Session 1: Exploratory spatial analysis in R

Duncan Lee

University of Glasgow

Overview

This session will cover the following:

- Mapping data.
- Quantifying spatial closeness.
- Assessing the presence of spatial correlation.

What is spatial data?

Each unit of data has an associated geographical identifier such as a coordinate:

- Latitude & longitude.
- UK Ordnance Survey grid reference.
- Easting & northing.

The identifier may also indicate membership of a region, for example:

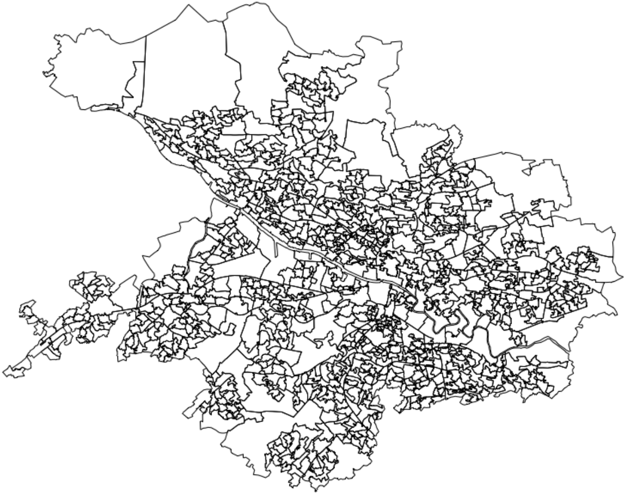
- Countries.
- UK Local health authorities.
- US Census tracts.
- Grid cells.

We focus on this second type of spatial data called **areal data**.

Spatial data analysis

- The data relate to K non-overlapping areal units, such as datazones.
- The visualisation of spatial data is typically in the form of a map.
- Spatial data typically display spatial correlation, with data points close together in space tending to have more similar values than data points further apart in space.
- This spatial correlation violates the assumption of independence in simple statistical models, requiring the need for more complex models.

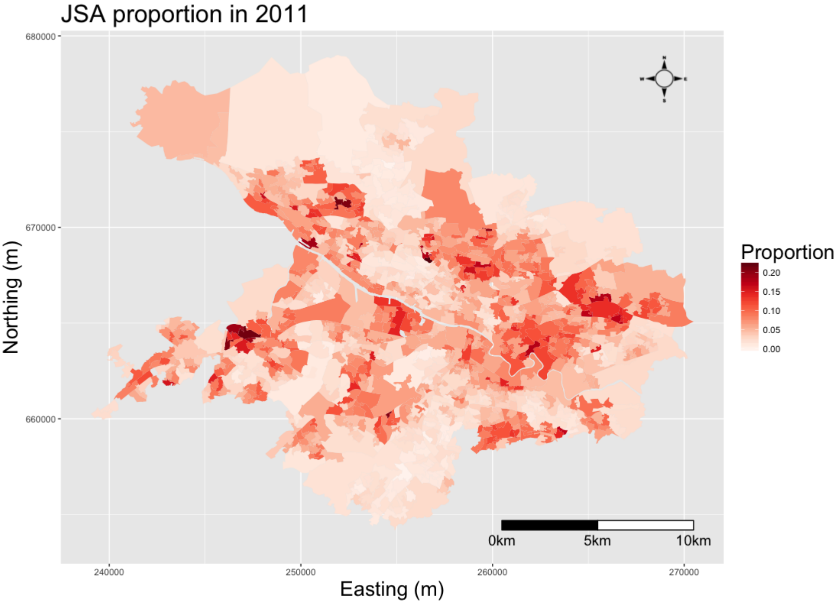
Example of datazones in Glasgow



What do I need to make a map?

- Geographically labelled data in `.xls`, `.txt`, `.csv` file.
- Access to computer software for mapping such as Geographic Information System (GIS) e.g. `MAPINFO`, `ARCGIS`, `QGIS`, or statistical software such as `R`.
- Access to underlying Geographic data e.g. standard boundary shapefiles such as `.shp`, `.dbf`, `etc`.

Example map - JSA in 2011



Defining spatial closeness

An important concept for defining and thus modelling spatial dependence in areal data is that of the neighbourhood or adjacency matrix, **W**, which is an $K \times K$ matrix that defines how the K areas (e.g. datazones, etc) are spatially located with respect to each other. The values in this matrix are typically binary.

