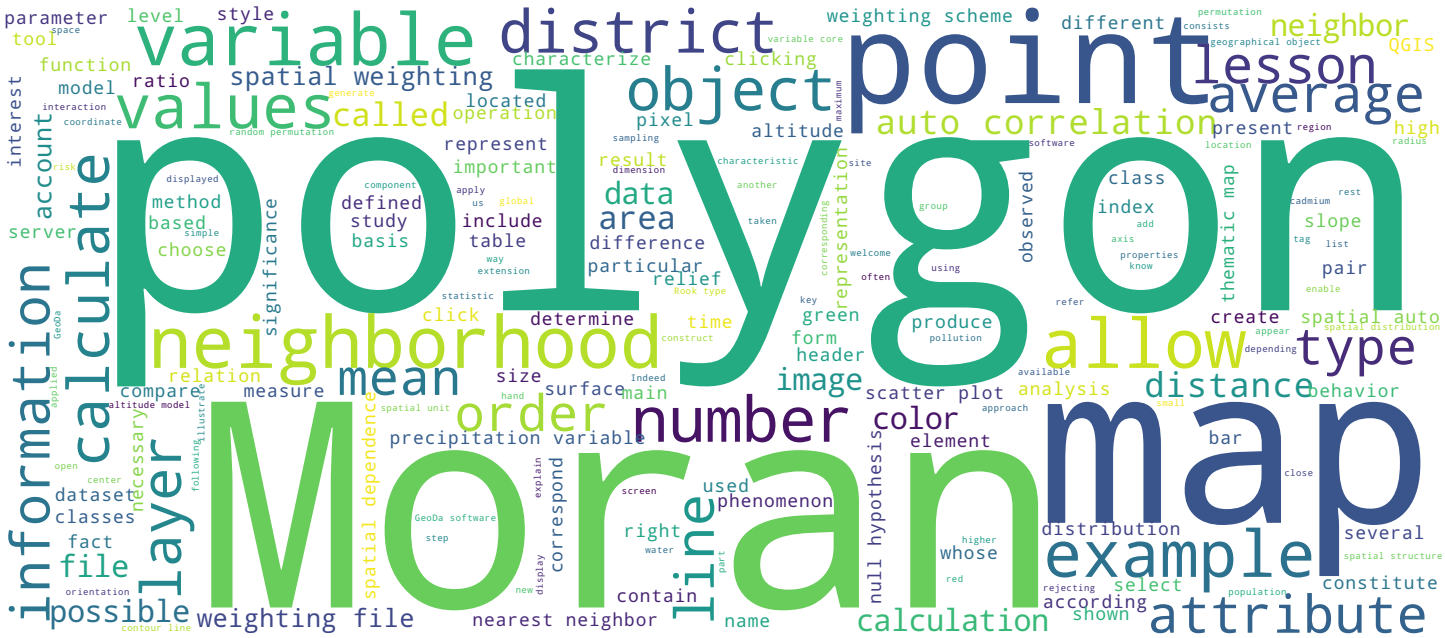
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

RESCIF
Réseau d'excellence des sciences de l'ingénieur de la Francophonie

Global Spatial Autocorrelation

An Introduction to Geographic Information Systems

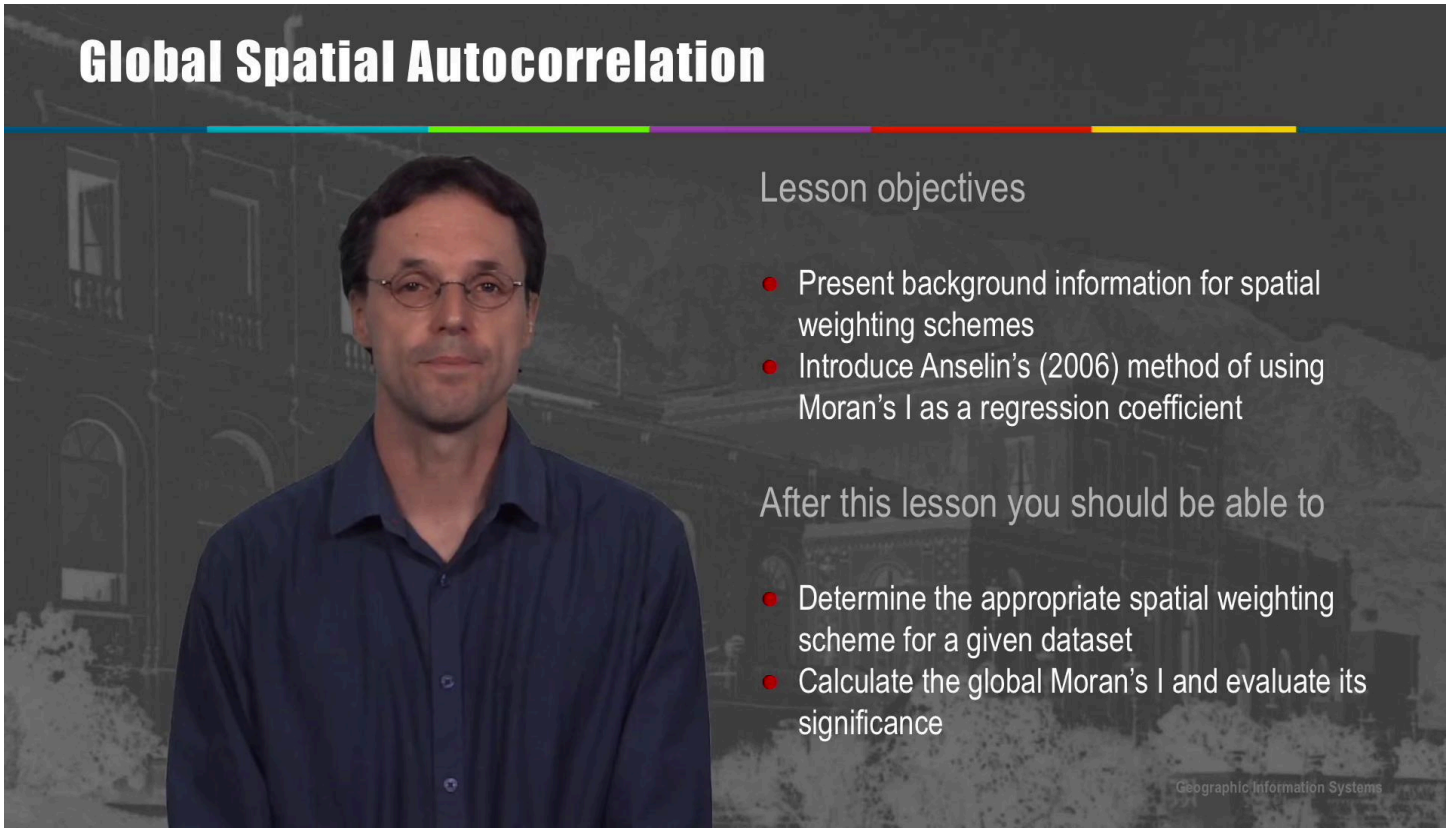
Stéphane Joost, Marc Soutter, Fernand Kouamé, Amadou Sall



