Week 6 wrap-up

Bayesian modelling for spatial and spatio-temporal data

MSc in Epidemiology

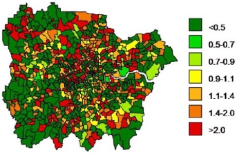
Week 6

- This week we introduced the field of spatial analysis, which is based on the non-independence of the observations.

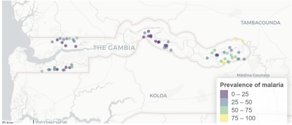
- We can use spatial analysis to solve a number of problems, when the data are characterized by geographical attributes.

- We started by classifying data into one of three basic types (based on the domain $\mathcal{D}$):
  - areal data,
  - point-referenced (or geostatistical) data,
  - random point pattern data.

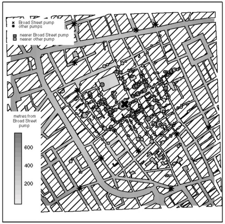
Type of spatial data



(a) **Areal data**
Map of SMR of leukaemia in children 0-15 yrs, 1985-96 for 872 wards in Greater London (Source: Best et al, 1991)



(b) **Geostatistical data**
Prevalence of malaria among a sample of village resident children in the Gambia (65 villages) (Source: Diggle et al, 2002)



(c) **Point pattern data**
The 1854 London cholera outbreak near Golden Square in London (source: Bivand et al, Applied Spatial Data Analysis with R, Springer, 2013)

Coordinate Reference System (CRS)

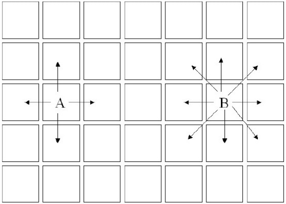
- We introduced the concept of Coordinate Reference System (CRS), which defines how the spatial elements of the data relate to the surface of the Earth.
- We introduced the difference between Geographic coordinate systems (GCS) and Projected coordinate systems (PCS):
  - The GCS identify any location on the Earth's surface using longitude and latitude, with units in decimal degrees or degrees.
  - The PCS provide mechanisms to project maps of the Earth's ellipsoid shape onto a two-dimensional Cartesian coordinate plan.
- Because the projections are a way to convert the Globe's curved surface into a two-dimensional plane, we introduce distorsion.

Spatial autocorrelation

- We focused on the key geographic concept of spatial autocorrelation, which implies that observations from units closer to each other are more similar than those recorded in units farther away.
- It quantifies the degree of which observations, at spatial locations, are similar to nearby observations.
- The detection of spatial autocorrelation is useful in spatial analysis for identifying underlying data structures, the degree of spatial randomness, or clustering in the data.
- We saw how to construct spatial proximity weights, that is weights that depend on the underline geography of the region.

How we define spatial neighbours?

- There are different ways to build the weights, according to the spatial neighborhood definition. We can define as neighbors areas that are adjacent according to rook criterion (A), or queen criterion (B):



In its simple form, $w_{ij} = 1$ if areas i and j are adjacent, 0 otherwise.

- We can expand the idea of neighborhood to include areas that are close, according to other definitions of proximity. Thus, we may take $w_{ij} = 1$ for all i and area j within a specified distance, or for a given area i, we may take $w_{ij} = 1$ if j is one of the k nearest neighbors of i.

Measuring spatial autocorrelation

- A number of approaches have been suggested for measuring spatial autocorrelation.
- Global measures of spatial autocorrelation share a common structure: calculate the similarity of values at locations *i* and *j*, then weight the similarity by the proximity of locations *i* and *j*.
- Null hypothesis: spatial randomness.
- We saw an example of global measure given by the Moran's I test, that can be applied to the data directly, or to the residuals from a regression model.

Starting a disease mapping analysis

- Collect (age and/or sex stratum-specific) morbidity or mortality data for each area i, $i = 1, \ldots, N$, and population data
- Compute the expected cases (i.e. number of disease cases would be expected if the study population had the stratum distribution of a reference population)
- Compute the SMR or SIR for each area i as:

$$SMR_i \text{ or } SIR_i = \frac{\text{Observed nb. cases in area } i}{\text{Expected nb. cases in area } i}$$

  - $SMR_i = 1$: the area i has equal number of observed and expected cases (equal i has the same risk that the standard population)
  - $SMR_i < 1$: the area i has less observed cases than expected cases (area i has lower risk than the standard population)
  - $SMR_i > 1$: the area i has more observed cases than expected cases (area i has higher risk than the standard population)

Problem with mapping the SMRs or SIRs

- A map of raw SMRs or SIRs can be unstable or misleading, particularly if the estimates are based on populations of very different sizes. In fact, since the variability in the estimated local risks depends on population size, some risks may be better estimated than others, and this may obscure spatial patterns in disease risk.
- For example, SMRs or SIRs based on small populations or on small numbers of disease cases (e.g. rare diseases) are very imprecise, as areas with small expected number of cases have high associated variance.
- They ignore possible spatial correlation between disease risk in nearby areas, due to possible dependence on spatially varying risk factors.
- They do not allow the inclusion of ecological covariates.

Smoothing

- Smoothed estimates of the RRs provide more robust estimates.

- The basic idea is to borrow information from neighbouring areas to produce a better (i.e. more stable and less noisy) estimate of the risk associated with each area and thus separate out the signal from the noise.

- This week we study a non-spatial smoothing of the estimates of RRs.

The Poisson log-Normal model (non-spatial smoothing)

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log(\lambda_i) &= \alpha + \theta_i \\ \theta_i &\sim \text{Normal}(0, \sigma_\theta^2) \\ \alpha &\sim \text{Normal}(0, 10^4) \\ 1/\sigma_\theta^2 &\sim \text{Gamma}(1, 0.01) \end{aligned}$$

- O_i and E_i are observed and expected number of cases in area i
- λ_i is the unknown area-specific relative risk of morbidity or mortality from the disease
- Parameter α is mean log relative risk, i.e. overall risk in the region of study
- Parameters θ_i: **area-specific random effects** to capture region-wide heterogeneity. They take into account the extra-Poisson variability, i.e. overdispersion, in the log-relative risks
- σ_θ^2: controls the magnitude of the θ_i

Disease mapping in action: our labs

- Tutorial 6.1 - Disease mapping study of COVID-19 mortality in England, March-July 2020
- Practical 6.1 - Disease mapping study of larynx cancer in West Yorkshire
- Practical 6.2 - Perform indirect standardization in R