- The ijth element $w_{ij} = 1$ if areas (i, j) are spatially close together, in which case they are said to be "neighbours".
- The ijth element $w_{ij} = 0$ if areas (i, j) are not spatially close together.

Always set $w_{ii} = 0$ as an area can't be a neighbour of itself.

Fife - One areal unit

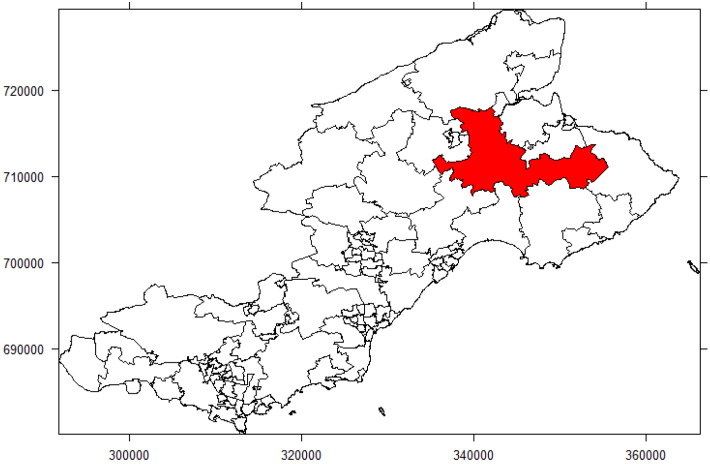


Figure: What are the neighbours of this region?

3 Common ways of specifying $\mathbf{W}$

There are 3 different approaches for specifying $\mathbf{W}$, which are that areas (i, j) are neighbours and hence $w_{ij} = 1$ if:

- they share a common border.
- their (population weighed) central points (centroids) are within a fixed distance d of each other.
- area i is one of the r closest areas to area j in terms of distance.

Otherwise $w_{ij} = 0$. The first of these is the most common as how to choose (d, r) is not clear.

Fife - Neighbours

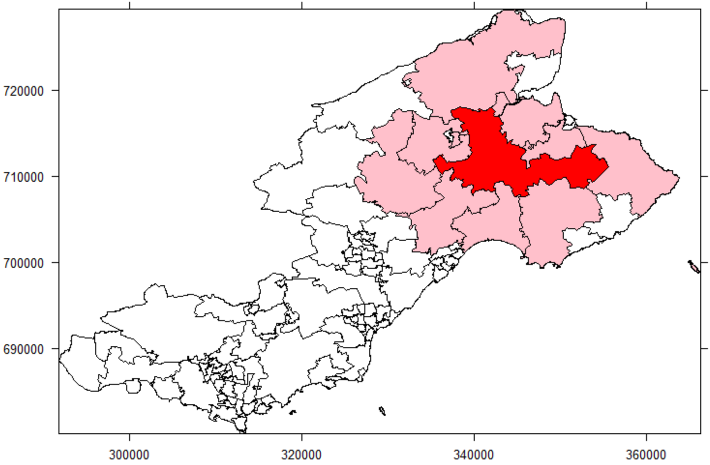


Figure: Neighbours share a common boundary

Implications

- There is not a lot of literature about choosing $\mathbf{W}$ in a model, typically people use the **sharing a common border** specification.
- The implication of choosing $\mathbf{W}$ is that if $w_{ij} = 1$ then data in areas (i, j) will be modelled as spatially correlated when we talk about modelling later on today, where as if $w_{ij} = 0$ they will be modelled as conditionally independent.
- All modelling results are thus dependent upon $\mathbf{W}$, although this is rarely stated explicitly.

Assessing if data are spatially correlated

- Standard regression models such as linear models, logistic regression models, etc assume that the errors (residuals) from the model are independent.
- This is typically unlikely in spatial areal unit data, where the residuals from any regression model are likely to be spatially correlated.
- Incorrectly assuming independence when it is not true will result in 95% uncertainty intervals that are too narrow.
- Thus we need a statistic for measuring the extent of the spatial correlation in a data set.

