PQStat Software
Statistical Computational Software

User Guide - PQStat

Spatial Analysis

Barbara Wieckowska

Version 1.8.4
P7909200222

www.pqstat.pl

Spis treści

1 SPATIAL ANALYSIS **2**
- 1.1 BASIC DEFINITIONS . . . 2
- 1.2 MAP OPENING . . . 4
- 1.3 MAP MANAGER . . . 5
  - 1.3.1 Map Viewing Tools . . . 6
  - 1.3.2 Selection Area Tools . . . 6
  - 1.3.3 Layers . . . 7
  - 1.3.4 Map Style Edition . . . 10
- 1.4 HOW TO REDUCE A WORKSPACE). . . . 11
- 1.5 GEOMETRIC CALCULATIONS . . . 12
- 1.6 SPATIAL WEIGHTS MATRIX . . . 13
  - 1.6.1 Weights matrix according to distance . . . 14
  - 1.6.2 Weights matrix according to contiguity . . . 14
- 1.7 SPATIAL SMOOTHING . . . 15

2 TESTING HYPOTHESES **19**

3 DESCRIPTIVE STATISTICS **21**

4 DENSITY ANALYSIS **27**
- 4.0.1 Quadrat Count Methods . . . 27
- 4.1 Kernel density estimator . . . 31
  - 4.1.1 Two-dimensional kernel estimator . . . 31
  - 4.1.2 Three-dimensional kernel estimator . . . 35

5 RANDOMNESS OF POINT DISTRIBUTION **36**
- 5.1 Nearest Neighbor Analysis . . . 36

6 SPATIAL AUTOCORRELATION **44**
- 6.1 Global Moran's I statistic . . . 44
- 6.2 Global Geary's C statistic . . . 52

7 LOCAL ESTIMATE OF SPATIAL CLUSTERING **57**
- 7.1 Local Moran's I statistic . . . 57
- 7.2 Local Getis-Ord's G statistic . . . 63
- 7.3 CutL . . . 71

1 SPATIAL ANALYSIS

Statistical spatial analysis is defined as a set of techniques for studying data which are located in space viewed in relationship with the surface of the Earth. Particular techniques of spatial analysis are used in diverse areas of science, from medicine (epidemiology), through logistics and physics, to economy (finding the best locations for plants, shops, etc.).

The development of the methods of spatial distribution analysis and of the analysis of interrelationships among objects has been and is, to a large extent, determined by the development of information technology. Computers with ever increasing computing power, together with GIS systems, allow the processing of large amounts of geographic data.

1.1 BASIC DEFINITIONS

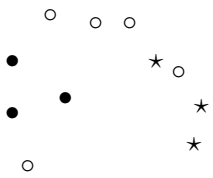
Geographic Information System - GIS –is a system for entering, storing, processing, and visualization of geographic data. From the technical point of view GIS is a tool which allows the analysis of interrelated:

- information about spatial locations of objects –represented by means of a **map**;
- descriptive characteristic concerning objects presented on a map –represented by means of a **database**.

Objects represented by means of a map are:

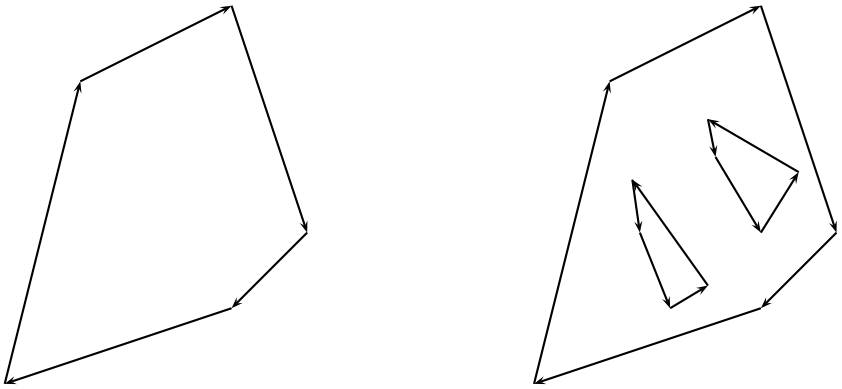
- **Points** –the location of which, in 2D, is defined with the help of two coordinates (x, y);
- **Multipoints** –they are points grouped in sets:

1)



1) An example of a multipoint in which each point is defined as belonging to one of 3 groups.

- **Lines** –they are created by linking subsequent points, in proper order (lines can intersect);
- **Polygons** –they are closed spaces, restricted by means of external rings (closed lines which do not intersect and which go through at least 3 points, in an appropriate order). Polygons can also contain internal rings constituting their internal boundaries. External rings are defined clockwise and the internal ones the other way round.



1) 2)

1) An example of a polygon which only has an external boundary (with no internal rings);
2) An example of a polygon which has both an external boundary and internal boundaries (areas defined by the internal rings constitute a part of an external area, i.e. they do not belong to the polygon).

Object attributes are entered into the base in the form of:
numbers –e.g. area, temperature,
texts –e.g. names of objects.

Map projection is a mathematical method of mapping the surface of the Earth onto a map surface. There is a number of methods for such mapping. The mappings can be based on a spheroid or on the surface of a ball (a sphere), or on a part of either of them. Each mapping forms the basis for defining an appropriate coordinate system. Because each projection of a surface entails certain distortions (distortions of angles, areas, lengths), the choice of a proper system depends on the aim for which the map is to be used.

Coordinate systems used in cartography are classified as:

- geographic coordinate systems (they define geographic latitude and longitude);
- cartesian coordinate systems;
- polar coordinate systems.

For a map to be loaded correctly, the program PQStat requires a vector map saved in a SHAPEFILE (shp) type of file and defined in **a proper Cartesian coordinate system**, with line scale.

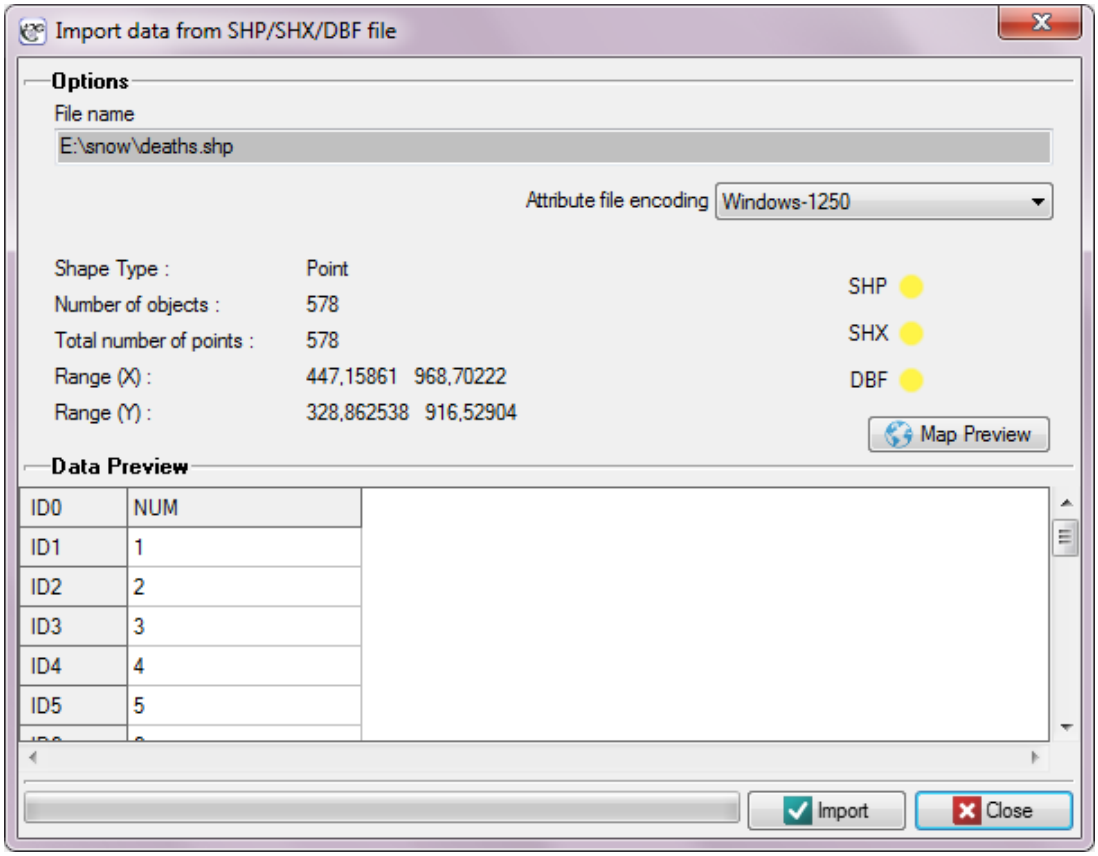
The program tries to automatically detect maps with geographic coordinates. If, while importing the map, the program detects a geographic coordinate system, it suggests converting the coordinates into a UTM system (**Universal Transverse Mercator**), on the basis of the WGS-84 system of reference. As conversion might be incorrect (due to the use of many geographic coordinate system and the lack of certainty with regard to the applied system), it is recommended that properly prepared maps be used –in a Cartesian coordinate system.

1.2 MAP OPENING

A map with the attribute file assigned to the map can be loaded via:

- import of a shapefile, SHP, into the datasheet,
- loading the PQS/PQX file which contains data from shapefiles (SHP).

Import of a Shapefile (SHP)
Import is made by choosing the menu option File→Import data...→SHP/SHX/DBF ESRI Shapefile (*.shp).



In the import window we can preview the imported map and its attributes saved in a DBF file. If the directory from which we import contains all files necessary for loading the map then the correct reading of appropriate files is confirmed in yellow by proper controls. Attributes ascribed to a shapefile, in the form of a DBF database, are not required for proper loading of a map. An attribute table can be completed after a map file has been loaded, by filling in proper cells of the datasheet linked with the map.

1.3 MAP MANAGER

The Map Manager is a tool for managing a map and the layers ascribed to the map. It is possible to view both maps imported to the PQStat program and maps opened directly from an SHP file.

The Map Manager is activated through the:

- menu Spacial analysis→Map Manager,
- button on the tool bar,
- context menu Map Manager on the name of a datasheet linked to the map.

A map can be opened with the help of the Map Manager:

- menu File→Open file... –if we open a map from a shapefile (SHP),
- menu Plik→Project maps or button on the tool bar –if we open a map from the PQStat program,

or in the Navigation tree of the PQStat program:

- context menu Map Manager on the name of the datasheet linked to the map.

The image which presents a map can be exported to a file in the BMP, PNG, or JPG format, by choosing, in the Map Manager window:

- menu File→Export view....

1.3.1 Map Viewing Tools

Zoom in –allows to view a map in a larger scale and see its details;

Zoom out –allows to view a map in a smaller scale and see all its parts;

Adjust to the window –allows such a view of a map that the whole image is displayed in a window;

Select –allows to choose a rectangular part of a map, which will be enlarged and adjusted to the window size;

Grabber –allows to move the image in the browser window so as to place a given part of the image in a chosen position.

As we are browsing the map we also get a tooltip concerning the ID and the name of the object we point to with the cursor. The name is loaded from the data sheet, it is the variable indicated as active in the Map manager. By default, during import the first variable of the text type is set as active.

We can get more information about the object pointed at by choosing the option Identify from the context menu. In the identification window it is also possible to Activate/Deactivate object.

1.3.2 Selection Area Tools

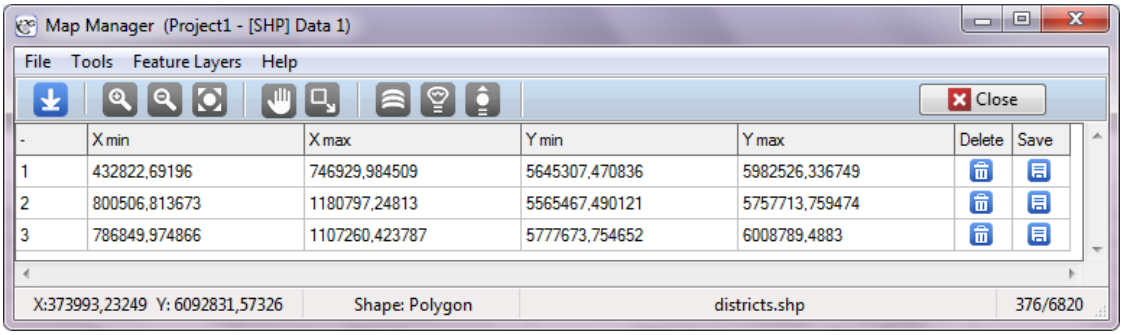
Creating and saving a selection area allows to choose parts of a map which can later be subjected to a separate analysis.

Creating a Selection Area

To select and save a selected area we choose Tools→Create selection area. Then, using the mouse or filling in the fields in the upper part of the Map Manager window, we select the chosen part of a map (elliptical or rectangular shape). The selection is saved with the use of the button Save.

Edition of a Selection Area

The placement of each selection area which was saved can be changed. In the edition window we can also delete a selection area.



That window is opened with the help of the menu Tools→Edit selection area, and closed with the use of the button Close.

Deleting All Selection Areas

All selection areas can be deleted with the use of the menu Tools→Delete all selection areas.

1.3.3 Layers

Both a map and elements added to it form layers. Layers are organized so that they only contain information about objects of one type. The use of layered organization enables easy modification of only selected objects.

The basic layer is the **base layer** containing a map. We can add new elements to that layer by creating new layers.

Adding Layers –to draw objects on subsequent layers choose the menu Feature Layers→Add Layer.

- **Layer –Result of statistical analysis**
  It is a layer created together with a report on the statistical spatial analysis. It presents the result of the statistical analysis, appended to the report. The information about the existence of layers one can draw on the map can be found at the bottom of the report (the button ->> + MAP <<-). The layer can also be added with the button in the Map Manager window.

  As long as there are no reports on statistical spatial analysis, the option window for analysis results is empty. When there are such reports the option window contains a list of layers. The names of the layers on the list consist of the name of the report from which a layer comes, together with the date and hour the report was created, and a description of the type of objects drawn there.
- **Layer –View of another map**
  That is a layer which presents a map related to another datasheet (the button in the Map Manager window). The map view can be a single layer or it can consist of a number of layers. It cannot be edited directly. The change of the look of the view is only possible when particular layers forming that view and placed in a real location (i.e. linked to another datasheet) are edited.

  As long as there is only one datasheet related to a map in the project, the option window

of another map is empty. When there are several datasheets the option window contains a list of layers. The names of the layers on the list comprise the number and name of the datasheet linked to the map and the name of the file from which the map was imported.

If the map is appended with a map view from another datasheet in such a way that a circular reference is made (e.g. map 2 is assigned a view of map 1, and map 1 is assigned a view of map 2), then a message is displayed about a circular reference. The reference will be managed but circular references are not advised.

- **Layer –Centroid of a polygon** –that is a layer of the point type.

  A centroid of a polygon is a point lying within a polygon and representing the center of mass (O'Rourke J (1998)[10]).

  Centroids can be drawn on the basis of calculations made on the map –in such a case we choose the option Draw and calculate based on map data, or on the basis of existing points the coordinates of which are in the datasheet –we then choose the option Draw based on the datasheet.

- **Layer –Center of a polygon** –that is a layer of the point type.

  A center is a point with coordinates of the X axis and the Y axis calculated as a mean from the coordinates of points constituting polygon vertices.

  Centers can be drawn on the basis of calculations made on the map –in such a case we choose the option Draw and calculate based on map data, or on the basis of existing points the coordinates of which are in the datasheet –we then choose the option Draw based on the datasheet.

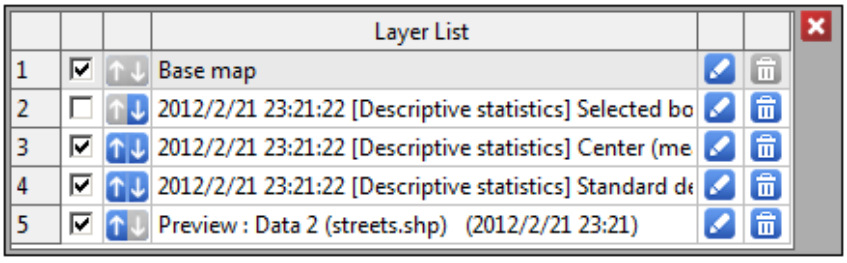
- **Layer –Label for an object** –that is a layer of the text type.

  A label is any text or number concerning objects presented on a map. Objects can be described by choosing from the datasheet a variable which contains proper labels.

- **Layer –Bounding types** –that is a layer of the polygonal type.

  Min. Bounding - convex hull –that is the smallest convex polygon in which analysed objects are enclosed (Yamamoto J.K. 1997 [17]);
  Min. Bounding - rectangle –that is the smallest rectangle in which analysed objects are enclosed;
  Min. Bounding - circle –that is the smallest circle in which analysed objects are enclosed;
  Rectangle from map bounding –that is a rectangle in which analysed objects are enclosed, with coordinates of the lower left-hand vertex = (min X, min Y) and of the upper right-hand vertex = (max X, max Y).

Layer List –the layer list allows to check how many visible layers constitute the received image (button on the tool bar).



The list also allows switching on and off visibility of a layer and changing the order in which layers are added, editing the layers, and deleting them. If the source of a layer (a report or

a map linked to the layer) is removed then such a layer is automatically removed from the layer list.

1.3.4 Map Style Edition

We can edit map layers by choosing the button  in the map list. The manner of editing depends on the type of objects presented on a given map (points, multipoints, lines, polygons). It is possible to select the style of lines, color fill, and the level of its transparency. By default the coloring utilizes one color only. In the case of layers representing the base map there is a number of coloring methods.

Coloring Methods:

Full color –when this method is used, all objects will be colored with the use of the same method –with the use of one color only (the button Fill).

Color gradation –when this method is used, objects will be colored according to the value assigned to them in a selected datasheet variable (the button Color gradation). For example, when coloring a map which shows altitude, color shade for points lying higher will be different from that for points lying lower. The variable according to which we will do the coloring should only contain numerical values. If that is not the case then the object for which there is no numerical value is not colored according to the coloring method chosen for that variable but has the default color for the map.

Methods for variable breaks used in color gradation:

- Natural Breaks (Jenks) –a method in which a variable is broken into such classes that variance in classes is minimized and variance among classes is maximized.
- Quantile Breaks –a method in which a variable is broken into classes with an equal number of units.

1.4 HOW TO REDUCE A WORKSPACE).

Workspace is limited for the purpose of indicating only those objects which will be subjected to the analysis. Such objects are indicated in the program by activating or deactivating them. Inactive objects are not subjected to statistical analyses.

Manual activation/deactivation of objects

- **Indicating a row in the data sheet** which describes the appropriate object and selecting the option Activate/Deactivate from the context menu on its name;
- **Indicating an object on the map** and selecting, from the context menu, the option Activate/Deactivate or Identify → Activate/Deactivate object.

Automatic activation/deactivation of objects

- **Selecting objects on the basis of data sheet** – for example, one can indicate as active only those shops which are groceries with an area not larger than 1000m2. In such a case, the setting of appropriate conditions for selecting objects takes place in the window of Activation/Deactivation available after selecting the Edit →Activate/Deactivate (filter)... menu. A detailed description of the manners of selection of that type can be found in the User Manual - PQStat (Chapter: How to Reduce Data Sheet Workspace).
- Selecting objects on the basis of a map – for example, one could only distinguish those shops which are within a rectangular or elliptical area marked on a map. We select the area with the use of the selection area tools (see Chapter 1.3.2) and later activate or deactivate in the window Activate/Deactivate in the selection available after selecting the Tools →Activate/Deactivate in the selection menu in the window of the Map manager.

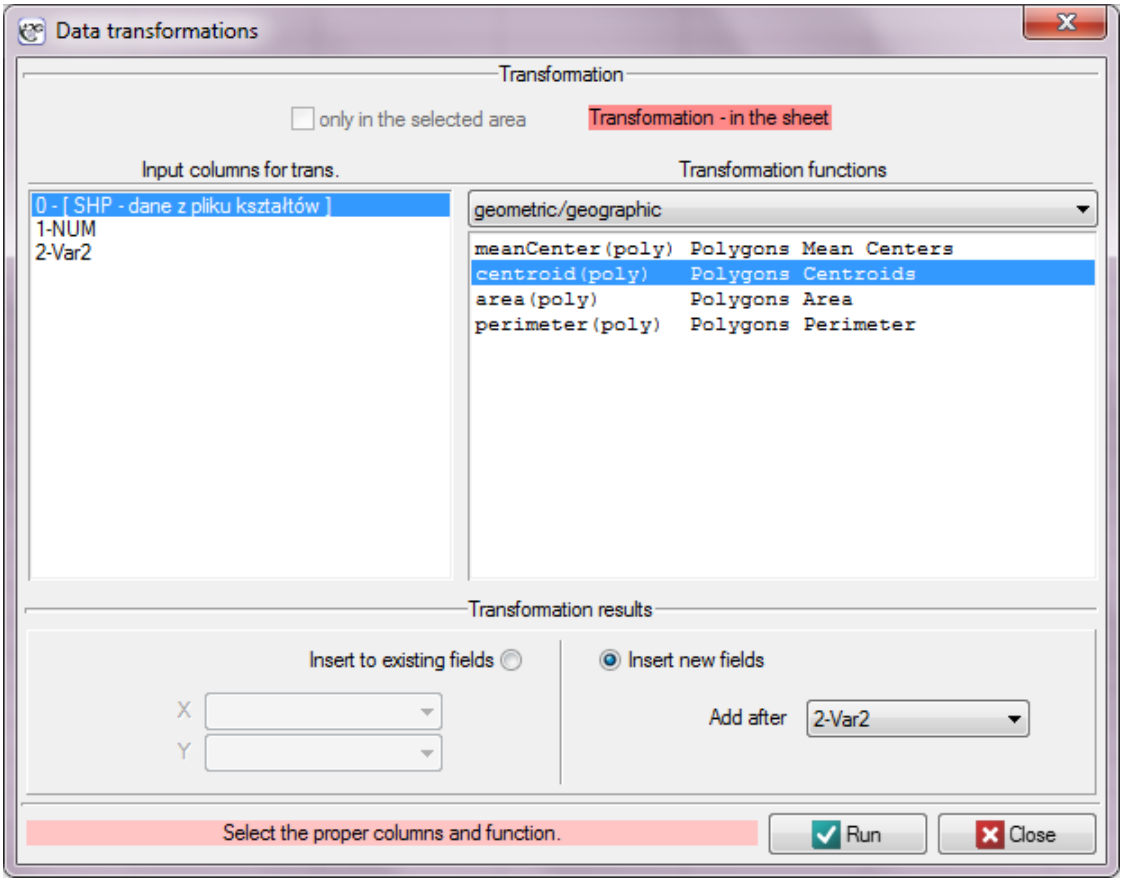
In order to activate all objects one should select the Tools →Activate all menu in the window of the Map manager or the Edit →Activate all menu in the window of PQStat.

1.5 GEOMETRIC CALCULATIONS

Geometric calculations are formulas (read the User Manual - PQStat (Chapter: Formulas)). The formulas can pertain data which describe map geometry and data visible in a datasheet.

- Formulas for data which describe map geometry - **geometric/geographic functions**

  Data for transformation are chosen from a shapefile (SHP)



  Available formulas:

  meanCenter (poly) - gives center coordinates for polygons,
  centroid (poly) - gives centroid coordinates for polygons,
  area (poly) - gives polygon areas,
  perimeter (poly) - gives polygon perimeters.

- Formulas for data visible in a datasheet - **creating maps**

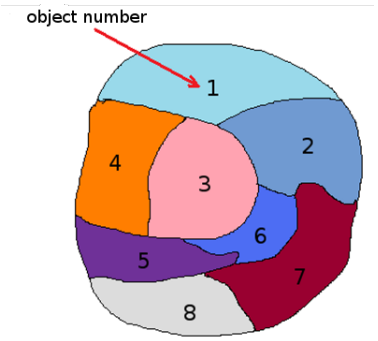
  Available formulas:

  map (points) - gives a vector map presenting points together with assigned datasheet.

