APTS 2007/08: Spatial and Longitudinal Data Analysis
Preliminary Material

Module leader: Peter J. Diggle (School of Health and Medicine, Lancaster University)

Aims: This module will introduce students to the statistical concepts and tools involved in modelling data which are correlated in time and/or space.

Learning outcomes: By the end of the module, students should have achieved:
• a clear understanding of the meaning of temporal and spatial correlation;
• a good working knowledge of standard models to describe both the systematic and the random parts of an appropriate model;
• the ability to implement and interpret these models in standard applications;
• an understanding of some of the key concepts which lie at the heart of current research in this area;
• appreciation of at least one substantial case study.

Prerequisites: Preparation for this module should establish familiarity with:
• standard models and tools for time series data, at the level of a typical undergraduate course on time series;
• standard models and tools for spatial data at its simplest level;
• inferential methods, including classical and Bayesian likelihood-based methods, to at least the level of the earlier APTS modules Statistical Inference and Statistical Modelling.

Topics:
• Introduction: motivating examples; the fundamental problem – analysing dependent data.
• Longitudinal data: linear Gaussian models; conditional and marginal models; why longitudinal and time series data are not the same thing.
• Continuous spatial variation: stationary Gaussian processes; variogram estimation what not to do and how to do it; likelihood-based estimation; spatial prediction.
• Discrete spatial variation: Markov random field models.
• Spatial point patterns: exploratory analysis; Cox processes and the link to continuous spatial variation; pairwise interaction processes and the link to discrete spatial variation.
• Spatio-temporal modelling: spatial time series; spatio-temporal point processes.
• Conclusion: review of available software (as preparation for mini-project); connections spatial and longitudinal data analysis as two sides of the same coin.

Assessment: One of:
• A critique, in essay form, of a specified research paper, including both modelling and application aspects;
• A mini-project involving the analysis of a data-set, selected by the student from several on offer (to allow students to focus on topics within the course which they find particularly interesting).

Preliminary Lecture 1: standard models and tools for time series data

Try working through the following material in your own time – without consulting any book on time series analysis. Don't worry if you can't complete all of the exercises – there will be an opportunity to ask about them during the week of the course itself.

1.1 A meteorological time series: maximum daily temperatures at Bailrigg, Lancashire, UK

Figure 1 shows a *time series* of daily maximum temperatures recorded at Bailrigg (the Lancaster University campus) over a period of one year, 1 September 1995 to 31 August 1996.

