When Populations and Hazards Collide: Modelling Exposures and Health Risks

Gavin Shaddick
University of Bath

&

James V. Zidek
University of British Columbia

12^{th} - 14^{th} November 2015

Course Contents

Outline

Thursday, November 12

- *09:30 - 10:00* Introduction
- *10:00 - 11:15* The need for Spatio-Temporal Modelling
- *11:15 - 11:30* Break
- *11:30 - 13:00* Spatial Lattice Processes and Applications
- *13:00 - 15:30* Lunch
- *15:30 - 17:00* Computer Labs

OUTLINE

Friday, November 13

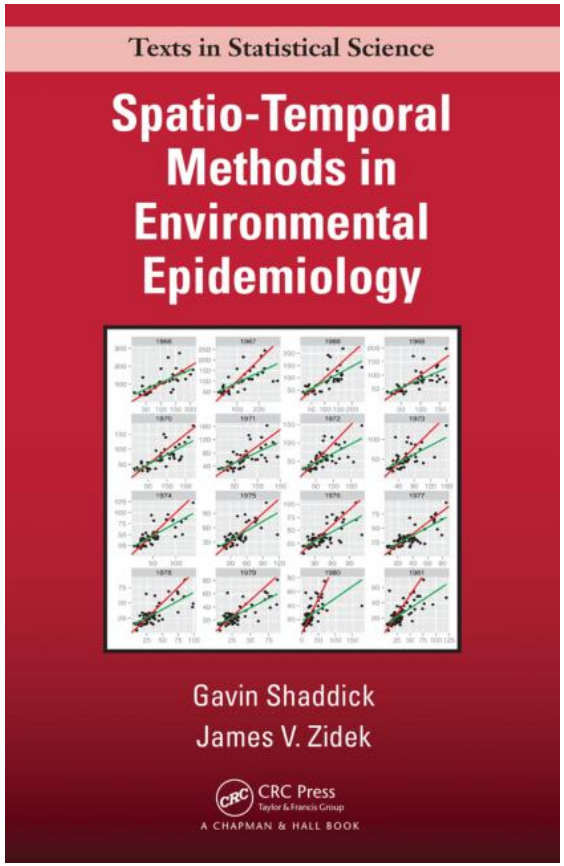
- *09:30 - 11:00* Point Referenced Spatial Processes
- *11:00 - 11:30* Break
- *11:30 - 13:00* Point Referenced Spatial Processes and Applications
- *13:00 - 15:30* Lunch
- *15:30 - 17:00* Spatio-Temporal Processes

Saturday, November 14

- *9:30 - 11:00* Computer Lab

COURSE TEXTBOOK

Title: Spatio-Temporal Methods in Environmental Epidemiology
Authors: Gavin Shaddick and Jim Zidek
Publisher: CRC Press
Resource Website:
`http://www.stat.ubc.ca/~gavin/STEPIBookNewStyle/`



Contact Information

Gavin Shaddick, University of Bath

- Email: G.Shaddick@bath.ac.uk
- Web: `http://www.bath.ac.uk/~masgs`

Jim Zidek, University of British Columbia

- Email: jim@stat.ubc.ca
- Webpage: `http://www.stat.ubc.ca/~jim`

The need for Spatio-Temporal modelling

THE NEED FOR SPATIO-TEMPORAL MODELLING

- Spatial epidemiology is the description and analysis of geographical data, specifically health data in the form of counts of mortality or morbidity and factors that may explain variations in those counts over space.
- These may include demographic and environmental factors together with genetic, and infectious risk factors.
- It has a long history dating back to the mid-1800s when John Snow's map of cholera cases in London in 1854 provided an early example of geographical health analyses that aimed to identify possible causes of outbreaks of infectious diseases.

EXAMPLE: JOHN SNOW'S CHOLERA MAP

Figure: John Snow's map of cholera cases in London 1854. Red circles indicate locations of cholera cases and are scaled depending on the number of reported cholera cases.Purple taps indicate locations of water pumps.

THE NEED FOR SPATIO-TEMPORAL MODELLING

- Advances in statistical methodology together with the increasing availability of data recorded at very high spatial and temporal resolution has lead to great advances in spatial and, more recently, spatio–temporal epidemiology.
- These advances have been driven in part by increased awareness of the potential effects of environmental hazards and potential increases in the hazards themselves.

THE NEED FOR SPATIO-TEMPORAL MODELLING

- Over the past two decades, population predictions based on conventional demographic methods have forecast that the world's population will rise to about 9 billion in 2050, and then level off or decline.
- However, recent analyses using Bayesian methods have provided compelling evidence that such projections may vastly underestimate the world's future population and instead of the expected decline, population will continue to rise.
- Such an increase will greatly add to the anthropogenic contributions of environmental contamination and will require political, societal and economic solutions in order to adapt to increased risks to human health and welfare.

THE NEED FOR SPATIO-TEMPORAL MODELLING

- In order to assess and manage these risks there is a requirement for monitoring and modelling the associated environmental processes that will lead to an increase in a wide variety of adverse health outcomes.
- Addressing these issues will involve a multi-disciplinary approach and it is imperative that the uncertainties that will be associated with each of the components can be characterised and incorporated into statistical models used for assessing health risks.

HEALTH-EXPOSURE MODELS

- An analysis of the health risks associated with an environmental hazard will require a model which links exposures to the chosen health outcome.
- There are several potential sources of uncertainty in linking environmental exposures to health, especially when the data might be at different levels of aggregation.
- For example, in studies of the effects of air pollution, data often consists of health counts for entire cities with comparisons being made over space (with other cities experiencing different levels of pollution) or time (within the same city) whereas exposure information is often obtained from a fixed number of monitoring sites within the region of study.

HEALTH-EXPOSURE MODELS

- Actual exposures to an environmental hazard will depend on the temporal trajectories of the population's members that will take individual members of that population through a sequence of micro-environments, such as a car, house or street.
- Information about the current state of the environment may be obtained from routine monitoring or through measurements taken for a specialised purpose.
- An individual's actual exposure is a complex interaction of behaviour and the environment.
- Exposure to the environmental hazard affects the individual?s risk of certain health outcomes, which may also be affected by other factors such as age and smoking behaviour.

ESTIMATING RISKS

- If a study is carefully designed, then it should be possible to obtain an assessment of the magnitude of a risk associated with changes in the level of the environmental hazard.
- Often this is represented by a relative risk or odds ratio, which is the natural result of performing log-linear and logistic regression models respectively.
- They are often accompanied by measures of uncertainty, such as 95% confidence (or in the case of Bayesian analyses, credible) intervals.

ESTIMATING RISKS

- However, there are still several sources of uncertainty which cannot be easily expressed in summary terms.
- These include the uncertainty associated with assumptions that were implicitly made in any statistical regression models, such as the shape of the dose-response relationship (often assumed to be linear).
- The inclusion, or otherwise, of potential confounders and unknown latencies over which health effects manifest themselves will also introduce uncertainty.

A NEW WORLD OF UNCERTAINTY

- The importance of uncertainty has increased dramatically as the twentieth century ushered in the era of post-normal science as articulated by Funtowicz and Ravetz.
- Gone were the days of the solitary scientist running carefully controlled bench-level experiments with assured reproducibility, the hallmark of good classical science.
- In came a science characterized by great risks and high levels of uncertainty, an example being climate science with its associated environmental health risks.

A NEW WORLD OF UNCERTAINTY

- Post-normal science called for a search for new approaches to dealing with uncertainty
- Ones that recognised the diversity of stakeholders and evaluators needed to deal with these challenges.
- That search led to the recognition that characterising uncertainty required a dialogue amongst this extended set of peer reviewers through workshops and panels of experts.
- Such panels are convened by the US Environmental Protection Agency (EPA) who may be required to debate the issues in a public forum with participation of outside experts (consultants) employed by interest groups such as in the case of air pollution the American Lung Association and the American Petroleum Producers Association.

DEPENDENCIES OVER SPACE AND TIME

- Environmental epidemiologists commonly seek associations between an environ- mental hazard Z and a health outcome Y .
- A spatial association is suggested if measured values of Z are found to be large (or small) at locations where counts of Y are also large (or small).
- A classical regression analysis might then be used to assess the magnitude of any associations and to assess whether they are significant.

Dependencies over space and time

- However such an analysis would be flawed if the pairs of measurements (of exposures), Z and the health outcomes, Y , are spatially correlated.
- This results in outcomes at locations close together being more similar than those further apart.
- In this case, or in the case of temporal correlation, the standard assumptions of stochastic independence between experimental units would not be valid.

Example: Spatial Correlation in the UK

- An example of spatial correlation can be seen in the next slide which shows the spatial distribution of the risk of hospital admission for chronic obstructive pulmonary disease (COPD) in the UK.
- There seem to be patterns in the data with areas of high and low risks being grouped together suggesting that there may be spatial dependence that would need to be incorporated in any model used to examine associations with potential risk factors.