1.6 SPATIAL WEIGHTS MATRIX

Spatial relations among objects presented on a map can be organized in the form of a matrix. The matrices are called **weights matrices**. Due to their large size and a great amount of detailed information, weights matrices are not carriers of knowledge which could be presented directly in the results of a conducted study but they form the basis of further analysis. The data included in the matrices are usually treated as weights in spatial analyses and, in this manner, allow to use the information from the map.

The simplest form of a weights matrix is a neighbor matrix. **Neighbor matrix** is a square table with zeros on the main diagonal, where the neighborhood of objects is marked with a binary value (1 – for neighboring objects, 0 – for non-neighboring objects).



no	1	2	3	4	5	6	7	8
1	0	1	1	1	0	0	0	0
2	1	0	1	0	0	1	1	0
3	1	1	0	1	1	1	0	0
4	1	0	1	0	1	0	0	0
5	0	0	1	1	0	1	1	1
6	0	1	1	0	1	0	1	1
7	0	1	0	0	1	1	0	1
8	0	0	0	0	1	1	1	0

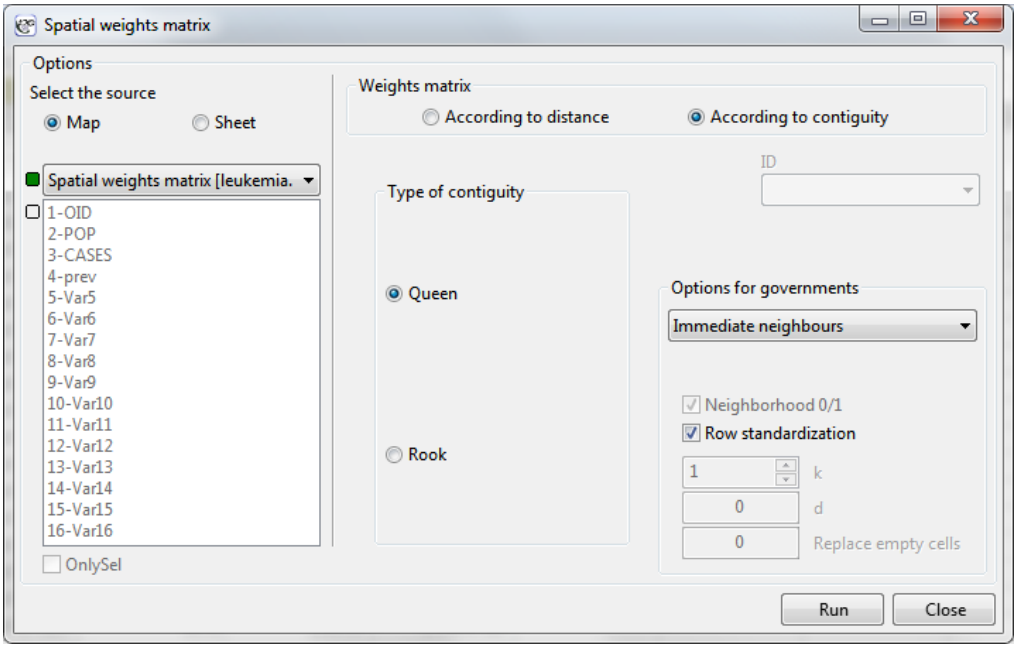
Tabela 1: Example of a neighbor matrix

In statistical analyses the most commonly used matrices are row-standardized matrices with the values of each of its rows summing to one. **Row standardization** means that each weight is divided by the sum of a row (the sum of the weights of all neighboring elements). As a result, the obtained weights are in the range from 0 to 1. The influence of objects with varying numbers of neighbors, in analyses based on a weights matrix standardized in this way, is balanced.

no	1	2	3	4	5	6	7	8
1	0	$1/3$	$1/3$	$1/3$	0	0	0	0
2	$1/4$	0	$1/4$	0	0	$1/4$	$1/4$	0
3	$1/5$	$1/5$	0	$1/5$	$1/5$	$1/5$	0	0
4	$1/3$	0	$1/3$	0	$1/3$	0	0	0
5	0	0	$1/5$	$1/5$	0	$1/5$	1/5	$1/5$
6	0	$1/5$	$1/5$	0	$1/5$	0	$1/5$	$1/5$
7	0	$1/4$	0	0	$1/4$	$1/4$	0	$1/4$
8	0	0	0	0	$1/3$	$1/3$	$1/3$	0

Selected weights matrices should reflect spatial relationships which connect the analyzed objects. The more realistic the reflection of the model of mutual influence of objects in space, the more exact results will be obtained.

The window with settings for weights matrices is accessed via the menu Spacial analysis→Tools→Spatial weights matrix.



1.6.1 Weights matrix according to distance

For creating a weights matrix based on the distances of points we should have at our disposal data from a map which contains objects such as a point, a multipoint, or a polygon. In the case of an analysis of polygons, calculations are based on centroids, and in the case of multipoints they are based on centers of objects.

Description you can find in User Guide - PQStat, section: the similarity matrix.

1.6.2 Weights matrix according to contiguity

For creating a weights matrix based on proximity of objects (contiguity) we should have at our disposal data from a map which contains objects such as a multipoint or a polygon.

Type of contiguity

The contiguity is usually understood as a common section with a non-zero length (i.e. a section longer than 1 point) – it is the **Rook** type neighborhood, or as any section (also of zero length, i.e. a point) – it is the **Queen** type neighborhood.

Weights matrix according to contiguity:

- **Direct neighbors** – it is a square symmetrical matrix in which on the main diagonal there are zeros, the elements outside the diagonal are:

$w_{ij} = 1$ –- if the objects are connected along a common border,

$w_{ij} = 0$ –- in the opposite case.

- **Neighbors (order of contiguity <=k)** – it is a square symmetrical matrix in which on the main diagonal there are zeros, the elements outside the diagonal are:

  $w_{ij} = 1$ –- if the objects are direct neighbors (they are connected along a common border),

  $w_{ij} = 2$ –- if the objects are the second nearest neighbors (the second degree of neighborhood, i.e. the so-called neighbor's neighbor)

  ...

  $w_{ij} = k$ –- if the objects are the k^{th} neighbors (k^{th} degree of neighborhood)

  $w_{ij} = 0$ –- neighborhood is farther than the k^{th} degree.

- **Neighbors (order of contiguity =k)** – it is a square symmetrical matrix in which on the main diagonal there are zeros, the elements outside the diagonal are:

  $w_{ij} = 1$ –- if the objects are the k^{th} neighbors (k^{th} degree of neighborhood)

  $w_{ij} = 0$ –- in the opposite case.

Weights matrices can be row standardized – it is the recommendation of some statistical analyses based on those matrices.

1.7 SPATIAL SMOOTHING

The idea of spatial smoothing is obtaining a better (more stable and less noisy) value of the variable. The most common methods of building such a variable are based on borrowing the information from neighboring areas or on using a larger amount of information from the studied region (L.A. Waller 2004 [14], Luc Anslin 2006 [2]). As a result, the values of the variable under study X with elements $x_1, x_2, ..., x_n$ will be transformed into a new, smoothed variable $smooth(X)$ with elements $smooth(x_1), smooth(x_2), ..., smooth(x_n)$.

The researcher can control the analysis by selecting the distance/neighborhood matrix of the objects, setting the eigenpotential for the object to be smoothed, and indicating the smoothing method.

Spatial weighting matrix

Information about the neighborhood of objects and their mutual distances is defined in the spatial weights matrix. If a neighborhood matrix is used for smoothing - carrying only information information about neighbors (1) or not (0), then only the objects neighbouring with the tested one will have an influence on the obtained result, and the size of this influence will be the same for all neighbors. When a researcher wants to gradate the size of this influence, he should choose a matrix with arbitrary positive values. At the same time, it is important to remember that a larger value in the weight matrix gives a greater influence on the smoothing result. Therefore, in order for closer objects to have a greater influence on the obtained result than distant objects, they

should have a higher weight in the matrix. Such an effect can be achieved by using, for example, an inverse Euclidean distance matrix inside a circle of radius d. Then, closer objects will have a greater influence on the resulting score than distant ones, and the influence of objects outside the circle will be zero.

For more extensive methods of constructing weight matrices, have a look at patial weight matrix and Similarity matrix.

Eigenpotential

The eigenpotential p of the smoothed object determines the amount of influence of information about the test object on the smoothed value for that object.

- **Eigenpotential value**
  The eigenpotential value sets the size of the elements placed on the main diagonal of the weight matrix. By default, the eigenpotential value is set to 1. If it is set to zero ($p = 0$), the smoothed value of the tested object is calculated based only on the information contained in neighboring objects. On the other hand, increasing the value of the eigenpotential increases its share in the calculation of the smoothed value for that object.
- **Potential value correction**
  The setting of the eigenpotential value alone determines the influence of the tested object on the obtained result, but it does not define by how much this influence is to be greater/lesser than that of the neighbouring objects (elements off the main diagonal of the weight matrix). The dependence of the value on the main diagonal of the matrix both on the given value of the potential and on the values of other elements of the matrix allows to determine the size of the influence of the tested object in relation to neighbouring objects. Correction of the potential value is given by the formula:

$$w_{ii} = p \cdot \sum_{j=1, j\neq i}^{n} w_{ij}$$

  As a result, selecting the potential value correction option and setting the potential value to magnitude 3, for example, ensures that the effect of information about the test object on the smoothed value for that object will be three times that of its neighboring objects.

Methods

- **Locally weighted average**
  This transformation consists in calculating the arithmetic mean of the values of the variable X for the object under study (according to the potential) and its neighboring objects (according to the given weight matrix). The observed value x_i is transformed to a smoothed value $smooth(x_i)$ according to the formula:

$$smooth(x_i) = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}$$

  where:

  n –- number of spatial objects (number of points or polygons),
  x_j –- are the values of the variable for the objects being compared,
  w_{ij} –- elements of the spatial weight matrix.

- **Locally weighted median**
  This transformation consists in calculating the median of the values of the variable X for the object under study (according to the potential) and its neighboring objects (according to a given matrix of weights). In order to determine it a neighborhood matrix is necessary, where weights are binary values. The value of one in the matrix means neighboring objects and zero means no neighboring.

- **Locally weighted average (corrected)**
  In the process of smoothing coefficients built on the basis of dividing two variables, determination of locally weighted average can be improved. The numerator and the denominator are smoothed and only then the quotient is created on the basis of these smoothed values. In this way, you can, for example, smooth the incidence rates determined in the course of epidemiological studies, where the numerator is the number of patients and the denominator is the size of the exposed population. In effect, objects with a larger population, will have a greater impact on the result of smoothing - therefore the denominator of the smoothed coefficient is called the adjustment variable.

  The observed value of the $\frac{x_i}{y_i}$ coefficient is converted to a smoothed value $smooth\left(\frac{x_i}{y_i}\right)$ according to the formula:

$$smooth\left(\frac{x_i}{y_i}\right) = \frac{\sum_{j=1}^{n} w_{ij}x_j}{\sum_{j=1}^{n} w_{ij}y_j}$$

  where:

  n –- number of spatial objects (number of points or polygons),
  w_{ij} –- elements of the spatial weight matrix.

- **Empirical Local Bayes Smoothing (corrected)**
  The method of local Bayes smoothing was developed as one way to deal with the instability of the coefficients associated with small data counts and was described in detail by Waller (2004 [14]). Smoothing aims to improve the locally weighted mean (adjusted) so as to reduce its variance.

  The observed value of the $\frac{x_i}{y_i}$ coefficient is converted to a smoothed value $smooth\left(\frac{x_i}{y_i}\right)$ according to the formula:

$$smooth\left(\frac{x_i}{y_i}\right)_{Bayes} = smooth\left(\frac{x_i}{y_i}\right) + C_i\left(\frac{x_i}{y_i} - smooth\left(\frac{x_i}{y_i}\right)\right)$$

  where:

  $smooth\left(\frac{x_i}{y_i}\right)$ – locally weighted average (adjusted),
  C_i –- shrink factor,
  $C_i = \frac{s^2 - \frac{x_i/y_i}{\bar{y}_i}}{s^2 - \frac{x_i/y_i}{\bar{y}_i} + \frac{x_i/y_i}{y_i}}$ jeśli $s^2 - \frac{x_i/y_i}{\bar{y}_i} > 0$
  $s_i^2 = \frac{\sum_{j=1}^{n} e_{ij}}{\sum_{j=1}^{n} y_j w_{ij}}$
  $e_{ij} = y_i\left(\frac{x_i}{y_i}w_{ij} - smooth\left(\frac{x_i}{y_i}\right)\right)$
  $\bar{y}_i = \frac{\sum_{i=1}^{n} y_i}{n}$ –- is the average population size,
  w_{ij} –- elements of the spatial weight matrix.

The shrinkage factor balances the local average $smooth(x_i/y_i)$ with the observed value of the coefficient x_i/y_i. When the population size of the adjustment variable y_i is small, then $C_i \to 0$ nd the estimated value is close to the locally weighted adjusted mean $smooth(x_i/y_i)$. When the population size is large, then $C_i \to 1$ and the estimated value approaches the true value observed at that site x_i/y_i.

2 TESTING HYPOTHESES

Verification of statistical hypotheses is checking certain assumptions formulated for parameters of a general population on the basis of results from a sample.

Formulation of hypotheses which will be verified with the help of statistical tests.

Each statistical test gives the general form of a null hypothesis –$\mathcal{H}_0$ and of an alternative hypothesis –$\mathcal{H}_1$:

$\mathcal{H}_0$: **in the studied population THERE IS NOT** a statistically significant
e.g. dependence,
e.g. difference,
...
between
e.g. spatial distribution,
e.g. presence of particular values,
...
in the analysed area,

$\mathcal{H}_1$: **in the studied population THERE IS** a statistically significant
e.g. dependence,
e.g. difference,
...
between
e.g. spatial distribution,
e.g. presence of particular values,
...
in the analysed area.

Example:

$\mathcal{H}_0$: THERE IS NOT a statistically significant dependence between the spatial distribution of chemist's shops in Wielkopolska –we assume that their distribution in the studied area is random.

If we do not know if the distribution of the shops can be more regular than random distribution, or the other way round –more clustered than random distribution, then the alternative hypothesis should be two-sided, i.e. we do not presume a particular direction:

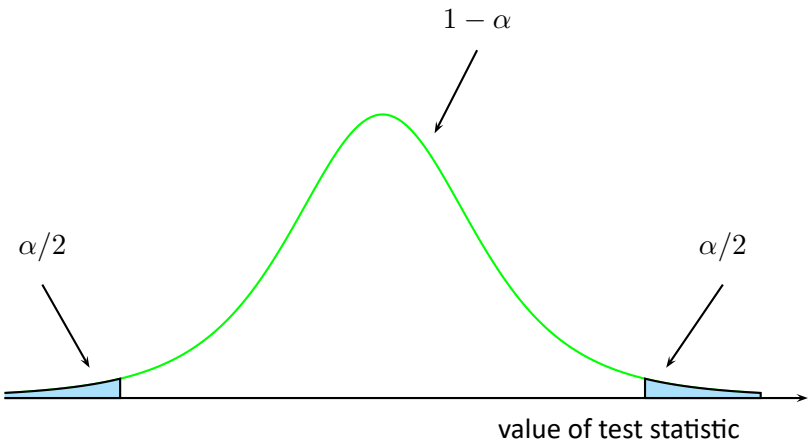
$\mathcal{H}_1$: THERE IS a statistically significant dependence between the spatial distribution of chemist's shops in Wielkopolska –we assume that their distribution in the given area is not random, i.e. we presume the presence of 2 directions: a distribution which is more regular than random distribution and a distribution which is more clustered than random distribution.

It may happen (in very rare cases) that we are certain that we know the direction in the alternative hypothesis. We can then utilize a one-sided alternative hypothesis.

Hypothesis Verification

To check which of the hypotheses, $\mathcal{H}_0$ or $\mathcal{H}_1$, is more probable, we select a proper statistical test.

Test statistic of a chosen test, calculated according to its formula, is subjected to the theoretical distribution appropriate for that statistic.



The program calculates the value of a test statistic and p-**value** for that statistic (that is the part of the area under the curve which corresponds to the value of the test statistic). Value p allows to choose which hypothesis, the null hypothesis or the alternative hypothesis, is more probable. The truth of the null hypothesis is always presumed and the proofs gathered in the data are to provide a sufficient number of arguments against that hypothesis:

$$\begin{aligned} &\text{if } p \leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ &\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Usually, **significance level** $\alpha = 0.05$ is chosen with the acceptance of the premise that in 5% of situations the null hypothesis will be rejected being a true one. In special cases a different significance level, e.g. 0.01 or 0.001, can be set.

3 DESCRIPTIVE STATISTICS

To conduct Descriptive Statistics on the basis of a Map data we should have at our disposal a point, multipoint, or polygonal file. In the case of an analysis of a polygonal file, calculations are based on centroids of polygons, and in the case of a multipoint file they are based on centers of objects.

Boundaries of an area in which analysed points are enclosed can be defined, depending on a particular need, with the help of: a convex hull, the smallest rectangle, a rectangle from from layer bounding, or the smallest circle. The studied area can also be defined only with the use of the size of its area.

The distance between the points is measured with the Euclidean metric.

The basic statistics made for point analysis:

- A –the area of a studied region,
- n –the size of a sample, i.e. the number of points lying within the studied region,
- $D = \frac{n}{A}$ –density,
- descriptive statistics of the distance matrix between points:
  - arithmetic mean with confidence interval,
  - standard deviation,
  - median,
  - quartiles,
  - minimum and maximum.

The analysis also gives a graph pertaining to a distance matrix and layers which can be drawn on the surface of a map. Layers pertain to centrographic measures: the measure of central tendency and the measure of dispersion:

- The center of point distribution: the mean of coordinates of the X axis and the Y axis ($\overline{x}$, $\overline{y}$),
- The area of standard deviations, built around the center, defined by:
  - Circle
    The radius of the circle is sdd –standard distance from the center (*standard distance deviation*) expressed with the formula:

$$sdd = \sqrt{\frac{\sum_{i=1}^{n} x_i'^2 + \sum_{i=1}^{n} y_i'^2}{n-2}},$$

    where:
    $x_i' = x_i - \overline{x}$,
    $y_i = y_i - \overline{y}$.

– Ellipse
The angle of the inclination of an ellipse axis (Y) with respect to the coordinate system (OY axis) is expressed with the formula:

$$\theta = \text{arc tg}\left(\frac{A+B}{C}\right),$$

where:
$A = \sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2,$
$B = \sqrt{\left(\sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2\right)^2 + 4\left(\sum_{i=1}^{n} x_i' y_i'\right)^2},$
$C = 2\sum_{i=1}^{n} x_i' y_i'.$

The lengths of the semiaxes of an ellipse:

$$\sigma_x = \sqrt{\frac{2}{n-2}\sum_{i=1}^{n}\left(x_i' \cos\theta - y_i' \sin\theta\right)^2}$$

$$\sigma_y = \sqrt{\frac{2}{n-2}\sum_{i=1}^{n}\left(x_i' \sin\theta + y_i' \cos\theta\right)^2}$$

– Rectangle
The lengths of rectangle sides are: $a = 2sd_x$, $b = 2sd_y$, where sd_x and sd_y are standard deviations for the coordinates of the X and Y axes

After the weights for particular objects have been defined, we calculate the weighted center of point distribution and the weighted circle representing the standard deviation area.

- The weighted center of point distribution: the weighted mean of coordinates of the X axis and the Y axis:

$$\overline{x_w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \qquad \overline{y_w} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

where:
w_i –weights representing the value of a feature in the ith object.

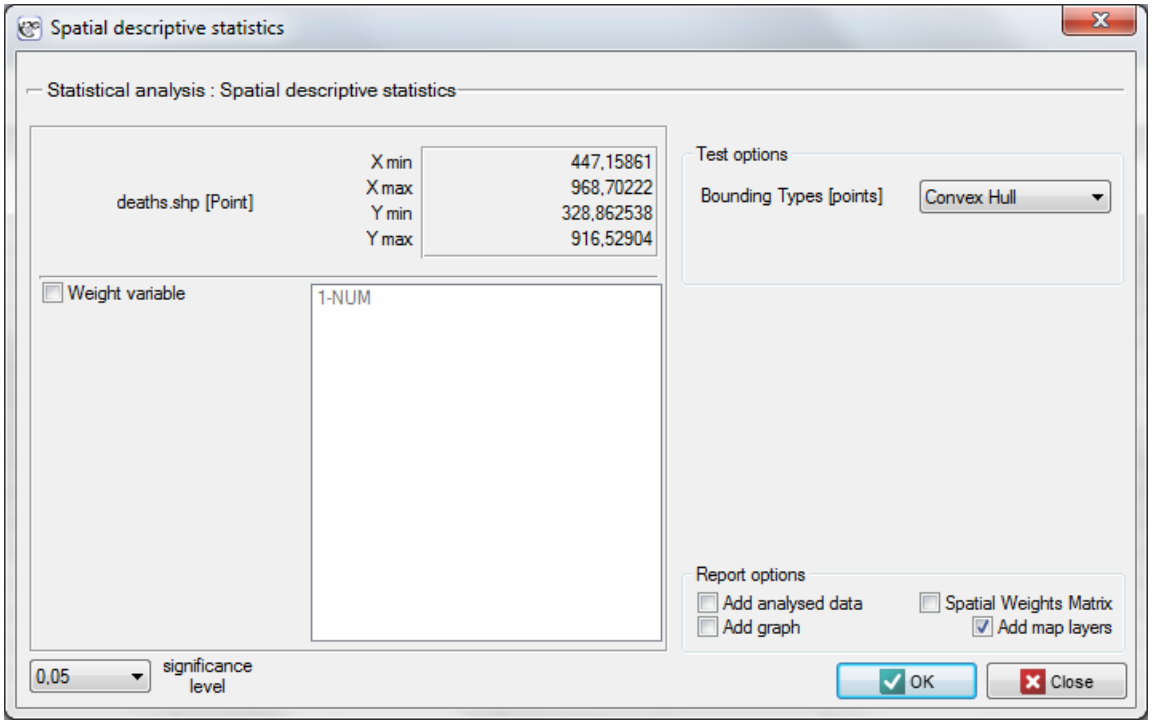
- Weighted circle
The radius of the circle is $wsdd$ –weighted standard distance from the center expressed with the formula:

$$wsdd = \sqrt{\frac{\sum_{i=1}^{n} w_i x_i^{*2} + \sum_{i=1}^{n} w_i y_i^{*2}}{\sum_{i=1}^{n} w_i - 2}},$$

where:
$x_i^* = x_i - \overline{x_w}$,
$y_i^* = y_i - \overline{y_w}$.

Note
In the formulas concerning the lengths of the radius of a circle and of a semiaxis of an ellipse, the denominator was decreased by value 2 –Buliung (2008)[3], Smith (2007)[11].

The window with settings for Descriptive statistics is accessed via the menu Spacial analysis → Spatial descriptive statistics.



EXAMPLE 3.1. (directory: snow, SHP files: deaths, pumps, streets)

Data for the analysis are probably the best known, classical example of the use of cartography in epidemiology. They present the epidemic of cholera in London in 1854. The map which presents the range of the epidemic was made by John Snow, a doctor and the discoverer of the cause of the epidemic, considered to be one of the founders of epidemiology. The coordinates of points which constituted the basis for drawing the maps come from the original John Snow's map which was digitalized by Rusty Dodson from the US National Center for Geographic Information Analysis (http://ncgia.ucsb.edu/Publications/Software/cholera/) and later presented in meters.