Search MOOC

Video

EPFL

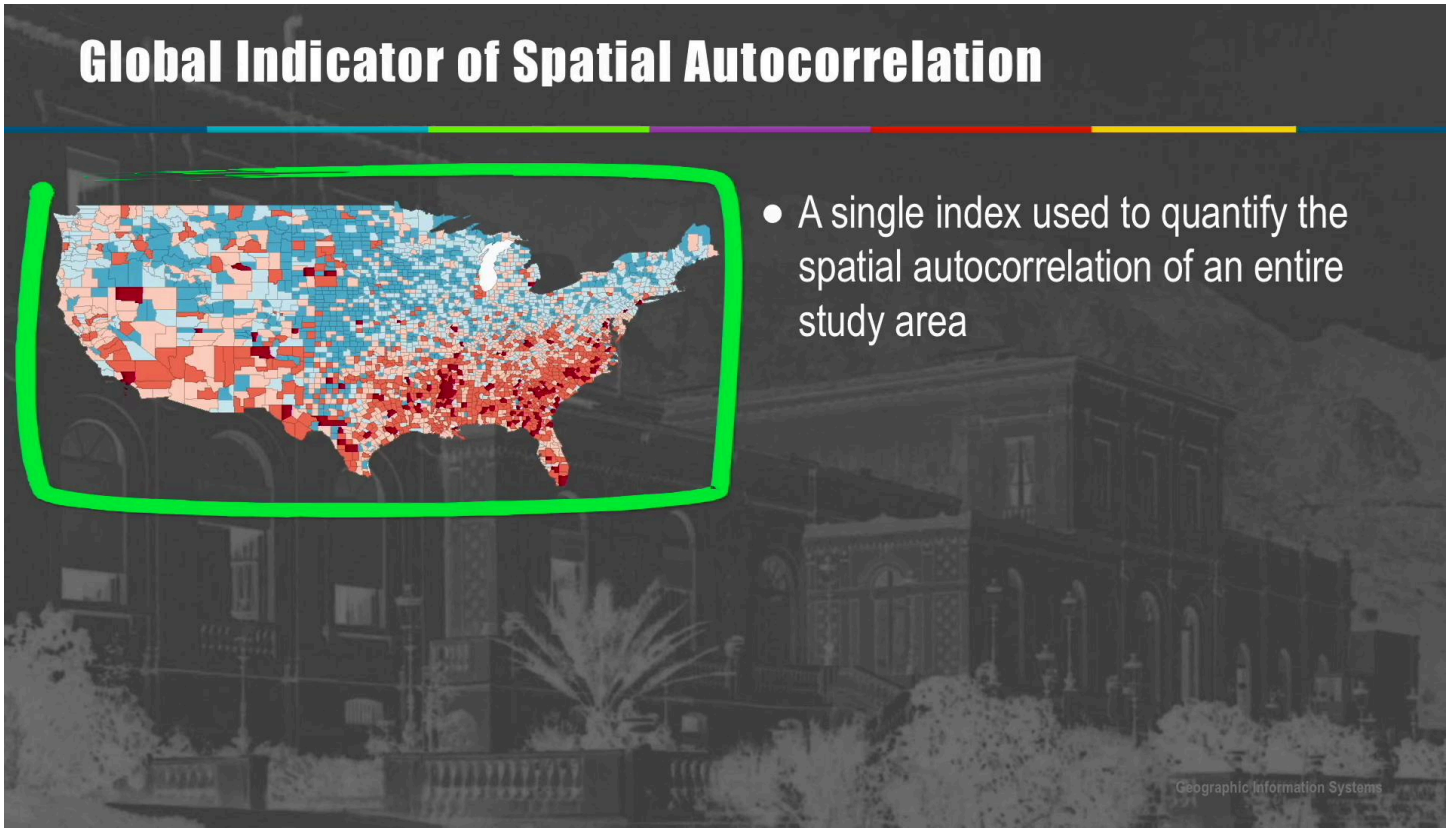


Hello and welcome to this lesson on a measure of the global spatial auto-correlation, the I of Moran. An auto-correlation index allows to quantify the regularity or the structure of a spatially distributed phenomenon by taking into account the neighborhood of each of the individuals which constitute the geodataset. We will see how to take this neighborhood into account in order to develop spatial weighting schemes. These diagrams are absolutely necessary to calculate an index such as the I of Moran and at the same time to estimate the scope of the spatial dependence. You will finally learn how to assess the significance of this index. The aims of this lesson are to pass on the notions necessary for the calculation of a spatial auto-correlation index. We will present the information useful for the definition of spatial weighting schemes and explain the I of Moran index, in particular its interpretation as a regression coefficient which provides an intuitive and easy to remember approach. After following our explanations, you should be able to determine a spatial weighting mode that is adapted to the analyzed datasets. And you should also be able to calculate the global I of Moran as well as its significance for any punctual or surface geodataset.

Notes

Summary

0m 31s



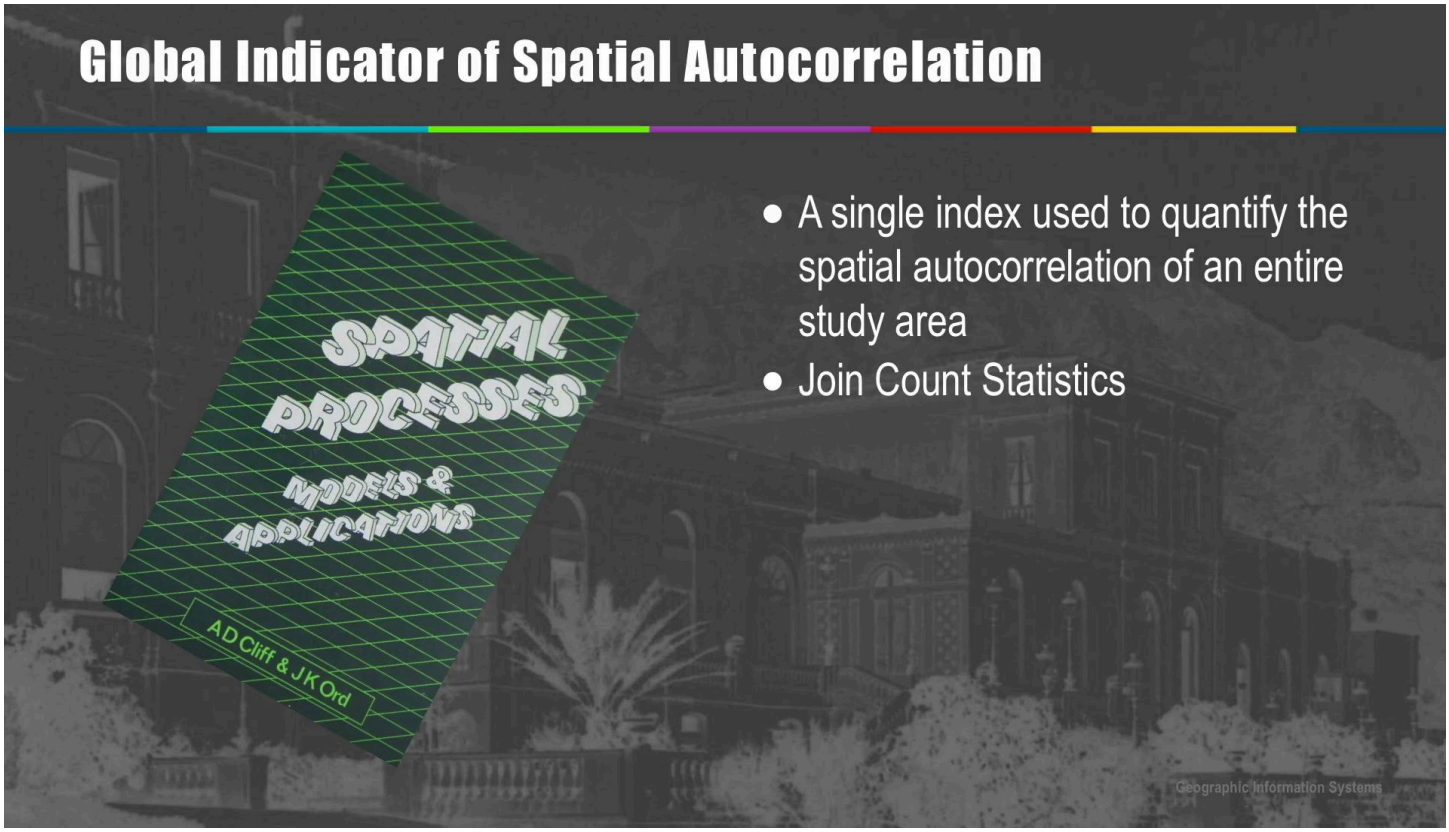
We present in this lesson the concepts required to calculate a global auto-correlation indicator. It is in fact a unique index which characterizes the spatial arrangement of the geographical units according to the values of a given attribute and this on the whole of an analyzed territory.

Notes

Summary

1m 55s

Global Indicator of Spatial Autocorrelation



- A single index used to quantify the spatial autocorrelation of an entire study area
- Join Count Statistics

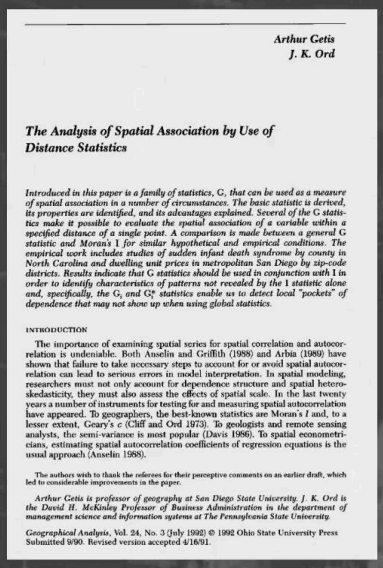
There are several measurement methods: the Join Count Analysis which is an enumeration statistic and has the particularity of being applicable only to polygons.

Notes

Summary

2m 14s

Global Indicator of Spatial Autocorrelation



Arthur Getis
J. K. Ord

The Analysis of Spatial Association by Use of Distance Statistics

Introduced in this paper is a family of statistics, G, that can be used as a measure of spatial association in a number of circumstances. The basic statistic is derived, its properties are identified, and its advantages explained. Several of the G statistics make it possible to evaluate the spatial association of a variable within a specified distance of a single point. A comparison is made between a general G statistic and Moran's I for similar hypothetical and empirical conditions. The empirical work includes studies of sudden infant death syndrome by county in North Carolina and dwelling unit prices in metropolitan San Diego by zip-code districts. Results indicate that G statistics should be used in conjunction with I *in order to identify characteristics of patterns not revealed by the* I *statistic alone and, specifically, the* G_i *and* G_i^* *statistics enable us to detect local "pockets" of dependence that may not show up when using global statistics.*

INTRODUCTION

The importance of examining spatial series for spatial correlation and autocorrelation is undeniable. Both Anselin and Griffith (1988) and Arbia (1989) have shown that failure to take necessary steps to account for or avoid spatial autocorrelation can lead to serious errors in model interpretation. In spatial modeling, researchers must not only account for dependence structure and spatial heteroskedasticity, they must also assess the effects of spatial scale. In the last twenty years a number of instruments for testing for and measuring spatial autocorrelation have appeared. To geographers, the best-known statistics are Moran's *I* and, to a lesser extent, Geary's *c* (Cliff and Ord 1973). To geologists and remote sensing analysts, the semi-variance is most popular (Davis 1986). To spatial econometricians, estimating spatial autocorrelation coefficients of regression equations is the usual approach (Anselin 1988).

The authors wish to thank the referees for their perceptive comments on an earlier draft, which led to considerable improvements in the paper.

Arthur Getis is professor of geography at San Diego State University. J. K. Ord is the David H. McKinley Professor of Business Administration in the department of management science and information systems at The Pennsylvania State University.

Geographical Analysis, Vol. 24, No. 3 (July 1992) © 1992 Ohio State University Press
Submitted 9/90. Revised version accepted 4/16/91.