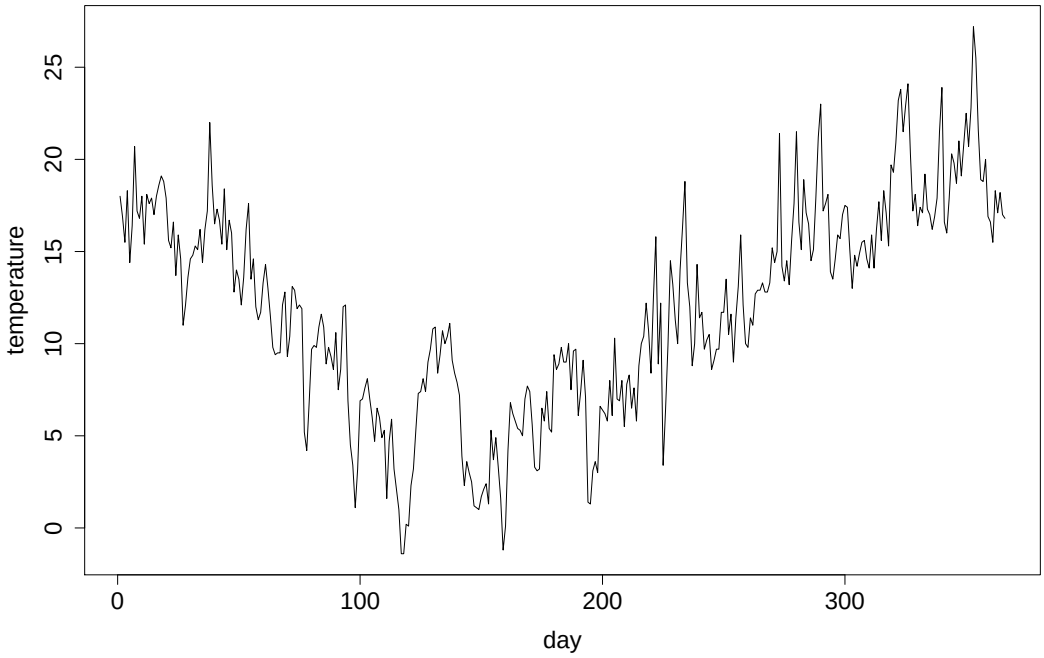

Figure 1: Daily maximum temperatures (degrees celsius) at the bailrigg field station, 1 September 1995 to 31 August 1996.

The goals of a statistical analysis of a data-set such as this could include:

- *describing* the historical pattern of variation
- *predicting* future maximum daily temperatures:
  - tomorrow?
  - next week?
  - next year?

In this preliminary lecture, we focus for the most part on the first of these goals.

Exercise 1.1

(a) describe the main features of this data-set in a few sentences
(b) download the data from `www.lancaster.ac.uk/staff/diggle/THING` and reproduce the plot shown in Figure 1
(c) fit a suitable linear multiple regression model to the data

1.2 Time series models

A *discrete-time real-valued stochastic process* is a collection of random variables, $\{Y_t : t = 1, 2, ...\}$, where t denotes time. A partial realisation of a process of this kind, for $t = 1, 2, ..., n$, is called a *time series*.

A simple class of time series models is given by the specification

$$Y_t = \mu_t + R_t \tag{1}$$

where $\mu_t = \mathrm{E}[Y_t]$ is the *trend* and R_t is the *residual*.

Note that in (1), $\mathrm{E}[R_t] = 0$ for all t by construction. If additionally, $\mathrm{Var}(R_t) = \sigma^2$ and $\mathrm{Cov}(Y_t, Y_{t-u}) = \sigma^2\rho(u)$ then Y_t is *second-order stationary*, and $\rho(u)$ is called the *autocorrelation function* of the process $\{Y_t : t = 0, 1, 2, ...\}$.

The term *second-order stationary* is sometimes abbreviated to *stationary*. A stronger condition than this is *strictly stationary*, whose definition is that, for all positive integers m and all sets of integers $\{u_1, ..., u_m\}$, the joint distribution of $\{Y_{t-u_k} : k = 1, ..., m\}$ does not depend on t.

A discrete-time process Y_t is *Gaussian* if, for all positive integers m and all sets of integers $\{u_1, ..., u_m\}$, the joint distribution of $\{Y_{u_k} : k = 1, ..., m\}$ is multivariate Normal.

Exercise 1.2

The standard assunption in linear multiple regression modelling is that the *residuals* are mutually independent with common variance.

(a) calculate the residuals from the regression model that you fitted in Exercise 1.1 (c)
(b) is it reasonable to assume that the residuals have a common variance?
(c) is it reasonable to assume that the residuals are independent?
(d) let $\rho(u)$ denote the autocorrelation function of the residual time series; suggest a way of estimating $\rho(u)$ for $u = 1, 2, 3, ...$
(e) plot your estimates $\hat{\rho}(u)$ against u; what does the plot suggest about the general shape of $\rho(u)$?

1.3 Two ways to define a first-order autoregressive process

The usual way to specify a *first-order autoregressive process* is

$$Y_t - \mu_t = \alpha(Y_{t-1} - \mu_{t-1}) + Z_t : t = 2, 3, ... \tag{2}$$

where Z_t is a sequence of independent random variables with common mean zero and common variance τ^2. We will also assume that the process is Gaussian, hence $Z_t \sim \mathrm{N}(0, \tau^2)$.

Exercise 1.3.1

(a) show that $\mathrm{E}[Y_t] = \mu_t$
(b) show that, if $|\alpha| < 1$ and we assume that Y_t is stationary with variance σ^2 then $\sigma^2 = \tau^2/(1-\alpha^2)$
(c) hence, what distribution must we specify for Y_1 so as to define a stationary Gaussian process Y_t?
(d) what happens when $\alpha = 1$?