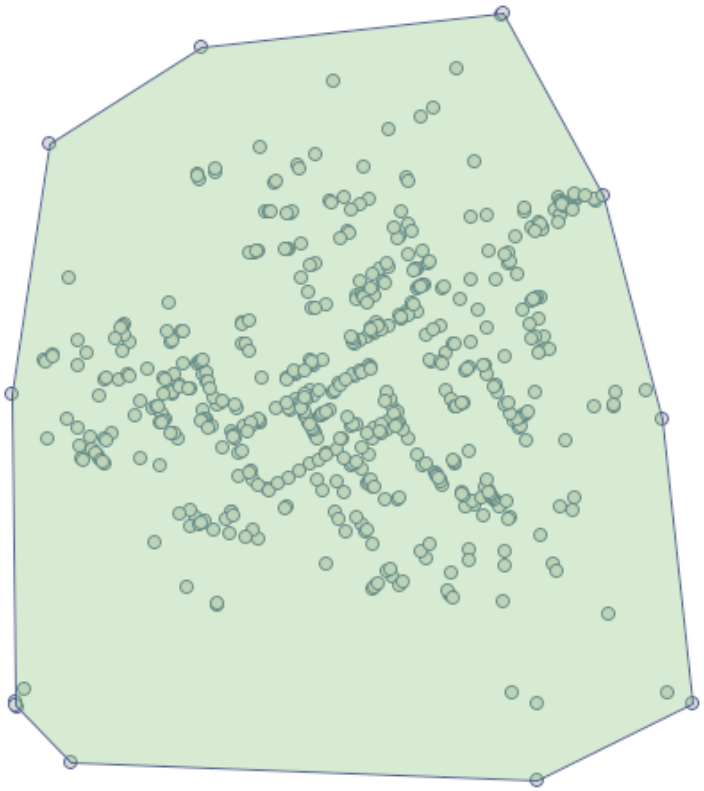
- The map deaths contains information about the location of 578 points (deaths due to cholera) in Soho –a London district.
- The map pumps contains information about the location of 13 points (water pumps) in Soho.
- The map streets contains information about the location of lines (streets) in Soho.

After importing the above shapefiles (SHP) we can view and edit each of them in the Map manager.

To conduct an analysis we select the deaths map and perform the Spatial descriptive statistics. Because we will utilize the map coordinates as data for the analysis, in the descriptive statistics window we select the option Use points from map coordinates and, as the bounding type, we select the Convex Hull.

Spatial descriptive statistics	->> + MAP <<-
Analysis time	0,19sec.
Analysed variables	SHP_X;SHP_Y
Significance level	0,05
Bounding Types	Convex Hull
Number of points	578
Area	257531,649115
Density	0,002244
Descriptive statistics of the distance matrix	
Arithmetic mean	171,909909
-95% CI for the group mean	171,465637
+95% CI for the group mean	172,354181
Standard deviation	92,562385
Median	160,616834
Lower quartile	102,745253
Upper quartile	229,246456
Minimum	0
Maximum	662,896352

The area in which there are the points (defined by the convex hull) is $0.257531 km^2$. We can draw them on the map by pressing the button ->> + MAP <<- and selecting the layer of object bounding.



There is on average over 2 points per $1000 m^2$ (density=0.002244 points per m^2).
The analysis of the point distance matrix allows a more exact evaluation of their density.

Some points are in the same place because the smallest distance is $0m$. There are also points at a far greater distance from each other –the greatest distance is $662.896352m$. We can also find information about the average distance and the standard deviation of the points here.

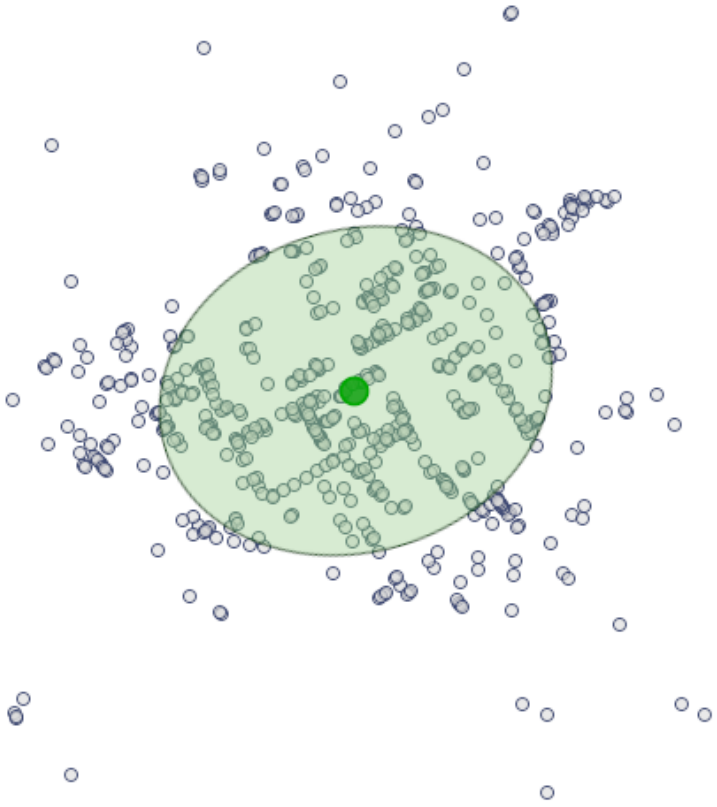
The most interesting information in the analysis of the deaths map is offered by the localized Center of point distribution (703.79, 631.65), together with the area of standard deviations which describe the the degree of concentration and the direction of dispersion (circle, ellipse, rectangle).

Circle (standard distance deviation)			
Mean X	Mean Y	r=sdd	Area
703,78827	631,64920	138,17912	59983,908

Rectangle (standard deviations)				
Mean X	Mean Y	a= 2sdX	b= 2sdY	Area
703,78827	631,64920	210,94859	178,16343	37583,326

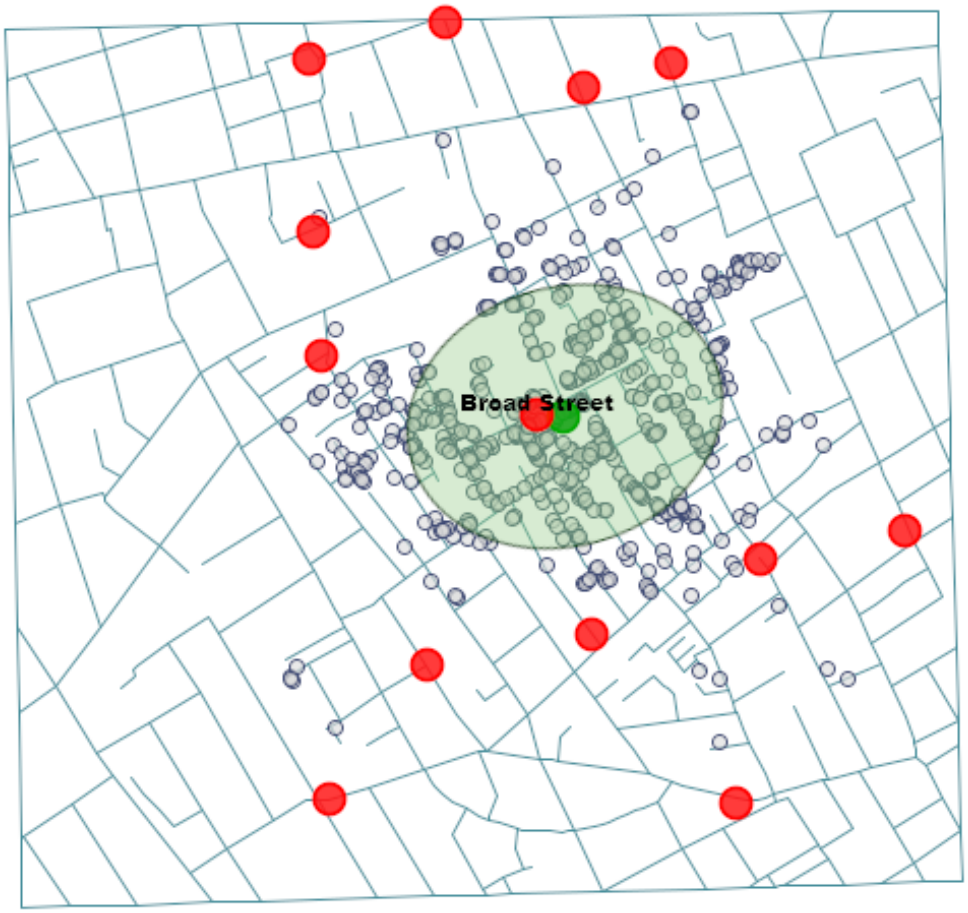
Standard deviational ellipse					
Mean X	Mean Y	Semiaxes	Semiaxes	Angle Y	Area
703,78827	631,64920	123,29712	151,60726	287,4	58724,971

The ellipse of standard deviations and the Center is drawn again by moving on to the map manager (on the layer list we uncheck the bounding).



As a result of conversations with local people, Snow suspected that water could have been the source of the epidemic. When the three maps are joined we can identify the water

pump the water from which turned out to be the cause of the epidemic. To find it we should first display the streets map in the Map Manager and next we should overlay the deaths map and the pumps onto it by pressing the button .



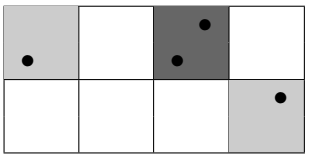
The source of the epidemic turned out to be the water pump on the Broad Street (we can display its label in the Map Manager). That is the only pump which was in the selected elliptical area, and its location (678.85, 633.27) and the location of the middle of the ellipse (703.79, 631.65), i.e. the place around which the deaths centered, are very close to each other.

4 DENSITY ANALYSIS

To perform a density analysis, we should have map data containing objects of type: point, multipoint or polygon. In the case of polygon analysis, calculations are based on centroids, and in the case of multipoints on the centers of objects.

4.0.1 Quadrat Count Methods

Graphically, this method is a generalization of a histogram, or one-dimensional analysis, to a two-dimensional case. Building a histogram we have one variable, which we divide into intervals of equal length and give the number of cases in each interval. When building a grid of squares, we have two variables on which we build the grid and give the number of cases in each grid square (DPS – Dot Per Square). The ratio of this number to the area of a square determines the intensity of the color in which a given grid square is colored.



Based on the number of casess in the grid squares, we can study their spatial distribution. If there are the same number of points in each square, it means perfectly uniform distribution. When the opposite is true, when the variation in the number of points in the squares is very large, it means that there are squares with a much larger number of points, that is, clusters are formed.

If we denote by n the number of points of the study area and by m the number of squares into which the study area is divided, then we can determine the mean, variance, and standard deviation of the number of points per square:

$$\overline{DPS} = \frac{n}{m}, \quad Var_{(DPS)} = \frac{\sum_{i=1}^{k} m_i(n_i - \mu)^2}{m-1}, \quad SD_{(DPS)} = \sqrt{var},$$

wherem_i – is the number of squares with the number of points equal to n_i.

Coefficient $VMR_{(DPS)}$

The most important information is provided by the variance-mean ratio – the coefficient, which is the quotient of the variance and the mean:

$$VMR_{(DPS)} = \frac{Var}{\overline{DPS}}$$

A value of $VMR_{(DPS)} < 1$ indicates too little variation in the number of points in squares which suggests a uniform dispersion effect, $VMR_{(DPS)} > 1$ indicates too much variation in the number of points in squares and therefore a clustering effect, and a value close to 1 indicates an average

variation in the number of points in squares which implies a random distribution of points.

The Index of Cluster Size (ICS) is often considered in the literature:

$$ICS_{(DPS)} = VMR_{(DPS)} - 1$$

The expected value of $ICS_{(DPS)}$ assuming random points is 0. A positive value indicates a clustering effect and a negative value indicates a regular distribution of points.

Significance of the coefficient $VMR_{(DPS)}$

The $VMR_{(DPS)}$ coefficient significance test is used to verify the hypothesis that the observed point counts n the squares are the same as the expected counts that would occur for a random distribution of points.
Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad VMR_{(DPS)} = 1, \\ \mathcal{H}_1: &\quad VMR_{(DPS)} \neq 1. \end{aligned}$$

The test statistic has the form:

$$\chi^2 = (m-1) \cdot VMR_{(DPS)}.$$

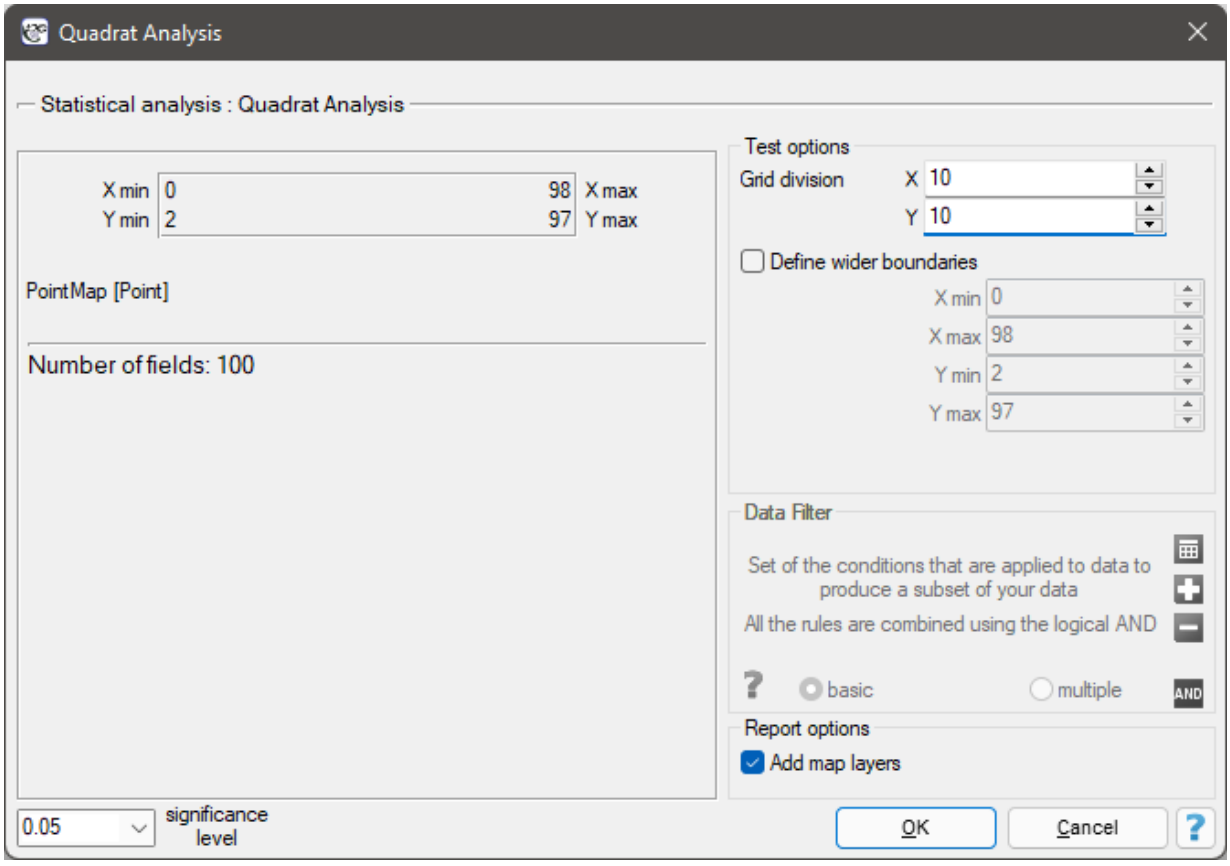
This statistic has an asymptotically χ^2 distribution with $df = m - 1$ degrees of freedom.

Value p, defined on the basis of test statistics, is compared with the significance level α :

$$\begin{aligned} &\text{if } p \leq \alpha \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ &\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Note
The result of the analysis depends to a large extent on the density of the grid and thus on the number/size of squares into which the analyzed area is divided. In the test options window, you can set the grid that will be used to divide the test area into squares by specifying the number of squares vertically and horizontally.

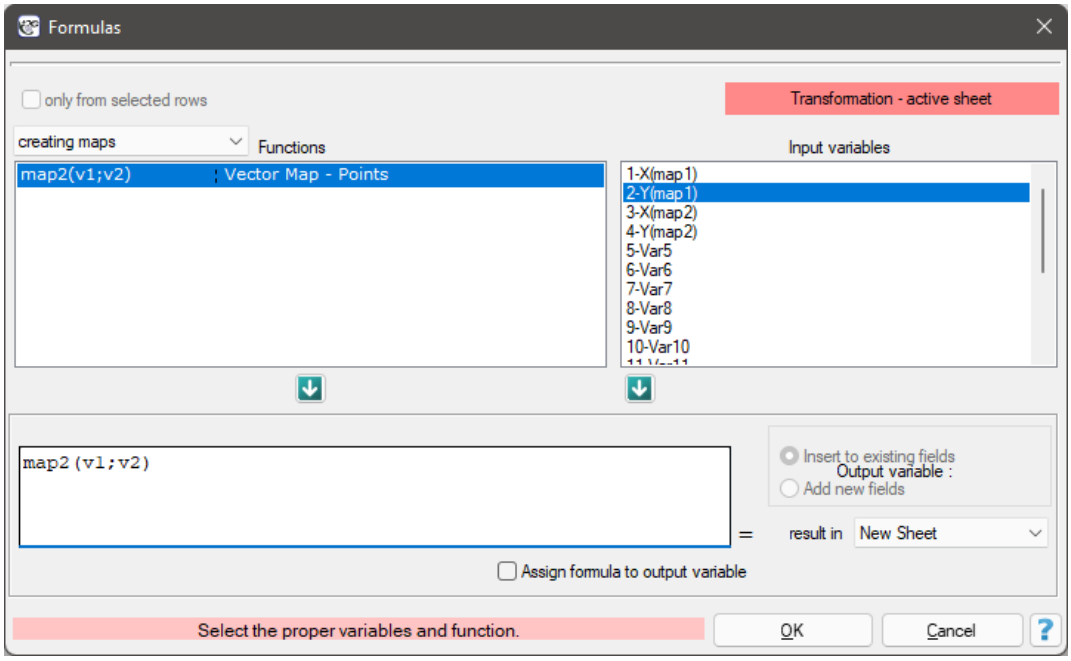
The window with the settings for the quadrat count method is launched via the menu Spatial Analysis→Spatial Statistics→Quadrat analysis



EXAMPLE 4.1. (file squares.pqs)

Using the datasheet, generate two point maps and perform a density analysis of these points. Answer the question: are the points randomly distributed in each of these maps?

You create the point maps using the formulas: menu Data→Formulas...



This will result in two new sheets containing maps. For each of these sheets, we perform a quadrat analysis.

Hypotheses:

$\mathcal{H}_0$: the distribution of points in the population from which the sample is drawn is random ,
$\mathcal{H}_1$: the distribution of points in the population from which the sample is drawn is not random.

The results for Map 1 indicate a significant variation in the number of points in the squares, that is, a clustering effect (value $p =< 0.00001$). This effect persists for different grid densities. For a grid density of 10:10 the VMR ratio is as high as 12.5, the entire report is included below:

Quadrat Analysis	[Add to Map]
Analysed variables	SHP_X;SHP_Y
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Bounding Types [points]	Bounding Rectangle
Number of points	100
Area	9310
Number of squares	100
Mean of the number of points per quadrat	1
Variance of the number of points per quadrat	12.5051
Standard deviation of the number of points per quadrat	3.5362
VMR (variance/mean ratio)	12.5051
Chi-square statistic	1238
Degrees of freedom	99
p-value	<0.0001
ICS (index of cluster size)	11.5051

For map 2, the situation is quite different. For the 10:10 density grid, we have a lack of statistical significance (value $p = 0.95847$) and the value of the coefficient $VMR = 0.77$ indicate that the distribution of points is random.

Quadrat Analysis	[Add to Map]
Analysed variables	SHP_X;SHP_Y
Number of unspecified	0
Number of missing data	0
Significance level	0.05
Bounding Types [points]	Bounding Rectangle
Number of points	100
Area	9801
Number of squares	100
Mean of the number of points per quadrat	1
Variance of the number of points per quadrat	0.7677
Standard deviation of the number of points per quadrat	0.8762
VMR (variance/mean ratio)	0.7677
Chi-square statistic	76
Degrees of freedom	99
p-value	0.9585
ICS (index of cluster size)	-0.2323

Using the ->> + MAP <<- button in the report, we move to the Map Manager to select the analysis grid from the displayed list of layers and obtain a graphic interpretation of the results.

Map1

0	0	0	1	1	0	0	1	0	1
2	0	1	0	1	0	0	1	0	0
0	2	0	1	0	2	1	1	2	0
0	1	0	1	0	0	0	0	0	0
0	1	2	0	0	0	0	0	0	0
1	1	1	0	1	0	1	0	0	0
0	0	1	1	0	1	28	1	0	0
0	1	0	0	0	0	2	0	0	0
20	1	0	1	0	0	1	0	0	0
10	0	0	1	1	0	0	1	1	0

Map2

1	2	1	1	0	3	1	1	1	3
1	3	1	0	1	1	1	1	2	0
0	1	2	0	0	1	0	1	0	0
2	0	0	0	1	0	0	1	3	0
1	1	1	1	1	3	1	1	0	0
1	0	0	1	1	0	1	0	1	1
3	1	1	2	0	1	1	1	1	1
0	0	1	2	2	1	1	1	2	1
0	1	0	2	1	1	0	0	2	1
2	1	0	2	2	1	0	3	2	3

4.1 Kernel density estimator

4.1.1 Two-dimensional kernel estimator

The two-dimensional kernel estimator (like the one-dimensional estimator) allows the distribution of the data, expressed by the method of squares, to be approximated by smoothing.

The two-dimensional kernel density estimator approximates the density of a data distribution by creating a smoothed density plane in a non-parametric manner. It produces a better density estimate than is given by the traditional method of squares, whose squares form a step function. As in the one-dimensional case, this estimator is defined based on appropriately smoothed summed kernel functions (see description in the PQStat User Manual). There are several smoothing methods to choose from and several kernel functions described for the one-dimensional estimator (Gaussian, uniform, triangular, Epanechnikov, quartic/biweight). While the kernel function has little effect on the resulting plane smoothing, the smoothing factor does.