- A single index used to quantify the spatial autocorrelation of an entire study area
- Join Count Statistics
- Geary's C
- Ripley's K
- Getis-Ord's G
- Moran's I

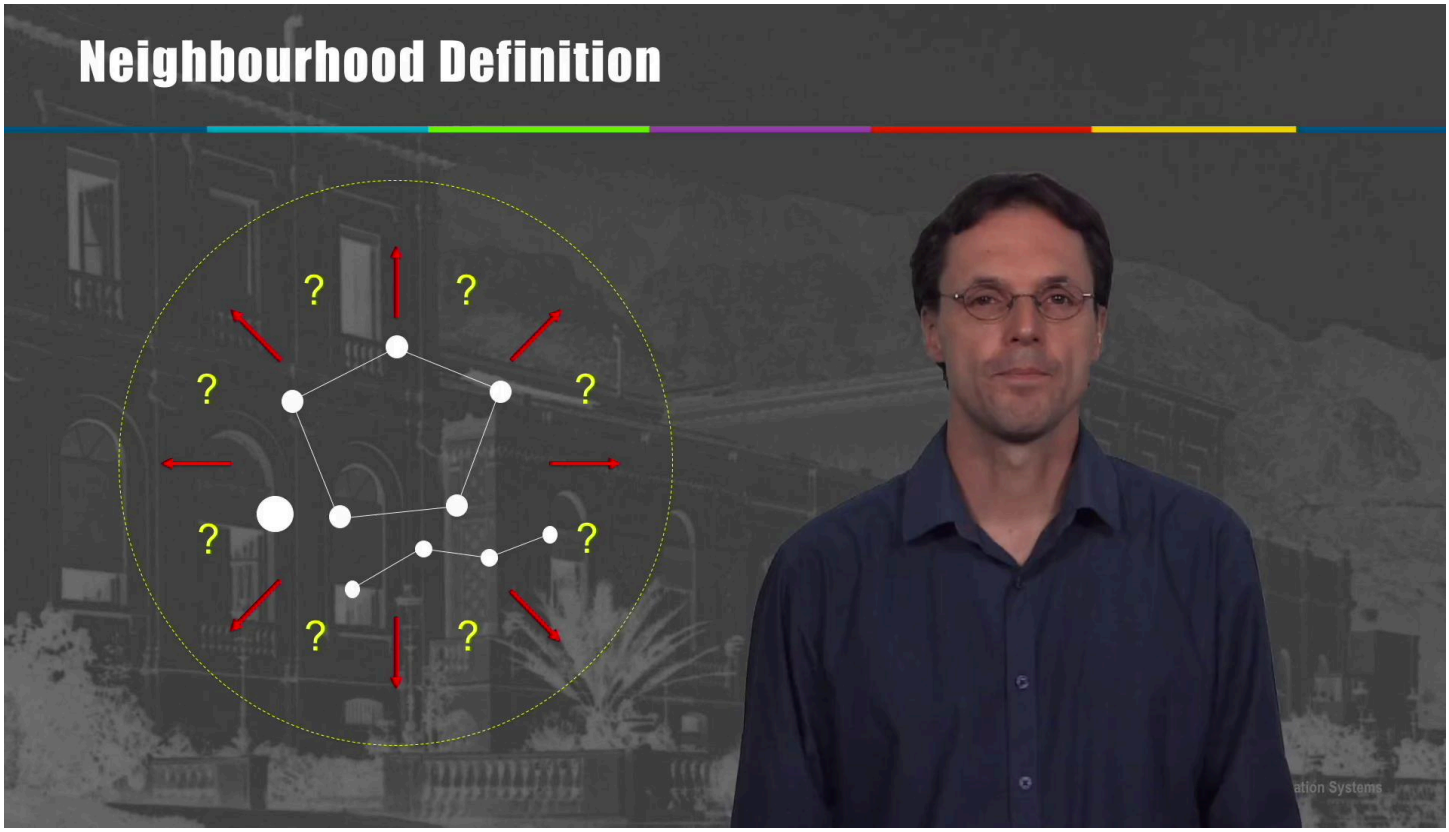
Getis, A., Ord, J.K. (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, **24**, 189–206.

And amongst the others, we find the C of Geary, the K of Ripley the G of Getis-Ord and finally the I of Moran which is the most widespread and which is the subject of this lesson.

Notes

Summary

2m 22s



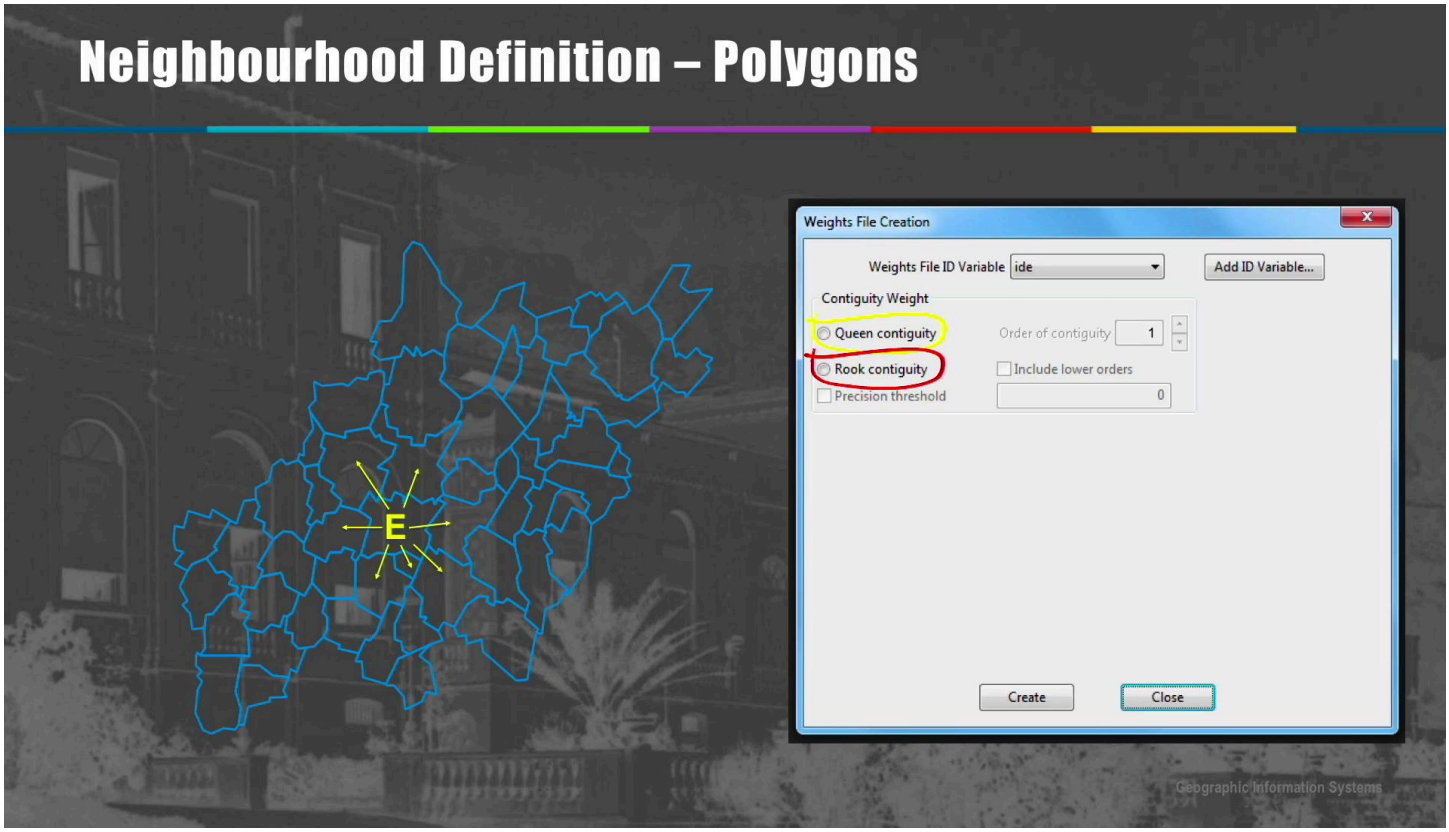
To quantify the spatial dependence and generate a global spatial auto-correlation measure, it is necessary to take the neighborhood of each of the geographical objects considered into account. Indeed the measure of the global auto-correlation consists in comparing the behavior of an object with the behavior of its neighbors and this on all the territory studied. The key being to define this neighborhood according to different possible criteria. Let's take for example this group of 54 points which represents in this case the mass centers or the district centroids for which we decide to define an neighborhood of 5km. The white circle identifies the object 1 and the yellow circle defines a radius of 5km around the latter. The neighborhood is defined here by this radius of 5 km. It then allows to compare the value of an attribute A of the object 1 with a statistic of this attribute A for the 17 objects highlighted in green and situated in this neighborhood. This statistic can be the average as is the case with the I of Moran. This operation is repeated so as to be able to compare the attribute A of each object with the average attribute A of its neighborhood, so 54 times here. There are several criteria on which to base ourselves to define the neighborhood of a geographic object. And we will use the GeoDa software interface here to illustrate the implementation of these criteria.