A less conventional way to specify a first-order autoregressive process is the following. Let $\mathcal{H}_t$ denote the *history* of the process Y_t, i.e. all values of the process at times $t' < t$ (this definition is worded so that it extends immediately to processes in continuous time, which we shall need later). Then, Y_t is a Gaussian first-order autoregressive process if

$$Y_t|\mathcal{H}_t \sim \mathrm{N}(\mu_t + \alpha(Y_{t-1} - \mu_{t-1}), \tau^2) : t = 2, 3, ... \tag{3}$$

A brief reflection should convince you that (2) and (3) are equivalent.

Exercise 1.3.2

(a) using whichever of (2) or (3) you consider the more natural, how would you define a non-Gaussian first-order autoregressive process?
(b) give an explicit construction for a Poisson first-order autoregressive process, and write an `R` function to simulate realisations of it
(c) does the unconditional, marginal distribution of Y_t have a simple form?
(d) why is the answer to (c) of no great interest?

1.4 Inference

The log-likelihood for a realisation $Y_t : t = 1, ..., n$ of the stationary Gaussian first-order autoregressive process takes the form

$$L(\mu, \alpha, \tau^2) = \log f(Y_1) + \sum_{t=2}^{n} g_t(Y_t, Y_{t-1}) \tag{4}$$

where

$$f(Y_1) = \{2\pi\tau^2/(1-\alpha^2)\}^{-0.5} \exp\{(Y_1 - \mu_1)^2(1-\alpha^2)/(2\tau^2)\}$$

and

$$g_t(y_t, y_{t-1}) = (2\pi\tau^2)^{-0.5} \exp[\{(Y_t - \mu_t - \alpha(Y_{t-1} - \mu_{t-1})\}^2/(2\tau^2)].$$

Exercise 1.4

Consider the above model in which $\mu = (\mu_1, ..., \mu_n)$ is specified by a linear model, $\mu = X\beta$.

(a) show that, if α is known and we condition on Y_1, the maximum likelihood estimates of β and τ^2 can be obtained by a weighted least squares calculation
(b) hence, again conditioning on Y_1, obtain the profile log-likelihood for α

(c) hence, fit the Gaussian first-order autoregressive process, with a suitable defined trend μ_t, to the Bailrigg maximum daily temperature data
(d) discuss, informally, how well (or how badly) this model fits the data.

1.5 Prediction (forecasting)

We will have more to say about prediction in the lecture course itself. For the time being, a good test of your understanding of this preliminary material is the following.

Exercise 1.5

(a) how would you predict the Bailrigg maximum temperature tomorrow? next week? next year?
(b) does your suggested method of prediction differ according to the required forecast lead-time, and if so why?

1.6 Further preliminary reading on time series

As suggested at the outset, you should have been able to work through this preliminary material without any previous knowledge of time series analysis. Having done so, you might like to read one or more introductory text-book accounts to get a different perspective on the subject. Two possibilities are Chatfield (2003, Chapters 1, 2) or Diggle (1990, Chapters 1, 2).

If you want a foretaste of two more advanced topics that I will discuss in the course itself, have a look at Diggle (1990, Chapters 3 and 4) on *spectral analysis* and Durbin and Koopman (2001, Chapter 1) on *state-space models*.

1.7 Longitudinal data

Longitudinal data are nothing more or less than *replicated* time series data. This apparently innocuous change has profound implications for how one approaches data analysis. Independent replication admits the possiblity of design-based, rather than model-based inference. Put another way, and depending on the scientific purpose of the investigation in hand, in longitudinal data analysis we may choose to accommodate the correlation between the different observations within each time series without explicitly modelling it.

A second distinction in practice is that time series analysis typically focuses on understanding the nature of the correlation structure with the series, whereas in longitudinal data analysis, ther scientific focus is more often on understanding the trend, $\mu(t)$, and in particular how this is affected by explanatory variables associated with each series.

