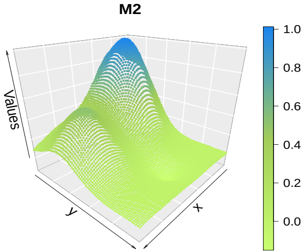
Statistical methods for spatial data analysis

Lecture 8: Areal data

Joaquin Cavieres

Geoinformatics, Bayreuth University



Outline

1. Introduction

2. Spatial neighbours
   - Neighbors based on contiguity
   - Neighbors based on k nearest neighbors
   - Neighbors based on distance
   - Neighborhood matrices

1. Introduction

Areal or lattice data arise when a fixed domain is partitioned into a finite number of subregions at which outcomes are aggregated. Examples of areal data are the number of cancer cases in counties, the number of road accidents in provinces, and the proportion of people living in poverty in census tracts.

Spatial data are often observed on polygon entities with defined boundaries (borders).

The borders of the polygon are defined by the researcher considering the problem under study, and it could consider administrative boundaries or maybe only arbitrary.

The observed data are frequently aggregations within the boundaries/borders, for example, population counts.

The areal entities may themselves constitute the units of observation, for example when studying local government behaviour where decisions are taken at the level of the entity, for example setting local tax rates.

In general, areal entities are aggregations, used commonly to consider measurements, like voting results at polling stations.

Areal entities also can be multiple geometrical entities, such as islands belonging to the same county; they may also surround other areal entities completely, and may contain holes, like lakes.

2. Spatial neighbours

The concept of spatial neighborhood is useful for the exploration of areal data to assess spatial autocorrelation and find out whether close areas have similar or dissimilar values. A very simple definition of this is:

Neighbors are areas that share a common border, perhaps a vertex.

From the previous definition we can build a spatial neighborhood matrix which will allow us to assess spatial autocorrelation. The elements of the spatial neighborhood matrix can be viewed as weights that spatially connect areas. The values of this matrix corresponding to close areas will have more weight than entries corresponding to areas that are farther apart.

2.1. Neighbors based on contiguity

Contiguity means that two spatial units share a common border of non-zero length. Operationally, we can further distinguish between a rook and a queen criterion of contiguity.

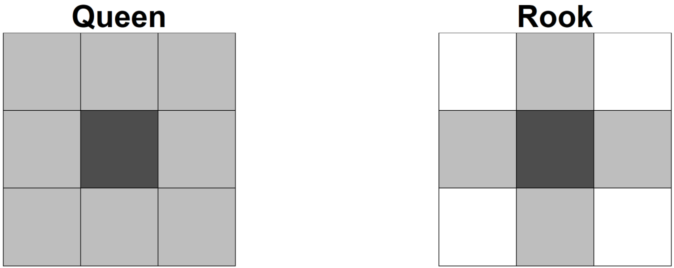
- Rook criterion: it define neighbors by the existence of a common edge between two spatial units
- Queen criterion: it is somewhat more encompassing and defines neighbors as spatial units sharing a common edge or a common vertex

The number of neighbors according to the queen criterion will always be at least as large as for the rook criterion.

In practice, to assess whether two polygons are contiguous requires the use of explicit spatial data structures to deal with the location and arrangement of the polygons.

If we create neighbors with the "sf" package, we will get a class of `sgbp` (sparse geometry binary predicate), which is similar to the standard `nb` class used in "spdep" package, but is not quite compatible. For th same reason, we have to convert from `sgbp` to `nb` (it is not too complicated).

It is important to keep in mind that the spatial weights are critically dependent on the quality of the spatial data source (GIS) from which they are constructed. Problems in the topology of the original file will result in inaccuracies for the neighbor relations included in the spatial weights.



