Modelling spatial data in R with CARBayes

Part 1: Introduction and exploratory analysis

Duncan Lee and Eilidh Jack

GEOMED 2017

Overview

- Introduction and motivation.
- Mapping spatial data.
- Data types.
- Defining spatial Closeness.
- Quantifying spatial Correlation.

What are spatial data?

Usually, 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.

Workshop looks at the second type of spatial data - areal data.

Motivation for modelling spatial data

- Visualise the spatial data on a map.
- Infer the underlying spatial pattern given a set of noisy spatial data.
- Ecological regression - what effect does a risk factor have on disease
- Risk estimation - which areas exhibit high-risks for a disease.

Example - Uptake of MMR vaccine

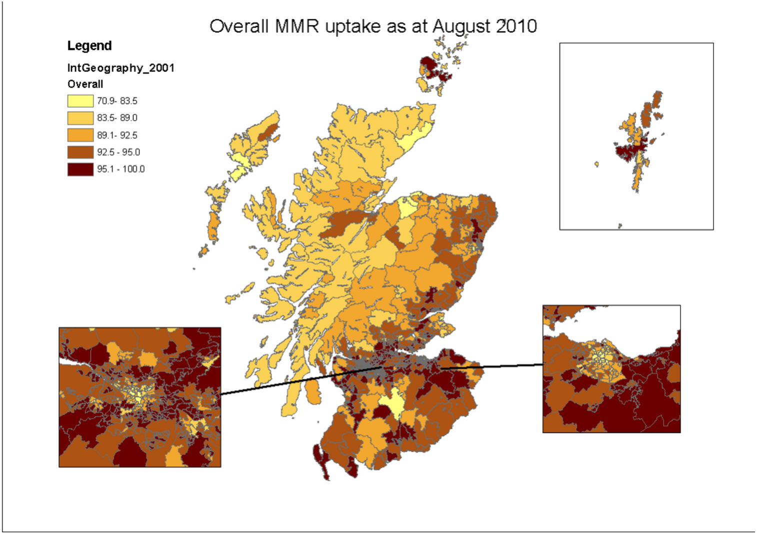


Figure: Percentage of children in each intermediate zone who receive the MMR vaccine by age 2

What do I need to make a map?

- Geographically labelled data such as:
  - Counts of disease incidence or prevalence.
  - Rates of disease incidence or prevalence.
- Shapefiles (`.shp`, `.dbf`, etc) giving the spatial outlines of the areas the data relate to.
- Geographic Information System (GIS) software such as MAPINFO, ARCGIS, QGIS, R.

Software

- Here we illustrate how to produce maps in `R`, so that you can perform mapping and modelling of the data in a single software environment.

- There are many different packages within `R` that can draw spatial maps. For example:
  - `spplot()` in the `sp` package.
  - `ggplot()` in the `ggplot2` package.

- Here we illustrate `ggplot()` because I think it produces nicer looking maps!

- It also allows one to overlay a map onto a Google map via the `ggmap` package.

Data types and models

Types of data measured

- continuous measurements - blood pressure.
- counts - number of hospital admissions.

`CARBayes` **can fit 3 main data models.**

- Binomial with logistic link function.
- Gaussian with identity link function.
- Poisson with log link function.