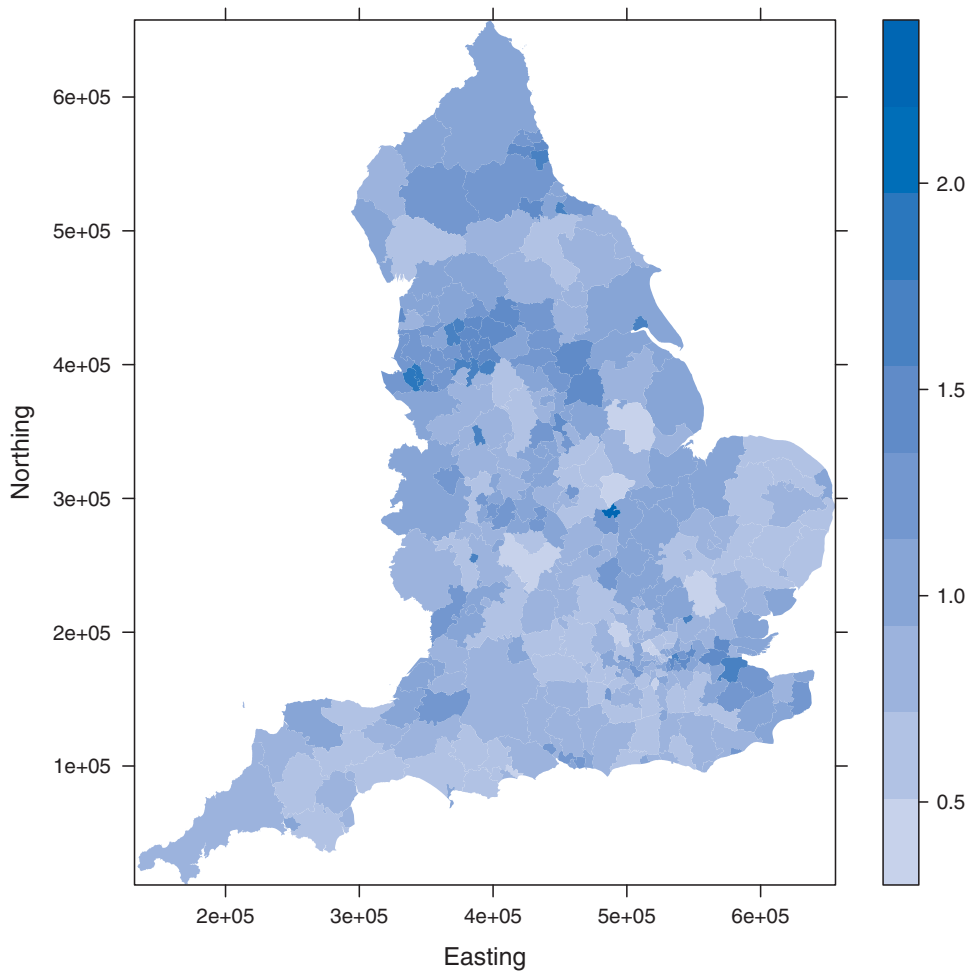


Figure: Map of the spatial distribution of risks of hospital admission for a respiratory condition, chronic obstructive pulmonary disease (COPD), in the UK for 2001. The shades of blue correspond to standardised admission rates, which are a measure of risk. Darker shades indicate higher rates of hospitalisation allowing for the underlying age–sex profile of the population within the area.

EXAMPLE: DAILY MEASUREMENTS OF PARTICULATE MATTER

An example of temporal correlation in exposures can be seen below, which shows daily measurements of particulate matter over 250 days in London in 1997. Clear auto-correlation can be seen in this series of data with periods of high and low pollution.

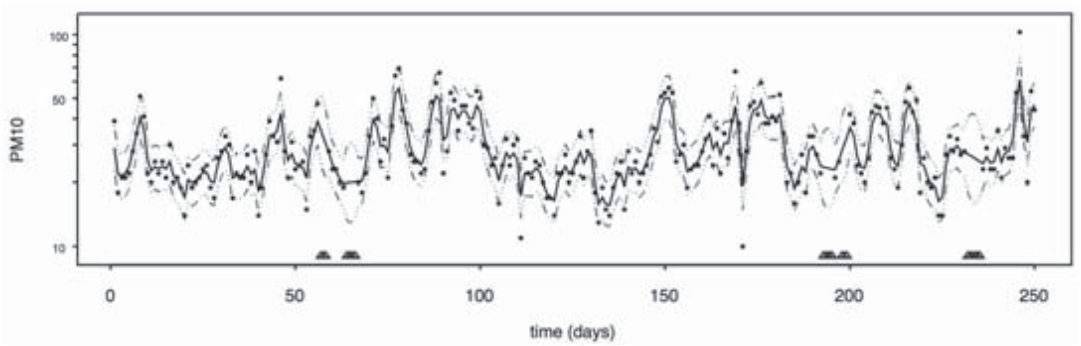


Figure: Time series of daily measurements of particulate matter (PM_{10}) for 250 days in 1997 in London. Measurements are made at the Bloomsbury monitoring site in central London. Missing values are shown by triangles. The solid black line is a smoothed estimate produced using a Bayesian temporal model and the dotted lines show the 95% credible intervals associated with the estimates.

Example: Spatial prediction of NO_2 concentrations in Europe

- In this example we see the result of using a spatial model to predict levels of nitrogen dioxide (NO_2) across Europe (Shaddick *etal.*, 2013).
- Measurements were available from monitoring sites at approximately 400 sites situated throughout Europe and these data were used to predict concentrations for every 1km $\times$ 1km geographical grid cell within the region.
- In this case, a Bayesian model was fit within WinBUGS and posterior predictions were imported (via R) to ESRI ArcGIS for mapping. The results can be seen in the next slide.

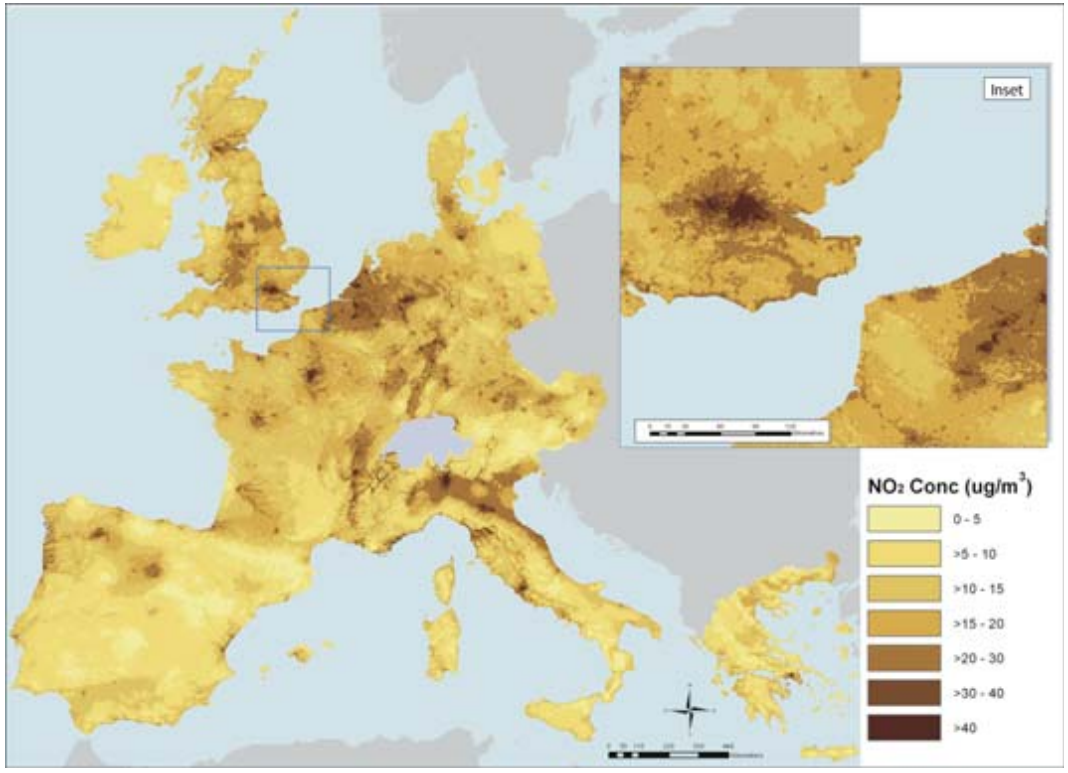


Figure: Predictions of nitrogen dioxide (NO_2) concentrations throughout Europe. The predictions are from a Bayesian spatial model and are the medians of the posterior distributions of predictions based on measurements from approximately 400 monitoring sites.

Dependencies over space and time

- Environmental exposures will vary over both space and time and there will potentially be many sources of variation and uncertainty.
- Statistical methods must be able to acknowledge this variability and uncertainty and be able to estimate exposures at varying geographical and temporal scales in order to maximise the information available that can be linked to health outcomes in order to estimate the associated risks.
- In addition to estimates of risks, such methods must be able to produce measures of uncertainty associated with those risks.
- These measures of uncertainty should reflect the inherent uncertainties that will be present at each of the stages in the modelling process.

DEPENDENCIES OVER SPACE AND TIME

- This has led to the application of spatial and temporal modelling in environmental epidemiology, in order to incorporate dependencies over space and time in analyses of association.
- The value of spatio–temporal modelling can be seen in two major studies:
  (1) The Children's Health Study in Los Angeles and
  (2) The MESA Air (Multi-Ethnic Study of Atherosclerosis Air Pollution) study.