The introductory chapter of either Diggle, Heagerty, Liang and Zeger (2002) or Fitzmaurice, Laird and Ware (2004) will give you more background information for this part of the course.

Exercise 1.7

Consider a study-design in which children's performance on an educational attainment

test is recorded at ages 5, 6 and 7. Let Y_{ij} denote the jth score for the ith child and assume a model

$$Y_{ij} = \alpha + \beta x_j + U_i + Z_{ij},$$

where x_j denotes age minus 6 (so $x_1 = -1, x_2 = 0, x_3 = 1$), the U_i are mutually independent $\mathrm{N}(0, \nu^2)$ and the Z_{ij} are mutually independent $\mathrm{N}(0, \tau^2)$.

(a) Deduce the distribution of $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})$

(b) Find the maximum likelihood estimators of α and β and their variances, assuming ν^2 and τ^2 are known.

(c) Compare your answers to (b) with the maximum likelihood estimators and their variances when the Y_{ij} are mutually independent, Normally distributed with means $\mathrm{E}[Y_{ij}] = \alpha + \beta x_j$ and common variances $\sigma^2 = \nu^2 + \tau^2$.

1.8 References

Chatfield, C. (2003). *The Analysis of Time Series: an Introduction (6th edition).* London: Chapman and Hall.

Diggle, P.J. (1990). *Time Series: a Biostatistical Introduction.*

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition).* Oxford: Oxford University Press.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis.* New Jersey: Wiley.

Preliminary Lecture 2: standard models and tools for spatial data

In this preliminary lecture, my main aim is to introduce you to the three different kinds of spatial data that have been most widely studied. In the course itself, I will say a little more about all three kinds, but will focus particularly on *real-valued continuous spatial variation*. As with Preliminary Lecture 1, try working through the material without using any specialised text-books but don't worry if you can't (or don't have time to) complete them all.

2.1 A simple taxonomy of spatial statistics

My preferred taxonomy is a minor variant on the one used by Cressie (1991) in his encyclopaedic coverage of the subject. Note that this is a taxonomy of spatial *processes* rather than of spatial *data*.

1. Discrete spatial variation
2. Continuous spatial variation:
   (a) real-valued processes
   (b) point processes

The primary distinction here is between a phenomenon that is defined on a finite (or countably infinite) set of locations, and one that is defined on a continuous spatial region, $A \in \mathbb{R}^2$. Within the second category, I distinguish *real-valued processes*, $\{S(x) : x \in \mathbb{R}^2\}$, from *point processes* whose realisations are countable sets of points, $\mathcal{X} = \{x_i \in \mathbb{R}^2 : i = 1, 2, ...\}$. The secondary distinction between spatially continuous real-valued processes and point processes is somewhat pragmatic, in that the data-analytic tools associated with the two types of process are somehwhat different in character.

2.2 Discrete spatial variation (Markov random fields)

A model for *discrete spatial variation* specifies the joint distribution of a random vector $Y = (Y_1, ..., Y_n)$, where the presumption is that each Y_i is associated with a spatial location x_i. The model has nothing to say about any other location, which at first sight casts some doubt on its relevance as a *spatial* model except in the rather rare circumstances that the spatial phenomenon being modelled is genuinely discrete; an example would be when the Y_i represent the yields of individual fruit-trees in an orchard (and even then, a sceptical response might be to plant an additional tree). In practice, the models are very useful as pragmatic approximations, either when a spatial continuum is approximated by a lattice or when data are derived from a spatial continuum by averaging over contiguous sub-areas. Examples of these two situations are image analysis and disease risk mapping, respectively; see Figure 2.

The core idea in the Markov random field approach to discrete spatial models is that the joint distribution of Y should be specified indirectly through its *full conditionals*, i.e.