Notes

Summary

2m 44s

Neighbourhood Definition – Polygons



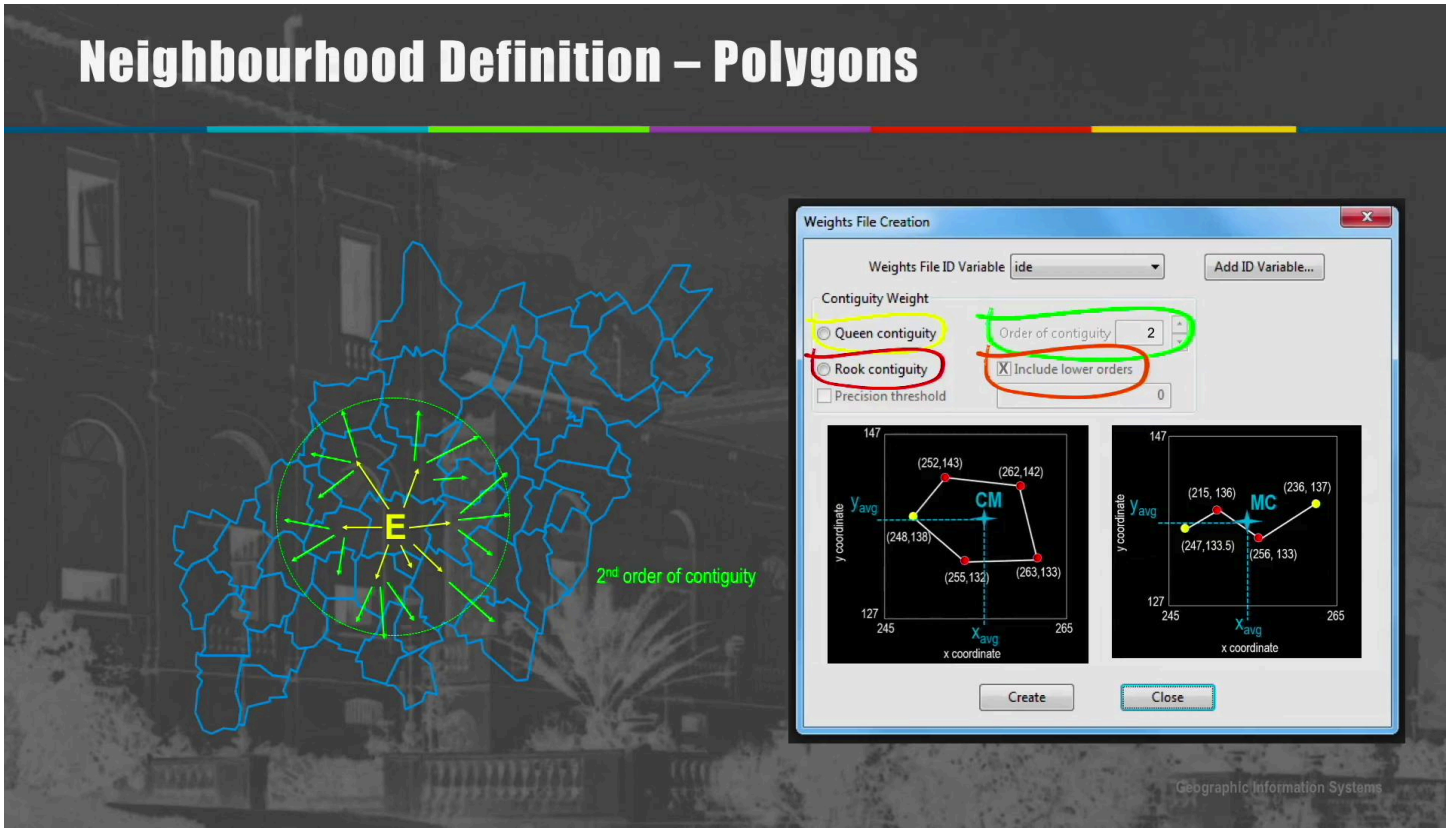
The options available to define a criterion depend partly on the type of object considered. When we want to define the neighborhood of punctual objects, The most frequently used criteria are to be based either on a distance or on a proximity criterion which consists of identifying the K nearest neighbors. Regarding the first option, the distance or radius used here is 5000 m and takes up the example used earlier. In this case, the neighborhood is defined only on the basis of the distance between the point and its neighbors and constitutes what is called a fixed core. The distance D of 5000 m is called the bandwidth. In the case of the nearest neighbors, here K is equal to 7, the neighborhood is adjusted depending on the density of objects around the point. And we then say that the associated function uses a variable core. The surface objects or polygons allow in addition to play on the adjacency or contiguity relations. There are 2 types of adjacency relations: the Queen and Rook types that refer to the movements of pieces in a chess game. The Queen type, so the queen moving in all directions, corresponds to a neighborhood that includes all the polygons that touch the polygon of interest, at least one pair of coordinates in common is required.

Notes

Summary

4m 12s

Neighbourhood Definition – Polygons

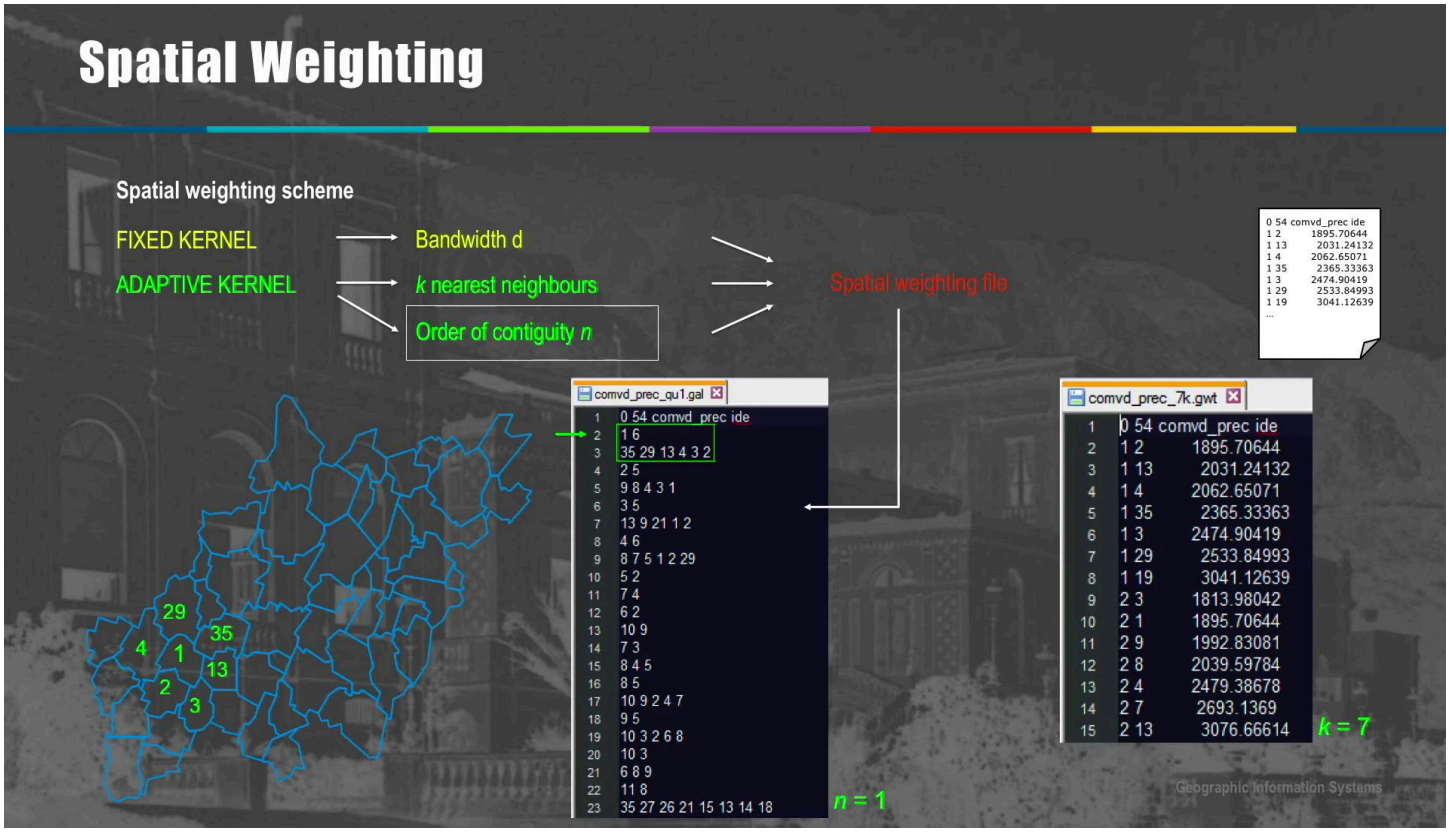


The Rook type, so the rook that moves only North, South, East and West, corresponds to a neighborhood that includes all the polygons that have at least 1 common side with the polygon of interest. This type of adjacency relation is mainly used in the case where geographic units are orthogonal polygons like many counties in the United States for example. There is also a parameter which is the order of contiguity. Indeed, we can take into account the immediately contiguous neighbors as is the case here in yellow around the polygon E which corresponds to an order of contiguity 1, but we can also define a neighborhood which rests on an order of contiguity 2, here designated by the green arrows, that is to say to include the polygons which, depending on the Queen or Rook type, are neighbors of the polygons of order 1. Finally, according to the characteristics of the current analysis, it is possible to include or not in the neighborhood the orders of contiguity of lower rank. The fixed cores like the distance, the variable cores like the K nearest neighbors are applicable to the polygons, but it is in this case the geographical coordinates of the geometric center of gravity of the object that are taken into account for the calculation. This is also the case when working with linear objects.

Notes

Summary

5m 37s



The neighborhood relations allow to determine the spatial weighting scheme that we want to apply to a dataset. This weighting scheme will be applied to create a file containing for each object the list of its neighbors. This is the spatial weighting file. This is what we can read when we edit this file. Here we take the example of the 54 districts used previously. The first line is the header and contains 4 elements separated by a space. The 0 that has no function in the current version of GeoDa, 54 which designates the number of polygons in the dataset, the file name and the name of the unique identifier. The rest of the file lists the neighbors of each polygon. In line 2, the number 1 designates the polygon number 1 and the number 6 indicates that this polygon has 6 neighbors according to the indicated criterion and that it is the polygons number 35 - 29 - 13 - 4 - 3 and 2 that are indicated in line 3. And so on, at line 4, polygon 2 has 5 neighbors that are polygons 9 - 8 - 4 - 3 and 1. With the criterion of the K nearest neighbors illustrated on the right, after the header, the file contains the 7 nearest neighbors of polygon 1 with on the right the distance that separates the centroids and then from line 9, the 7 nearest neighbors of polygon 2, and so on.