Queen

Rook

Figure 1: Neighbors based on contiguity. Area of interest is represented in black and its neighbors in gray. Source [1]

See the example in `R`

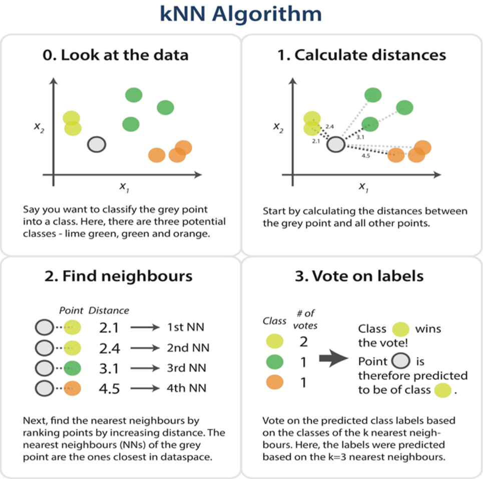
2.2. Neighbors based on k nearest neighbors

The k nearest neighbor algorithm (knn) is a non-parametric method (also known as supervised Machine Learning algorithm). It is used for both classification and regression, and predicts a target variable using one or multiple independent variables. The knn method stores all the available data and classifies a new data point based on the similarity.

Note: Non-parametric means that it does not make any assumptions on the underlying data distribution.

The algorithm works in the following way:

- Select the k value: number of nearest neighbors
- Calculate the Euclidean distance from k value to data points.
- Take the k nearest neighbors considering the calculated Euclidean distance.
- Among these k neighbors, count the number of the data points in each category.
- Classify the new data points to that category for which the number of the neighbor is maximum.



Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

Start by calculating the distances between the grey point and all other points.

Point	Distance	
○ ·· ●	2.1	→ 1st NN
○ ·· ●	2.4	→ 2nd NN
○ ·· ●	3.1	→ 3rd NN
○ ·· ●	4.5	→ 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

Class	# of votes
●	2
●	1
●	1

Class ● wins the vote! Point ○ is therefore predicted to be of class ● .

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Figure 2: Diagram source: Cambridge Coding (`https://cambridgecoding.wordpress.com/`).

See the example in R

2.3. Neighbors based on distance

Using the `dnearneigh()` function of the "spdep" package, we can calculate a list of of neighbors based on a distance between specific lower and upper bounds.

The more simple mathematical expression of this algorithm is using the Euclidean distance:

$$\text{D(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{1}$$

where D(x,y) is the Euclidean distance.

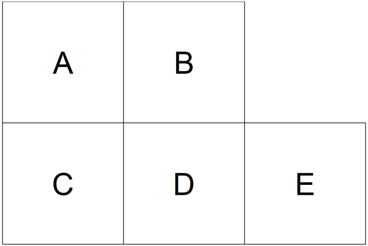
See the example in `R`

2.4. Neighborhood matrices

A spatial neighborhood matrix $\boldsymbol{W}$ defines a neighborhood structure over the entire study region, and its elements can be viewed as weights.

The (i, j)th element of a spatial neighborhood matrix W, denoted by $w_{i,j}$, spatially connects areas i and j in some way, where $i, j \in \{1, \ldots n\}$. The matrix W defines a neighborhood structure over the entire study region, and its elements can be viewed as weights.

If we are working with neighbors based on contiguity, we can build a binary matrix with $w_{ij} = 1$, if regions i and j share a common boundary, and $w_{ij} = 0$ otherwise.



	A	B	C	D	E	Sum
A	0	1	1	1	0	3
B	1	0	1	1	1	4
C	1	1	0	1	0	3
D	1	1	1	0	1	4
E	0	1	0	1	0	2

Figure 3: Left: Areas of the study region. Right: Spatial weight matrix calculated by assuming neighboring areas share a common boundary, and the sum of weights for each area. Source [1]

We can propose other definitions assuming that $w_{ij} = 1$ for all i and j whiting a specified distance, or, we can assume $w_{ij} = 1$ if j is one of the k nearest neighbors of i.

On the other hand, we can define the weights w_{ij} as the inverse distance between areas.

Considering the previous, we will build two types of neighborhood matrices:

1. Spatial weights matrix based on a binary neighbor list.
2. Spatial weights matrix based on inverse distance values.

See the example in `R`

Thanks!...

- Moraga, P., 2023. Spatial Statistics for Data Science: Theory and Practice with R
- Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. CRC Press.
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press.
- Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.