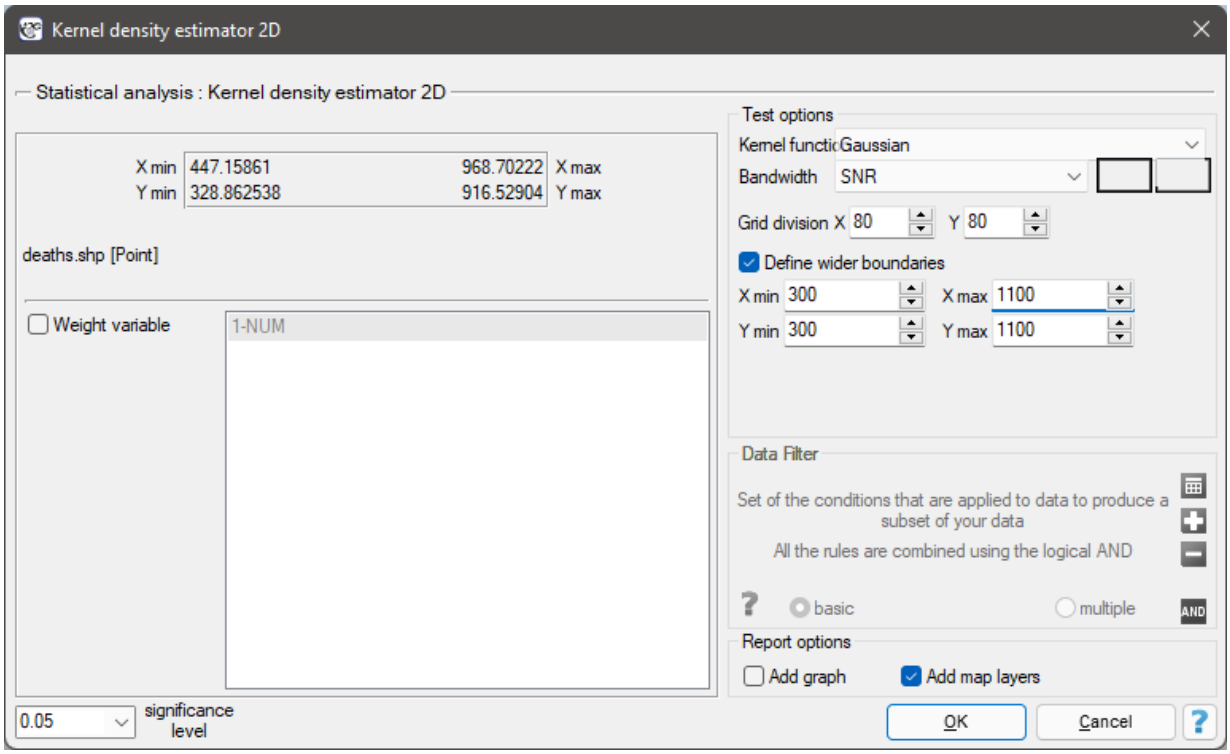
For each point x in the range defined by the data, the density or kernel estimator is determined. It is obtained by summing the product of the kernel function values at that point:

$$\hat{f}_K(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_h(t_i)K_h(s_i)$$

If we give the individual cases weights w_i, then we can construct a weighted nuclear density estimator defined by the formula:

$$\hat{f}_K(x,y) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i K_h(t_i)K_h(s_i)$$

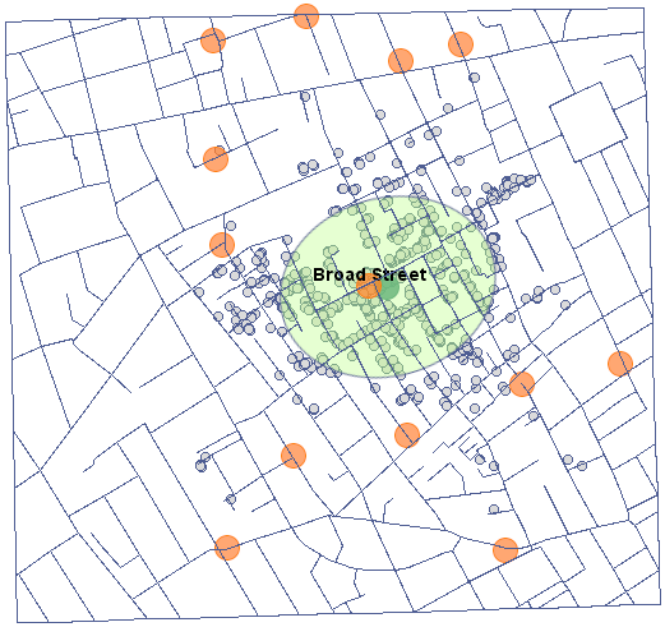
The window with settings for the kernel 2D density estimator ptions is launched via the menu Spatial analysis→Spatial statistics→Kernel density estimator 2D



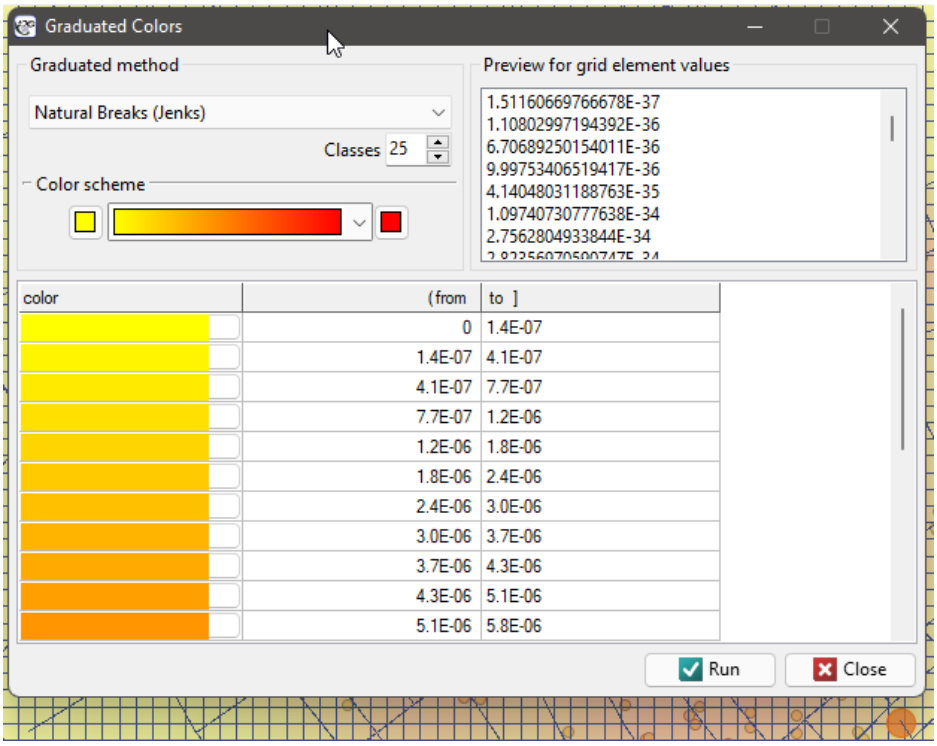
EXAMPLE (3.1) c.d. *(snow.pqs file)*

> Currently, the main problem in presenting point data on the location of people is the need to protect them. Data protection prohibits publishing research results in such a way, that it would be possible to recognize a given person on their basis. A good solution in this case is a point density estimator.

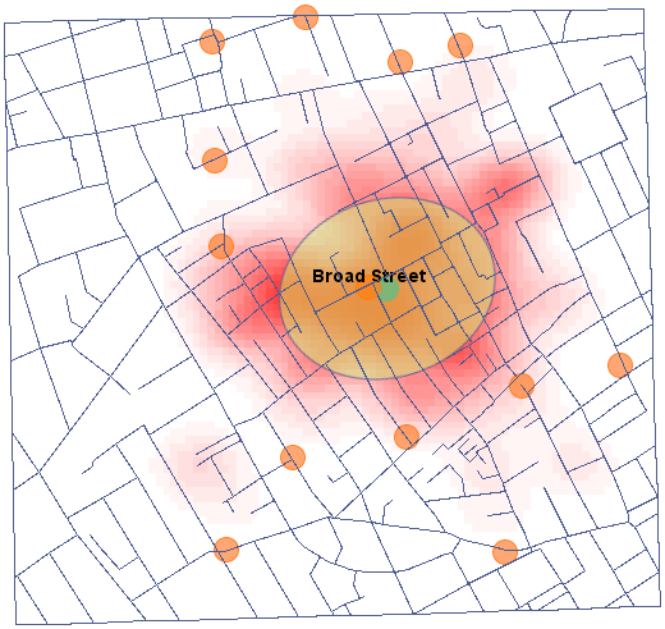
We will present point data illustrating the cholera epidemic in London in 1854 using such an estimator. To do so, we will use a map of points (deaths due to cholera) with layers already overlaid to illustrate both streets and water pumps, and the result of an analysis by physician John Snow.



In the analysis window for the point map, we will stay with the Gaussian (normal) distribution kernel and the SNR smoothing factor. The grid density will be set to 80:80 and the boundaries will be increased so that the edges do not have a sharp edge by entering 300 as the minimum value for the X and Y coordinates and 1100 as the maximum value.
Using the ->> + MAP <<- button in the report, we go to the Map Manager, where we can add a layer representing this estimator (the last item in the list of layers).



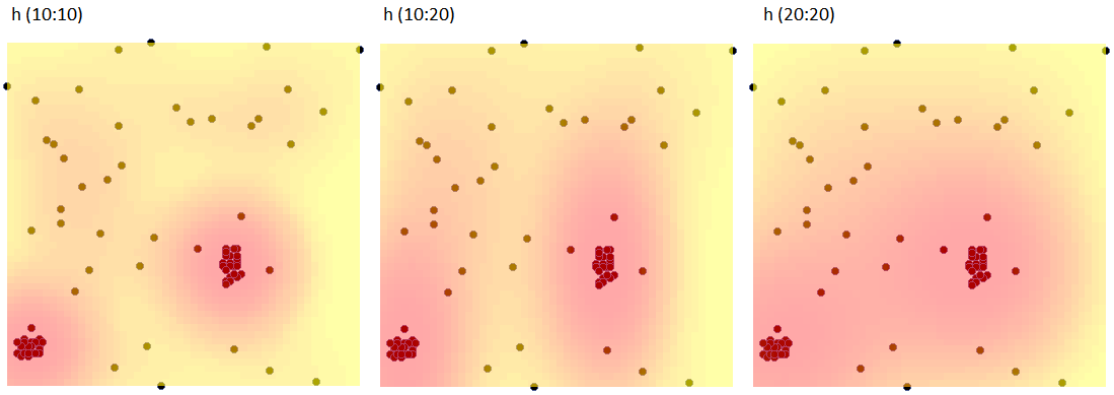
After applying the nuclear density estimator layer, edit it o remove the grid lines and change the yellow color to the natural background color (white in this case). The layer thus obtained is moved up, so that it is drawn at the beginning. We turn off the points layer (Base Map).

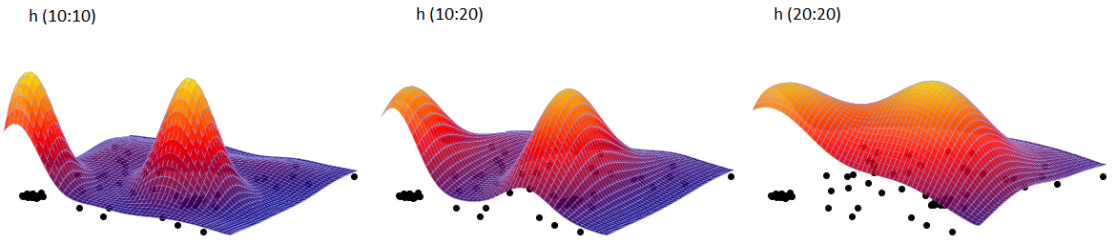


EXAMPLE (4.1) cont. *(squares.pqs file)*

Using the kernel estimator, we represent the point density for map 1 - obtained in the earlier part of the task.

In the analysis window, we set the grid density to 50:50 and the kernel type as normal distribution and include a graph. We perform the analysis three times while changing the User smoothing factor: h (10:10), then h (10:20) and h (20:20). The obtained results presented on the map (via Map Manager) and on the 3D graph are shown below:





4.1.2 Three-dimensional kernel estimator

The three-dimensional kernel estimator (like the one-dimensional estimator and the two-dimensional estimator) allows you to approximate the distribution of the data by smoothing it.

The three-dimensional kernel density estimator approximates the density of the data distribution by creating a smoothed density plane in a non-parametric way. Graphically, we can represent it by plotting the first two dimensions in layers created by the third dimension. As in the one-dimensional case (see description in the PQStat User's Guide) and the two-dimensional estimator, this estimator is defined based on appropriately smoothed summed kernel functions. There are several smoothing methods to choose from and several kernel functions described for the one-dimensional estimator (Gaussian, uniform, triangular, Epanechnikov, quartic/biweight). While the kernel function has little effect on the resulting plane smoothing, the smoothing factor does.

For each point x in the range defined by the data, the density that is the kernel estimator is determined. It is formed by summing the product of the kernel function values at that point:

$$\hat{f}_K(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} K_h(t_i)K_h(s_i)K_h(r_i)$$

If we give the individual cases weights w_i, then we can construct a weighted kernel density estimator defined by the formula:

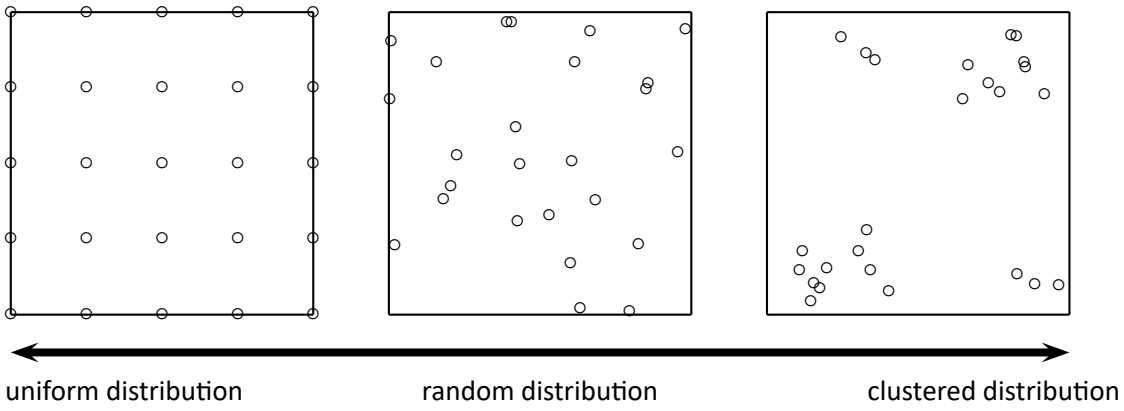
$$\hat{f}_K(x,y,z) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i K_h(t_i)K_h(s_i)K_h(r_i)$$

The window with settings for the kernel 3D density estimator options is launched via the menu Spatial analysis→Spatial statistics→Kernel 3D density estimator **Note**
Displaying subsequent layers of the estimator, determined by the third dimension, is possible by editing the layer in the map Manager window and selecting the appropriate layer index.

5 RANDOMNESS OF POINT DISTRIBUTION

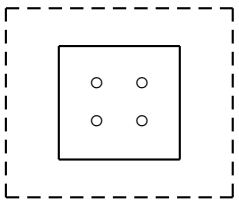
To conduct analysis of the randomness of point distribution on the basis of a Map data we should have at our disposal a point, multipoint, or polygonal file. In the case of an analysis of a polygonal file, calculations are based on centroids of polygons, and in the case of a multipoint file they are based on centers of objects.

The effect of uniform dispersion appears when points are distributed more regularly than the possible result of random distribution. If the spatial distribution is as probable as any other distribution we speak about spatial randomness. When the points come in groups we speak about clustered distribution.



5.1 Nearest Neighbor Analysis

In the Nearest Neighbor Analysis the boundaries of the area in which the analysed points are enclosed have the crucial influence on the result. The example below illustrates regularly distributed points and their clustered distribution when bounded by a large rectangle.



Depending on the needs, the bounding can be defined with the help of: a convex hull, the smallest rectangle, a rectangle from layer bounding, or the smallest circle. The studied area can also be defined only with the use of the size of its area.

The distance between the points is measured with the Euclidean metric.

The first stage of the nearest neighbor analysis is calculating the distance among all points. Next, for each point we search for the nearest point, i.e. for the nearest neighbor (NN).

Note
The distances between all points are defined by a spatial weight matrix. In Moran's analysis window we can choose matrix generated previously by using menu Spatial analysis → Tools → Spatial weights matrix or indicate the neighbor matrix according to contiguity – Queen, row standardized, that is proposed by the program.

The basic statistics for the analysis of the nearest neighbors are:

- $d_i(NN)$ –the distance of each point from its nearest neighbor,
- $\overline{NN}$ –the mean nearest neighbor distance:

$$\overline{NN} = \sum_{i=1}^{n} \frac{d_i(NN)}{n}$$

- $SD_{(NN)}$ –standard deviation of the nearest neighbors distance,
- $\overline{ran}$ –mean random nearest neighbor distance:

$$\overline{ran} = \frac{0.5}{\sqrt{\frac{n}{A}}}$$

Nearest Neighbor Index

Nearest Neighbor Index (NNI) is based on a method described by botanists: Clark and Evans (1954) [4]. NNI compares distances observed between the nearest points, and distances which would appear for a random distribution of points.

$$NNI = \frac{\overline{NN}}{\overline{ran}}$$

When the compared distances are the same then $NNI = 1$. When the observed distances between the nearest points are smaller than expected then the points are nearer to one another than in a random distribution, and $NNI < 1$. In such a case, clusters occur. When the situation is reverse, then $NNI > 1$, which points to the occurrence of the effect of uniform distribution, i.e. points are distributed more regularly than in a case of random distribution.

Significance of the Nearest Neighbor Index

The test for checking the significance of the Nearest Neighbor Index NNI serves the purpose of verifying the hypothesis that the distances observed between the nearest points are the same as the expected distances which would appear in a random distribution of points.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad NNI = 1, \\ \mathcal{H}_1: &\quad NNI \neq 1. \end{aligned}$$

The test statistic has the form presented below:

$$Z = \frac{\overline{NN} - \overline{ran}}{SE_{\overline{ran}}},$$

where:

$SE_{(ran)} = \sqrt{\frac{4-\pi A}{4\pi n^2}}$ –standard error of the mean random nearest neighbor distance

Statistics Z asymptotically (for a large sample size) has the normal distribution.

Value p, defined on the basis of test statistics, is compared with the significance level α :

$$\begin{array}{lll} \text{if } p \leq \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

Analysis of Subsequent Nearest Neighbors

To analyse subsequent nearest neighbors one takes into account the distance to the second nearest neighbor, the third nearest neighbor, and so on, to the k-order nearest neighbor. For the neighborhood of each order (from the nearest neighbor to the k-order neighbor) subsequent Nearest Neighbor Indexes $k_{ordered}NNI$ are calculated:

$$k_{ordered}NNI = \frac{k_{ordered}\overline{NN}}{k_{ordered}\overline{ran}}$$

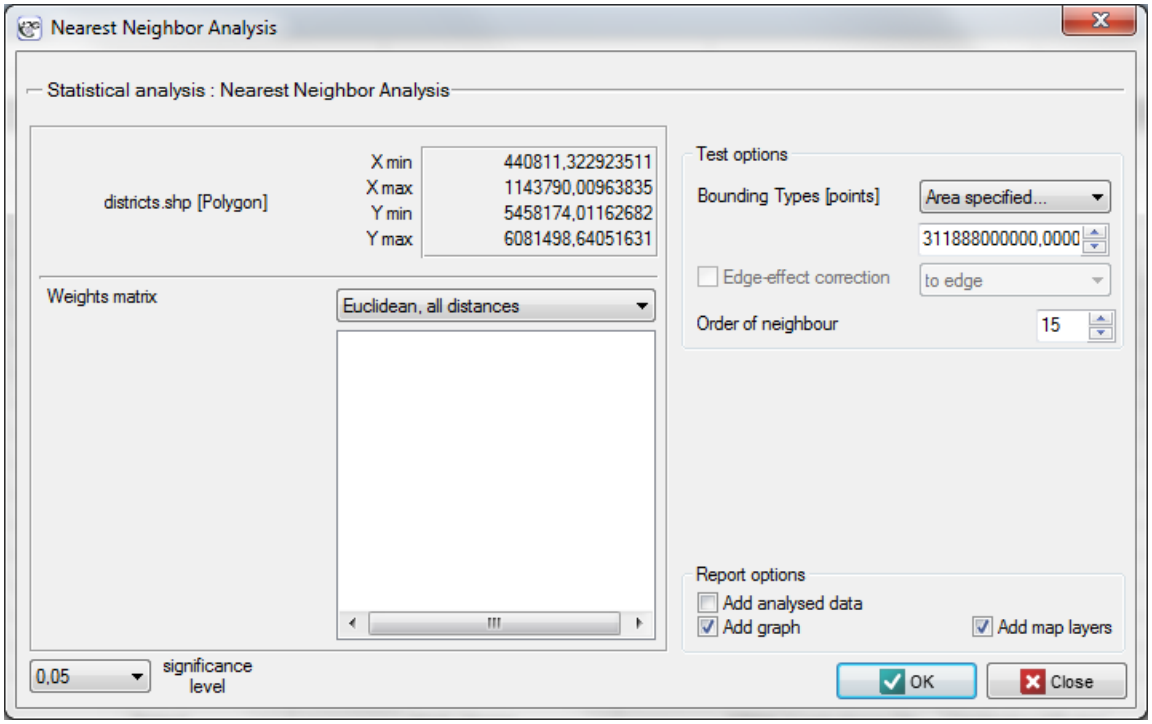
where:
$k_{ordered}\overline{NN}$ –mean distance from neighbors of k-order,
$k_{ordered}\overline{ran} = \frac{k(2k)!}{(2^k k!)^2 \sqrt{\frac{n}{A}}}$ –mean random distance from neighbors of k-order.

The results of the point density analysis conducted for subsequent neighbors can be presented on a graph so as to illustrate the placement of NNI in reference to the line which shows the random point structure and so as to check if a growing or falling trend has been achieved for the indexes.

Edge Effect

Objects placed near the bounding show a tendency to be further away from their nearest neighbors than other objects within the analysed area. The reason for it is the simple fact that the nearest neighbors of the objects near the border can be objects outside the studied area. In such a situation we can conduct an analysis with an adjustment for the edge effect. In such a case the distance of a point from its nearest neighbor ($d_i(NN)$) is calculated as the minimum distance of the point from its neighbors and from the boundary. Thus, if the distance of the point from the boundary will be smaller than the distance from its neighbors, then the distance from the boundary is considered to be $d_i(NN)$. However, such a calculation of the nearest neighbor requires an assumption that there will always be a neighboring point on the border.

The window with settings for Nearest Neighbor Analysis is accessed via the menu Spacial analysis → Spatial Statistics → Nearest Neighbor Analysis.



EXAMPLE 5.1. (directory: districts, SHP files: districts)

The admistrative division of Poland into powiats should, by definition, be uniform. With the use of NNI we will check if that is the case.

- The districts map contains information about locations of polygons (Polish powiats).

The nearest neighbor analysis will be based on centroids representing powiats. We can draw them (add the centroid layer to the map of powiats) with the use of the Map manager.

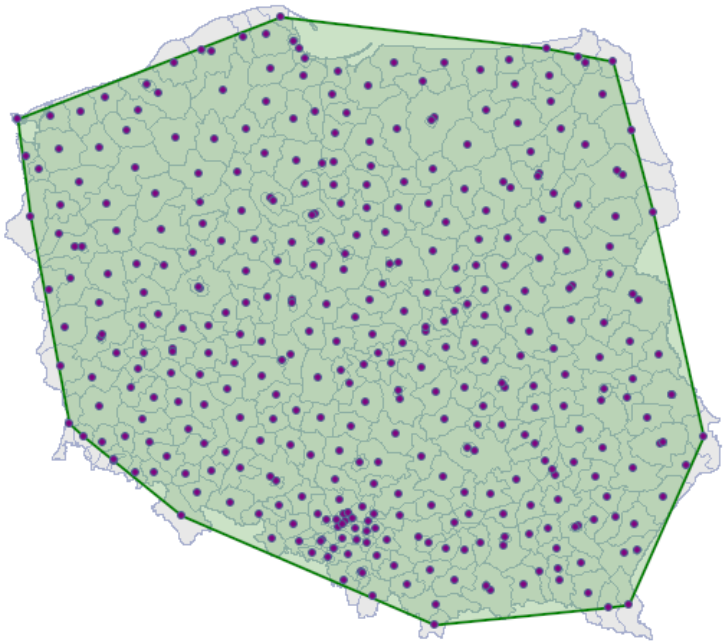
The nearest neighbor analysis will be made with the use of information about the size of the area of Poland –it is $311888000000m^2$. Apart from the nearest neighbor index we will also calculate the indices of subsequent neighbors, up to 15.

Nearest Neighbor Analysis	->> + MAP <<-
Analysis time	0.04sec.
Analysed variables	SHP_X;SHP_Y
Significance level	0.05
Spatial weights matrix	Euclidean - wszystkich eleme
Bounding Types [points]	Area specified
Number of points	379
Area	311888000000
Density	0
The nearest neighbor (NN)	
Mean Distance (NN)	19668.564923
Standard Dev. Distance (NN)	9102.84769
Expected Mean Distance (NN)	14343.321467
Nearest Neighbor Index (NNI)	1.37127
SE	385.125171
Z statistic	13.827306
p-value	<0.000001

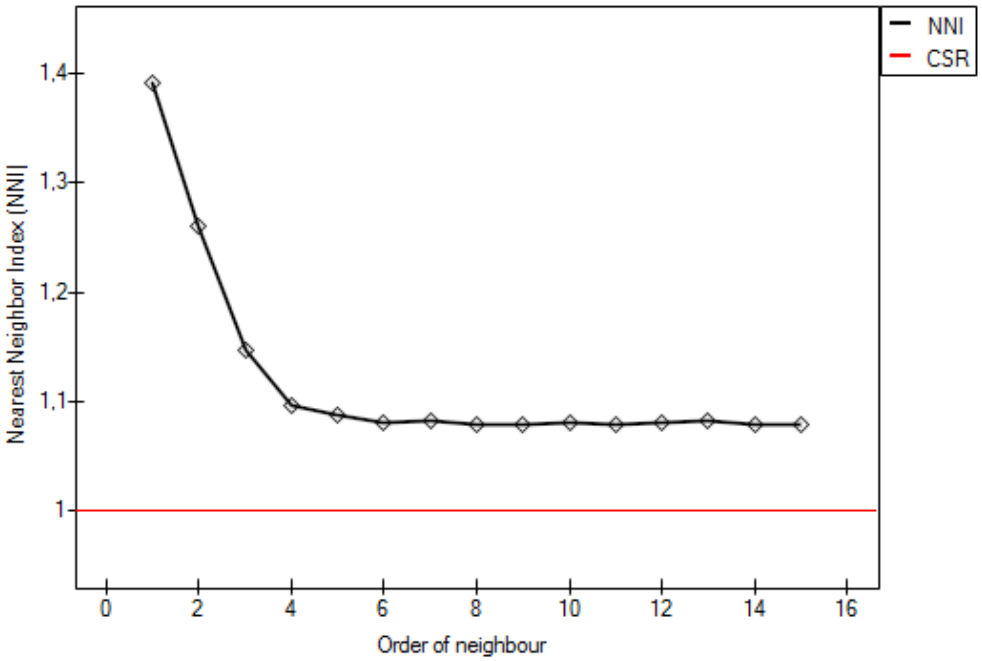
After entering the size of the area in the analysis window, a nearest neighbor index amounting to 1.37127 was received. Its statistical significance was ($p < 0.000001$), greater than value 1. The mean distance between the nearest neighboring centroids is $19668.564923m$ and the standard deviation is $9102.84769m$. We will receive a very similar result when we return to the analysis (the button) and choose the convex hull as the bounding ($NNI = 1.382828$, $p < 0.000001$).

