results (ESRI 2018d).

The output of this indicator is the so-called z-score for each analysed object. The higher (positive) the z-score value, the higher the intensity of clustering of high values in the area (so-called hotspot), and vice versa – the smaller (negative) the z-score is, the higher the intensity of clustering with a cold spot.

An example demonstrating the use of spatial autocorrelation methods is described in the analysis of the economically strongest and also the weakest regions in Europe. The monitored variable is GDP, the spatial unit is NUTS 3 regions.

The GDP is expressed in purchasing power standard per inhabitant in the year 2015. In Fig. 3.35a, a choropleth map is used to display the GDP in regions. By this visualisation, areas with the highest or lowest values can be defined, especially the big difference between east and west are visible. But can be said with certainty which regions are the strongest and which are the weakest? In many cases, when the data doesn't have a clear pattern, or inappropriate visualisation is used, it might be a difficult task. For that reason, spatial autocorrelation is calculated. Figure 3.35b

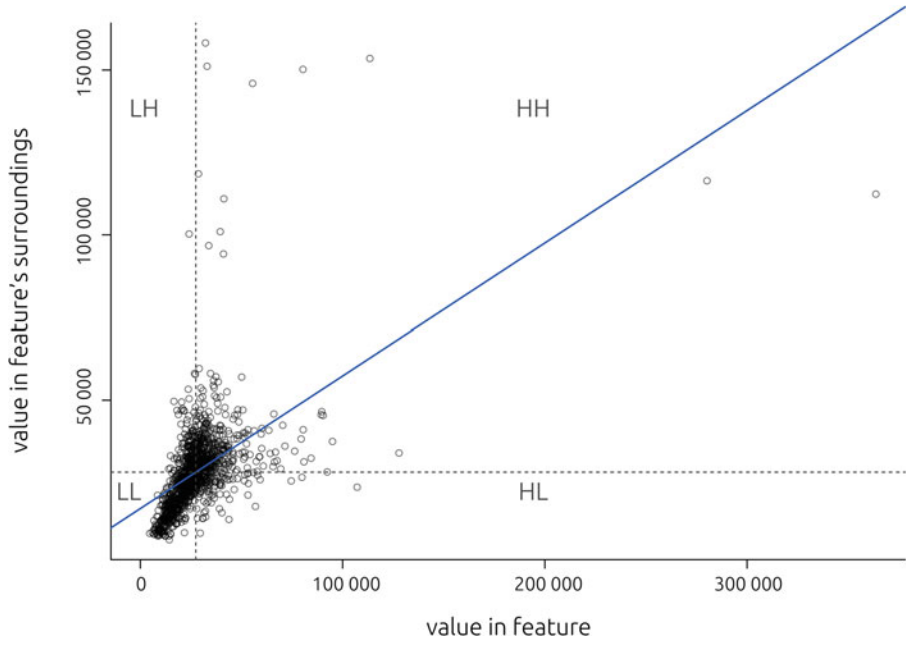

Fig. 3.34 Moran's plot. (Source: Authors)

shows the distribution of spatial autocorrelation calculated for the GDP data. Only the shades of green are statistically significant; darker shade stands for higher autocorrelations = clusters of low or high values are present. The final step is the derivation of cluster type based on the value of autocorrelation in every region and its neighbourhood. This can be done by LISA analysis (Fig. 3.35c) or Getis-Ord G∗ (Fig. 3.35d). See the difference between this methods caused by the different approach, how they calculate the membership to any clusters.

Now user can state that regarding the spatial distribution of GDP, there is a great cluster of low values in the eastern European and several small clusters of high value in the central Europe, Sweden and UK. In the rest of the area of interest, the GDP value has a random distribution without statistically significant patterns.

3.4.4 Geostatistics

As mentioned in the introduction to the chapter, the term spatial statistics is often confused with the term geostatistics. In the narrower sense, geostatistics is used only to define a set of interpolation algorithms – algorithms used to estimate the values of the continuous phenomenon or its intensity in any location of the controlled area where no measurements have been made. The continuous character is typical of environmental phenomena such as temperature, air pressure or soil concentration. In the context of economic