Example: Children's health study – Los Angeles

- Children may suffer increased adverse effects to air pollution compared to adults as their lungs are still developing.
- They are also likely to experience higher exposures as they breathe faster and spend more time outdoors engaged in strenuous activity.
- The effects of air pollution on children?s health is therefore a very important health issue.

Example: Children's health study – Los Angeles

- The Children's Health Study began in 1993 and is a large, long-term study of the effects of chronic air pollution exposures on the health of children living in Southern California.
- Approximately 4000 children in twelve communities were enrolled in the study although substantially more have been added since the initiation of the study.
- Data on the children's health, their exposures to air pollution and many other factors were recorded annually until they graduated from high school.
- While the study was observational in nature, children were selected to provide good contrast between areas of low and high exposure.
- Spatio–temporal modelling issues had to be addressed in the analysis since data were collected over time and from a number of communities which were distributed over space

Example: Children's health study – Los Angeles

Current levels of air pollution have chronic, adverse effects on lung growth leading to clinically significant deficit in 18-year-old children. Air pollution affects both new onset asthma and exacerbation. Living in close proximity to busy roads is associated with risk for prevalent asthma. Residential traffic exposure is linked to deficit in lung function growth and increased school absences. Differences in genetic makeup affect these outcomes. (http://hydra.usc.edu/scehsc/about-studies-childrens.html)

EXAMPLE: AIR POLLUTION AND CARDIAC DISEASE

- The MESA Air (Multi-Ethnic Study of Atherosclerosis and Air Pollution) study involves more than 6000 men and women from six communities in the United States.
- The study started in 1999 and continues to follow participants' health.
- The central hypothesis for this study is that long-term exposure to fine particles is associated with a more rapid progression of coronary atherosclerosis (hardening of the heart arteries)
- The problems caused by the smallest particles is their capacity to move through the gas exchange membrane into the blood system.
- Particles may also generate anti-inflammatory mediators in the blood that attack the heart.

EXAMPLE: AIR POLLUTION AND CARDIAC DISEASE

- Data are recorded both over time and space and so the analysis has been designed to acknowledge this.
- The study was designed to ensure the necessary contrasts needed for good statistical inference by taking random spatial samples of subjects from six very different regions.
- The study has yielded a great deal of new knowledge about the effects of air pollution on human health.
- In particular, exposures to chemicals and other environmental hazards appear to have a very serious impact on cardiovascular health.

EXAMPLE: AIR POLLUTION AND CARDIAC DISEASE

Results from MESA Air show that people living in areas with higher levels of air pollution have thicker carotid artery walls than people living in areas with cleaner air. The arteries of people in more polluted areas also thickened faster over time, as compared to people living in places with cleaner air. These findings might help to explain how air pollution leads to problems like stroke and heart attacks. (http://depts.washington.edu/mesaair/)

Bayesian Hierarchical Models

Bayesian hierarchical models are an extremely useful and flexible framework in which to model complex relationships and dependencies in data and they are used extensively throughout the course. In the hierarchy we consider, there are three levels;

(1) The observation, or measurement, level; $Y|Z, X_1, \theta_1$.
Data, Y, are assumed to arise from an underlying process, Z, which is unobservable but from which measurements can be taken, possibly with error, at locations in space and time. Measurements may also be available for covariates, X_1. Here θ_1 is the set of parameters for this model and may include, for example, regression coefficients and error variances.

Bayesian Hierarchical Models

(2) The underlying process level; $Z|X_2, \theta_2$.
The process Z drives the measurements seen at the observation level and represents the true underlying level of the outcome. It may be, for example, a spatio–temporal process representing an environmental hazard. Measurements may also be available for covariates at this level, X_2. Here θ_2 is the set of parameters for this level of the model.

(3) The parameter level; $\theta = (\theta_1, \theta_2)$.
This contains models for all of the parameters in the observation and process level and may control things such as the variability and strength of any spatio–temporal relationships.

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- A spatial–temporal random field, $Z_{st}, s \in \mathcal{S}, t \in \mathcal{T}$, is a stochastic process over a region and time period.
- This underlying process is not directly measurable, but realisations of it can be obtained by taking measurements, possibly with error.
- Monitoring will only report results at N_T discrete points in time, $T \in \mathcal{T}$ where these points are labelled $T = \{t_0, t_1, \ldots, t_{N_T}\}$.
- The same will be true over space, since where air quality monitors can actually be placed may be restricted to a relatively small number of locations, for example on public land, leading to a discrete set of N_S locations $S \in \mathcal{S}$ with corresponding labelling, $S = \{s_0, s_1, \ldots, s_{N_T}\}$.

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- There are three levels to the hierarchy that we consider.
- The observed data, $Y_{st}, s = 1, ..., N_S, t = 1, ..., N_T$, at the first level of the model are considered conditionally independent given a realisation of the underlying process, Z_{st}.

$$Y_{st} = Z_{st} + v_{st}$$

  where v_{st} is an independent random, or measurement, error term
- The second level describes the true underlying process as a combination of two terms: (i) an overall trend, μ_{st} and (ii) a random process, ω_{st}.

$$Z_{st} = \mu_{st} + \omega_{st}$$

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- The trend, or mean term, μ_{st} represents broad scale changes over space and time which may be due to changes in covariates that will vary over space and time.
- The random process, ω_{st} has spatial–temporal structure in its covariance.
- In a Bayesian analysis, the third level of the model assigns prior distributions to the hyperparameters from the previous levels.

Dealing with 'big' data

- Due to both the size of the spatio–temporal components of the models that may now be considered and the number predictions that may be be required, it may be computationally impractical to perform Bayesian analysis using packages such as WinBUGS or MCMC in any straightforward fashion.
- This can be due to both the requirement to manipulate large matrices within each simulation of the MCMC and issues of convergence of parameters in complex models.
- During this course, we will show examples of recently developed techniques that perform 'approximate' Bayesian inference.
- This is based on integrated nested Laplace approximations (INLA) and thus do not require full MCMC sampling to be performed.
- INLA has been developed as a computationally attractive alternative to MCMC.

Dealing with 'big' data

- In a spatial setting such methods are naturally aligned for use with areal level data rather than the point level.
- This is available within the R-INLA package and an example of its use can be seen in the Figure on the next slide
- This shows a triangulation of the locations of black smoke (a measure of particulate air pollution) monitoring sites in the UK.
- The triangulation is part of the computational process which allows Bayesian inference to be performed on large sets of point-referenced spatial data.

Dealing with 'big' data

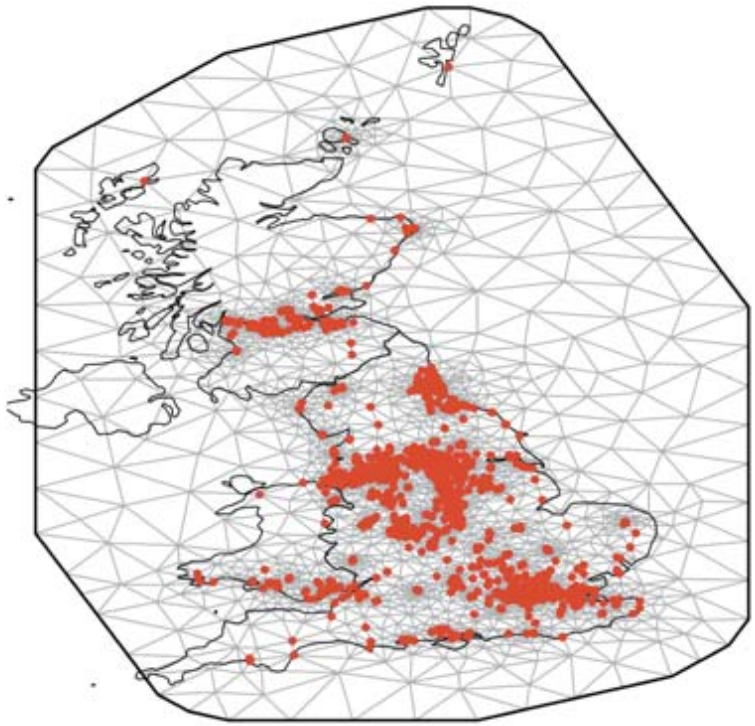


Figure: Triangulation for the locations of black smoke monitoring sites within the UK for use with the SPDE approach to modelling point-referenced spatial data with INLA. The mesh comprises 3799 edges and was constructed using triangles that have minimum angles of 26 and a maximum edge length of 100 km. The monitoring locations are highlighted in red.

Spatial Data

Three main types of spatial data are commonly encountered in environmental epidemiology. They are:

(i) Lattice

(ii) Point-Referenced

(iii) Point-Process Data

Spatial Data: Lattices

- Lattices refer to situations in which the spatial domain consists of a discrete set of 'lattice points'.
- These points may index the corners of cells in a regular or irregular grid.
- Alternatively, they may index geographical regions such as administrative units or health districts.
- We denote the set of all lattice points by $\mathcal{L}$ with data available at a set of N_L points, $l \in L$ where $L = l_1, ..., l_{N_L}$.
- In many applications, such as disease mapping, L is commonly equal to $\mathcal{L}$. A key feature of this class is its neighbourhood structure; a process that generates the data at a location has a distribution that can be characterised in terms of its neighbours.