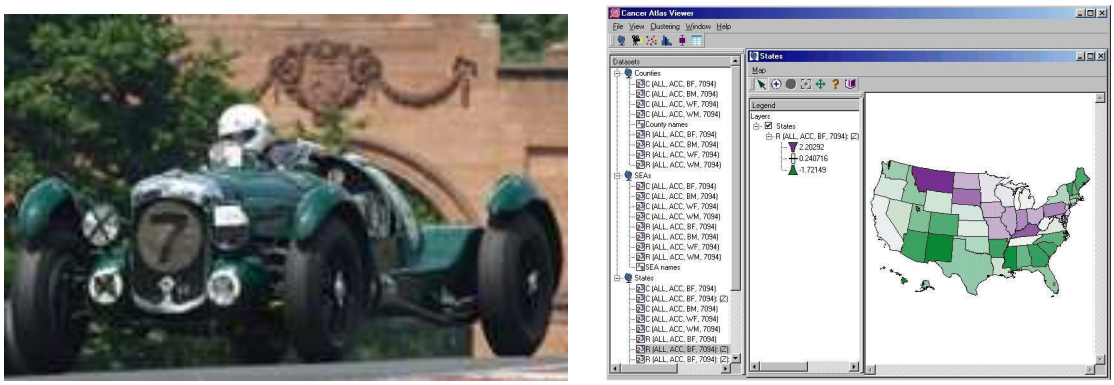

Figure 2: A digital image (left-hand-panel) and a cancer risk map (right-hand panel)

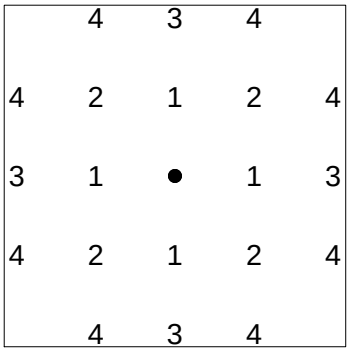

Figure 3: First-order to fourth order neighbours of a central location in a regular square lattice

the set of n univariate conditional distributions of Y_i given $\{Y_j : j \neq i\}$. The following exercise shows that the full conditionals do indeed specify the joint distribution.

Exercise 2.2.1

Show that, for any random vector Y with joint density $f(y)$ and full conditionals $f_i(y_i|\{y_j : j \neq i\})$ and any two feasible realisations y and z,

$$\frac{f(y)}{f(z)} = \prod_{i=1}^{n} \frac{f_i(y_i|y_1, ..., y_{i-1}, z_{i+1}, ..., z_n)}{f_i(z_i|y_1, ..., y_{i-1}, z_{i+1}, ..., z_n)}$$

If you are finding this difficult, see Besag (1974).

To be useful, a model needs to impose some structure on the full conditionals. To do this, we define the *neighbours* $\mathcal{N}_i$ of location i to be those locations j such that the full conditional of Y_i depends on Y_j, hence $f_i(y_i|\{y_j : j \neq i\}) = f_i(y_i|\{y_j : j \in \mathcal{N}_i\})$. For a process on a regular lattice this gives a natural hierarchy of Markov random field models, as illustrated in Figure 3. For a process on an irregular set of locations, it is not obvious how the neighbourhoods should be defined.

Note that the probabilistic structure of a Markov random field is identical to that of a *graphical model* for general multivariate data; the graph with edges between pairs of neighbours in a Markov random field is the conditional independence graph of the equivalent graphical model. For an introduction to graphical models see, for example, Whittaker (1990).

In the early days of spatial statistics (i.e. the early 1970's), the specification of models through their full conditionals was considered somewhat controversial (see, for example, the Discussion of Besag, 1974). To a statistician brought up in the age of hierarchical models and the ubiquitous Gibbs sampler, this might seem rather quaint, but it gives me an excuse to show though a simple example why space is not like time.

As we have seen in Section 1 of this preliminary material, a first-order autoregressive time series model Y_t (with zero mean for simplicity) can equally be defined as

$$Y_t = \alpha Y_{t-1} + Z_t : Z_t \mathrm{N}(0, \tau^2) \tag{5}$$

or

$$Y_t | \{Y_s : s < t\} \sim \mathrm{N}(\alpha Y_{t-1}, \tau^2). \tag{6}$$

The one-dimensional spatial analogues of (5) and (6) are

$$Y_i = \alpha(Y_{i-1} + Y_{i+1}) + Z_t : Z_t \sim \mathrm{N}(0, \tau^2) \tag{7}$$

or

$$Y_i | \{Y_j : j \neq i\} \sim \mathrm{N}(\alpha(Y_{i-1} + Y_{i+1}), \tau^2), \tag{8}$$

but these are *not* equivalent.