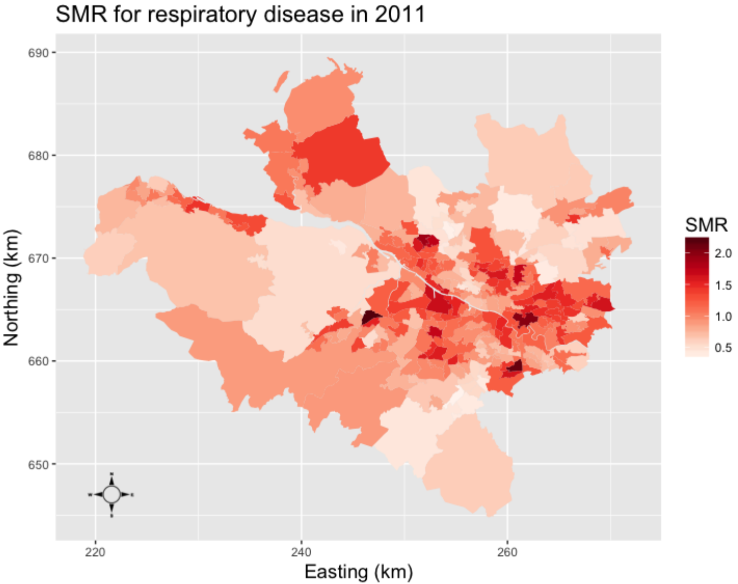
In this workshop we focus on count data and Poisson models.

Count data

- We have data for $k = 1, \ldots, K$ non-overlapping areal units.
- For the kth areal unit we have data (Y_k, E_k), which denote the observed and expected numbers of disease cases in area k.
- The expected number of disease cases E_k controls for population sizes and demographic structures, and is computed via indirect standardisation.
- The simplest measure of disease risk is the standardised morbidity (morality) ratio (SMR) which for area k is computed as:

$$\text{SMR}_k = \frac{Y_k}{E_k}.$$

Example - respiratory hospitalisation in Glasgow



Defining spatial closeness

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

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

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

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

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

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

Otherwise $w_{kj} = 0$. The first of these is the most common as how to choose d in the latter two cases is not clear.

Example - Fife

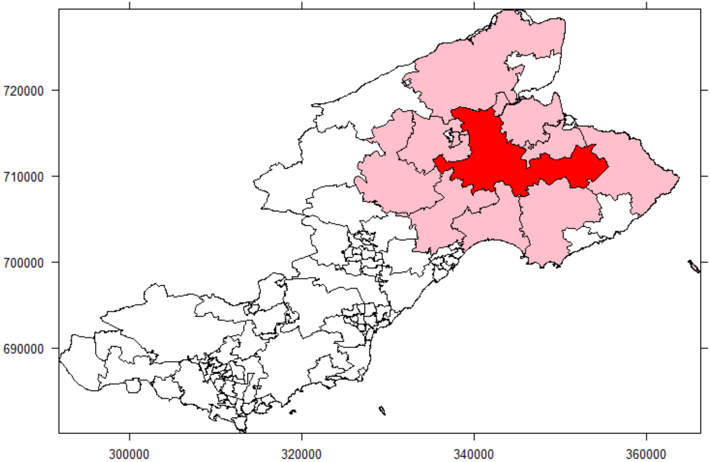


Figure: Neighbours share a common boundary

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 $w_{kj} = 1$ then data in areas (k, j) will be modelled as spatially correlated, where as if $w_{kj} = 0$ 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.

Moran's I statistic

The data on disease are denoted by $\mathbf{y} = (y_1, \ldots, y_K)$ measured at K locations, which could be the SMR or residuals from a model. We want to know if y_k is correlated with itself at nearby locations. Moran's I statistic is:

$$I = \frac{K \sum_{k=1}^{K} \sum_{j=1}^{K} w_{kj}(y_k - \bar{y})(y_j - \bar{y})}{(\sum_{k=1}^{K} \sum_{j=1}^{K} w_{kj}) \sum_{k=1}^{K}(y_k - \bar{y})^2}.$$

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:

$$\sum_{k=1}^{K} \sum_{j=1}^{K} w_{kj}(y_k - \bar{y})(y_j - \bar{y})$$

- If the data are positively spatially correlated then this quantity will have positive values, because spatially close data points (y_k, y_j) (where $w_{kj} = 1$) will both be either above or below the mean (they will be similar).

- In contrast, under independence then (y_k, 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.

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

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

- The test statistic for this test is Moran's I statistic.
- The p-value is computed via a permutation testing idea.
- The test can be implemented in `R` using the `moran.mc()` function.