Notes

Summary

7m 07s

Correlation Between Neighbouring Measurements

comvd_prec_qu1.gal

```
1   0 54 comvd_prec ide
2   1 6
3   35 29 13 4 3 2
4   2 5
5   9 8 4 3 1
6   3 5
7   13 9 21 1 2
8   4 6
9   8 7 5 1 2 29
10  5 2
11  7 4
12  6 2
```

id	neighbours ($\rightarrow\omega$)	C	z	$\bar{z}$
1	35, 29, 13, 4, 3, 2	6	10291.14	10244.17
2	9, 8, 4, 3 , 1	5	10166.64	10085.33
3	13, 9, 21, 1, 2	5	10494.71	10334.34
…	…	…	…	…

$$I_m = \frac{Covariance}{Variance} = \frac{1/C \sum \omega_{i,j}(z_i-\bar{z})(z_j-\bar{z})}{1/n \sum (z_i-\bar{z})^2} = \frac{n \sum (z_i-\bar{z})(z_j-\bar{z})}{C \sum (z_i-\bar{z})^2}$$

Where n: number of spatial units; C: number of neighbours; z_i: variable value for unit i; z_j: variable value for unit j; ω_{ij}: weight of connection, 1 if adjacent, otherwise 0

Geographic Information Systems

The spatial auto-correlation is quantified by the calculation of a correlation between the measurements geographically adjacent to a measured phenomenon. With the I of Moran, we calculate the correlation between the analyzed attribute of a geographical object and the mean of this attribute calculated within the defined neighborhood. We will therefore use the contiguity spatial weighting file described earlier to calculate the mean of this attribute in the neighborhood of each of the objects constituting the dataset. The variable Z that we use in this example represents the monthly average of the sum of precipitations expressed in tenths of a millimeter per district. This variable Z is the third column of the table after the IDE identifier and the list of adjacent polygons which allows to calculate the average of Z in the last column. The neighbors listed in the second column allow to determine ω, so the weight attributed to the polygons in the calculation of the average of Z in the neighborhood. In the case of the contiguity criterion, this weight is 1 if a polygon is adjacent and 0 if it is not. The I of Moran auto-correlation coefficient Is an extension of the Bravais-Pearson correlation coefficient.

Notes

Summary

8m 41s

Correlation Between Neighbouring Measurements

comvd_prec_qu1.gal

```
1   0 54 comvd_prec ide
2   1 6
3   35 29 13 4 3 2
4   2 5
5   9 8 4 3 1
6   3 5
7   13 9 21 1 2
8   4 6
9   8 7 5 1 2 29
10  5 2
11  7 4
12  6 2
```

id	neighbours ($\rightarrow\omega$)	C	z	$\bar{z}$
1	35, 29, 13, 4, 3, 2	6	10291.14	10244.17
2	9, 8, 4, 3 , 1	5	10166.64	10085.33
3	13, 9, 21, 1, 2	5	10494.71	10334.34
…	…	…	…	…

$$I_m = \frac{Covariance}{Variance} = \frac{1/C \sum \omega_{i,j}(z_i-\bar{z})(z_j-\bar{z})}{1/n \sum (z_i-\bar{z})^2} = \frac{n \sum (z_i-\bar{z})(z_j-\bar{z})}{C \sum (z_i-\bar{z})^2}$$

Where n: number of spatial units; C: number of neighbours; z_i: variable value for unit i; z_j: variable value for unit j; ω_{ij}: weight of connection, 1 if adjacent, otherwise 0

Value of I varies between +1 (positive net correlation) and -1 (negative net correlation); 0 indicates an *absence of autocorrelation*, *no spatial dependence*, or *geographically neutral space*

Geographic Information Systems

It expresses the importance of the difference in the values of a variable between all the pairs of neighboring geographical objects. It is defined as shown here in the formula by the ratio between the covariance of a variable, in relation to the average of this variable and its variance on all the studied area. It is a measure of the linear correlation between 2 variables, so Z and Z bar here, and produces a value between +1 and -1 where 1 means that there is a total positive correlation, 0 no correlation and -1 negative or inverse total correlation.

Notes

Summary

9m 57s

Moran's I as a Regression Coefficient

ide >	commune	z	z_barre
1	Bettens	10291.14	10244.1741
2	Bournens	10166.64	10085.3390
3	Boussens	10494.71	10399.5093
4	Daillens	9708.88	9904.0585
5	Lussery-Villars	9642.24	9583.5026
6	Mex (VD)	9983.16	9897.9559
7	Penthalaz	9458.12	9636.0535
8	Penthaz	9557.04	9825.9107
9	Sullens	10374.92	9924.5102
10	Vufflens-la-Ville	9421.00	9971.7072
11	Assens	10763.34	10953.1723
12	Bercher	10413.63	10655.9440
13	Bioley-Orjulaz	10432.68	10526.8663
14	Bottens	11705.89	11404.4981
15	Bretigny-sur-Morrens	11519.39	11430.8364
16	Cugy (VD)	11901.36	11436.0724
17	Dommartin	11519.02	11617.8423
18	Echallens	10583.54	10857.6636
19	Eclagnens	10281.59	10325.2212
20	Essertines-sur-Yverdon	10406.37	10680.8931
21	Etagnières	10732.18	10760.8724
22	Fey	10626.72	10719.9970
23	Froideville	12821.19	12162.2771

...

In 1996, Luc Anselin proposed an interpretation of the I of Moran as a regression coefficient. This interpretation allows to fully understand the calculation of the I of Moran via the exploratory statistical tools implemented in the GeoDa software. We will implement it by calculating the I of Moran on the precipitation variable which characterizes the 54 districts of our dataset. We will use the same weighting scheme, so a variable core with a Queen contiguity order of 1. The I of Moran formula is equivalent to calculating the correlation between the precipitation value for each district, it is the variable z in the table, and the average of the precipitations for the neighbors of each district, it is the variable z bar in the table.

Notes

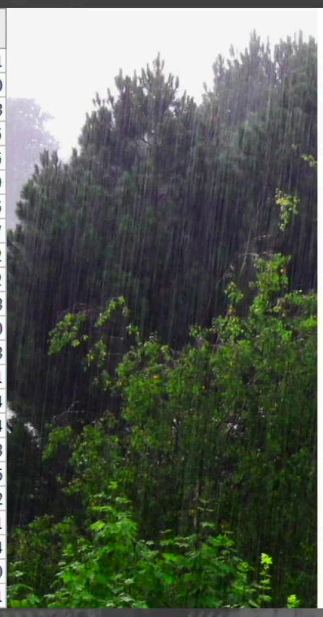
Summary

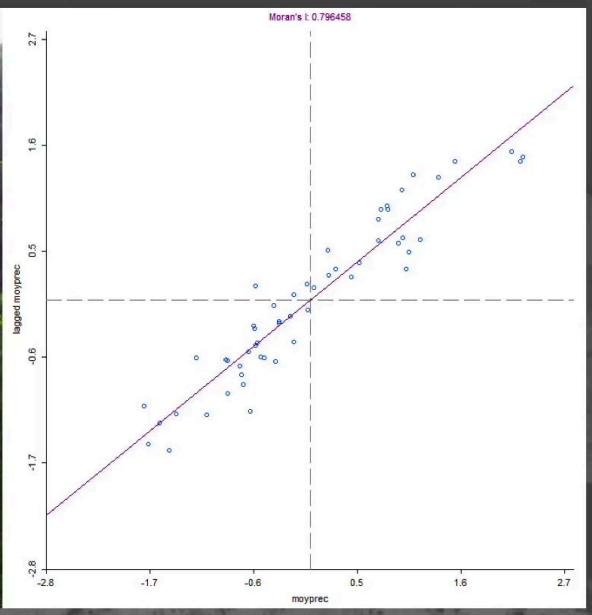
10m 35s

Moran's I as a Regression Coefficient

ide >	commune	z	z_barre
1	Bettens	10291.14	10244.1741
2	Bournens	10166.64	10085.3390
3	Boussens	10494.71	10399.5093
4	Daillens	9708.88	9904.0585
5	Lussery-Villars	9642.24	9583.5026
6	Mex (VD)	9983.16	9897.9559
7	Penthalaz	9458.12	9636.0535
8	Penthaz	9557.04	9825.9107
9	Sullens	10374.92	9924.5102
10	Vufflens-la-Ville	9421.00	9971.7072
11	Assens	10763.34	10953.1723
12	Bercher	10413.63	10655.9440
13	Bioley-Orjulaz	10432.68	10526.8663
14	Bottens	11705.89	11404.4981
15	Bretigny-sur-Morrens	11519.39	11430.8364
16	Cugy (VD)	11901.36	11436.0724
17	Dommartin	11519.02	11617.8423
18	Echallens	10583.54	10857.6636
19	Eclagnens	10281.59	10325.2212
20	Essertines-sur-Yverdon	10406.37	10680.8931
21	Etagnières	10732.18	10760.8724
22	Fey	10626.72	10719.9970
23	Froideville	12821.19	12162.2771

...