Exercise 2.2.2

Show that for the model (7) the full conditional of Y_i depends on Y_{i-2}, Y_{i-1}, Y_{i+1} and Y_{i+2}.

The so-called simultaneous autoregressive construction (7) turned out to be a dead-end, not least because it cannot be adapted either to non-lattice or non-Gaussian models.

2.3 Real-valued continuous spatial variation (geostatistics)

Real-valued continuous spatial variation is arguably the most widely encountered form of spatial process in nature. Our interest in this Section is with processes that are measured at a finite set of sampled locations, leading to data of the form $(Y_i, x_i) : i = 1, ..., n$, where Y_i is a measured value at a location x_i in a spatially continuous region of interest A. Probably the earliest systematic development of a statistical methodology for analysing data of this kind is the Fontainebleau school of geostatistics led by the late Georges Matheron. The name *geostatistics* derives from its origins in mineral exploration, where Y_i represents the grade of mineral extracted from a sample of material at a location x_i in a region A under consideration for future exploitation.

Geostatistical models are naturally hierarchical, consisting of a first-level model for an underlying spatially continuous process $\{S(x) : x \in \mathbb{R}^2\}$, and a second-level model for the sampling distribution of the Y_i conditional on $S(\cdot)$. The simplest, and therefore most

widely used, model specifies $S(\cdot)$ to be a Gaussian process and the Y_i as conditionally independent Gaussian with $\mathrm{E}[Y_i|S(\cdot)] = S(x_i)$ and $\mathrm{Var}\{Y_i|S(\cdot)\} = \tau^2$. Equivalently,

$$Y_i = S(x_i) + Z_i, \tag{9}$$

where the Z_i are independent $\mathrm{N}(0, \tau^2)$.

Note that (9) is somewhat reminiscent of the simultaneous autoregressive specification (7). However, the model (9), unlike a Markov random field model, does adjust itself automatically to any augmentation of the data by measurements at additional locations. Nevertheless, I still favour the conditional formulation over (9) because it extends easily to non-Gaussian processes. For example, a geostatistical model for count data can be obtained by retaining the Gaussian specification for $S(\cdot)$ and making the conditional distribution of Y_i given $S(\cdot)$ Poisson with expectation $\exp\{S(x_i)\}$.

To specify the Gaussian process $S(\cdot)$, we need only specify its mean and covariance structure. For simplicity, assume that the mean is constant. For a *stationary* process, we require $\mathrm{Var}\{S(x)\} = \sigma^2$ for all x and $\mathrm{Corr}\{S(x), S(x')\} = \rho(x - x')$ for all x and x'. If $S(\cdot)$ is also *isotropic*, then $\rho(x - x') = \rho(||x - x'||)$ where $||\cdot||$ denotes distance.

The Matérn family of correlation functions, named after Matérn (1986) is given by

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi), \tag{10}$$

in which $K_{\kappa}(\cdot)$ denotes a modified Bessel function of order κ, $\phi > 0$ is a scale parameter with the dimensions of distance, and $\kappa > 0$ is a shape parameter which determines the mean-square differentiability of $S(\cdot)$.

Exercise 2.3

For a variety of good and bad reasons, geostatisticans traditionally do not work directly with the covariance function. Instead, they characterise the second-moment structure of their models in terms of the *variogram*,

$$V(u) = \frac{1}{2}\mathrm{var}\{S(x) - S(x - u)\}, \tag{11}$$

if this is well-defined.

(a) Show that for any stationary process $S(\cdot)$, $V(u) = \sigma^2\{1 - \rho(u)\}$.

(b) Find a non-stationary process $S(\cdot)$ whose variogram is well-defined.