Nearest Neighbor Analysis	->> + MAP <<-
Analysis time	0.04sec.
Analysed variables	SHP_X;SHP_Y
Significance level	0.05
Spatial weights matrix	Euclidean - wszystkich eleme
Bounding Types [points]	Convex Hull
Number of points	379
Area	306696110047.008
Density	0
The nearest neighbor (NN)	
Mean Distance (NN)	19668.564923
Standard Dev. Distance (NN)	9102.84769
Expected Mean Distance (NN)	14223.436336
Nearest Neighbor Index (NNI)	1.382828
SE	381.906196
Z statistic	14.257764
p-value	<0.000001

We add the boundaries defined by the convex hull by pressing the button ->> + MAP <<- and choosing the layer of bounding.



The correction of the effect of a boudary defined in this way lowers the value of NNI to 1.340503 but leaves the general tendency of the subsequent nearest neighbor indices unchanged.

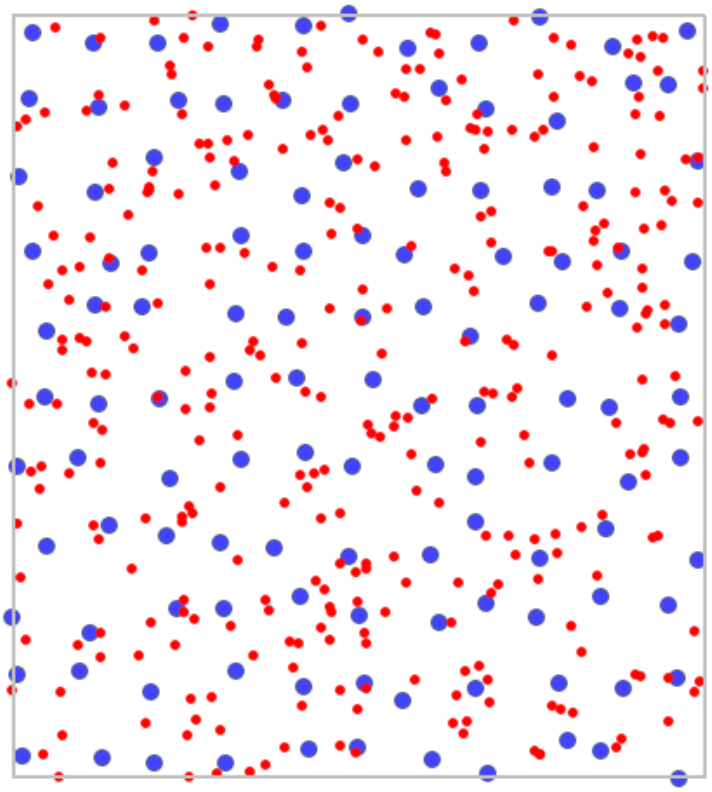


In each of the analysis described above the subsequent neighbor indices are greater than 1 and, although they initially approximate 1, from order 5 they stabilize at the level of about 1.1. The result, then, confirms the uniform distribution of Polish powiats.

EXAMPLE 5.2. (directory: poplar, SHP files: T-poplar, S-poplar)

Competition among species has an influence on the changes in the distribution of particular species of plants and on their density. Competition within a species is usually stronger than that among different species as members of the same species have almost identical demands and compete for the same resources. The intensity of competition within a species increases with the growth of the population. To check the influence of the competition on a certain species of balsamic poplar, a wooded area not regulated by man was studied. Locations of young trees and of old ones were studied.

- Mapa T-poplar contains fictitious information about the locations of 121 points (old balsamic poplars) in a rectangular wooded area.
- Mapa S-poplar contains fictitious information about the locations of 326 points (young balsamic poplars) in a rectangular wooded area.



On the map young poplars were marked in red and old poplars were marked in blue.

On the basis of the nearest neighbor indices, the structure of poplar density was compared in the area defined by a rectangle of layer bounding.

Nearest Neighbor Analysis	->> + MAP <<-
Analysis time	0,05sec.
Analysed variables	SHP_X;SHP_Y
Significance level	0,05
Bounding Types	Bounding Rectangle
Number of points	326
Area	13714634,9573852
Density	0,00002377
The nearest neighbor (NN)	
Mean Distance (NN)	103,116479003
Standard Dev. Distance (NN)	57,129039778
Expected Mean Distance (NN)	102,55417152
Nearest Neighbor Index (NNI)	1,005483029
SE	2,969041767
Z statistic	0,189390223
p-value	0,849786986

Nearest Neighbor Analysis	->> + MAP <<-
Analysis time	0,04sec.
Analysed variables	SHP_X;SHP_Y
Significance level	0,05
Bounding Types	Bounding Rectangle
Number of points	121
Area	13683234,2560387
Density	0,000008843
The nearest neighbor (NN)	
Mean Distance (NN)	282,270974905
Standard Dev. Distance (NN)	50,395872027
Expected Mean Distance (NN)	168,140254422
Nearest Neighbor Index (NNI)	1,678782846
SE	7,990073817
Z statistic	14,284063339
p-value	<0.000000001

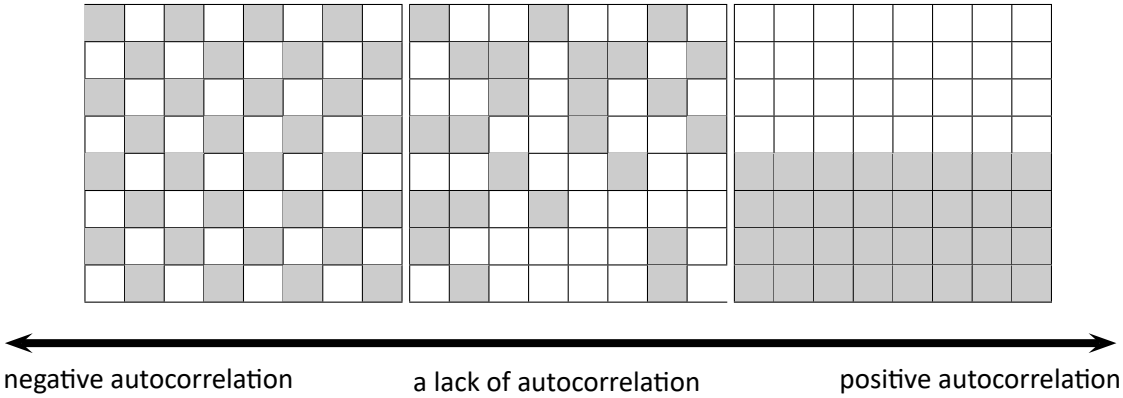
Young poplars have greater density than old ones. Their mean nearest neighbor distance is $103.12m$ whereas for old poplars the value is $282.27m$. Due to competition in the development of the structure of forest stand the spatial pattern for old trees is more regular ($NNI = 1.68, p < 0.000001$) than the one for young poplars ($NNI = 1.01, p = 0.8498$).

6 SPATIAL AUTOCORRELATION

To conduct spatial autocorrelation on the basis of a Map data we should have at our disposal a point, multipoint, or polygonal file. In the case of an analysis of a polygonal file based on the calculation of objects distances, calculations are based on centroids of polygons, and in the case of a multipoint file they are based on centers of objects.

An analysis of the phenomenon of autocorrelation is based on values assigned to spatial objects. Spatial autocorrelation means that the values of geographically near objects are more similar to one another than those of remote objects. The phenomenon causes the creation of spatial clusters with similar values.

Spatial autocorrelation may not occur – we then speak of spatial randomness. The obtained spatial distribution is as probable as any other distribution. When the neighboring values are similar to one another we can speak about positive autocorrelation. Negative autocorrelation occurs when the values of neighboring areas are more varied than in the case of random distribution.



When analyzing autocorrelation we can consider a dichotomous variable (i.e. the presence or absence of a given feature) or a variable with many categories, pointing to the degree of intensity of the analyzed feature.

For a dichotomous variable the analysis of positive autocorrelation consists of searching for clusters with the same value. Usually, objects in which the studied phenomenon occurs are marked in black color on the map, and the ones in which the phenomenon does not occur are marked in white color. Clusters of objects of the same color – the so-called ”black-black”,”white-white” – are looked for.

For a variable which describes the degree of intensity of a studied feature the analysis of positive autocorrelation consists of searching for clusters with similar values. Usually, objects on the map are colored in accordance with the degree of intensity of the studied phenomenon, from the lightest (low values) to the darkest (high values). Clusters of objects with a similar shade are looked for.

6.1 Global Moran’s I statistic

It is an analysis of the degree of intensity of a given feature in spatial objects.

We use two pieces of information for the construction of a coefficient which will allow to check if the neighboring objects form clusters with similar values of the variable:

1. information about the values of a variable for particular objects x_i,
2. information about which objects are neighbors – weights matrix with elements w_{ij}.

Note
The objects neighborhood is defined by a spatial weight matrix. In Moran's analysis window we can choose any matrix generated previously by using menu Spatial analysis → Tools → Spatial weights matrix or indicate the neighbor matrix according to contiguity – Queen, row standardized, that is proposed by the program.

Note
It is not recommended to conduct Moran's analysis for objects without neighborhood (objects described in the weight matrix only with the 0 value). Such objects can be excluded from the analysis by deactivating them or an analysis can be made with the use of a different manner of defining neighborhood (a different weight matrix).

Moran's I coefficient – introduced by Moran in 1948 [9].

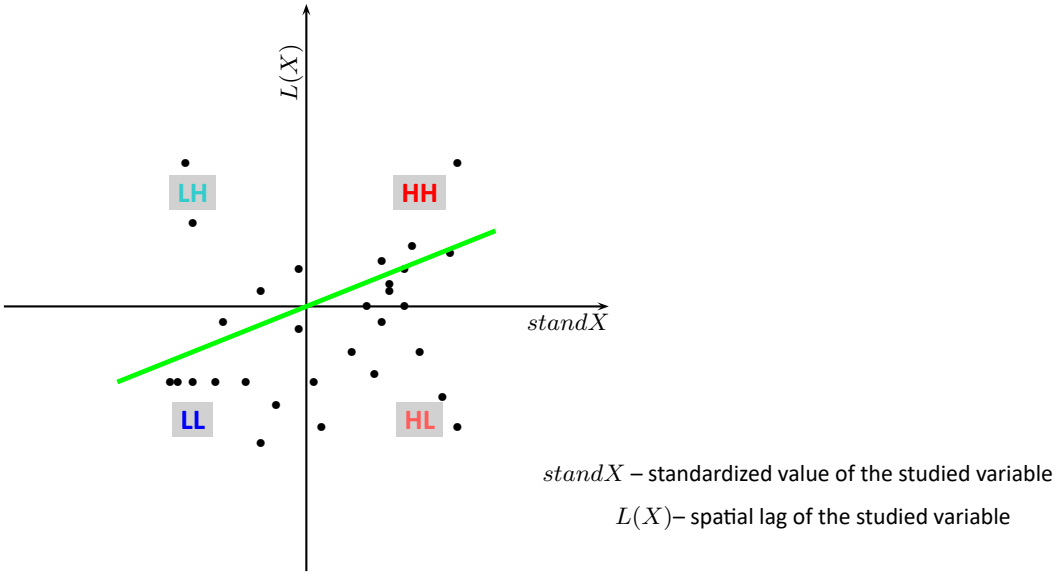
In order to check if the selected objects are characterized by similar values of the variable one can use the multiplying rule which says that multiplying 2 positive numbers gives a positive result and multiplying 2 different numbers (1 positive and 1 negative) gives a negative result. With the use of this rule we calculate $\sum\sum x_i x_j$. Unfortunately, as the results of that rule are only obtained when there are both positive and negative values, the simple rule must be modified so as to ensure the presence of different signs. The values of the variable will, then, be replaced in the earlier formula with the differences of the values of the variable and of its mean value. In this way the objects with values smaller than the mean will be negative and those with values greater than the mean will be positive: $\sum\sum(x_i - \overline{x})(x_j - \overline{x})$. Obviously, the summation should concern neighboring objects, which means that, at this point, information from weights matrices must be used:

$$\sum\sum w_{ij}(x_i - \overline{x})(x_j - \overline{x})$$

In this way non-neighboring objects obtain the weight value 0, for which reason the values of those objects are not added. Further operations which change the formula obtained in this manner are made with the view to making the obtained coefficient I independent from the number of analyzed objects and to standardizing it so that its values are limited to the interval $< -1; 1 >$. As a result, Moran's autocorrelation coefficient is expressed with the formula:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(x_i - \overline{x})\,(x_j - \overline{x})}{S_0\sigma^2}$$

where:
n – the number of spatial objects (the number of points or polygons),
x_i, x_j – are the values of the variable for the compared objects,
$\overline{x}$ – it is the mean value of the variable for all objects,
w_{ij} – elements of the spatial weights matrix (weights matrix row standardized),
$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$,
$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ – variance



Rysunek 1: Moran's diagram

Moran's linear autocorrelation coefficient I studies the strength of the linear relationship between the standardized variable X ($stand(x_i)$) and the spatial lag of the variable X ($L(x_i)$). Spatial lag is the weighted mean from the standardized values of neighboring objects

$$L(x_i) = \sum_{j=1}^{N} w_{ij} stand(x_j).$$

A graphic presentation of spatial autocorrelation is Moran's scatter plot. Points in the first quarter (**HH**) and in the third quarter (**LL**) are objects surrounded by similar neighbors: **HH** (high-high) – objects with high values, surrounded by objects with high values; **LL** (low-low) – objects with low values, surrounded by objects with low values. Points in the second quarter (**LH**) and the fourth quarter (**HL**) are objects surrounded by neighbors not similar to them. **LH** (low-high) – objects with low values, surrounded by objects with high values; **HL** (high-low) – objects with high values, surrounded by objects with low values.

The belonging to and distribution of points in the four quarters of Moran's diagram indicates the type of autocorrelation. If points are distributed mainly in the second quarter (**LH**) and fourth (**HL**) – it is a sign of negative correlation, if they belong mainly to the first quarter (**HH**) and third (**LL**) – it is a sign of positive correlation. If the points are distributed evenly in all four quarters then spatial autocorrelation does not exist.

On the Moran's diagram there is a regression line, the direction of which also allows to interpret Moran's coefficient I:

- $I > 0$ indicates the presence of clusters of similar values – positive autocorrelation, i.e. measurement points lie near the straight line, and the increase of the variable $standX$ is reflected in the increase of the variable $L(X)$;
- $I < 0$ indicates the presence of the so-called hot spots, i.e. decidedly different values in neighboring areas – negative autocorrelation, i.e. measurement points lie near the straight line but the increase of the variable $standX$ is accompanied by a decrease of the variable $L(X)$;

- $I \approx 0$ indicates random distribution of the studied value in space – a lack of autocorrelation, i.e. the obtained spatial distribution is as probable as any other distribution.

The square of Moran's coefficient I^2 informs about the degree (it is a percentage) to which the value of the variable in the object i is explained by the value of that variable in neighboring objects.

Note
When the values of a studied feature are characterized by a great variability of variance then it is desirable to stabilize that variability. The basic information about smoothing variables have been described in the Chapter 1.7 SPATIAL SMOOTHING

Significance of Moran's autocorrelation coefficient

A test for checking the significance of Moran's autocorrelation coefficient serves the purpose of verifying the hypothesis about a lack of correlation between $standX$ and spatial lag $L(X)$.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad I = 0, \\ \mathcal{H}_1: &\quad I \neq 0. \end{aligned}$$

The test statistic has the form presented below:

$$Z = \frac{I - E(I)}{\sqrt{var(I)}},$$

where:
$E(I) = \dfrac{-1}{n-1}$ – the expected value,
$var(I)$ – variance.

Depending on the assumption concerning the distribution of the population from which the sample has been taken, the manner of selecting variance is chosen (Cliff and Ord (1981)[5], and Goodchild (1986)[8]). If it is normal distribution, then:

$$var(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{S_0^2(n^2-1)} - E(I)^2,$$

where:
$S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2$,
$S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2$.

If it is random distribution, then:

$$var(I) = \frac{n\left((n^2-3n+3)S_1 - nS_2 + 3S_0^2\right)}{(n-1)^{(3)}S_0^2} - \frac{K_2\left((n^2-n)S_1 - 2nS_2 + 6S_0^2\right)}{(n-1)^{(3)}S_0^2} - E(I)^2,$$

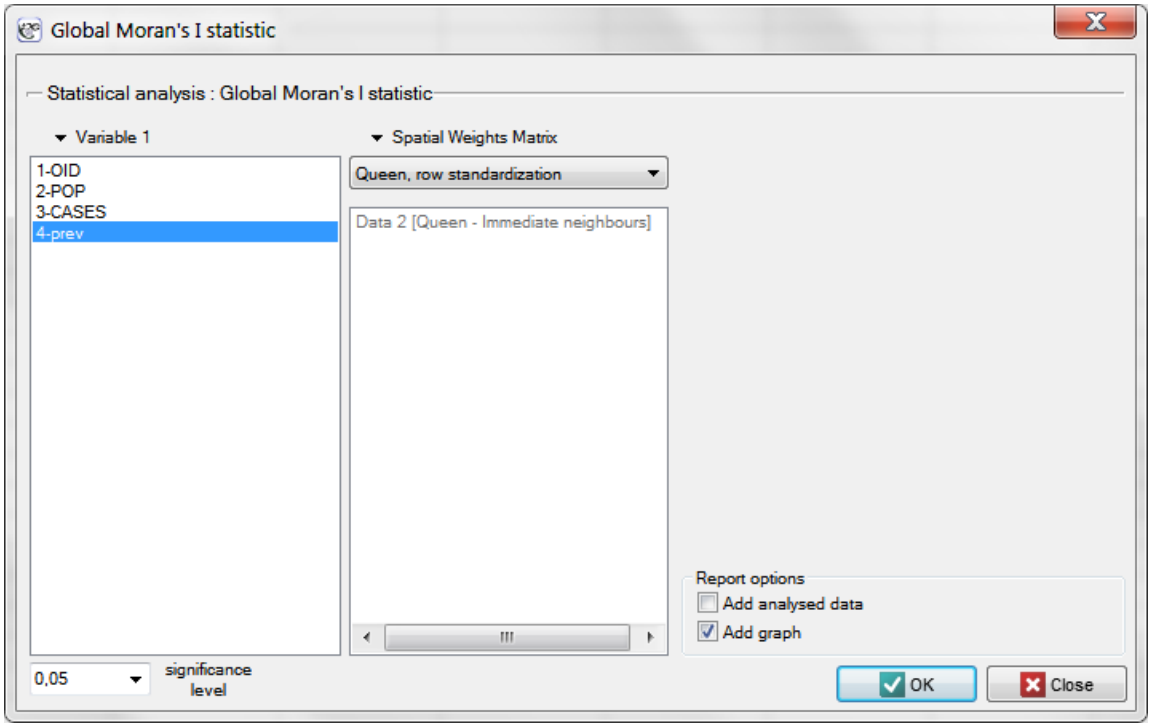
where:
$K_2 = \frac{n\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}$,
$n^{(b)} = n(n-1)(n-2)...(n-b+1)$.

Statistics asymptotically (for a large sample size) has the normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

if $p \leq \alpha \implies$ we reject $\mathcal{H}_0$ and accept $\mathcal{H}_1$,
if $p > \alpha \implies$ there is no reason to reject $\mathcal{H}_0$.

The window with settings for Moran's analysis is accessed via the menu Spacial statistics→Tools→Moran's global I statistic.



EXAMPLE 6.1. (catalog: leukemia, file: leukemia.pqs)

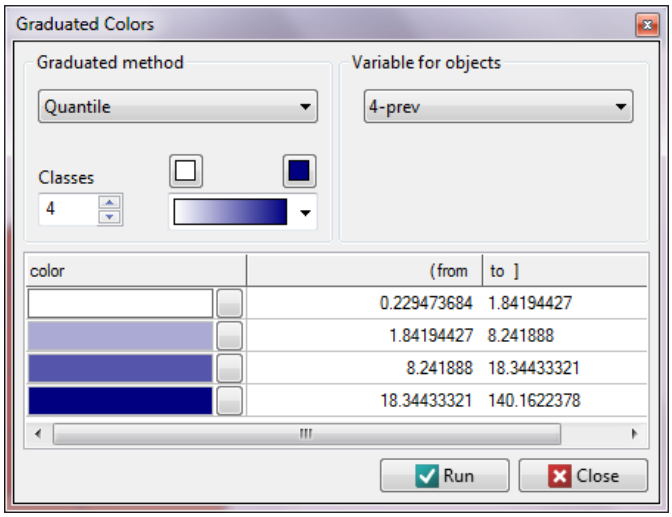
The analysis will concern the data gathered and analyzed by L.A. Waller and others in 1992[12] and 1994[13], described on 281 objects in 2004[14].

- The map leukemia contains information about the location of 281 polygons (census tracts) in the northern part of the state of New York. The map is prepared in the set of flat rectangular coordinate system UTM 18N and is based on the data of the file BNA (Boundary File) available on the server CIESIN ftp.ciesin.columbia.edu
- Data for the map leukemia:
  - Column CASES – the number of cases of leukemia in the years 1978-1982, ascribed to particular objects (census tracts). The value should be an integral number, however, in agreement with Waller's (1994) description, some cases which could not be objectively ascribed to a particular region have been divided proportionately. Hence, the numerousnesses of the cases ascribed to the 281 objects are not integral numbers.
  - Column POP – population size in particular objects.
  - Column prev – the frequency coefficient of leukemia per 100000 people, for each object in one year: prev=(CASES/POP)*100000/5

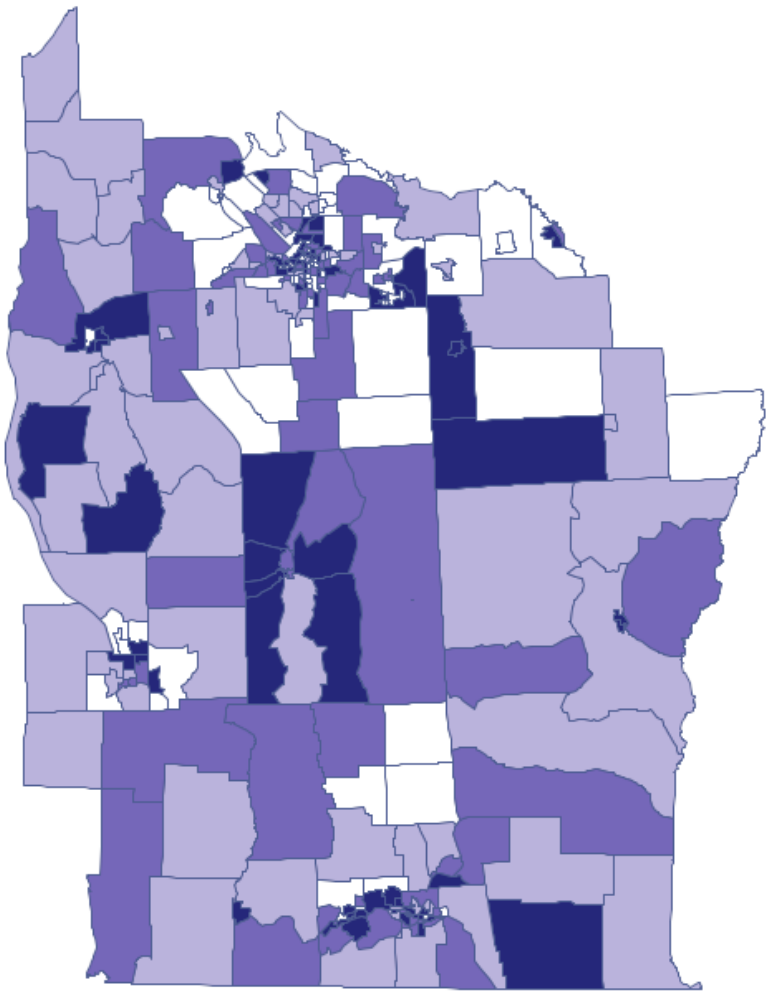
Epidemiologically interesting are the regions in which the prevalence of leukemia is higher, as their grouping could indicate the existence within their boundaries of environmental teratogens causing an increased frequency of occurrence of leukemia.