Spatial Data: Point-Referenced

- Point-referenced data are measured at a fixed, and often sparse, set of 'spatial points' in a spatial domain or region.
- That domain may be continuous, $\mathcal{S}$ but in the applications considered in this course the domain will be treated as discrete both to reduce technical complexity and to reflect the practicalities of siting monitors of environmental processes.
- For example, when monitoring air pollution, the number of monitors may be limited by financial considerations and they may have to be sited on public land.
- Measurements are available at a selection of N_S sites, $s \in S$ where $S = s_1, ..., s_{N_S}$.
- Sites would usually be defined in terms of their geographical coordinates such as longitude and latitude, i.e. $s_l = (a_l, b_l)$.

Spatial Data: Point Processes

- Point-process data consists of a set of points, S, that are randomly chosen by a spatial point process.
- These points could mark, for example, the incidence of a disease such as childhood leukaemia.
- Despite the importance of spatial point process modelling we do not cover this topic and its range of applications in this course.

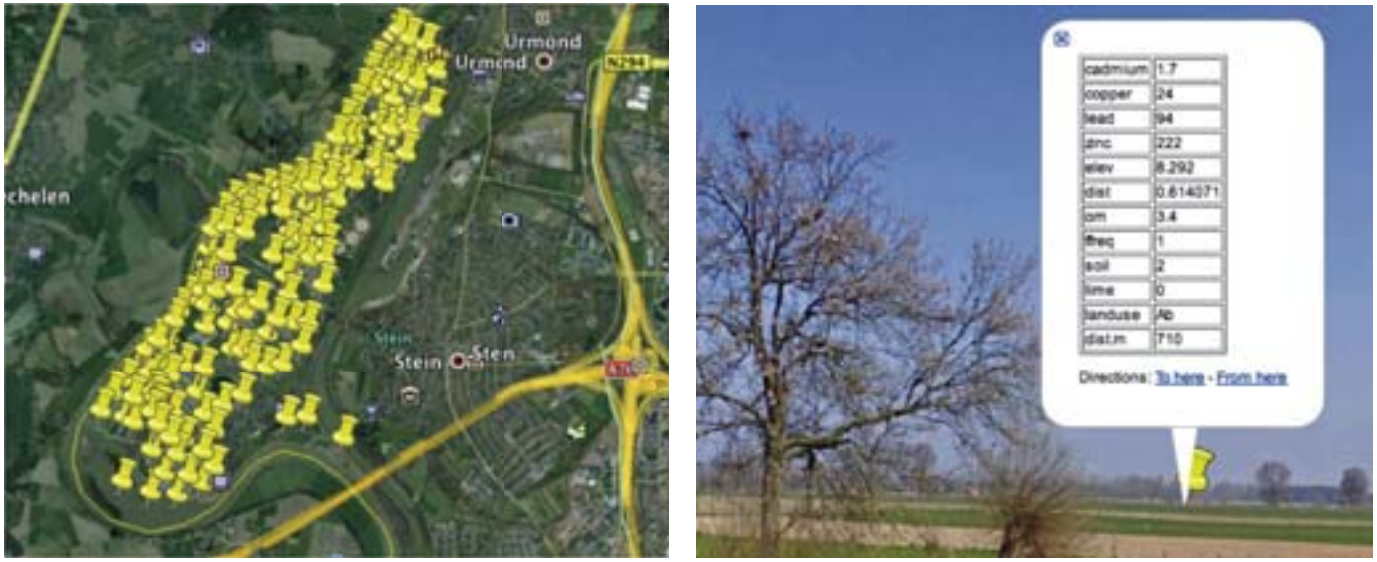
EXAMPLE: VISUALISING SPATIAL DATA

- Data visualisation is an important topic which encompasses aspects of model building, including the assessment of the validity of modelling assumptions, and the presentation of results.
- We illustrate this by mapping measurement of lead concentrations in the Meuse River flood plain.
- The Meuse River is one of the largest in Europe and the subject of much study.
- A comprehensive dataset was collected in its flood plain in 1990 and provides valuable information on the concentrations of a variety of elements in the river.
- The information is measured at 155 sampling sites within the flood plain.

EXAMPLE: VISUALISING SPATIAL DATA

- The figure on the next slide shows the result of using Google maps to visualise data. It shows the sampling sites marked with map tacks.
- Google's Street View then lets an observer see the map tacks. Clicking on one of the visible map tacks reveals the sample data record for that site within Street View.

EXAMPLE: VISUALISING SPATIAL DATA



(a) Sampling sites near Meuse River (b) Map tack opens to show sample

Figure: Google Earth and Google Street Map provide useful ways of visualising spatial data. Here we see (a) the location at which samples were taken in the Meuse River flood plain and (b) the information that was collected.

Good Approaches to Spatio-Temporal Modelling

- Often spatio–temporal models are purpose-built for a particular application and then presented as a theoretical model.
- It is then reasonable to ask what can be done with that model in settings other than those in which it was developed.
- More generally, can it be extended for use in other applications?

Good Approaches to Spatio-Temporal Modelling

There are a number of key elements which are common to good approaches to spatio–temporal modelling. The approaches should do the following:

- Incorporate all sources of uncertainty. This has led to the widespread use of Bayesian hierarchical modelling in theory and practice.
- Have an associated practical theory of data-based inference.
- Allow extensions to handling multivariate data. This is vital as it may be a mix of hazards that cause negative health impacts. Even in the case where a single hazard is of interest, the multivariate approach allows strength to be borrowed from the other hazards which are correlated with the one of concern.

Good Approaches to Spatio-Temporal Modelling

- Be computationally feasible to implement. This is of increasing concern as we see increasingly large domains of interest. One might now reasonably expect to see a spatial domain with thousands of sites and thousands of time points.
- Come equipped with a design theory that enables measurements to be made optimally for estimating the process parameters or for predicting unmeasured process values. Good data are fundamental to good spatio–temporal modelling, yet this aspect is commonly ignored and can lead to biased estimates of exposures and thus risk.

Good Approaches to Spatio-Temporal Modelling

- Produce well calibrated error bands. For example, a 95% band should contain predicted values 95% of the time, i.e. they have correct *coverage probabilities*. This is important not only in substantive terms, but also in model checking.
- There may be questions about the formulation of a model, for example of the precise nature of the spatio–temporal process that is assumed, but that may be of secondary importance if good empirical performance of the model can be demonstrated.

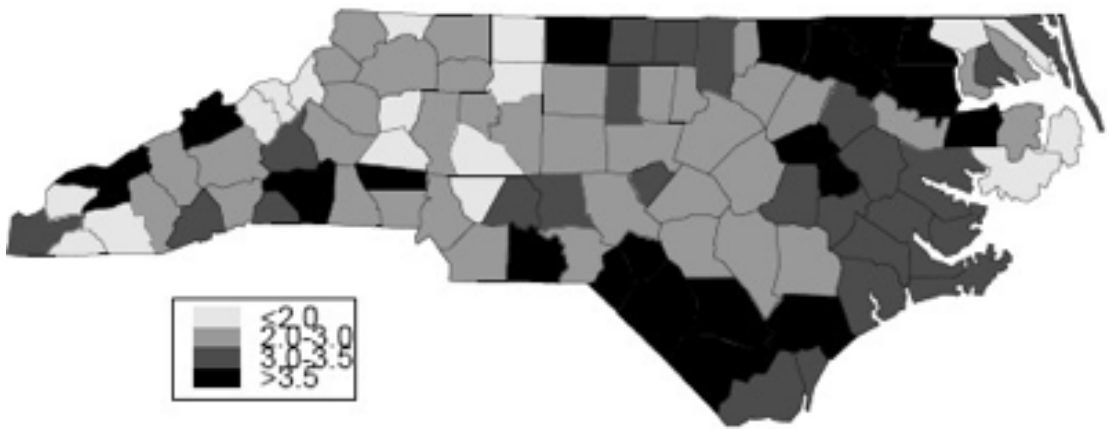
Spatial Lattice Processes

WHY MODEL LATTICE PROCESSES?

- To spot spatial patterns such as elevated disease counts near hazardous waste sites.
- To smooth data across space by borrowing strength - small units may not contain much data

EXAMPLE: SIDS DATA

This well known data were treated in Cressie's 1993 text on spatial statistics. The represent counts of the sudden death infant syndrome. A plot of the counts and their counts is given in the figure. This exemplifies data obtained from records representing administrative regions like cities. Concerns about cause in high count regions.



PROXIMITY MATRICES

Play fundamental role in analyzing such data. Form: $W = \{w_{ij}\}$ with $w_{ii} = 0$ represents the proximity to one another of two locations or regions i, j.