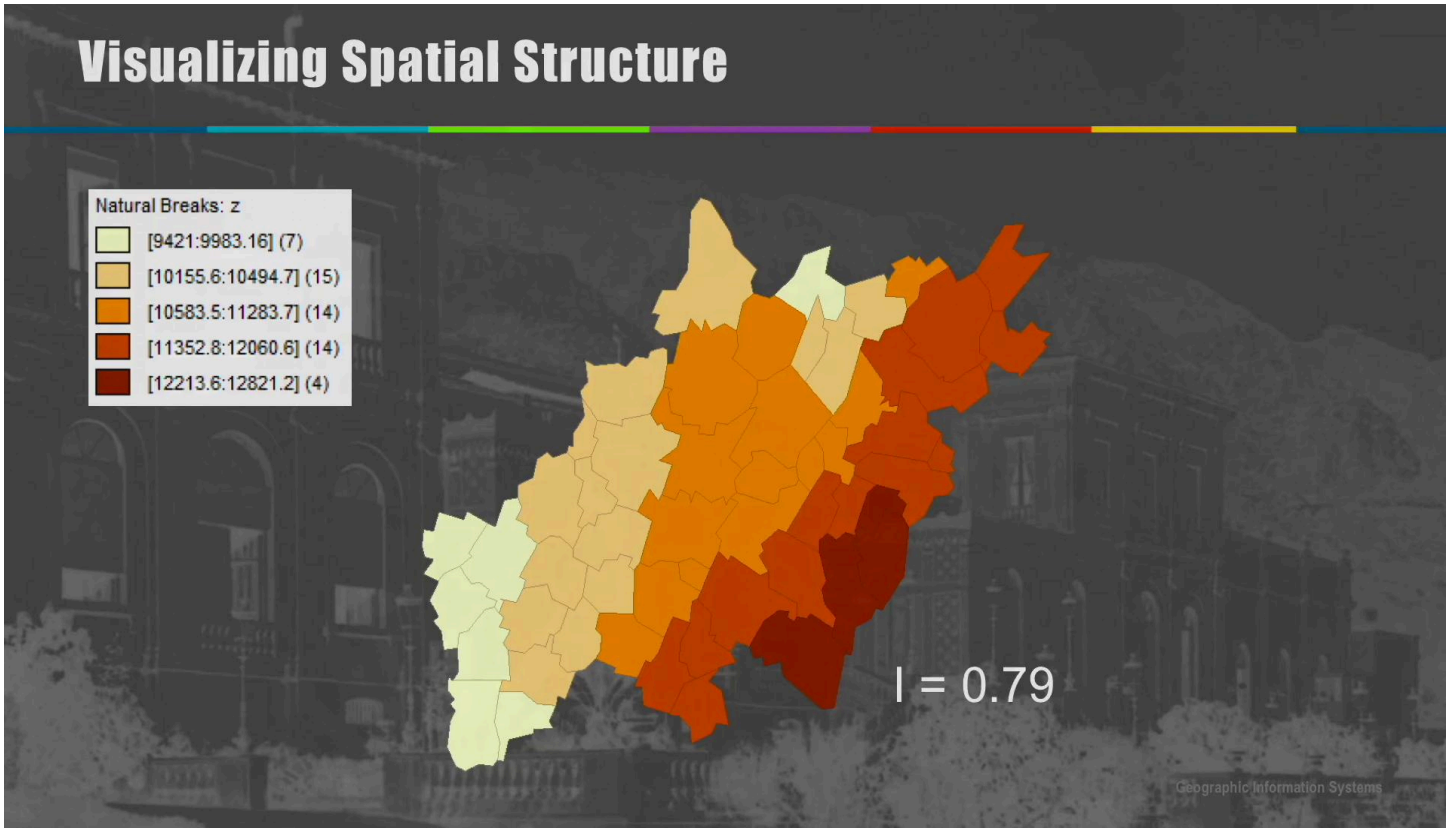
Geographic Information Systems

By applying a regression of the independent variable Z on the dependent variable Z bar the slope of the regression line is equal to the I of Moran. By means of a bivariate scatter plot, this interpretation provides a means of visualizing the linear relation between the studied variable and the average value of this variable in its neighborhood. Here, the beta parameter, so the slope, is equal to 0.79. The scatter plot of Moran shows the same relation but on the centered and reduced values. This high I of Moran shows that the precipitation values on the districts of this territory resemble the average of their neighbors and that there is therefore a spatial auto-correlation, that the precipitation variable used is therefore spatially dependent.

Notes

Summary

11m 23s



This thematic map of the precipitation variable in 5 places highlights the spatial structure revealed by the I of Moran. Indeed, the intensity of precipitations declines progressively from East to West.

Notes

Summary

12m 12s

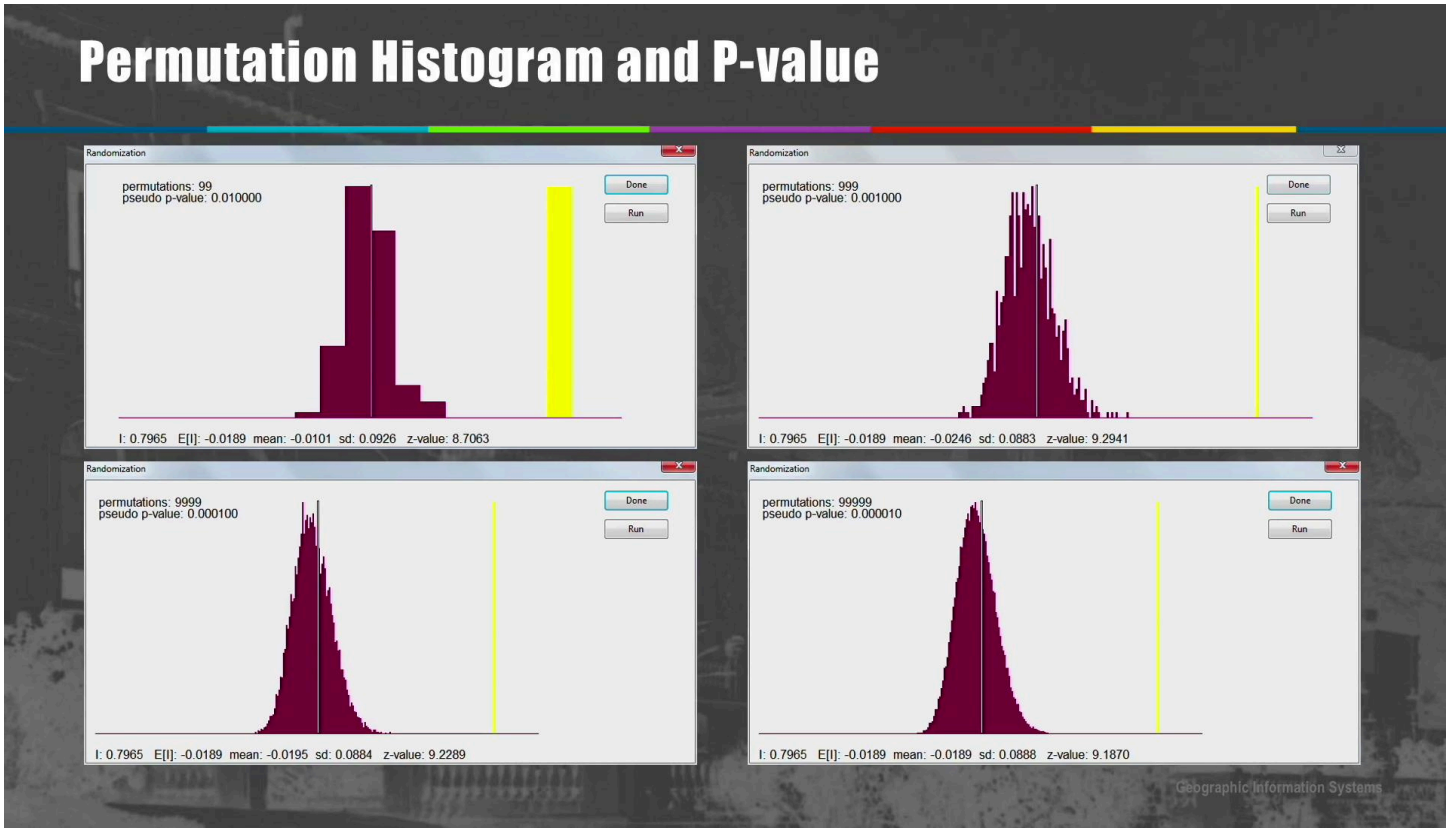
Significance of Moran's I and Random Permutations

But is this spatial auto-correlation value statistically significant? Indeed, to what extent is the spatial arrangement of the precipitation variable not due to chance?