We start from presenting the geographic distribution of the frequency coefficient (prev)

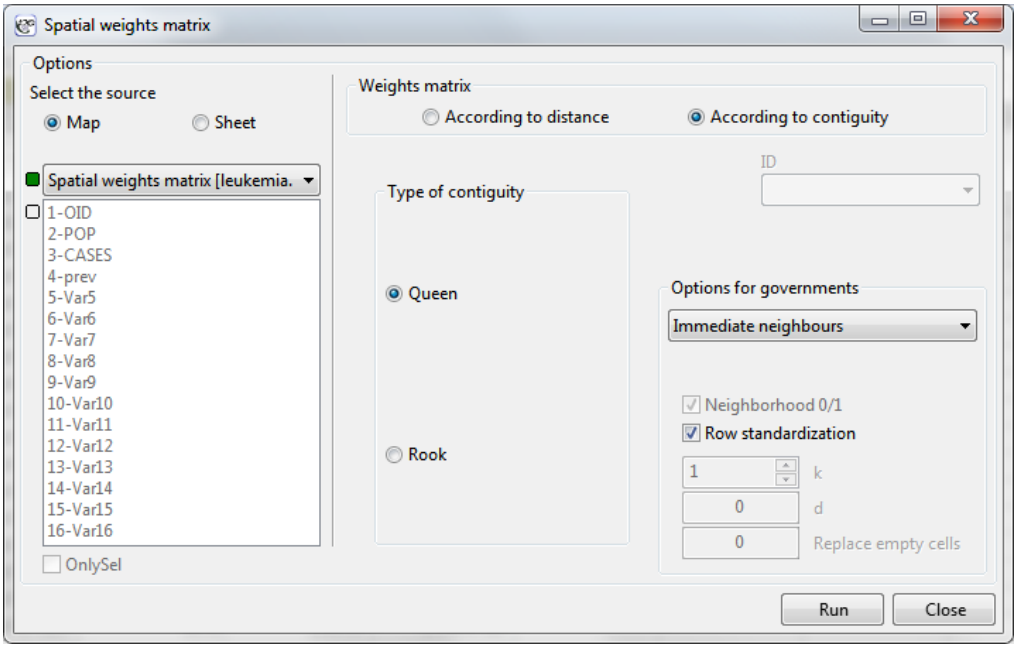
on the map. For that purpose we draw a map in the Map Manager and edit the layer, choosing Graduated colors:



We have at our disposal several ways of coloring a map – we choose coloring in accordance with the values of the variable prev, dividing it into quartiles:



Dark colors on the map present the places with a higher frequency coefficient of leukemia, whereas light places signify a lower frequency coefficient. In order to learn if their geographic distribution is random or if they forms clusters, we will calculate Moran's coefficient. Before calculating that coefficient we should determine the manner of defining neighborhood of regions and it is advisable to create an appropriate weights matrix. In Moran's analysis window we can choose any matrix generated previously by using menu Spatial analysis → Tools → Spatial weights matrix or indicate the neighbor matrix according to contiguity – Queen, row standardized, that is proposed by the program.

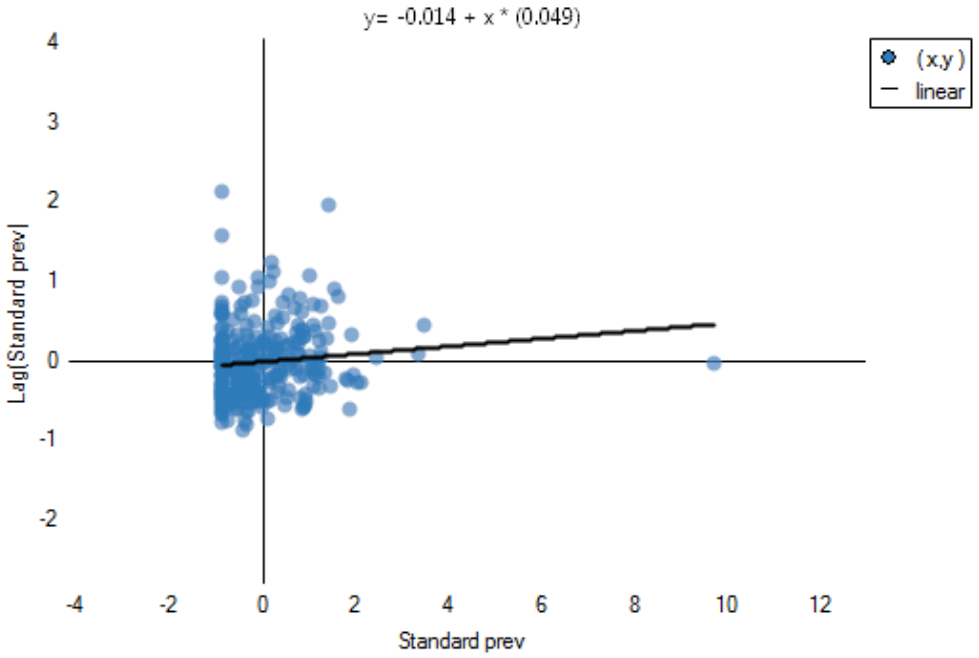


Having generated the weights matrix we select the file leukemia and start Moran's analysis by selecting the menu Spatial analysis → Spatial statistics → Global Moran's I statistic. In the analysis window we select the variable Prev and the neighbor matrix Queen, and select the option Add graph.

Moran's correlation coefficient obtained in the analysis is small and has the value $I = 0.048577$:

Global Moran's I statistic	
Analysis time	0.27sec.
Analysed variables	prev
Significance level	0.05
Spatial weights matrix	Queen - Immediate neighbou
Number of objects	281
Moran's I	0.048577
Expected I	-0.003571
Under normality assumption	
Variance I	0.001395
Z statistic	1.3962
p-value	0.162654
Under randomness assumption	
Variance I	0.001241
Z statistic	1.480333
p-value	0.138784

When we test the significance of Moran's coefficient we study the randomness of the distribution of the frequency coefficient of leukemia in the studied region. We check if similar shades on the map are located close to one another or not. In other words: we check if the odds of having leukemia in the studied population depends on geographic location or not. The value p calculated with the assumption of randomness, as in the case of the assumption of normality, is greater than the standard assumed significance level 0.05, which means that there is no evidence for autocorrelation. Thus, we assume that the distribution of the variable prev is a random distribution. Moran's diagram confirms that assumption:



The existence of positive autocorrelation, in which we are the most interested, would result in the distribution of the points of the Moran's diagram in quarters I and III. Here, however, we see that the points are as frequent in quarters I and III as in II and IV.

6.2 Global Geary's C statistic

Similarly to Moran's analysis, global Geary's statistic studies the degree of the intensity of a given feature in spatial objects.

Note
It is not recommended to conduct Geary's analysis for objects without a neighborhood (objects described in a weight matrix only with the value 0). Such objects can be excluded from the analysis by deactivating them (Chapter **??**), or the analysis can be made with the use of a different manner of defining neighborhood (a different weight matrix).

Geary's autocorrelation coefficient – introduced by Geary in 1954 [7].

It is one of the possible alternatives for the global Moran's statistic. Similarly to Moran's analysis, Geary's statistic studies the degree of intensity of a given x_i feature in spatial objects described with the use of a weight matrix with w_{ij} elements. This time, instead of computing the sum of quotients: $\sum\sum w_{ij}(x_i-\overline{x})(x_j-\overline{x})$ we compute the sum of the difference squares:

$$\sum\sum w_{ij}(x_i-x_j)^2$$

As a result, Geary's autocorrelation coefficient is expressed with the formula:

$$c=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-x_j)^2}{2S_0sd^2}$$

where:
n – the number of spatial objects (the number of points or polygons),
x_i, x_j – are the values of the variable for the compared objects,
w_{ij} – elements of the spatial weights matrix (weights matrix row standardized),
$S_0=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}$,
$sd^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$ – variance,
$\overline{x}$ – it is the mean value of the variable for all objects.

The interpretation of Geary's coefficient:

- $c<1$ and $c\approx 0$ means the occurrence of clusters with similar values – a positive autocorrelation;
- $c>1$ means the occurrence of the so-called hot spots, i.e. distinctly different values in neighboring areas – a negative autocorrelation;
- $c\approx 1$ means a random spatial distribution of the studied variable – a lack of autocorrelation.

Note
When the values of a studied feature are characterized by a great variability of variance then it is desirable to stabilize that variability. The basic information about smoothing variables have been described in the Chapter 1.7 SPATIAL SMOOTHING

Significance of Geary's autocorrelation coefficient

A test for checking the significance of Geary's autocorrelation coefficient serves the purpose of

verifying the hypothesis about a lack of spatial autocorrelation

Hypotheses:

$$\begin{aligned} \mathcal{H}_0: &\quad C = 1, \\ \mathcal{H}_1: &\quad C \neq 1. \end{aligned}$$

The test statistic has the form presented below:

$$Z = \frac{C - E(C)}{\sqrt{var(C)}},$$

where:
$E(C) = 1$ – the expected value,
$var(C)$ – variance.

Depending on the assumption concerning the distribution of the population from which the sample has been taken, the manner of selecting variance is chosen (Cliff and Ord (1981)[5], and Goodchild (1986)[8]). If it is a normal distribution, then:

$$var(C) = \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_0^2},$$

where:
S_1 and S_2 are defined as for Moran's analysis.

If it is a random distribution, then:
$var(CS) = \frac{(n-1)S_1\left(n^2-3n+3-(n-1)b_2\right)-(n-1)S_2\left(n^2+3n-6-(n^2-n+2)b_2\right)\frac{1}{4}+S_0^2\left(n^2-3-(n-1)^2b_2\right)}{n(n-2)^{(2)}S_0^2},$

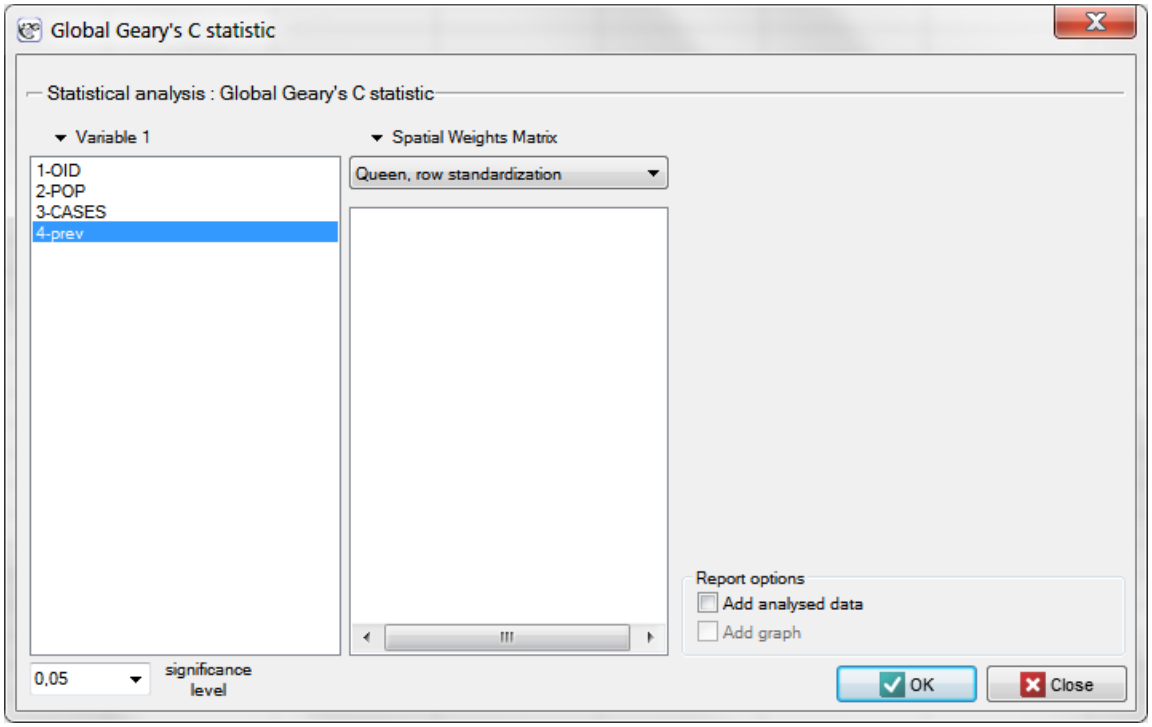
where:
$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2},$
$n^{(b)} = n(n-1)(n-2)...(n-b+1).$

Statistics Z has, asymptotically (for large sample sizes), normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{aligned} &\text{if } p \leq \alpha \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ &\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The window with settings for Geary's analysis is accessed via the men Spacial analysis → Spacial statistics → Global Geary's C statistic.



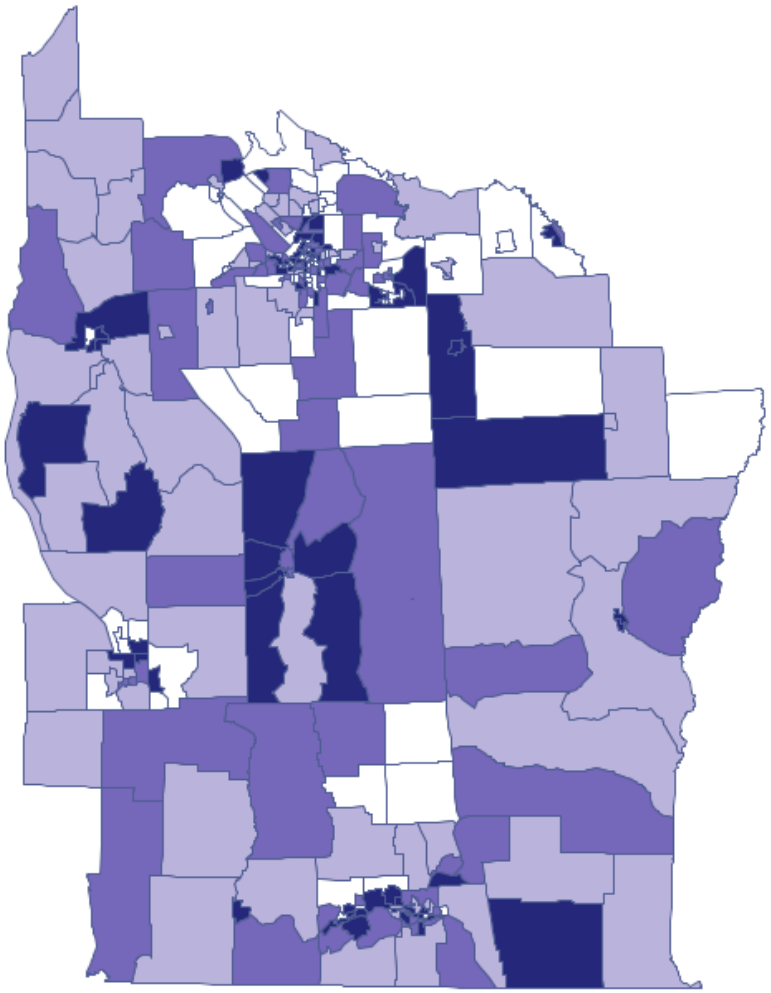
EXAMPLE 6.1 cont. (catalog: leukemia, file: leukemia)

We will analyze the data concerning leukemia.

- The map leukemia contains information about the location of 281 polygons (census tracts) in the northern part of the state of New York.
- Data for the map leukemia:
  - Column CASES – the number of cases of leukemia in the years 1978-1982, ascribed to particular objects (census tracts). The value should be an integral number, however, in agreement with Waller's (1994) description, some cases which could not be objectively ascribed to a particular region have been divided proportionately. Hence, the numerousnesses of the cases ascribed to the 281 objects are not integral numbers.
  - Column POP – population size in particular objects.
  - Column prev – the frequency coefficient of leukemia per 100000 people, for each object in one year: prev=(CASES/POP)*100000/5

Global Moran's analysis has pointed to a lack of spatial autocorrelation. This time, in order to check if in the studied area of the northern part of the state of New York it is possible to localize clusters of leukemia we will compute the global Geary's C statistic.

We start from the presentation of the geographic distribution of the prevalence coefficient (prev) on the map, according to the values of the prev variable, dividing it into quartiles:



Dark colors on the map present the places with a higher prevalence of leukemia, whereas light places signify a lower prevalence. Geary's correlation coefficient obtained in the analysis equals: 0.884986.

Global Geary's C statistic	
Analysis time	0,50sec.
Analysed variables	prev
Significance level	0,05
Spatial weights matrix	Queen - Immediate neighl
Number of objects	281
Geary C	0,884986
Expected C	1
Under normality assumption	
Variance C	0,001665
Z statistic	-2,818827
p-value	0,00482
Under randomness assumption	
Variance C	0,005827
Z statistic	-1,506738
p-value	0,131878

The obtained result, assuming a random distribution of data, is different from the result obtained with the assumption of a normal distribution. That can be indicative of an instability of the results and point to the need of further analyses based on smoothed variables.

7 LOCAL ESTIMATE OF SPATIAL CLUSTERING

In the local analysis we try to define clusters according to their placement, size, and intensity. A cluster is understood as a limited gathering of objects of certain intensity, placed in space and/or time, an accidental appearance of which is highly improbable. If we identify such a gathering which is not accidental but is a statistically significant cluster we can infer the reasons for its occurrence.

7.1 Local Moran's I statistic

Local Moran's I statistic is the most popular analysis from those defined as LISA (Local Indicators of Spatial Association) (Luc Anselin 1995 [1]). In contrast to Global Moran's I statistic it defines the local spacial autocorrelation, i.e. defines the similarity of a spatial unit to its neighbors and studies the statistical significance of that dependence.

Local Moran's I coefficient

The local form of Moran's I coefficient for the i observation is defined with the formula:

$$I_i = \frac{(x_i - \overline{x}) \sum_{j=1}^{n} w_{ij} (x_j - \overline{x})}{\sigma^2}$$

where:
n – the number of spatial objects (the number of points or polygons),
x_i, x_j – are the values of the variable for the compared objects,
$\overline{x}$ – it is the mean value of the variable for all objects,
w_{ij} – elements of a spacial weight matrix (it is recommended that the matrix is row standardized),
$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ – variance

The interpretation of the local Moran's coefficient is analogous to its global counterpart, however, it largely depends on the selected weight matrix. Most often non-zero matrices are ascribed only to neighboring objects. As a result the local coefficient only describes the similarity of objects in the zone of neighborhood. Row standardization makes it easier to compare the values of coefficients obtained for various objects as the expected value for each coefficient is then the same.

High values of a coefficient point to the occurrence of clusters with similar values while low values of a coefficient point to the occurrence of the so-called hot spots, and values near the expected value $E(I_i)$ point to the random distribution in space of the studied variable.
The expected value is defined with the formula:

$$E(I_i) = \frac{-\sum_{j=1}^{n} w_{ij}}{n-1}$$

The significance of Moran's autocorrelation coefficient

By testing the statistical significance of the relationship among the neighboring objects the following hypotheses are studied:

$$\begin{aligned} \mathcal{H}_0: &\quad I_i = E(I_i), \\ \mathcal{H}_1: &\quad I_i \neq E(I_i). \end{aligned}$$

The test statistic has the form presented below:

$$Z_i = \frac{I_i - E(I_i)}{\sqrt{var(I_i)}},$$

where:

$var(I_i) = \frac{w_{i(2)}(n-b_2)}{n-1} + \frac{2w_{i(kh)}(2b_2-n)}{(n-1)(n-2)} - \frac{\left(\sum_{j=1}^{n} w_{ij}\right)^2}{(n-1)^2}$ – variance in a random distribution,

$b_2 = \frac{(n-1)\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}$,

$w_{i(2)}$ – the sum of weights square for the i row,

$2w_{i(kh)}$ – the sum of possible weights ratios for the i row, after the exclusion of ratios with the same indexes.

The Z_i statistics has, asymptotically (for large sample sizes), normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{aligned} &\text{if } p \leq \alpha \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ &\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

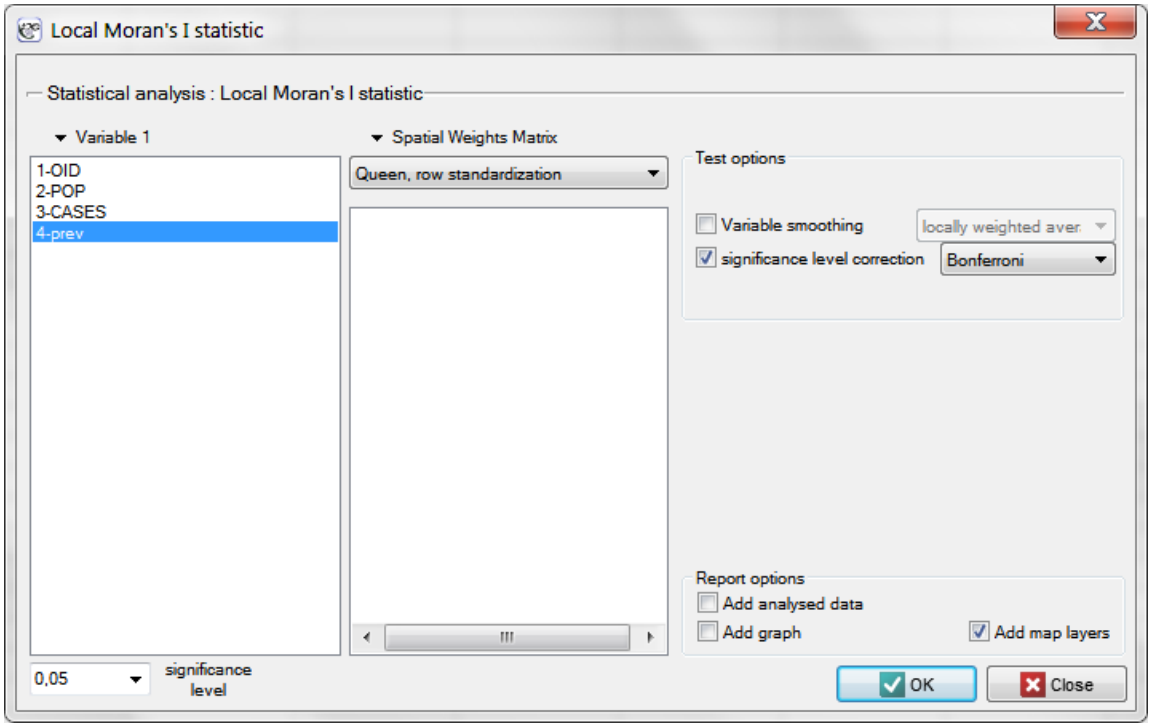
Due to the problem of a lack of independence of coefficients computed for neighboring objects it is suggested to use a corrected significance level α. The suggested corrections are: Bonferroni correction: $\alpha_1 = \alpha/k$ or Šidák correction: $\alpha_1 = 1-(1-\alpha)^{1/k}$, where k is the arithmetic mean number of the neighbors.

Map layers

The combination of information from Moran's scatter plot (the division of objects into: High-High, Low-Low, Low-High, High-Low) and from the significance of the local Moran's statistics presents on a map the so-called **spatial regimes:**

- Statistically significant **High-High** objects (objects with high values surrounded by objects with high values) are marked in red on the map;
- Statistically significant **Low-Low** objects (objects with low values surrounded by objects with low values) are marked in blue on the map;
- Statistically significant **Low-High** objects (objects with low values surrounded by objects with high values) are marked in light blue on the map;
- Statistically significant **High-Low** objects (objects with high values surrounded by objects with low values) are marked in light red on the map.