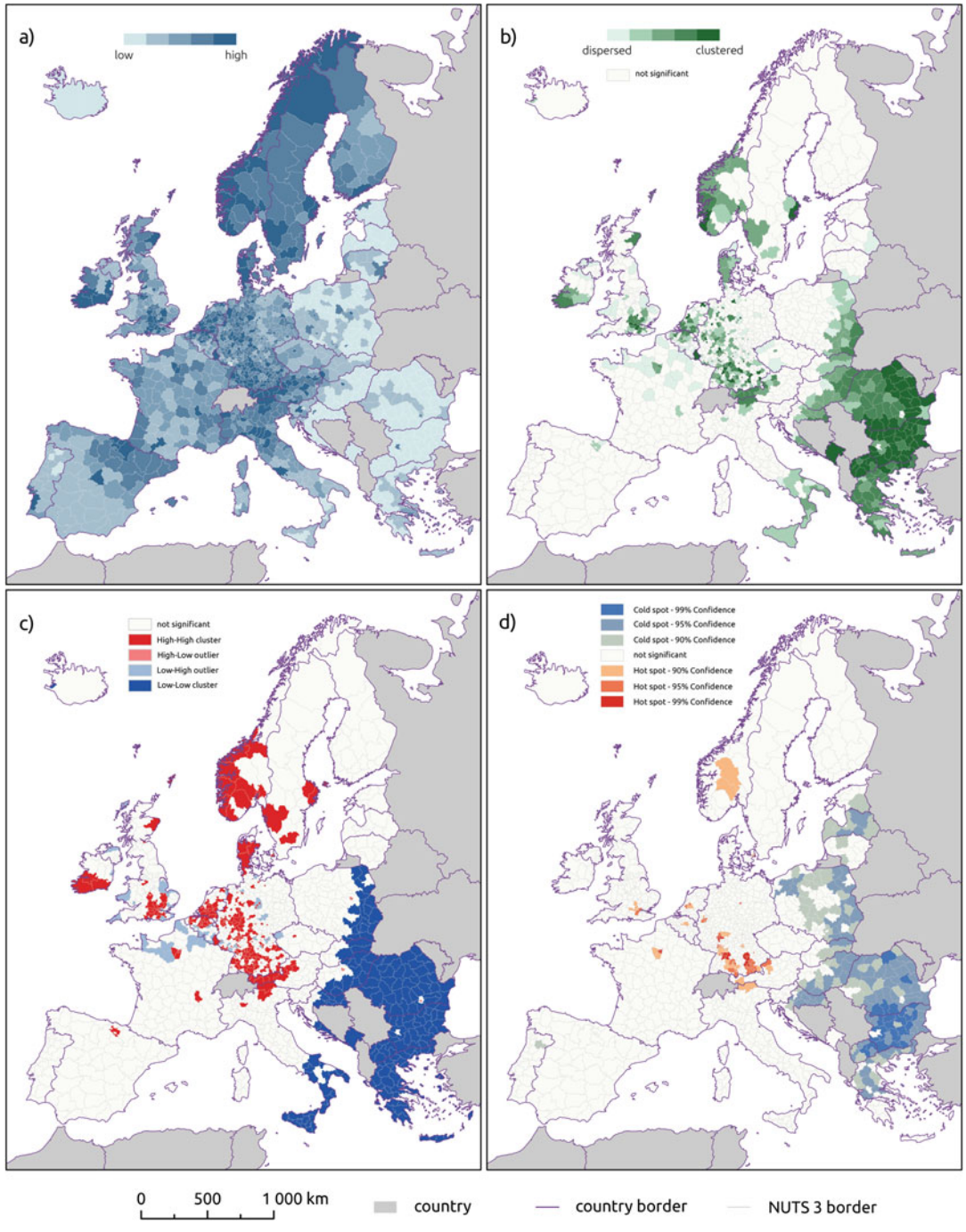

Fig. 3.35 Analysis of spatial autocorrelation of GDP in Europe. (Source: Authors)

data, there would be a lack of applications, so this topic will not be further discussed.

References

Albrecht, J. (2005). *Geographic information science*. http://www.geography.hunter.cuny.edu/~jochen/gtech361/.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis, 27*, 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial analytical perspectives on GIS* (pp. 111–125). London: Taylor and Francis.

Anselin, L. (2005). Spatial statistical modeling in a GIS environment. In D. J. Maguire, M. Batty, & M. Goodchild (Eds.), *GIS, spatial analysis, and modeling* (1st ed.). ESRI Press.

Bennett, L., Vale, F., D'Acosta, J. (2017). *Spatial statistics: Simple ways to do more with your data*. https://www.esri.com/arcgis-blog/products/product/analytics/spatial-statistics-resources/

Chang, K. T. (2010). *Introduction to geographic information systems* (5th ed.). New York: McGraw-Hill.

Cliff A. D., & Ord J. K. (1973). *Spatial autocorrelation*. London: Pion Ltd.

Curtin, K. (2007). Network analysis in geographic information science: Review, assessment, and projections. In *Cartography and geographic information science* (Vol. 34, pp. 103–111). London: Taylor and Francis.

DeMers, M. (2008). *Fundamentals of geographic information systems* (4th ed.). Hoboken: Wiley.

Dixon, P. M., El-shaarawi, A. H., & Piegorsch, W. W. (2002). Ripley's K function. *Encyclopedia of Environmetrics, 3*, 1796–1803.

ESRI. (2018a). *ArcGIS Desktop Help*. http://desktop.arcgis.com/en/arcmap/10.3

ESRI. (2018b). *What is a network dataset?* https://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/what-is-a-network-dataset.htm. Accessed 28 Dec 2018.

ESRI. (2018c). *How Kernel density works—Help | ArcGIS desktop*. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/how-kernel-density-works.htm. Accessed 22 Nov 2018.

ESRI. (2018d). *How hot spot analysis (Getis-Ord Gi∗) works*. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm. Accessed 23 Oct 2018.

Getis, A. (2005). Spatial statistics. In *New developments in geographical information systems: Principles, techniques, management and applications* (2nd ed., pp. 239–252). New York: Wiley.

Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x .

Gillan, J., Gonzalez, L. *Ripley's K Function and pair correlation function*. http://wiki.landscapetoolbox.org/doku.php/spatial_analysis_methods:ripley_s_k_and_pair_correlation_function. Accessed 22 Oct 2018.

Haining, R. (2003). *Spatial data analysis: Theory and practice*. Cambridge: Cambridge University Press.

INSEE Eurostat. (2018). *Handbook of spatial analysis*.

Longley PA, Goodchild MF, Maguire DJ, Rhind DW (1999) Geographical information systems: Principles and technical issues.

Longley, P. A., Goodchild, M., Maguire, D. J., & Rhind, D. W. (2010). *Geographic information: Systems and science* (3rd ed.). Wiley Publishing.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Monogr Stat Appl Probab*.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234–240.

Tomlin, D. (1994). Map algebra: One perspective. *Landscape and Urban Planning, 30*, 3–12. https://doi.org/10.1016/0169-2046(94)90063-9.

Tomlin, D., & Berry, J. K. (1979). A mathematical structure for cartographic modeling in environmental analysis. In *Proceedings of the 39th symposium of the American conference on surveying and mapping*.

w3schools.com. (2018) Retrieved December 30, 2018, from, https://www.w3schools.com/sql/sql_where.asp