Pearson's correlation coefficient

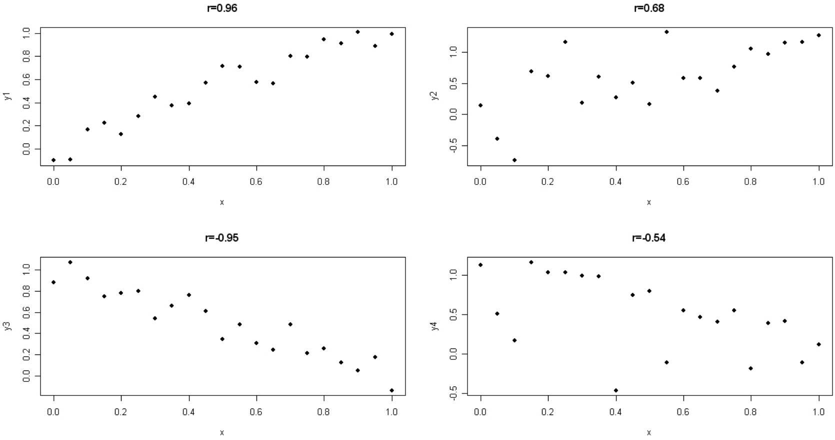
If we have two sets of non-spatial data, $\mathbf{x} = (x_1, \ldots, x_K)$ and $\mathbf{y} = (y_1, \ldots, y_K)$ then Pearson's correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{K}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{K}(x_i - \bar{x})^2 \sum_{i=1}^{K}(y_i - \bar{y})^2}}$$

where $(\bar{x}, \bar{y})$ are the means of $(\mathbf{x}, \mathbf{y})$ and

- Positive values indicate positive correlation (as x_i increases so does y_i).
- Negative values indicate negative correlation (as x_i increases then y_i decreases).
- Close to zero values indicate no correlation.

Examples



Moran's I statistic

In the spatial case we only have one data set, say $\mathbf{y} = (y_1, \ldots, y_K)$ measured at K locations, and we want to know if y_i is correlated with itself at nearby locations. Thus we alter Pearson's correlation coefficient to get Moran's I statistic as follows:

$$I = \frac{K \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij}) \sum_{i=1}^{K} (y_i - \bar{y})^2}.$$

As before positive values represent positive spatial correlation (the closer two data points are the more similar their values will be, while a value close to zero represents independence.

Spatial correlation is quantified by the top part of Moran's I, namely:

$$\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij}(y_i - \bar{y})(y_j - \bar{y})$$

- If the data are positively spatially correlated then this quantity will have positive values, because spatially close data points (y_i, y_j) (where $w_{ij} = 1$) will both be either above or below the mean (they will be similar).
- In contrast, under independence then (y_i, y_j) could be similar (both above or both below the mean) yielding a positive contribution to the above or very different (one above and one below the mean), yielding a negative contribution to the above. Thus overall the above sum will be close to zero.

Values for Moran's I

In theory Moran's I takes the same set of values as any correlation coefficient, namely the interval between -1 and 1, where:

- $I = -1$ strong negative spatial correlation - data points close together in space have very different values.
- $I = 0$: Independence - no spatial correlation.
- $I = 1$: strong positive spatial correlation - data points close together in space have very similar values.

However, Moran's I values above 0.5 are relatively rare, so a value of 0.2 would indicate positive spatial correlation.

Assessing significant spatial correlation

The significance of the spatial correlation can be assessed by a statistical hypothesis test. The hypotheses for this test are:

$$H_0 - \text{no spatial association}$$

$$H_1 - \text{some spatial association}$$

- Under the alternative hypothesis H_1 it could be that the spatial autocorrelation is positive $(I > 0)$ or negative $(I < 0)$. But that will be obvious from the value of Moran's I.
- The test statistic for this test is Moran's I statistic.

Calculating the p-value against independence

- The p-value is computed via a permutation testing idea.
- First compute Moran's I statistic for the data.
- Then randomly rearrange the data values to areas and re-compute Moran's I for the random arrangement. This should give a value close to zero.
- Repeat the process say 1,000 times and you have 1,000 values of Moran's I statistics generated under independence (no spatial correlation).
- Then the p-value essentially quantifies how extreme the observed value of Moran's I from the real data is compared to the 1,000 values generated under independence from the randomly permuted data.

The test can be implemented in R using the `moran.mc()` function.