Notes

Summary

12m 34s

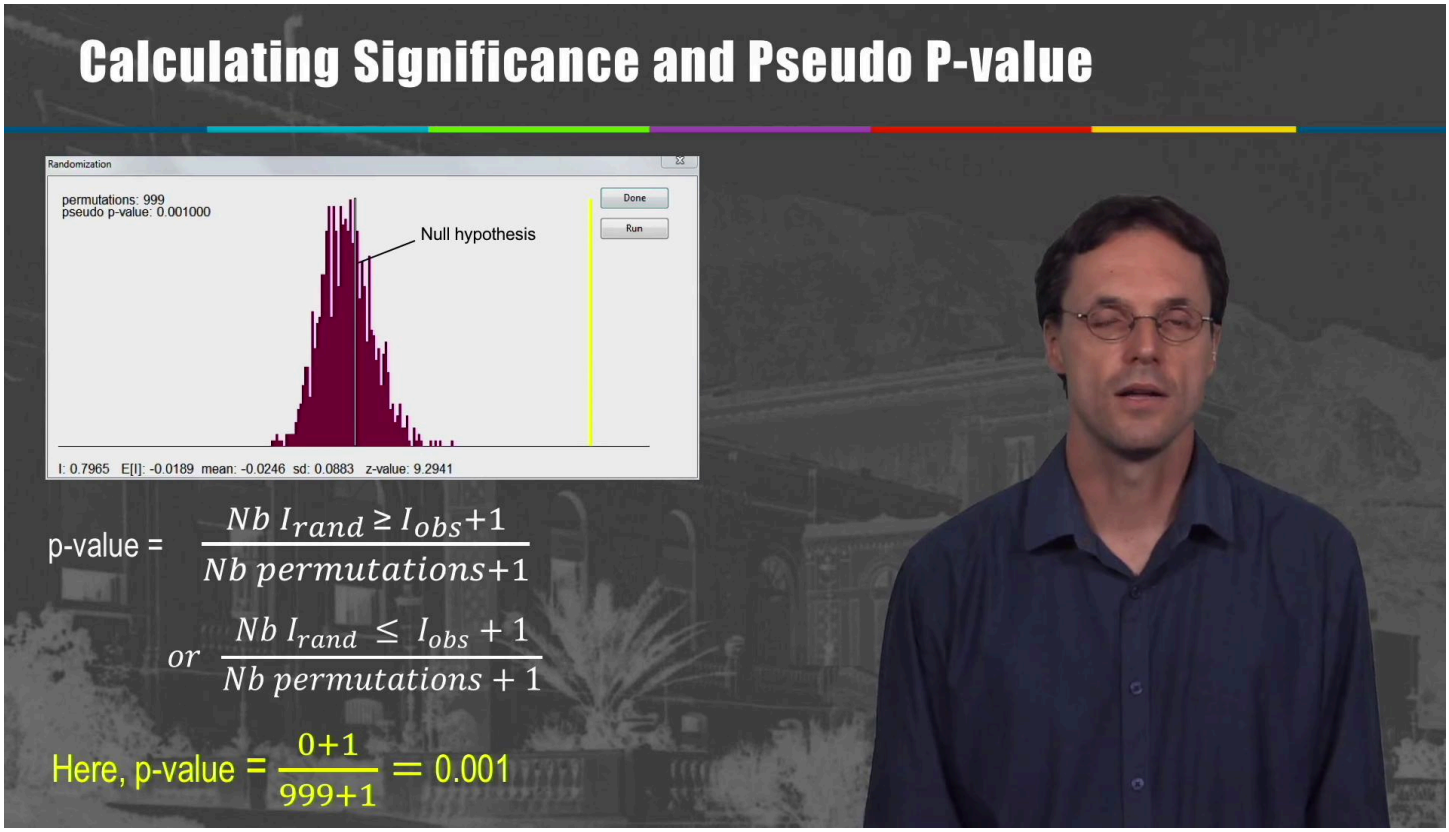


How does the situation observed, so the spatial structure revealed by the I of Moran of 0.79, behave if we compare it, if not to all possible configurations, in any case to a large part of them? By possible configuration, we think about changing the values randomly, we have used here a fictitious variable with values between 1 and 54 to illustrate the method, between all the other possible locations in the analyzed dataset. Each configuration corresponds to a random draw. And we proceed to several thousand of permutations in the case of the Monte Carlo method. with each draw, we calculate the I of Moran which we compared with the I of Moran of the observed situation. After 999 prints, we have 999 configurations and as many I of Moran values to compare with the observed situation, which allows to know if the latter resembles the random configurations or if, on the contrary, it is clearly different. The Geoda software will store the I of Moran corresponding to all the random configurations generated and use them to construct a histogram. On the basis of actual precipitation data per district, here is the histogram generated on the basis of 99 prints, then 999, of 9,999 and finally of 99,999 random draws out of a possible total of 54 factorials.

Notes

Summary

12m 47s



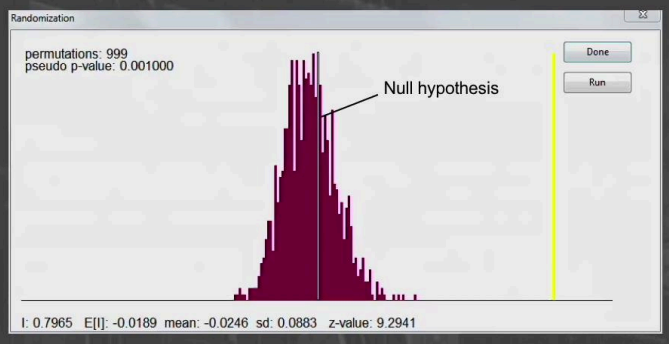
The higher the number of prints, the more normal the distribution appears and the more the standard deviation and the mean are supposed to approach their theoretical values. Let's now look in detail at the histogram generated on the basis of 999 permutations. The random distribution represents what can be called the neutral space. Around the average, which is shown here and which is close to 0, the precipitation values of the districts do not resemble the average of their neighbors. And we see that the value of the I of Moran for the situation observed here indicated by the yellow line, and this yellow line is not a histogram bar, is clearly distinguished from the rest of the distribution. It is therefore apparently significant. A non-significant value would appear as here in green, for example, so in the middle of the neutral distribution, where there is no spatial dependence. This significance is translated numerically by a probability of rejecting the null hypothesis. The p-value is a limit value of rejection of the null hypothesis. And the null hypothesis here, is that the value observed is the fact of chance and it resembles the other values generated by random permutation.

Notes

Summary

14m 14s

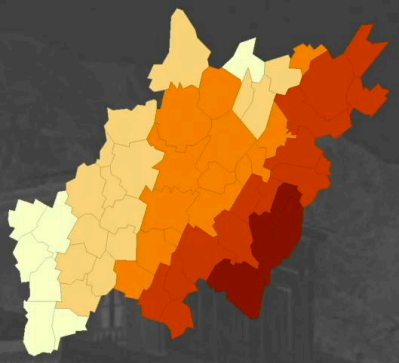
Calculating Significance and Pseudo P-value



$$\text{p-value} = \frac{Nb\ I_{rand} \geq I_{obs} + 1}{Nb\ permutations + 1}$$

$$or\ \frac{Nb\ I_{rand} \leq I_{obs} + 1}{Nb\ permutations + 1}$$

$$\text{Here, p-value} = \frac{0+1}{999+1} = 0.001$$



A Moran's I of 0.79 corresponds to a spatial structure that is significantly different from one that is randomly distributed

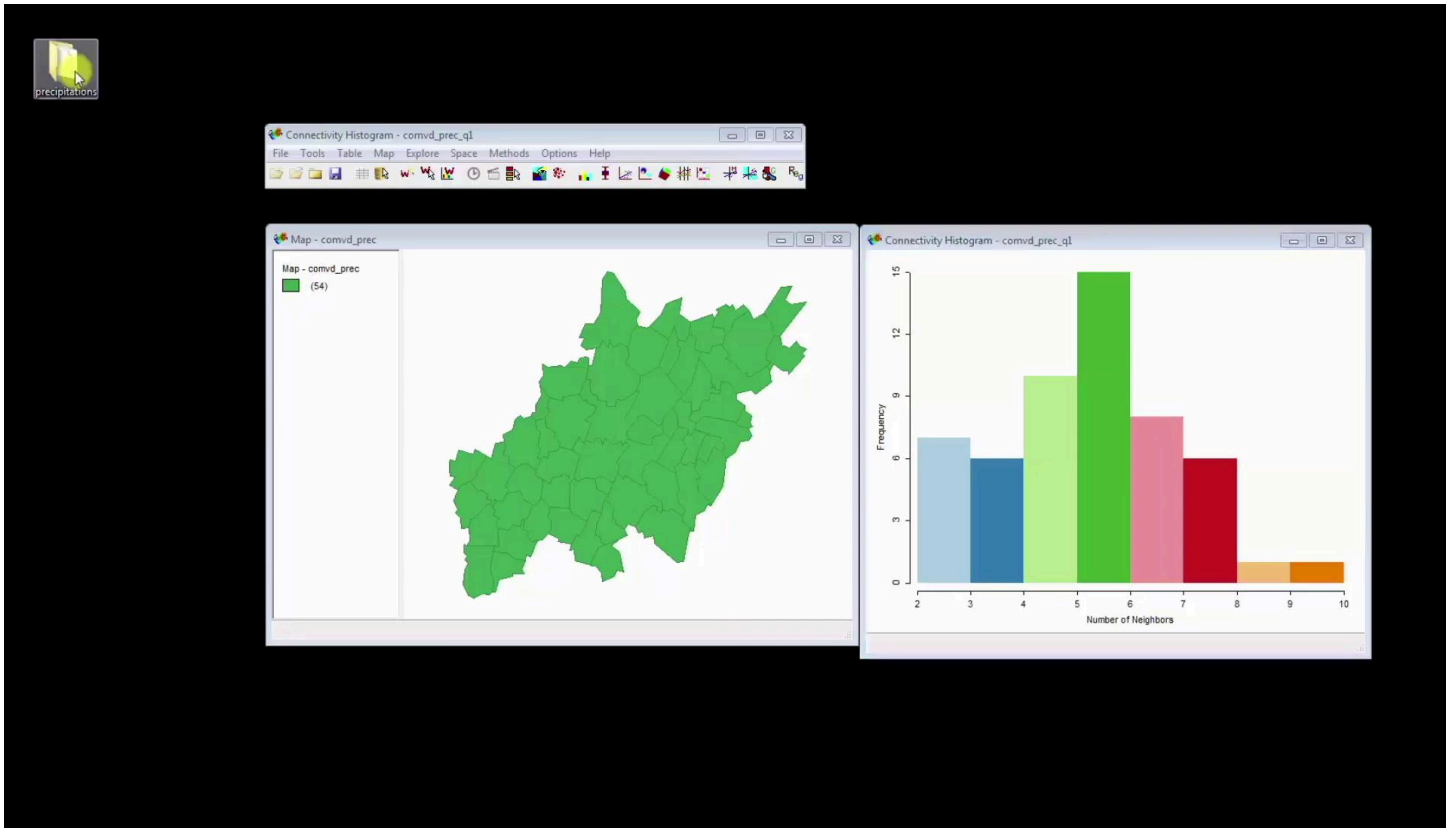
Geographic Information Systems

The lower the p-value, the weaker the risk of making a mistake by rejecting this null hypothesis, so, the higher the observed value is significantly different from a random distribution. This p-value, is the ratio between the randomly generated number of I of Moran which are larger or equal than the observed I of Moran plus 1 and the total number of random permutations plus 1. As the I of Moran can also be negative, the determination of the p-value will in this case be based on the number of randomly generated I of Moran which are smaller or equal than the observed I of Moran. We speak in the case of this test of pseudo p-value because the threshold of significance depends on the number of permutations. In the present case, we can conclude that the analyzed precipitation variable is spatially or globally significantly auto-correlated.