- Examples:
  - $w_{ij} = 1$ if and only they have common boundary.
  - $w_{ij} =$ inverse distance between units
  - $w_{ij} = 1$ if distance between units is $\leq K$
  - $w_{ij} = 1$ for all m of i's nearest neighbours j
- W is typically symmetric, but need not be
- $\tilde{W}$: is standardized so rows sum to one but symmetry lost
- W's elements called "weights"
- Can be used to define neighbours of i

MORAN'S I

W can be used to define clustering indices such as Moran's I for n regions:

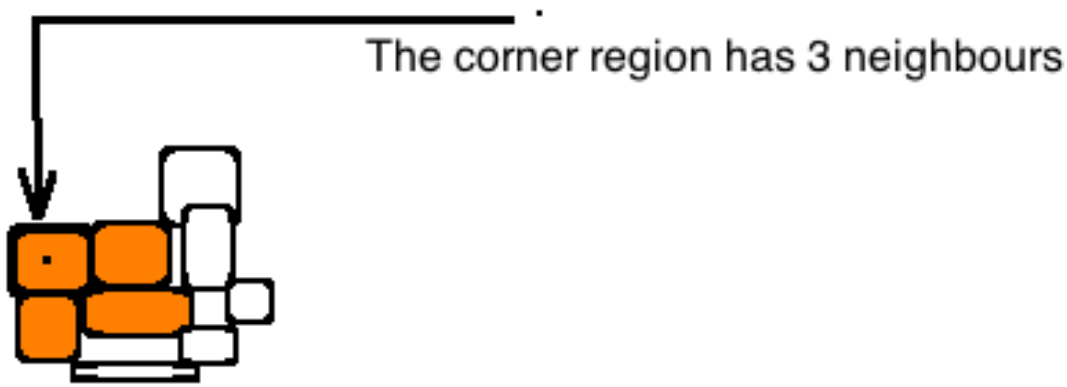
$$I = \frac{\sum_i \sum_j w_{ij}/w_{\{-i\}.}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i (Z_i - \bar{Z})^2/n}.$$

Here $w_{\{-i\}.} = \sum_{i \neq j} w_{ij}$ & I large means that nearby points are similar. Good exploratory tool for cluster detection.

NOTE: W can be used to construct smoothers.

MARKOV RANDOM FIELD (MRF)

Markov random fields focus on local modelling of spatial relationships through conditional distributions.



NEIGHBOURHOODS

- $D = \{s_1, \ldots, s_m\}$ be the lattice indices (e.g centroids)
- $Y(s_i)$ be a process of interest
- $\mathbf{Y}_{-i}$: all responses but $Y(s_i)$
- Define $N(s_i) \subset \{s_1, \ldots, s_m\}$ as s_i neighbourhood if $[Y(s_i) \mid Y_{-i}] = [Y(s_i) \mid Y(s_j),\ s_j \in N(s_i)]$

Local dependence

Specify local spatial dependencies by:

$$[Y(s_i)|Y(s_j),\ s_j \in N(s_i)] \text{ for all } i$$

Do these determine joint distribution $[Y(s_1), \ldots, Y(s_m)]$?

If yes field is MRF.

BROOK'S LEMMA

Brook's lemma says "YES" if $N(s_i) \equiv D_{-i}$ for all i. More precisely it says if $m = 2$ for simplicity and we pick fix (y_{10}, y_{20}), for any (y_1, y_2).

$$[y_1, y_2] = \frac{[y_1 \mid y_2][y_2 \mid y_{10}]}{[y_{10} \mid y_2][y_{20} \mid y_{10}]}[y_{10}, y_{20}]$$

Left hand side proper means integration determines normalizing constant. But doesn't answer question for all MRFs. Need some new

concepts.

General result

- **Definition:** A **clique** is a set of cells or lattice indices such that each element is a neighbour of every other element
- **Definition:** A **potential function** Q of order k is a function of k arguments that is exchangeable in these arguments
  - **Example:** For binary (i.e. 0,1) data and k = 2, we take $Q(y_i, y_j) = I\{(y_i = y_j\} = y_i y_j + (1 - y_i)(l - y_j)$
- **Definition:** $p(y_1, \ldots, y_m)$ is a **Gibbs distribution** if [as function of $\{y_i\}$] it's product of potentials on cliques

LOCAL MODELLING

- All cliques of size 1 $\Leftrightarrow$ implies independence
- For cliques of size 2 $\Leftrightarrow$ common choice is

$$p(y_1, \ldots, y_m) \propto \exp\left[-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2 I\{i \sim j\}\right]$$

and therefore $[y_i \mid y_{-i}] = N(\sum_{j\in N(s_i)} y_j/m_i, \tau^2/m_i)$ where $m_i = \mid N(s_i) \mid$ is the number of neighbours of i

- **Hammersley-Clifford Theorem:** If MRF (i.e. $[y_i \mid y_j\ , j \in N(s_{-i})\ ]$ uniquely determines $p(y_1, \ldots, y_m))$ then the latter must be a Gibbs distribution.
- **Geman and Geman:** The converse: if we have a joint Gibbs distribution, then we have an MRF.

MARKOV RANDOM FIELDS - EXAMPLE

Features:

- $Y(s_i)$ = probability a health event for any individual in a region i with $m(s_i)$ susceptibles.
- $Z(s_i)$ = # of infecteds $\sim Bin(m(s_i), Y(s_i))$.
- $N(s_i)$ = all regions within fixed distance (e.g.48 km) of i. Conditional on $N(s_i)$, $Y(s_i)$ has beta distribution with parameters depending on counts in neighbours.
- parsimonious model but unclear how to include time

MARKOV RANDOM FIELDS: NOTES

PROS:

- elegant, simple mathematics + computational power
- may be useful component in hierarchical model

CONS:

- compatible joint distribution may not exist
- neighbours may be hard to specify
- a new site may not have neighbours for spatial prediction!
- conditional distributions may be hard to specify when "sites" are regions

CONDITIONAL AUTOREGRESSIVE MODEL (CAR)

Space not ordered like time. The conditional autoregressive approach (CAR) tries to emulate the AR approach. An MRF form. As before:

- $D = \{s_1, \ldots, s_m\}$ be the lattice
- $Y(s_i)$ be a response of interest
- $\mathbf{Y}_{-i}$ be all responses but $Y(s_i)$
- $N(s_i)$ be s_i neighbourhood

CAR model (Gaussian case):

$$Y(s_i) \sim N\left(\mu_i, \sigma_i^2\right), \text{ for all i}$$

with

$$E(Y(s_i)|\mathbf{Y}_{-i}) = \sum_{s_j \in N(s_i)} b_{ij} Y(s_j, t), \; Var(Y(s_i)|\mathbf{Y}_{-i}) = \tau_i^2$$

CONDITIONAL AUTOREGRESSIVE MODEL (CAR)

Does CAR necessarily determine a joint distribution

$$[Y(s_i), \ldots, Y(s_m)]?$$

Answer: Yes under reasonable conditions.

IMPLICATIONS

Brook's lemma implies:

$$p(y) = e^{-\frac{1}{2}y'D^{-1}(I-B)y}$$

with $y = y_{1:m}$ where $D = diag\{\tau_1^2, \ldots, \tau_m^2\}$ & $B = \{b_{ij}\}$. Note that

$D^{-1}(I - B)$ must be symmetric so for all i, j

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}$$

meaning that B is not symmetric! Also note that $Cov(Y) = (I - B)^{-1}D$.

IAR: INTRINSIC AUTOREGRESSION

Much flexibility exists in choice of B. But natural choice is $B = W$ with $w_{ij} = 0$ *or* 1 for an adjacency matrix. Yet that would not be an allowable. Curiously $b_{ij} = w_{ij}/w_{i+}$ works & gives

$$p(y_i \mid y_{-i}) = N(\sum_j w_{ij} y_j / w_{i\cdot}, \tau_i^2 / w_{i+})$$

with $w_{i\cdot} = \sum_j w_{ij}$ while

$$p(y) = e^{-\frac{1}{2} y'(D_w - B)y}$$

where $D = diag\{w_{1+}, \ldots, w_{m+}\}$ and hence $Cov(Y)^{-1} = D_w - B$

IAR: INTRINSIC AUTOREGRESSION

However

$$(D_w - B)\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 0$$

so:

1. the inverse of the covariance matrix is singular
2. the covariance is undetermined
3. the probability distribution is not integrable.

IAR: INTRINSIC AUTOREGRESSION

More explicitly,

$$p(y) \propto \exp\left[-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2 w_{ij}\right]$$

which is non-integrable. An example where natural & proper local dependence models do not yield proper joint distribution. Meaning Y does not have stochastic generator, MCMC cannot be used, and so on. This model has been called the *intrinsic autoregression model* which de facto means a model concentrated on a lower dimensional space say where $Y_\cdot = 0$. Modellers use it despite issues.

FIXING THE IAR

Choose $D_w - \rho W$ instead with $\rho < 1$.

But now,

$$p(y_i \mid y_{-i}) = N(\rho \sum_j w_{ij} y_j / w_{i+}, \tau_i^2 / w_{i+})$$

so conditional mean is fraction of neighbourhood mean. Makes interpretation and inference challenging (ρ is an extra parameter). Further even with ρ large say 0.95, Moran's I is small (around 0.25) in simulated samples. So fix is unappealing.

Situation resembles AR(1) as the autocorrelation goes to 1 - model flips from AR (a stationary process) to a random walk (a non-stationary process).