The window with the settings of the local Moran's analysis option is accessed via the menu Spatial analysis → Spatial statistics → Local Moran's I statistic.



EXAMPLE 6.1 cont. (catalog: leukemia, file: leukemia)

We will analyze data about leukemia.

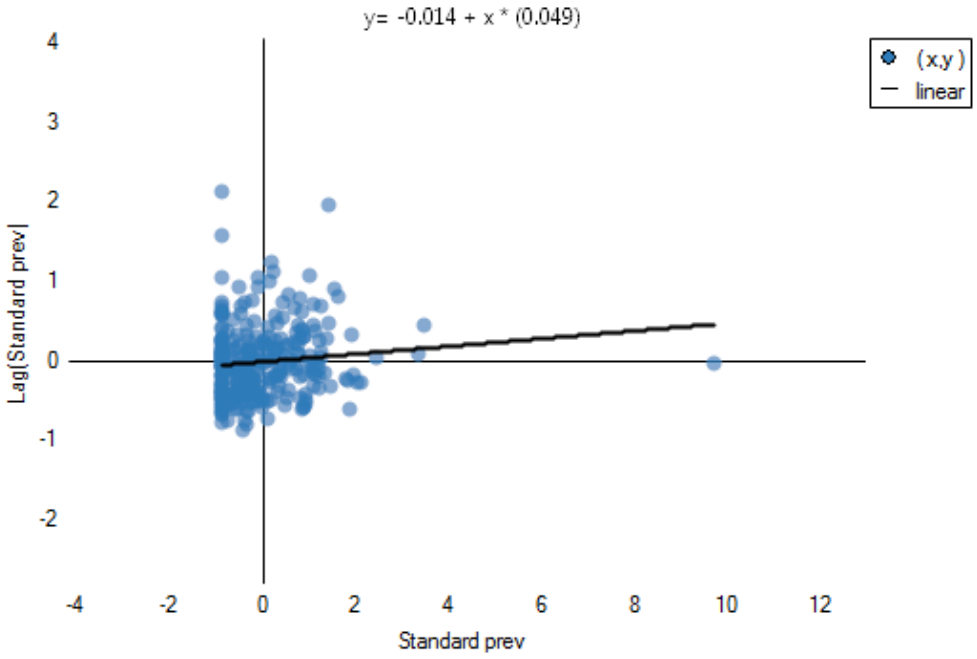
- The map leukemia contains information about the location of 281 polygons (census tracts) in the northern part of the state of New York.
- Data for the map leukemia:
  - Column CASES – the number of cases of leukemia in the years 1978-1982, ascribed to particular objects (census tracts). The value should be an integral number, however, in agreement with Waller's (1994) description, some cases which could not be objectively ascribed to a particular region have been divided proportionately. Hence, the numerousnesses of the cases ascribed to the 281 objects are not integral numbers.
  - Column POP – population size in particular objects.
  - Column prev – the frequency coefficient of leukemia per 100000 people, for each object in one year: prev=(CASES/POP)*100000/5

The global analysis has not yielded an unambiguous answer as to the occurrence of spatial autocorrelation. We will, then, check if we can find regions in which the prevalence of leukemia is significantly higher.

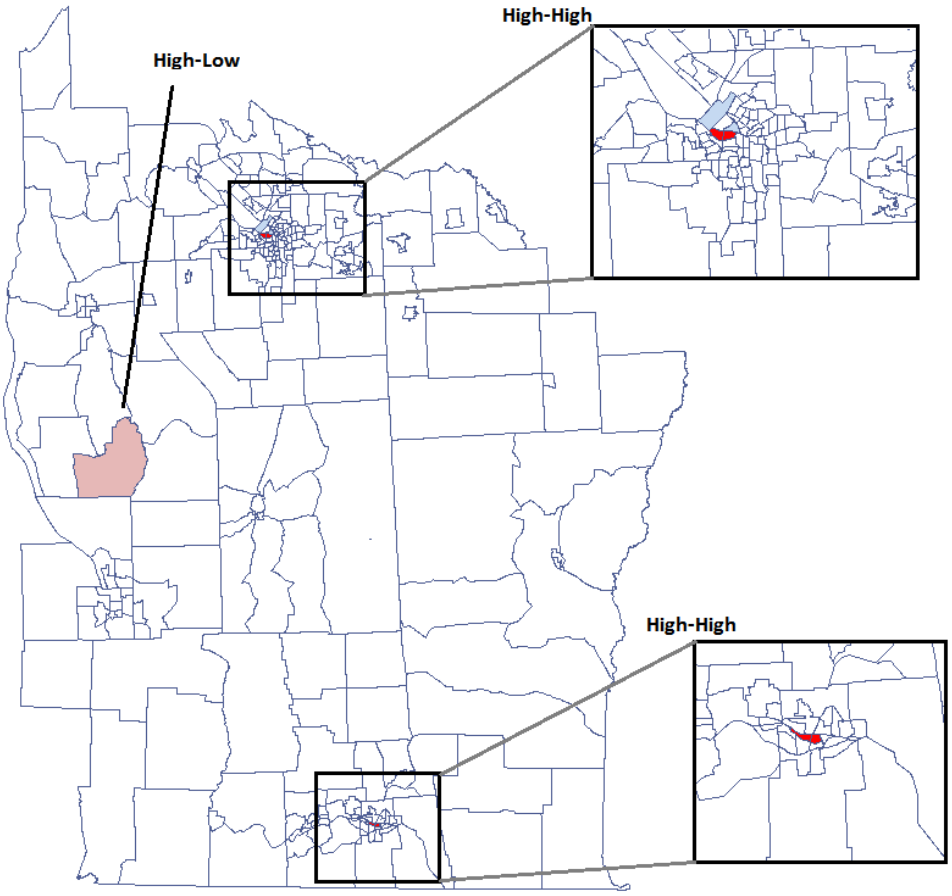
In order to localize clusters of leukemia and regions which contrast with the environment with respect to the prevalence of that disease we will compute the local Moran's coefficient. For the analysis we will use the prev variable and the neighborhood matrix – Queen, row standardized (according to the contiguity) which is suggested by the program. In order to use a different matrix one has to generate it first, see chapter: Spatial weight matrix. We also select one of the corrections of the significance level.

Local Moran's I statistic	->> + MAP <<-
Analysis time	0,53sec.
Analysed variables	prev
Significance level	0,05
Corrected significance level (Bonferroni)	0,009147
Average number of neighbors	5,466192
Spatial weights matrix	Queen - Immediate neighl
Number of objects	281
Mean Ii	0,048404
Standard deviation Ii	0,380885
Frequency (High-High 1)	4
Frequency (Low-low 3)	0
Frequency (Low-High 2)	2
Frequency (High-Low 4)	1

The obtained report presents the values of local coefficients, the values of test statistics, and the corresponding values of test probability. We will also find here the information about the number of regions defining the spatial regimes (High-High, Low-Low, Low-High, High-Low).

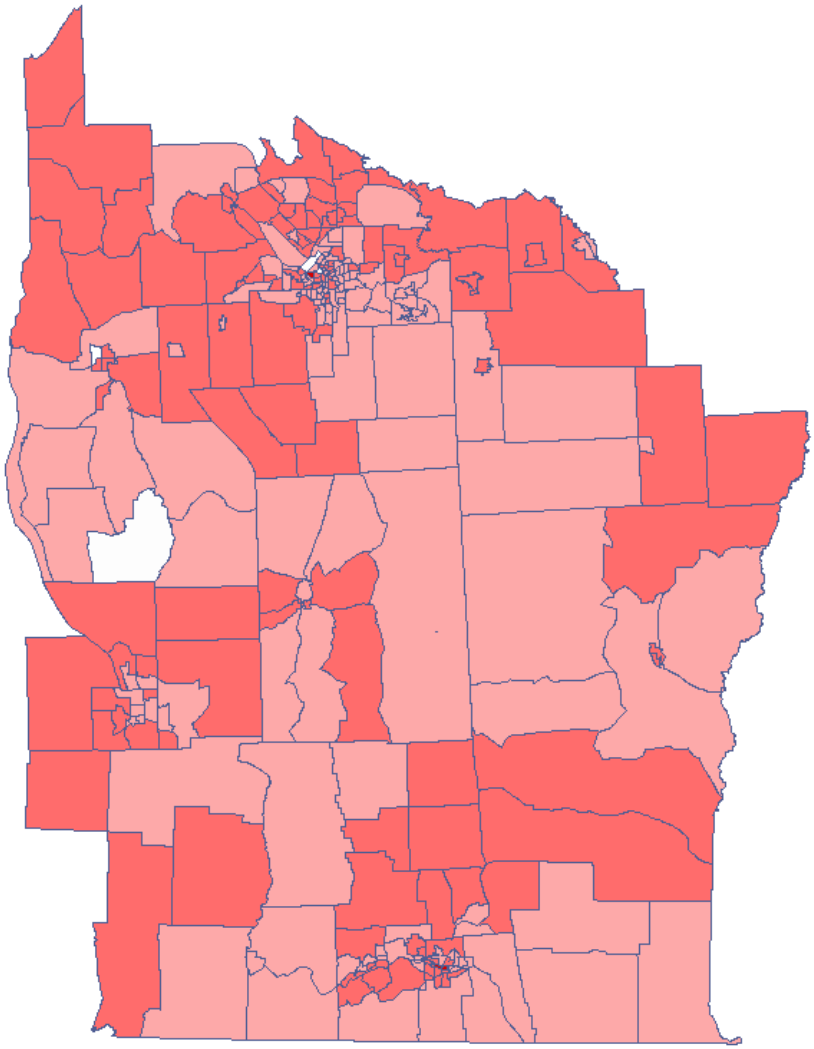


Also, a result is ascribed to the analysis, which we can draw on the map (button ->> + MAP <<-) – those are spatial regimes described in the report with the use of the color column.



We have been able to localize small but significant clusters in which the prevalence of leukemia is higher. The red color is used for the 2 clusters (4 register regions) lying in smaller and more populated regions – they are the centers of the clusters with high leukemia values. The light red color is used for the census tract with high values of the coefficient describing the prevalence of leukemia. The region contrasts with the neighboring census tracts which are characterized by a relatively low coefficient.

The obtained results can be further illustrated when the map is colored with the values of the local Moran's I_i coefficient or the values of a test statistic, or p values. One just has to copy the appropriate columns from the report into a datasheet. In this example we will use the values of the $Z(I_i)$ test statistic for coloring. Having pasted it into an empty column of a datasheet, in the map manager we color the base map according to the values of that column, selecting the standard deviation with the coefficient 3 as a way of gradiating colors. Positive and high values of the Z_i statistics point to the occurrence of clusters of similar values while negative and low values of that statistic point to the occurrence of the so-called hot spots. Values close to 0 point to a random distribution of the studied value in space.



By analyzing the smoothed variable textsfprev we strengthen the clusterization effect. We obtain a similar result but this time we localize 3 clusters (19 census tracts) which are cluster centers.

Local Moran's I statistic	->> + MAP <<-
Analysis time	0,52sec.
Analysed variables	prev
Significance level	0,05
Corrected significance level (Bonferroni)	0,009147
Average number of neighbors	5,466192
Spatial weights matrix	Queen - Immediate neighl
Number of objects	281
Variable smoothing	locally weighted average
Mean Ii	0,430168
Standard deviation Ii	1,004429
Frequency (High-High 1)	19
Frequency (Low-low 3)	8
Frequency (Low-High 2)	0
Frequency (High-Low 4)	0

7.2 Local Getis-Ord's G statistic

Getis and Ord's G_i statistic (Getis and Ord 1992, Ord and Getis 1995) allows the detection of a local concentration of high and low values in neighboring objects and studies the statistical significance of that dependence. Getis and Ord have also defined a G_i^* statistics, very similar to the G_i statistics. The only difference between them is that in the case of the former the object for which the study is made also takes part in the analysis. In a weight matrix, then, the so-called potential is defined for that object, i.e. the neighborhood with itself (values on the axis are greater than 0).

Getis-Ord's G coefficient

The local form of Getis and Ord's G coefficient for the i observation is defined with the formula:

$$G_i = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} x_j}, \quad \text{where: } i \neq j.$$

The G_i^* coefficient is defined with the same formula but the computations are also made for the studied object, that is the object for which the i and the j indexes are equal.

As the coefficient is based on a quotient of two sums of the values of the (x_j) objects, in order to interpret the coefficient correctly it is important that the analyzed phenomenon is described with the use of positive numbers. The interpretation of Getis and Ord's local coefficient, similarly to local Moran's coefficient, depends, to a great degree, on the selected weight matrix (row standardization of the matrix is recommended). High values of the G_i or G_i^* coefficients point to a concentration of objects with high values of the analyzed phenomenon, whereas low values point to a clustering of objects with low values. When the values are close to the expected value then the spatial distribution of the studied value is random.
The expected value is defined with the formula:

$$E(G_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n-1}, \quad \text{where: } i \neq j;$$

$$E(G_i^*) = \frac{\sum_{j=1}^{n} w_{ij}}{n}.$$

The significance of Getis and Ord's coefficient

By testing the statistical significance of the relationship among the neighboring objects the following hypotheses are studied:

$$\begin{array}{ll} \mathcal{H}_0: & G_i = E(G_i) \\ \mathcal{H}_1: & G_i \neq E(G_i), \end{array} \qquad \begin{array}{ll} \mathcal{H}_0: & G_i^* = E(G_i^*) \\ \mathcal{H}_1: & G_i^* \neq E(G_i^*). \end{array}$$

The test statistic has the form presented below:

$$Z_i(G) = \frac{\sum_{j=1}^{n} w_{ij}x_j - \overline{x}(i)\sum_{j=1}^{n} w_{ij}}{s(i)\sqrt{\frac{(n-1)\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-2}}}, \quad \text{where: } i \neq j;$$

$$Z_i(G*) = \frac{\sum_{j=1}^{n} w_{ij}x_j - \overline{x}^*\sum_{j=1}^{n} w_{ij}}{s^*\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}}.$$

where:
$\overline{x}(i)$ i $\overline{x}^*$ – the mean of the variable X,
$s(i)^2$ i s^{*2} – the variance of the X variable.

The Z_i statistics has, asymptotically (for large sample sizes), normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{array}{lll} \text{if } p \leq \alpha & \Longrightarrow & \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \Longrightarrow & \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

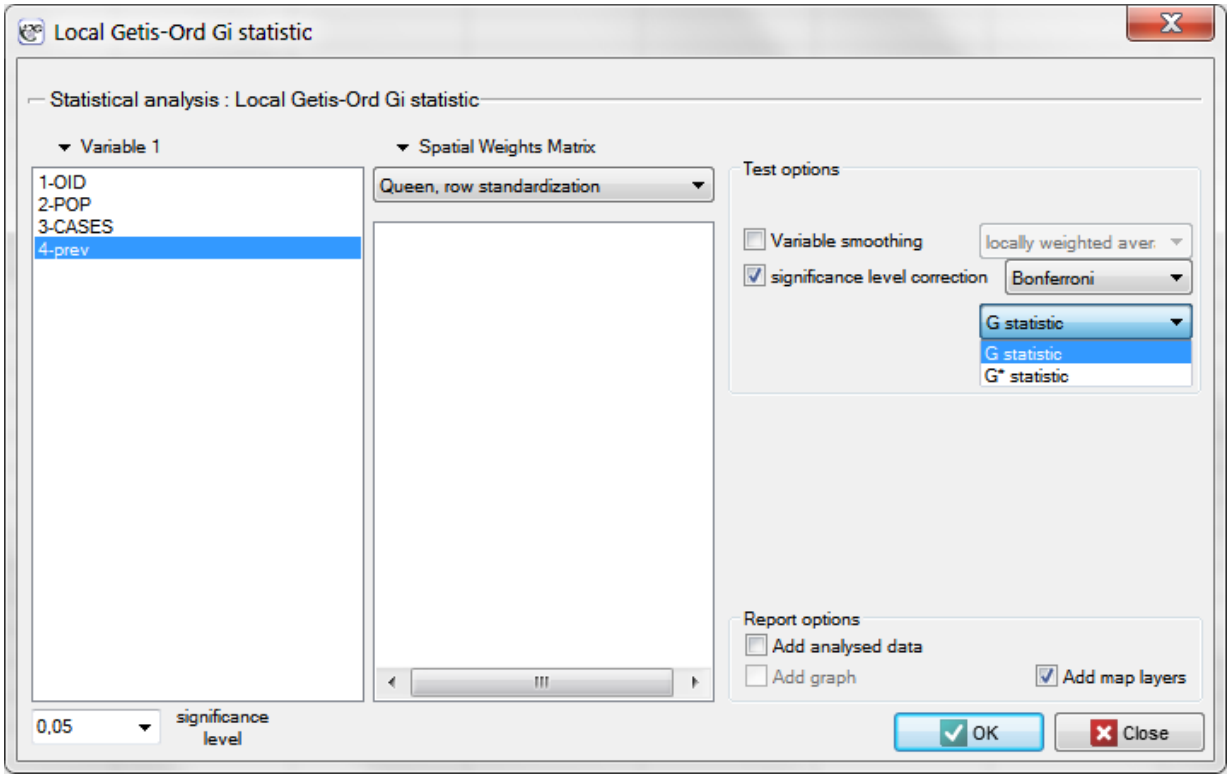
Due to the problem of a lack of independence of coefficients computed for neighboring objects it is suggested to use a corrected significance level α. The suggested corrections are: Bonferroni correction: $\alpha_1 = \alpha/k$ or Šidák correction: $\alpha_1 = 1 - (1-\alpha)^{1/k}$, where k is the arithmetic mean number of the neighbors.

Map layers

The combination of the information from the value of the Z_i statistic and its significance presents the so-called **spacial regimes** on the map:

- Statistically significant objects with high values of the Z_i statistic are marked as **High-High** (objects with high values surrounded by objects with high values) and marked in red on the map;
- Statistically significant objects with low values of the Z_i statistic are marked as **Low-Low** (objects with low values surrounded by objects with low values) and marked in blue on the map;

The window with settings for Local Getis and Ord's analysis is accessed via the menu Spatial analysis $\rightarrow$ Spatial statistics $\rightarrow$ Getis-Ord G_i statistic.



EXAMPLE 6.1 cont. (catalog: leukemia, file: leukemia)

We will analyze the data concerning leukemia.

- The map leukemia contains information about the location of 281 polygons (census tracts) in the northern part of the state of New York.
- Data for the map leukemia:
  - Column CASES – the number of cases of leukemia in the years 1978-1982, ascribed to particular objects (census tracts). The value should be an integral number, however, in agreement with Waller's (1994) description, some cases which could not be objectively ascribed to a particular region have been divided proportionately. Hence, the numerousnesses of the cases ascribed to the 281 objects are not integral numbers.
  - Column POP – population size in particular objects.
  - Column prev – the frequency coefficient of leukemia per 100000 people, for each object in one year: prev=(CASES/POP)*100000/5

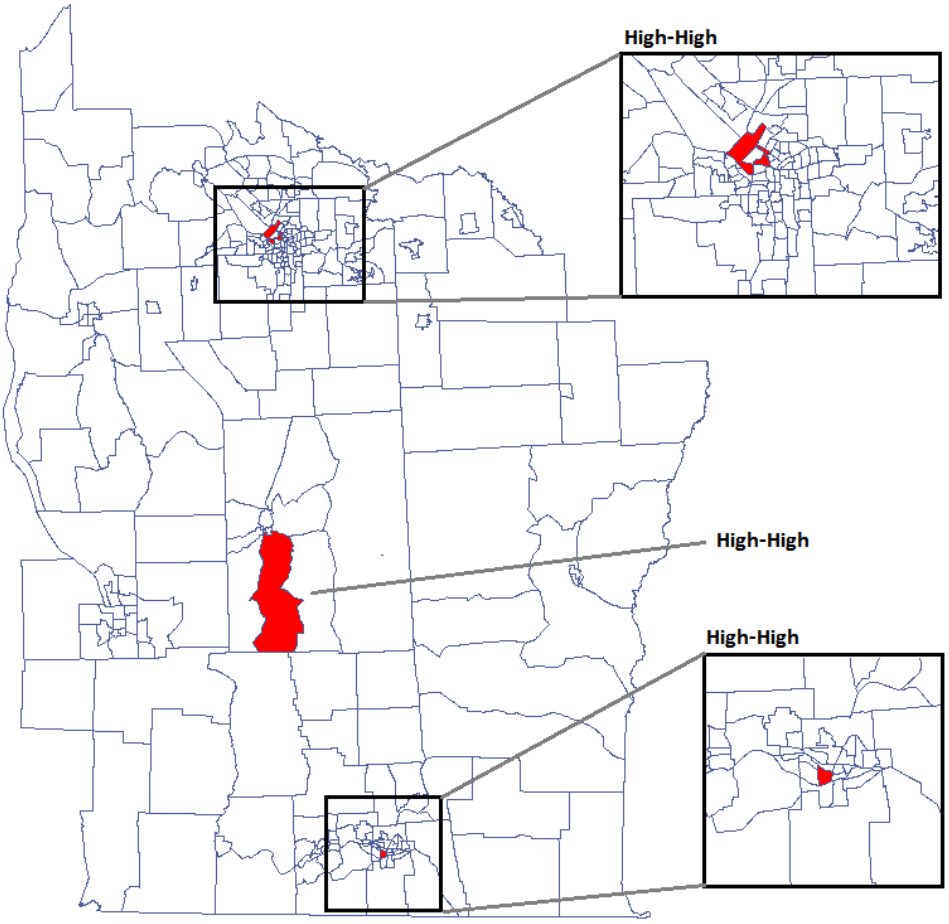
The global analysis has not yielded an unambiguous answer as to the occurrence of spatial autocorrelation. We will, then, check if we can find regions in which the prevalence of leukemia is significantly higher.

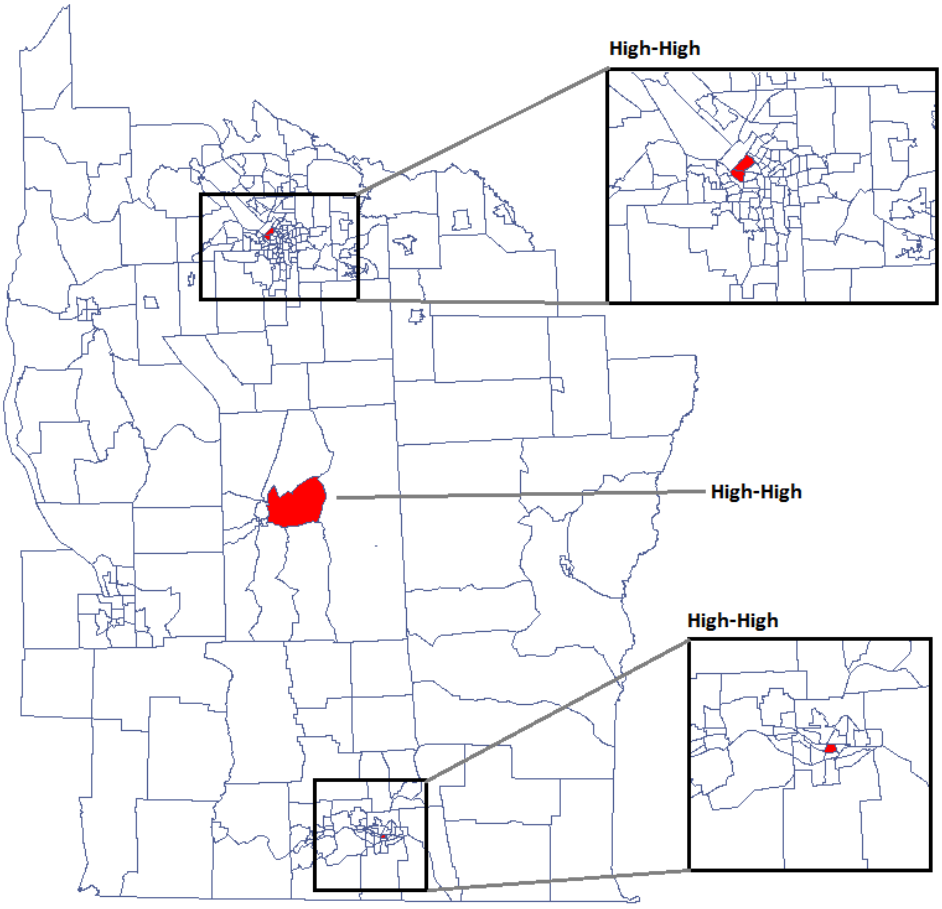
In order to localize leukemia clusters we will compute the G_i and G_i^* coefficients. The analysis will be conducted with the prev variable and the neighborhood matrix – Queen, row standardized (according to contiguity) which is suggested by the program. In order to use a different matrix one has to generate it first, see chapter: Spatial weight matrix. We also select one of the corrections of the significance level.

