Introduction to Spatial Statistics – overview

• Course aims
  – What probability and statistics do we need?
  – What about mathematics?
• Spatial analyses and spatial processes
• The three types of spatial process
  1. Geostatistical processes
  – Example 1: Monthly temperature over the UK
  2. Areal processes
  – Example 2: Scottish Lip Cancer dataset
  3. Point processes
  – Example 3: Biological cell locations
  – Example 4: Student seating locations
• Objectives of a spatial analysis
Course aims

• To introduce the **three main types of spatial data**, and discuss a range of descriptive methods for identifying trends and spatial correlation.

• To describe a number of **statistical models** for representing **spatial correlation**.
  
  – To describe the method of **Kriging**, a commonly used tool for **spatial prediction**.

• To show how to **apply the techniques** from the course to **real spatial datasets** in the statistical package R.

What probability and statistics do we need?

A knowledge of Probability, Statistical Inference, Regression models, and R (at the masters level). In particular:

• We need the concepts of random variables (RVs) and vectors, along with their expectations and variances.

• We need to manipulate covariances and correlations between random variables.

• We need to write down conditional distributions.

• We use normal (univariate and multivariate), chisquared, $t$, and $F$ distributions.

• We will use likelihood methods.

• We need to be able to build confidence intervals and carry out hypothesis tests.

• A knowledge of how to fit and interpret regression models with independent errors.

• We will spend a lot of time thinking about building statistical models that can help us to investigate scientific questions.
What about mathematics?

This course also requires knowledge of linear algebra and calculus at the level of 2nd year Mathematics.

• **Linear algebra**
  - We need vectors and matrices.
  - We need to be able to add, subtract, and multiply vectors and matrices (Sometimes it helps to think of these operations in terms of the individual elements).
  - We need the transpose of a vector or matrix.
  - We need to know about the rank of a matrix.
  - We need to invert matrices (when possible).

• **Calculus** (For prediction, least squares, and maximum likelihood)
  - We have to take partial derivatives, set equal to zero, and solve to maximize or minimize a quantity.
  - We can check it is the maximum or minimum using the second derivative.
  - We need the chain rule for derivatives.

Spatial analyses and spatial processes

• A **spatial analysis** involves studying and modeling the dependence of data, collected in space.
  - We move away from processes being independent and identically distributed (IID).

• Depending on the form of the spatial process, the spatial data are typically **indexed** by locations/points or regions.
  - The locations or observed values may be random.

• We imagine the spatial data to be a realisation of a stochastic process.
  - A **stochastic process** is a family of random variables, 
    \[ \{Z(s) : s \in D\}, \]
    indexed by \( s \in D \), defined on a probability space.
  - The **spatial domain** \( D \) of the stochastic process defines the type of spatial process.
The three types

1. **Geostatistical processes:**
   A stochastic process that is defined continuously over a spatial domain.

2. **Areal processes:**
   A stochastic process that is defined on a finite collection of regions in a spatial domain.

3. **Point (pattern) processes:**
   A stochastic process where the points themselves are random.

Geostatistical processes

(Also known as a point-referenced process).

- A **geostatistical process** is the stochastic process
  \[ \{ Z(s) : s \in D \}, \]
  where \( D \) is a fixed subset of the \( p \)-dimensional space \( \mathbb{R}^p \).
- Common choices are \( p = 2 \) or \( p = 3 \).
- The location \( s \) varies **continuously** over \( D \).
- Typically we observe the geostatistical process at a **finite** number of locations.
  - So, why allow the process to vary continuously over a spatial domain?
  - Answer:
Example 1: Monthly temperature over the UK

- We study the monthly mean temperature for June 2012, at 29 locations on the UK mainland.

(Data source: http://www.metoffice.gov.uk/climate/uk/stationdata/)

- Questions of interest:
Areal processes

- Let
  \[ \{B_i : i = 1, \ldots, m\} \]
  denote a partition of \( m \) distinct regions, such that \( \bigcup_{i=1}^{m} B_i = D \) and \( B_i \cap B_j = \emptyset \) for each \( i \neq j \).