THE REVISED IAR

PROS:

- makes distribution proper
- adds parametric flexibility
- $\rho = 0$ interpretable as independence

CONS:

- hard to rationalize model with Y_i's conditional expectation a fraction of neighbour average – spatial interpretation?
- interpretation of ρ? As correlation seems tenuous since
  - $\rho = 0.80$ yields $0.1 < Moran's I < 0.15$
  - $\rho = 0.90$ yields $0.2 < Moran's I < 0.25$
  - $\rho = 0.99$ yields $Moran's I < 0.5$

CAR Note:

Spatial prediction with CAR is ad hoc using:

$$p(y_0 \mid y) = N(\sum_j w_{0j} y_j / w_{0+}, \tau^2 / w_{0+})$$

Well defined but not a CAR! That is it could not arise by application of Brook's lemma.

CAR IN THE NON-GAUSSIAN CASE

The CAR theory extends to the non–Gaussian case as the following example shows.

The following hierarchical model induces a CAR structure.

- **Measurement model:**

$$Z(s_i) \sim \; ind \; Poi(\exp[Y(s_i)])$$

- **Process model:**

$$[\mathbf{Y}|\boldsymbol{\beta}, \tau^2, \phi] = Gau(\mathbf{X}\boldsymbol{\beta}, \Sigma[\tau^2, \phi])$$

  where $\mathbf{X}$ represents site specific covariates or factors & $\Sigma[\tau^2, \phi]$ the CAR neighbourhood structure.

- **Parameter model:** $[\boldsymbol{\beta}, \tau^2, \phi]$

SIMULTANEOUS AUTOREGRESSION (SAR)

This natural model is like a CAR:

$$Y(s_i) - \mu(s_i) = \sum_j b_{ij}(Y(s_j) - \mu(s_j)) + \epsilon_i$$

where $\epsilon_i \sim indN(0, \sigma_i^2)$. In vector matrix form:

$$Y - \mu = B(Y - \mu) + \epsilon$$

or

$$Y = \mu + \epsilon^*$$

where $\epsilon^* \sim N_m(0, (I - B)^{-1}\Sigma(I - B')^{-1})$ with $\Sigma = diag\{\sigma_1^2, \ldots, \sigma_m^2\}$.

This model capture spatial independence through the mean structure - a moving average of the $\{\epsilon_i\}$.

SAR IN EXTENDED FORM

- **Data model:** $[Z(s_i) \mid Y(s_i), \sigma_\epsilon] = indN(Y(s_i), \sigma_\epsilon)$
- **Process model:** $[Y \mid \beta, \sigma^2, \rho] = N(X\beta, \sigma^2(I - \rho W')(I - \rho W))$ where W has zeros down the diagonal but need not be the adjacency matrix.
- **Parameter model:** Prior distribution on the parameters.

A large class of models. Can see the affect of covariates on the process Y. CAR can also incorporate the $X\beta$ type model.

NOTE ON MISALIGNED DATA

Different responses measured at monitoring sites in a systematic way. We call unmeasured complements at each site.

systematically missing. Often these unmeasured values are predicted from the others at different sites.

Change of support means data measured at different resolutions, e.g. some at a county level, some at point locations.

NOTES ON AREAL DATA

Sometimes areal data can profitably be modelled as an aggregate of individual data.

- Can reflect greater uncertainty due to variation within areas.
- Was used to explore the ecological effect and develop model that avoids it.

Disease Mapping

Disease mapping has a long history in epidemiology, and may be defined as the estimation and presentation of summary measures of health outcomes.

The aims of disease mapping include

- simple description,
- hypothesis generation,
- allocation of health care resources, assessment of inequalities, and
- estimation of background variability in underlying risk in order to place epidemiological studies in context.

Disease Mapping

Unfortunately there are well-documented difficulties with the mapping of raw estimates since, for small areas and rare diseases in particular, these estimates will be dominated by sampling variability.

For the model

$$Y_i \sim \text{Poisson}(E_i\theta_i)$$

the MLE is

$$\widehat{\theta}_i = \text{SMR}_i = \frac{Y_i}{E_i}$$

with variance

$$\text{var}(\widehat{\theta}_i) = \frac{\theta_i}{E_i}$$

so that areas with small E_i have high associated variance.

EXAMPLE: SCOTTISH LIP CANCER

The Figure on the next slide shows the SMRs for the Scottish lip cancer data, and indicates a large spread with an increasing trend in the south-north direction.

The variance of the estimate is $\text{var}(\text{SMR}_i) = \text{SMR}_i/E_i$, which will be large if E_i is small.

For the Scottish data the expected numbers are highly variable, with range 1.1–88.7. This variability suggests that there is a good chance that the extreme SMRs are based on small expected numbers (many of the large, sparsely-populated rural areas in the north have high SMRs).

EXAMPLE: SCOTTISH LIP CANCER

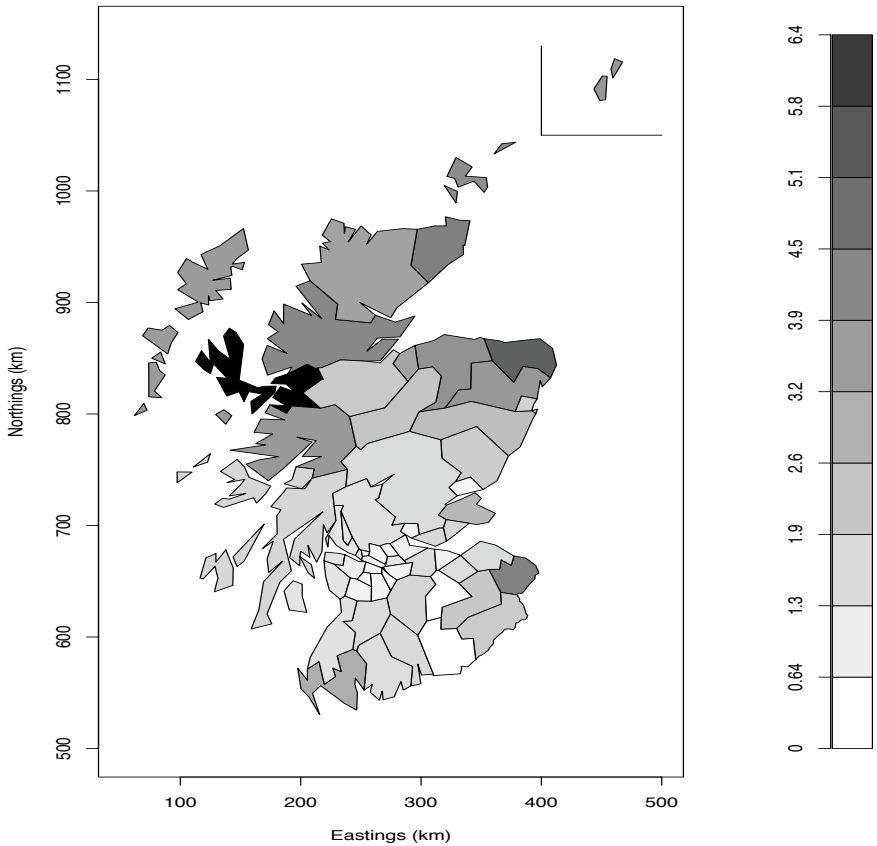


Figure: SMRs in 56 counties of Scotland.

Smoothing Models

The above considerations led to methods being developed to *smooth* the SMRs using hierarchical/random effects models that use the data from the totality of areas to provide more reliable estimates in each of the constituent areas.

We first describe models that do not use spatial information before turning to models that exploit both spatial and non-spatial information.

Poisson-Gamma Model Without Covariates

We begin by describing a simple Poisson-Gamma two-stage model that offers analytic tractability and ease of estimation.

We assume there are no covariates and assume the first stage likelihood is given by

$$Y_i|\theta_i, \beta \sim_{ind} \text{Poisson}\left(\mu E_i \theta_i\right), \tag{1}$$

where μ is the overall relative risk, and reflects differences between the reference rates and the rates in the study region.

At the second stage the random effects θ_i are assigned a distribution. We initially assume that across the map the deviations of the relative risks from the mean, μ, are modelled by

$$\theta_i|\alpha \sim_{iid} \text{Ga}(\alpha, \alpha), \tag{2}$$

a gamma distribution with mean 1, and variance $1/\alpha$.

Poisson-Gamma Model Without Covariates

The advantage of this Poisson-gamma formulation is that the marginal distribution of $Y_i|\mu, \alpha$ (obtained by integrating out the random effects θ_i), is negative binomial.

Marginally, the mean and variance are given, respectively, by

$$\begin{aligned} E[Y_i|\mu,\alpha] &= E_i\mu \\ \text{var}(Y_i|\mu,\alpha) &= \text{E}[Y_i|\mu,\alpha](1+\text{E}[Y_i|\mu,\alpha]/\alpha), \end{aligned} \tag{3}$$

so that the variance increases as a quadratic function of the mean, and the scale parameter α can accommodate different levels of "overdispersion".

This form is substantively more reasonable than the naive Poisson model; it is important to consider excess-Poisson variability resulting from unmeasured confounders, data anomalies in numerator and denominator, and model misspecification.

EXAMPLE: DISEASE MAPPING FOR SCOTLAND

We make use of a mapping function that is on the course website:
`PrettyPoly <- function(y, poly, nrepeats, ncut=1000, nlevels=10, lower=NULL, upper=NULL )` with arguments:

- `y` the variable to be mapped
- `poly` the $x - y$ coordinates of the polygons, with different polygons separated by `NA`s.
- `nrepeats` a vector of the same length as `y` with each entry containing the number of repeats of the appropriate entry in `y`.
- `ncut` The number of grey-scale levels to convert `y` to.
- `nlevels` The number of grey levels to plot.
- `lower` The value (on the same scale as `y`) that white is assigned to.
- `upper` The value (on the same scale as `y`) that black is assigned to.