Local Getis-Ord Gi statistic	->> + MAP <<-
Analysis time	0,47sec.
Analysed variables	prev
Significance level	0,05
Corrected significance level (Bonferroni)	0,009147
Average number of neighbors	5,466192
Spatial weights matrix	Queen - Immediate neighl
Number of objects	281
Mean Gi	0,003517
Standard deviation Gi	0,001815
Frequency (Low-Low 1)	0
Frequency (High-High 2)	6

Local Getis-Ord Gi statistic	->> + MAP <<-
Analysis time	0,46sec.
Analysed variables	prev
Significance level	0,05
Corrected significance level (Bonferroni)	0,009147
Average number of neighbors	5,466192
Spatial weights matrix	Queen - Immediate neighl
Number of objects	281
Potential value	1
Mean Gi*	0,007063
Standard deviation Gi*	0,004523
Frequency (Low-Low 1)	0
Frequency (High-High 2)	4

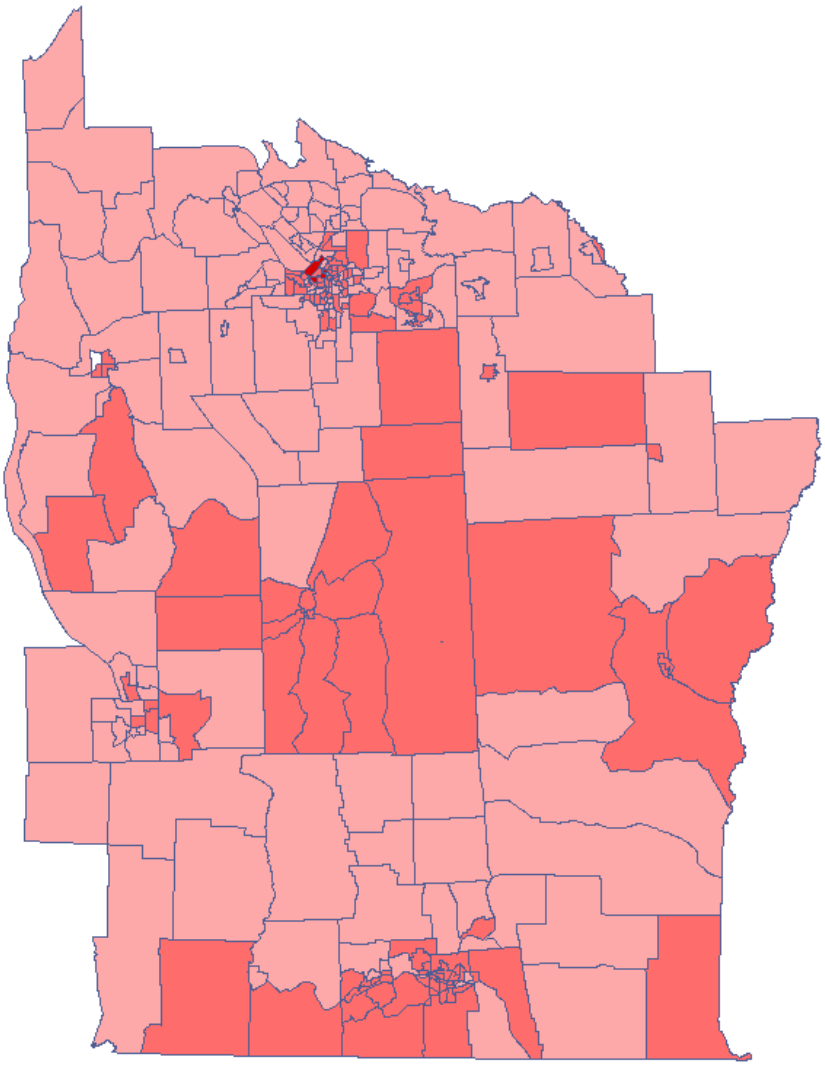
The obtained report presents the values of local coefficients, the values of test statistics, and the corresponding values of test probability. We will also find the information about the number of regions defining the spatial regimes (High-High, Low-Low). Also, a result is ascribed to the analysis, which we can draw on the map (button ->> + MAP <<-) – those spatial regimes are defined in the report with the use of the color column.



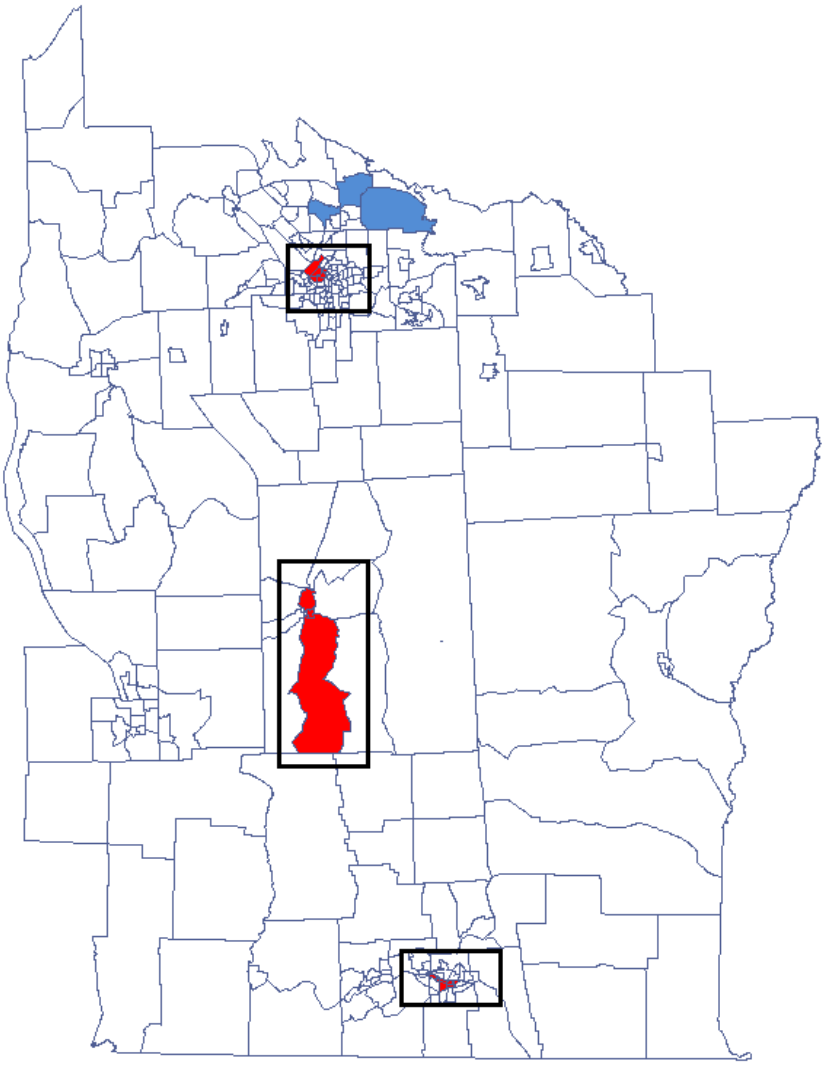


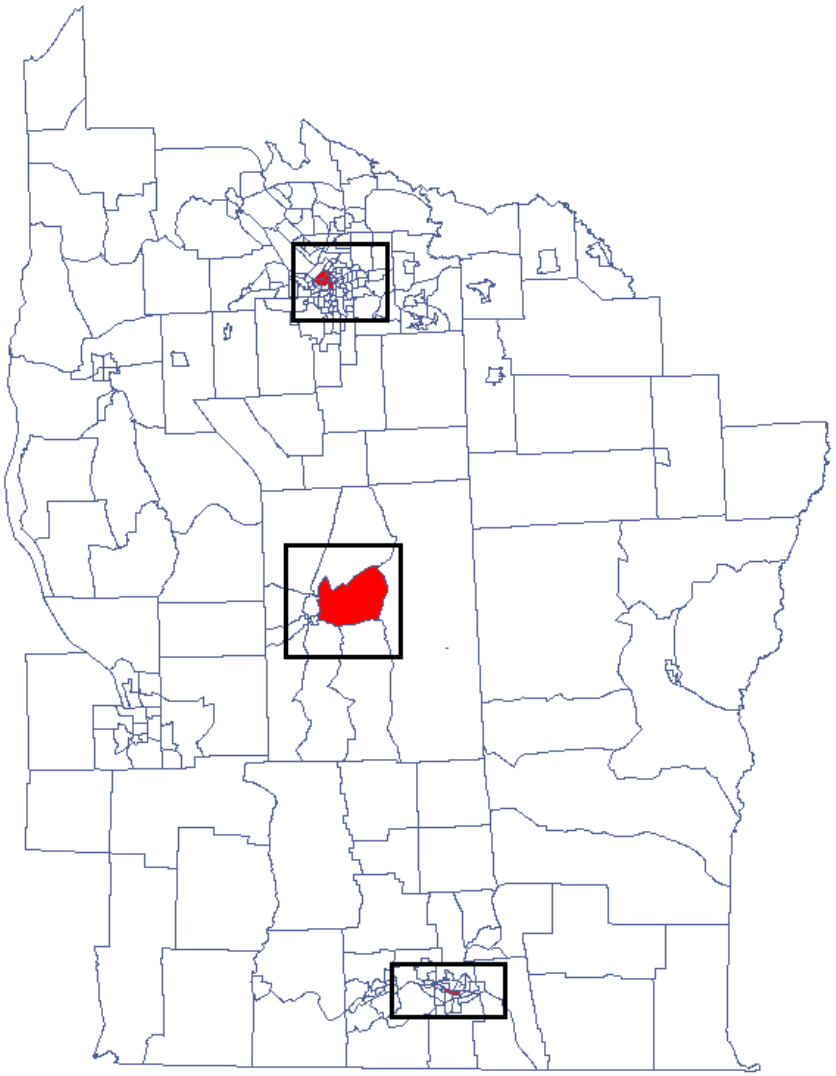
We were able to localize 3 clusters (6 census tracts in the analysis of the G_i coefficient and 4 tracts in the analysis of the G_i^* coefficient) in which the prevalence of leukemia is significantly higher. They are the centers of clusters with high values of leukemia, marked in red on the map.

The obtained results can be additionally illustrated by coloring the map so as to present the values of the local Getis and Ord's coefficient or the values of the test statistic, or the p values. One just has to copy the appropriate columns from the report and paste them into a datasheet. In this example we will use the values of the $Z(G_i)$ test statistic for coloring. Having pasted it into an empty column of a datasheet, in the map manager we color the base map according to the values of that column, selecting the standard deviation with the coefficient 3 as a way of gradiating colors. Positive and high values of the Z_i statistic point to a concentration of objects with high values, whereas negative and low values point to a concentration of objects with low values, and the values near zero point to a random spatial distribution of the studied variable (a map can be added with means and confidence intervals).



By analyzing the smoothed variable textsfprev we strengthen the clusterization effect. We obtain a similar result, i.e. 3 clusters (15 census tracts in the analysis of the G_i coefficient and 9 tracts in the analysis of the G_i^* coefficient) which are cluster centers.





7.3 CutL

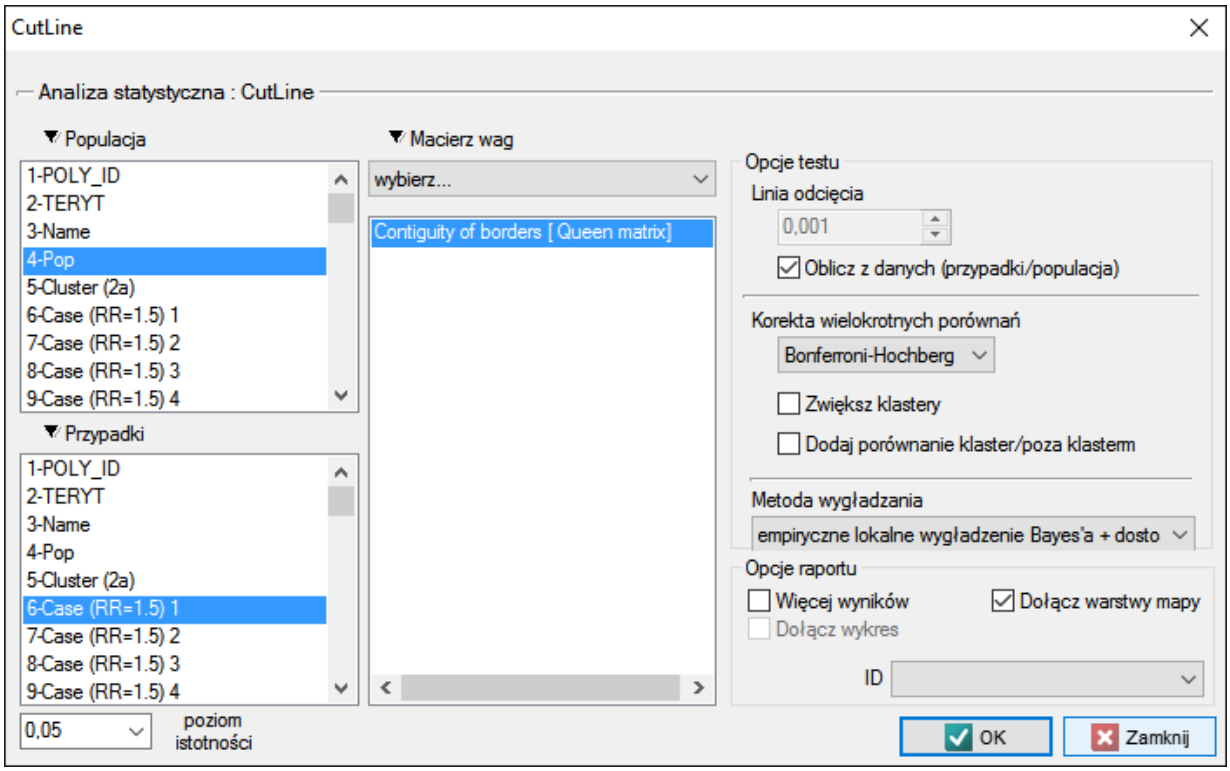
CutL method was developed to detect clusters that have significantly higher prevalence than that specified by the investigator[15]. As a result, the program locates clusters, examines their statistical significance, and plots them on a map.

Note
Analysis is based on the commonly used Binomial test for equal proportions.

Analyses are conducted on aggregated data (polygons map). In each of the two columns of the data sheet, enter the population size and number of cases for EACH object – these are the data required for proper analysis.

ID	Pop	Cases
1	548028	505
2	4896	2
3	3981	5
4	5658	7
5	9591	4
6	3011	2
7	4938	7
8	8664	11
...	...	...
...	...	...

The settings window of CutL statistic can be opened in Spatial analysis→Spatial statistics→CutL



The analysis is based on data from the population size and number of cases, as well as the neighborhood matrix on the map.

Using the **neighborhood matrix**:
The Queen neighborhood matrix is the default chosen during the analysis. Other matrices may be used in this analysis, but this requires prior preparation and selection in the analysis window.

The **cut-off line** is the value above which statistically significant clusters can be detected and may be set in the analysis window. If the investigator does not specify this value, then it is set to the average prevalence calculated for the area under study.

Options

Corrections for multiple comparisons

The following corrections for multiple comparisons may be used:

- Bonferroni-Hochberg

- Sidak-Hochberg
- Benjamini-Hochberg

Compare cluster/outside In addition, each cluster may be compared with the area outside of the cluster. The Binomial test for one proportion compares the prevalence within the cluster to the specified prevalence value, which is the prevalence outside of the cluster. Hypothesis testing is one-sided resulting from the higher prevalence inside the cluster than outside.

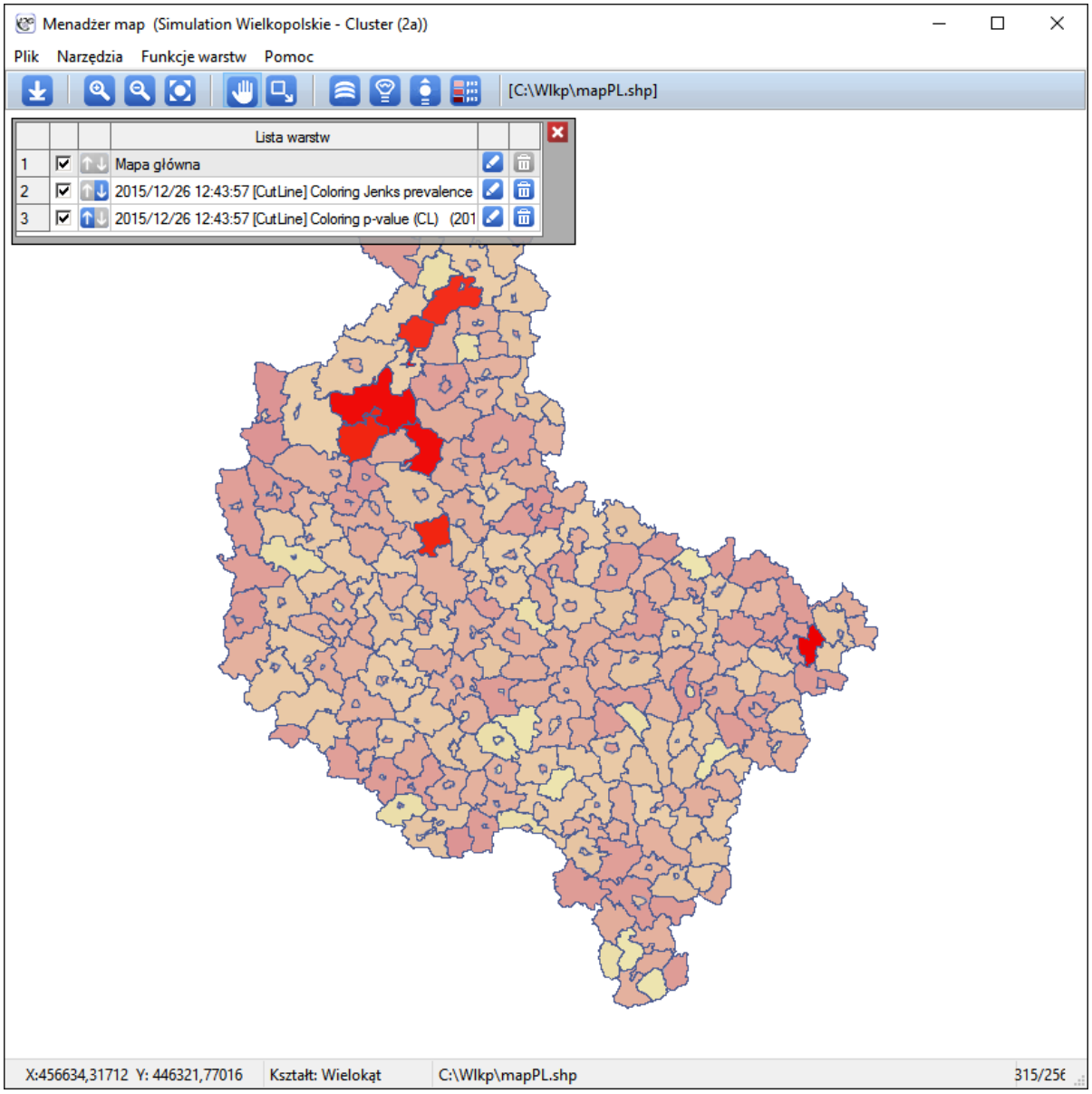
Results of Analysis

The results of the analysis are presented in the form of a report and map with superimposed layers.

CutLine	[Dodaj do mapy]
Czas analizy	0,22 sek.
Analizowane zmienne	Pop;Case (RR=1.5) 1
Poziom istotności	0,05
Macierz wag przestrzennych	Queen matrix
Korekta wielokrotnych porów	Bonferroni-Hochberg
Zwiększone klastery	Nie
Liczba obiektów	315
Suma populacji	3467016
Suma przypadków	3467
Współczynnik ogółem	0,001
Współczynnik CutLine	0,001
	automatyczne
Liczba wykrytych klasterów	4
Liczba istotnych klasterów	4

Klastery CutLine

ID	Liczba obie	Pop	Przypadki	Wsp	Wsp/CutLir	RR	Wartość p
Piła	3	82190	118	0,001436	1,435704	1,451056	0,000356
Czarnków	4	37367	66	0,001766	1,766273	1,781143	0,000058
Suchy Las	1	15971	27	0,001691	1,690572	1,695992	0,007268
Grzegorzew	1	5733	15	0,002616	2,616443	2,623467	0,001791



Spatial-Temporal CutL

sing the CutL method, it is also possible to determine temporal-spatial clusters (Więckowska B. 2019 [16]), i.e., clusters that do not persist for the entire time range under study, but only for a shorter period.

Individual time layers are added to the datasheet by selectingr Edit Timeline from the project tree, after indicating the appropriate map.

The spatio-temporal analysis window is obtained by selecting the menu Spatial analysis→Spatial statistics→CutL space-time.

Literatura

[1] Anselin L., (1995), *Local Indicators of Spatial Association – LISA; Geographical Analysis, 27(2): 93–115*

[2] Anselin L., Lozano N., Koschinsky J. (2006) *Rate Transformations and Smoothing. GeoDa Center Research Report https://geodacenter.asu.edu*

[3] Buliung R.N., Remmel T.K. (2008), *Open source, spatial analysis, and activity-travel behaviour research: capabilities of the aspace package. Journal of Geographical Systems 10, 191-216*

[4] Clark P.J., Evans F.C. (1954), *Distance to nearest neighbour as a measure of spatial relationships in populations. Ecology 35, 445-453.*

[5] Cliff A.D., Ord J.K., (1981), *Spatial Processes: Models and Applications. Pion: London.*

[6] Fisher R.A. (1936), *The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188.*

[7] Geary R.C., (1954), *The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 5, 115-45*

[8] Goodchild M.F., (1986), *Spatial Autocorrelation, CATMOG 47, Geobooks: Norwich UK*

[9] Moran P.A.P., (1947), *The Interpretation of Statistical Maps. Journal of the Royal Statistical Society, B10, 243-51*

[10] O'Rourke J. (1998), *Computational Geometry in C (2nd ed). Massachusetts: Smith College*

[11] De Smith M.J., Goodchild M.F., Longley P.A. (2007) *Geospatial Analysis, A Comprehensive Guide to Principles, Techniques and Software Tools (2nd ed). Matador*

[12] Waller L.A., Turnbull B.W., Clark L.C., Nasca P., (1992), *Chronic disease surveillance and testing of clustering of disease and exposure : Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. Environmetrics, 3, 281-300*

[13] Waller L.A., Turnbull B.W., Clark, L.C., Nasca P., (1994), *Spatial pattern analyses to detect rare disease clusters, in Case Studies in Biometry, N. Lange, et al., Editors. , John Wiley and Sons: New York, 3-23*

[14] Waller L.A., Gotway C.A., (2004) *Applied Spatial Statistics for Public Health Data. New York: John Wiley and Sons*

[15] Więckowska B., Marcinkowska J. (2017), *CutL: an alternative to Kulldorff's scan statistics for cluster detection with a specified cut-off level. Geospatial Health, 12(2): 556*

[16] Więckowska B., Górna I., Trojanowski M., Pruciak A., Stawińska-Witoszyńska B. (2019), *Searching for space-time clusters: The CutL method compared to Kulldorff's scan statistic 14(2)*

[17] Yamamoto J.K. (1997), *A Pascal program for determining the convex hull for planar sets. Computers and Geosciences 23, n. 7, 725-738*