2.4 Spatial point processes

The simplest model for a spatial point process is the *homogeneous planar Poisson process*. One way to define this process is through the following postulates:

PP1. The number of points of the process in any planar region A follows a Poisson distribution with mean $\lambda|A|$ where $|\cdot|$ denotes area and the constant $\lambda > 0$ is the *intensity* of the process.

PP2. Numbers of points of the process in disjoint regions are stochastically independent.

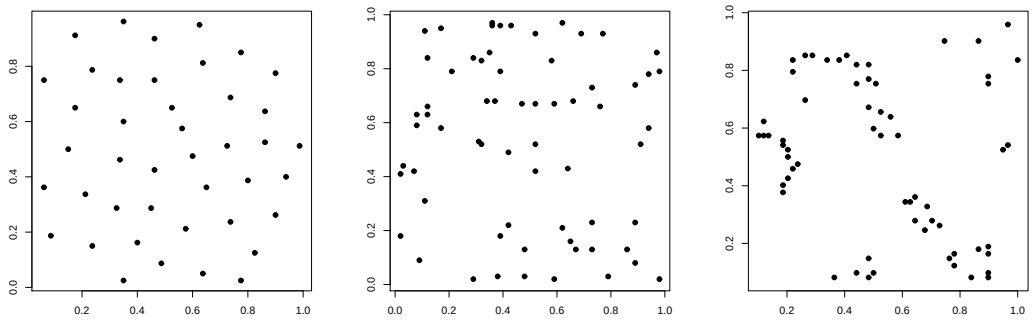

Figure 4: Locations of 42 biological cells (left-hand panel), 65 Japanese black pine saplings (centre panel) and 62 redwood seedlings (right-hand panel).

One consequence of PP1 and PP2, which is sometimes used as part of the definition of the Poisson process, is:

PP3. Conditional on there being n points of the process in a region A, the positions of the points form an independent random sample from the uniform distribution on A.

The Poisson process is rarely satisfactory as a model for naturally occurring point processes, but it is the foundation on which many more interesting models are built. It is not immediately obvious that PP1, PP2 and PP3 are mutually consistent, and they certainly would not be if the Poisson distribution in PP1 were replaced by an arbitrary discrete distribution.

Exercise 2.4.1

(a) Prove that PP1 and PP2 imply PP3.

(b) Derive the distribution of the distance from the origin to the closest point in a homogeneous planar Poisson process of intensity λ

(c) Generalise the result of (b) to obtain the joint distribution of the distances $X_1, ..., X_k$ from the origin to the first, second, ... kth nearest points of the process.

The Poisson process provides a standard of *complete spatial randomness*. We can use this as a dividing hypothesis, and so characterise spatial point process data as regular, completely random or aggregated. Figure 4 shows three data-sets that exemplify this classification.

Exercise 2.4.2

Think of a simple stochastic model that would generate regular spatial point patterns, and one that would generate aggregated patterns. Write an R function to simulate each of your models on the unit square.

2.5 Further preliminary reading on spatial statistics

Cressie (1991) remains the single most wide-ranging account of spatial statistical models and methods. Possibly more accessible accounts of each of the three sub-areas I have described in this preliminary material are the introductory chapters of Rue and Held (2005) on discrete spatial variation, Diggle and Ribeiro (2007) on geostatistics and Diggle (2003) on point processes.

2.6 References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192–225.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York : Wiley.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns (second edition)*. London : Edward Arnold.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based geostatistics*. New York: Springer.
Whittaker, J.C. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester : Wiley.

Matérn, B. (1986). *Spatial variation (second edition)*. Berlin : Springer-Verlag.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: CRC Press.

Computing resources

I shall be using `R` throughout the course. Useful `R` packages that you might like to look at beforehand (all available from CRAN) include (amongst many others):

For longitudinal data analysis: `gee`, `lme4`, `nlme`

For spatial analysis: `spBayes`, `geoR`, `geoRglm`, `spatstat`, `splancs`

Note that CRAN also offers `R2WinBUGS`, a package that allows the WinBUGS and OpenBUGS software to be accessed within `R`

Note also that I claim no special expertise in any of these!

PJD,22 July 2008