EXAMPLE: DISEASE MAPPING FOR SCOTLAND

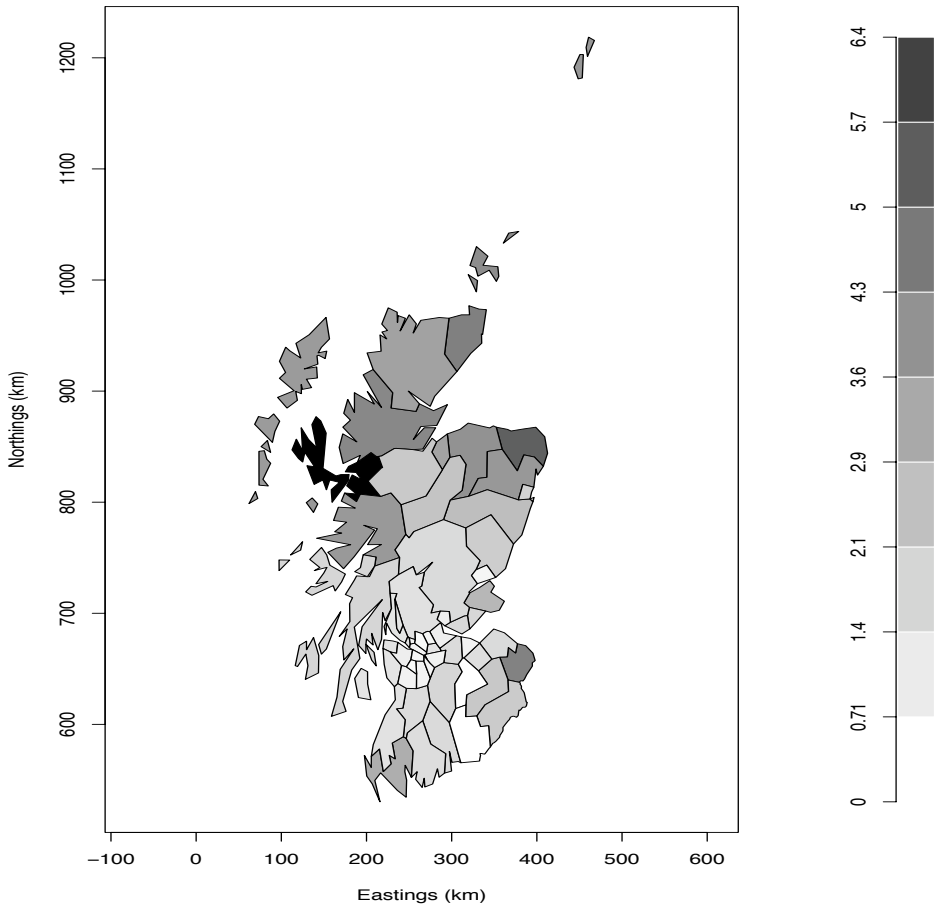


Figure: SMRs for Scottish counties.

Example: Disease Mapping for Scotland

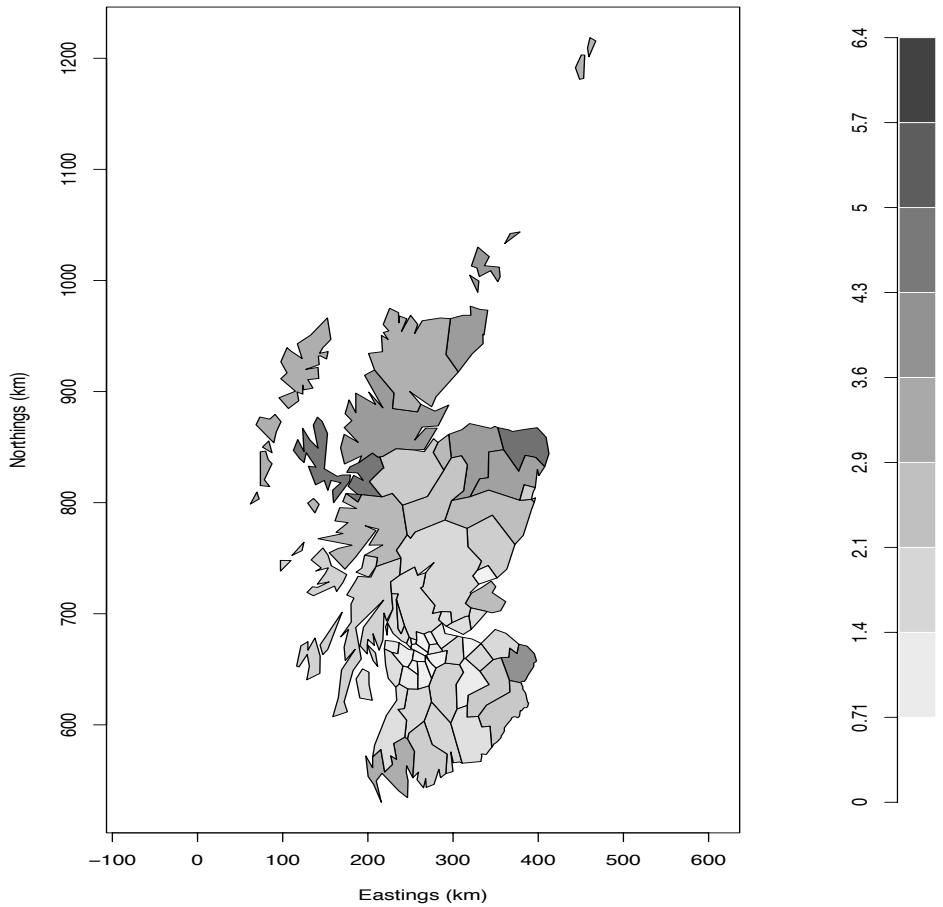


Figure: Posterior mean estimates for Scottish counties.

Poisson-Gamma Model

We now carry out a fully Bayesian analysis of the model for which empirical Bayes was used previously:

$$\begin{aligned} Y_i|\theta_i, \beta_0 &\sim \text{Poisson}(E_i e^{\beta_0}\theta_i) \\ \theta_i &\sim \text{Ga}(\alpha, \alpha) \end{aligned}$$

We require priors for β_0 and α. For example:

$$\begin{aligned} \beta_0 &\sim \text{N}(m, v) \\ \alpha &\sim \text{Ga}(a, b) \end{aligned}$$

with m, v, a, b picked to reflect beliefs about β_0 and α.

MCMC ANALYSIS OF THE POISSON-GAMMA MODEL

In the example that follows we specify a flat prior for β_0, and a Ga(1,1) prior for α.

The iterative algorithm is run for 10,000 iterations, with the first 4,000 discarded as "burn-in".

We summarize the posteriors for the relative risks:

$$\text{RR}_i = \exp(\beta_0)\theta_i$$

and for β_0 and α. The posterior mean for β_0 is 0.36, compared to 0.35 under empirical Bayes, and the posterior mean for α is 1.79, compared to 1.88 under empirical Bayes.

Similarly the posterior means and posterior medians agree very closely.

POISSON-LOGNORMAL MODEL

The Poisson-gamma model offers analytic tractability, but does not easily allow the incorporation of spatial random effects.

A Poisson-lognormal non-spatial random effect model is given by:

$$Y_i|\beta, V_i \sim_{ind} \text{Poisson}(E_i\mu_i e^{V_i}) \quad V_i \sim_{iid} N(0, \sigma_v^2) \tag{4}$$

where V_i are area-specific random effects that capture the residual or unexplained (log) relative risk of disease in area i, $i = 1, ..., n$. Whereas in the Poisson-Gamma model we have $\theta \sim \text{Ga}(\alpha, \alpha)$, here we have $\theta = e^{V_i} \sim \text{LogNormal}(0, \sigma^2)$.

Model (**??**) does not give a marginal distribution of known form, but does naturally lead to the addition of spatial random effects.

The marginal variance is of the same quadratic form as (**??**). Empirical Bayes is not so convenient for this model, and so we resort to a fully Bayesian approach for which we need to specify prior distributions.

Prior Choice for Non-Spatial Model

We need to specify priors for:

- The regression coefficients $\boldsymbol{\beta}$.
- The variance of the random effets σ_v^2.

For a rare disease, a log-linear link is a natural choice:

$$\log \mu(\mathbf{x_i}, \boldsymbol{\beta}) = \beta_\mathbf{0} + \sum_{\mathbf{j=1}}^{\mathbf{J}} \beta_\mathbf{j} \mathbf{x_{ij}},$$

where x_{ij} is the value of the j-th covariate in area i.

For regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)$, an improper prior

$$p(\boldsymbol{\beta}) \propto 1$$

may often be used, but in very circumstances such a choice may lead to an improper posterior.

If there are a large numbers of covariates, or high dependence amongst the elements of $\mathbf{x}$, then more informative priors will be beneficial.