- The regions \( B_i \) can be regular or irregular.

- An areal process is the stochastic process
  \[ \{Z(B_i) : i = 1, \ldots, m\}. \]

- Equivalently, we could represent this process as
  \[ \{Z(s_i) : i = 1, \ldots, m\}, \]
  where each \( s_i \in D \) is a location representative of the region \( B_i \).

  Often the centroid of \( B_i \) is used for \( s_i \).

- Areal processes are also known as lattice processes – this is a misnomer as it implies a regular structure to the spatial data.

Example 2: Scottish Lip Cancer dataset

(Based on a description by Waller and Gotway)

- Clayton and Kaldor [1987] report the number of lip cancer cases registered from 1975-1986 for 56 districts in Scotland (“the observed”), classified according to a collection of political regions that were defined in 1973.

  - These districts are no longer in use.

- In addition they report the estimated expected number of cases per district accounting for different age distributions in each district (“the expected”).

- They calculate the standardized morbidity rate (SMR), which assuming that the expected numbers are constant, is defined as the observed divided by the expected.

- Additional covariate: the percentage of the workforce in each district employed by agriculture, fishing and forestry.
Example 2: Scottish Lip Cancer dataset, cont.

- Questions of interest:

Spatial maps of the observed and expected cases

- Discussion:
Spatial maps of the SMR and % in agg/fish/forest

<table>
<thead>
<tr>
<th>SMR</th>
<th>% employed in agg/fish/forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,2)</td>
<td>[-2.4,0.5)</td>
</tr>
<tr>
<td>(2,4)</td>
<td>[0.5,4)</td>
</tr>
<tr>
<td>(4,6)</td>
<td>[4,8.5)</td>
</tr>
<tr>
<td>(6,8]</td>
<td>[8.5,13)</td>
</tr>
<tr>
<td>(13,20)</td>
<td>[13,20)</td>
</tr>
<tr>
<td>(20,26.4]</td>
<td>[20,26.4]</td>
</tr>
</tbody>
</table>

• Discussion:

Point processes

• A point process is the stochastic process \( \{Z(s) : s \in D\} \), where \( D \) is a random set in \( \mathbb{R}^p \).

• It sufficient to let \( Z(s) = 1 \) at all \( s \).

• Thus point processes are quite different in nature from geostatistical and areal processes.
  – The interest here is in studying where (and possibly how) the points lie, not in learning about the value of the process itself.

• More complicated version of this process:
  – A marked point process relaxes the assumption that \( Z(s) = 1 \), allowing the stochastic process to take on random values, in addition to \( D \) being random.
**Example 3: Biological cell locations**

- Crick and Lawrence [1975] recorded the locations of the centres of 42 biological cells observed under optical microscopy in a histological section. The microscope field-of-view has been rescaled to the unit square.
- The data were analysed in Ripley [1977], Ripley [1981], and Diggle [1983].

**Discussion:**

**Example 4: Student seating locations**

- Discussion:
Objectives of a spatial analysis

1. Analysis

• To find a statistical model that adequately explains the dependence observed in spatial data.

  Ex 1: For the UK monthly mean temperatures, we could find a statistical model that accounts for the spatial trends (smooth changes in the mean), but also accounts for the fact that temperatures of locations nearby are more similar than those far apart (there is spatial dependence).

  Ex 2: For the biological cell locations example, we can estimate the intensity of the process in space. A region with a higher intensity will tend to have more spatial points in it.

• Model can be used for further statistical inference.

  Ex: In June 2012, was there evidence of decrease in temperatures in the North of Scotland, relative to the South of England?

• “Essentially, all models are wrong, but some are useful.” [?]"
Objectives of spatial analysis, continued

4. Simulation

- Using a spatial model which adequately describes a physical process, we can simulate repeatedly from this model and investigate how the process behaves.

- Useful if the physical process is very complicated.

  Ex: Modeling turbulence in the air, or modeling water flows.

References