Notes

Summary

15m 26s

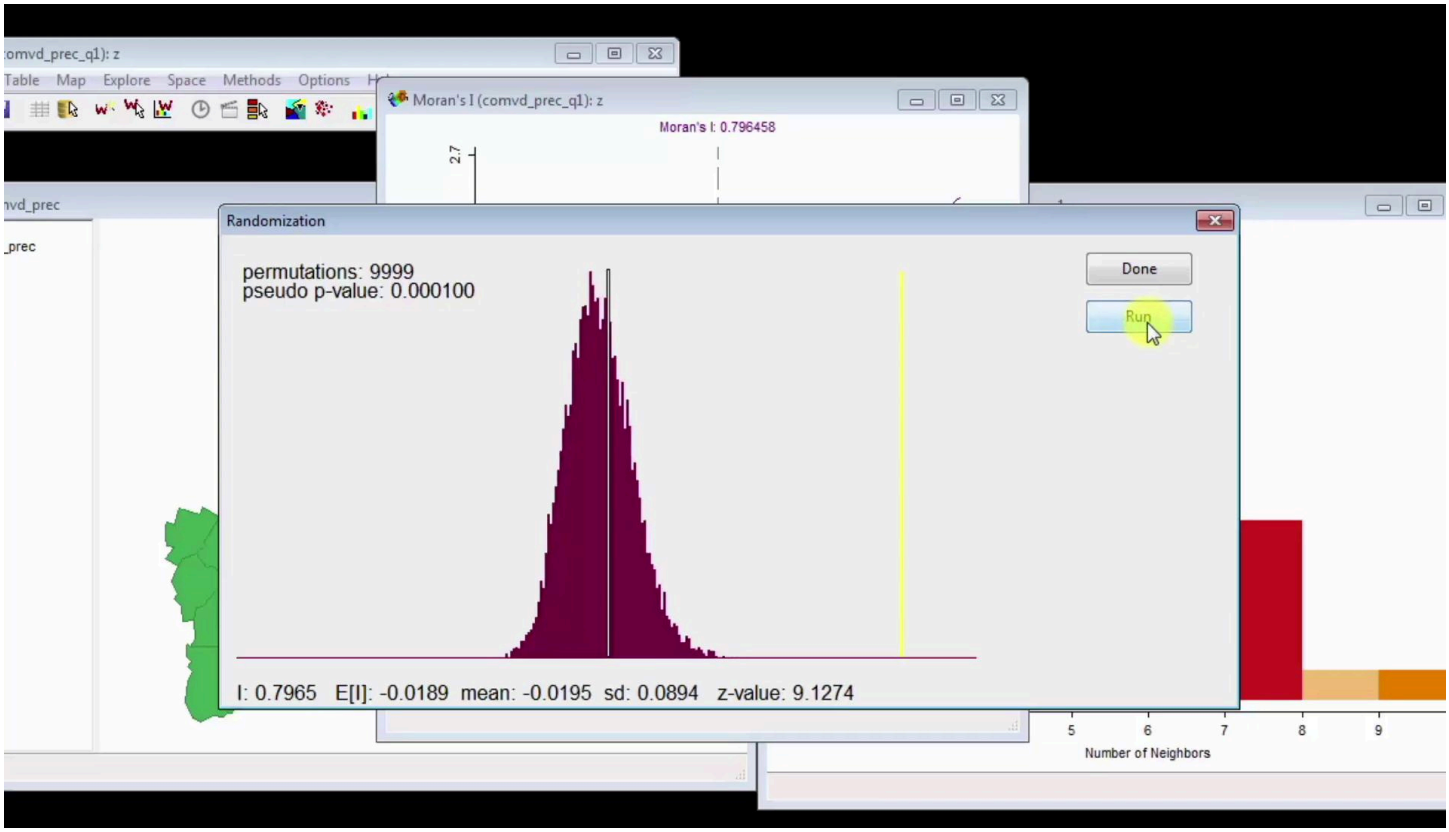


We will see now how to generate a spatial weighting file, to calculate the global I of Moran and to evaluate its significance with GeoDa. First, we need to open a file in shape format which contains the polygons of the districts and their attributes including our precipitation variable. Next, we have to create the spatial weighting file by clicking on the create weights button. And in the corresponding dialogue box, we must first select an unique identifier, here IDE, then select the desired weighting scheme, so a Queen contiguity criterion of 1. After clicking the create button, we have to choose where to store this weighting file and give it a name to identify it. Then we will inspect the connectivity. We will use the connectivity histogram button which allows us to inspect the connectivity histogram, so to give us an indication of the frequency of polygons having the number of neighbors indicated in the X axis. Thus, it is possible to highlight the best connected district here or the seven least connected districts which are located in the edges of the zone. Let's now take a look in the weighting file.

Notes

Summary

16m 31s

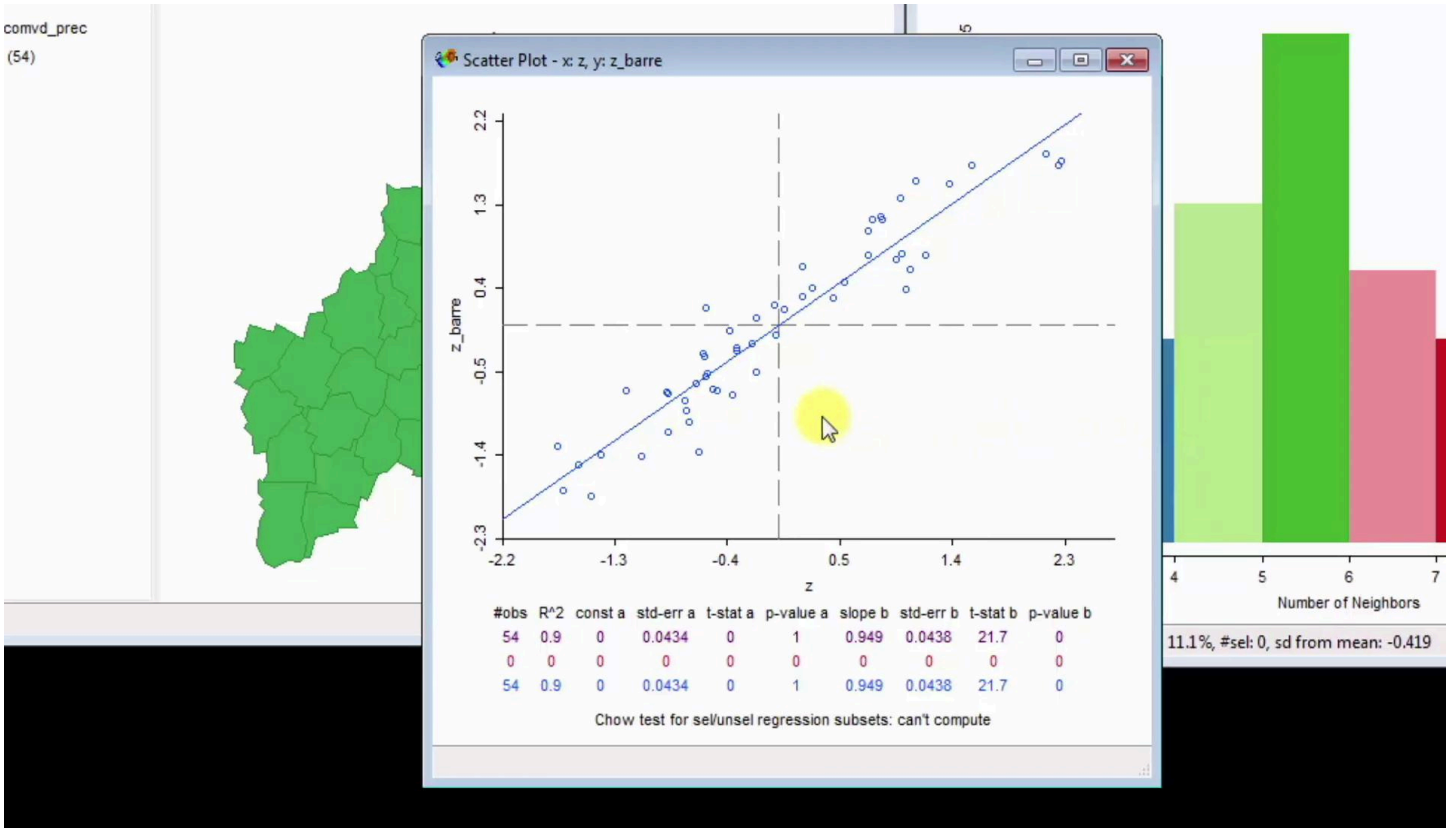


So this weighting file, we can open it with a text editor and we find the structure presented earlier with the header on the first line then the description of the neighborhood of the polygon 1 which has 6 neighbors, then the polygon 2, which counts 5, and the polygon 3, which also counts 5, and so on. We will now start calculating the I of Moran. We must select the variable of interest to begin with and it is Z which refers to the monthly precipitation. The scatter plot of Moran is generated immediately since we had already generated the weighting file earlier and it was loaded. It shows us the cloud of centered and reduced values and above an I of Moran of 0.79. Then, it is quickly possible to test the significance by right-clicking on the scatter plot and by clicking on randomization. And we will generate 9,999 random spatial configurations. The corresponding histogram shows us here a very small p-value which means that we are unlikely to be wrong by rejecting the null hypothesis, and that the spatial structure observed and characterized by an I of Moran of 0.79 is significantly different from a random spatial distribution.

Notes

Summary

18m 01s



Each time we click the run button, we regenerate a series of 9,999 random permutations with different statistics. But we realize that the pseudo p-value remains very small. Finally, we will calculate the I of Moran again, but according to Luc Anselin's interpretation, so to construct a linear regression between the precipitation variable Z on the X axis and the weighted precipitation variable Z bar on the Y axis. And we note that indeed the beta coefficient which gives the slope of the regression line is equal to 0.79 and that by standardizing the two distributions of precipitation expressed in tenth of millimeters of water, we reach the scatter plot of Moran obtained directly earlier.

Notes

Summary

19m 29s