- In general we might expect residual relative risks in areas that are "close" to be more similar than in areas that are not "close".
- We would like to exploit this information in order to provide more reliable relative risk estimates in each area.
- This is analogous to the use of a covariate x, in that areas with similar x values are likely to have similar relative risks.
- Unfortunately the modelling of spatial dependence is much more difficult since spatial location is acting as a surrogate for unobserved covariates.
- We need to choose an appropriate spatial model, but do not directly observe the covariates whose effect we are trying to mimic.

We first consider the model

$$Y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, U_i, V_i \sim_{ind} \text{Poisson}(E_i \mu_i e^{U_i + V_i})$$

with

$$\log \mu_i = g(\mathbf{S}_i, \boldsymbol{\gamma}) + f(\mathbf{x_i}, \boldsymbol{\beta}), \tag{5}$$

where

- $\mathbf{S}_i = (S_{i1}, S_{i2})$ denotes spatial location, the centroid of area i,
- $f(\mathbf{x_i}, \boldsymbol{\beta})$ is a regression model,
- $g(\mathbf{S}_i, \boldsymbol{\gamma})$ is an expression that we may include to capture large-scale spatial trend – the form

$$f(\mathbf{S}_i) = \gamma_1 S_{i1} + \gamma_2 S_{i2},$$

  is a simple way of accommodating long-term spatial trend.
- The random effects $V_i \sim_{iid} N(0, \sigma_v^2)$ represent non-spatial overdispersion,
- U_i are random effects with spatial structure. We describe two forms.

A Joint Model

- Assume that $\mathbf{U} = (U_1, ..., U_n)$ arise from a zero mean multivariate normal distribution with variances $\text{var}(U_i) = \sigma_u^2$ and correlations $\text{corr}(U_i, U_j) = \exp(-\phi d_{ij}) = \rho^{d_{ij}}$ where d_{ij} is the distance between the centroids of areas i and j, and $\rho > 0$ is a parameter that determines the extent of the correlation.
- This model is *isotropic* since it assumes that the correlation is the same in all spatial directions. We refer to this as the *joint* model, since we have specified the joint distribution for $\mathbf{U}$.
- More generally the correlations can be var as $\text{corr}(U_i, U_j) = \exp(-(\phi d_{ij})^\kappa)$.

A Conditional Model

- An alternative approach is to specify the distribution of each U_i as if we knew the values of the spatial random effects U_j in "neighbouring areas"
- We need to specify a rule for determining the "neighbours" of each area.
- Spatial models that start with the n area-specific residual spatial random effects all suffer from a level of arbitrariness in their specification – in an epidemiological context the areas are not regular in shape (as opposed to images for example, which are on a regular grid).
- To define *neighbours*, a number of authors have taken the neighbourhood scheme to be such that areas i and j are taken to be neighbours if they share a *common boundary*. This is reasonable if all regions are of similar size and arranged in a regular pattern (as is the case for pixels in image analysis where these models originated), but is not particularly attractive otherwise.

- Various other neighbourhood/weighting schemes are possible.
- We could take the neighbourhood structure to depend on the distance between area centroids and determine the extent of the spatial correlation (i.e. the distance within which regions are considered neighbours).
- In typical applications it is difficult to assess whether the spatial model chosen is appropriate, which argues for a simple form, and to assess the sensitivity of conclusions to different choices

THE ICAR MODEL

- A common model is to assign the spatial random effects an intrinsic conditional autorgressive (ICAR) prior.
- Under this specification it is assumed that

$$U_i | U_j, j \in \partial_i \sim N\left(\overline{U}_i, \frac{\omega_u^2}{m_i}\right),$$

where ∂i is the set of neighbours of area i, m_i is the number of neighbours, and $\overline{U}_i$ is the mean of the spatial random effects of these neighbours.

- The parameter ω_u^2 is a conditional variance and its magnitude determines the amount of spatial variation.
- The variance parameters σ_v^2 and ω_u^2 are on different scales, σ_v is on the log odds scale while ω_u is on the log odds scale, *conditional* on $U_j, j \in \partial i$; hence they are not comparable (in contrast to the joint model in which σ_u is on the same scale as σ_v).

THE ICAR MODEL

- Notice that if ω_u^2 is "small" then although the residual is strongly dependent on the neighbouring value the overall contribution to the residual relative risk is small.
- This is a little counterintuitive but stems from spatial models having two aspects, strength of dependence and total amount of spatial dependence, and in the ICAR model there is only a single parameter which controls both aspects.
- In the joint model the strength is determined by ρ and the total amount by σ_u^2. A non-spatial random effect should always be included along with the ICAR random effect since this model cannot take a limiting form that allows non-spatial variability; in the joint model with U_i only, this is achieved as $\rho \to 0$. If the majority of the variability is non-spatial, inference for this model might incorrectly suggest that spatial dependence was present.

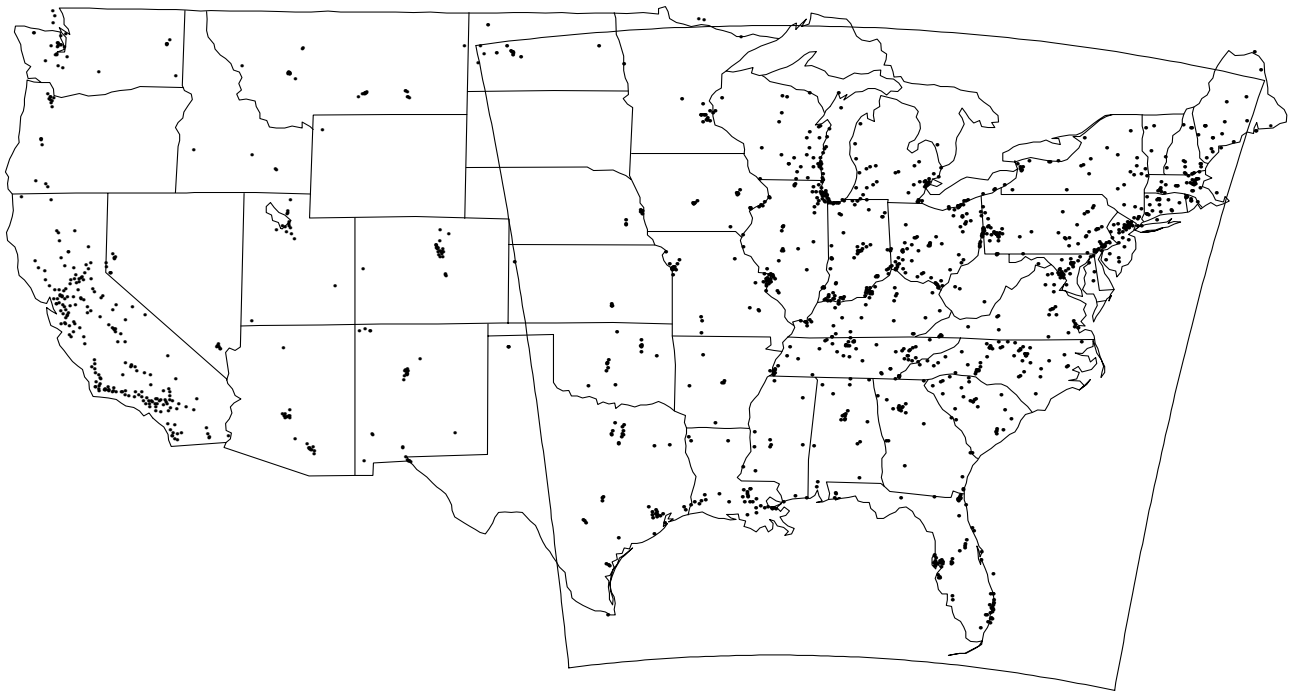
(a) SMR estimates

(b) Smoothed estimates

Figure: Raw and smoothed estimates in 56 counties of Scotland.

Point Referenced Spatial Processes

EXAMPLE: US OZONE MONITORING SITES



RANDOM FIELD

A random process $Y(s),\ s \in D \subset R^d$ **for some** d

Usually $d = 2$. Interest focuses on stochastic inter-site spatial dependence (correlation in the case of Gaussian fields) between $Y(s_1)$ and $Y(s_2)$.

Why?

CORRELATION: IS YOUR ENEMY!

Suppose $Y(s_i) = \mu + W(s_i),\ i = 1, \ldots, p$ where for any two sites $corr[W(s_1), W(s_2)] = 0.97$. A naive statistician might take

$$\bar{Y} \pm 1.96 \frac{s}{\sqrt{p}}$$

as a 95% CI. But strong correlation effectively reduces the sample size to $p = 1$. It makes the CI much larger. Of particular concern in spatial regression where $Y = X\beta + \epsilon$ where Y is sample of measurements made at various locations in a random field.

CORRELATION IS YOUR FRIEND!

Strong intersite correlation enables strength to be "borrowed". Measurements at a few sites can be used to predict the rest.

Bad and good correlation has thus led to an explosion of interest in stochastic models for random fields.

Spatial modelling: start with EDA! Case study 1

The Maas or Meuse: major European river. Rises in France. Flows through Belgium & the Netherlands. Draining into North Sea. ITotal length of 925 km. Has been monitored over time.

Spatial modelling: start with EDA! Case study 1

Various chemical measurements are stored in a dataset found in the gstat package.

```
> library(gstat)
> data(meuse)
> str(meuse)
```

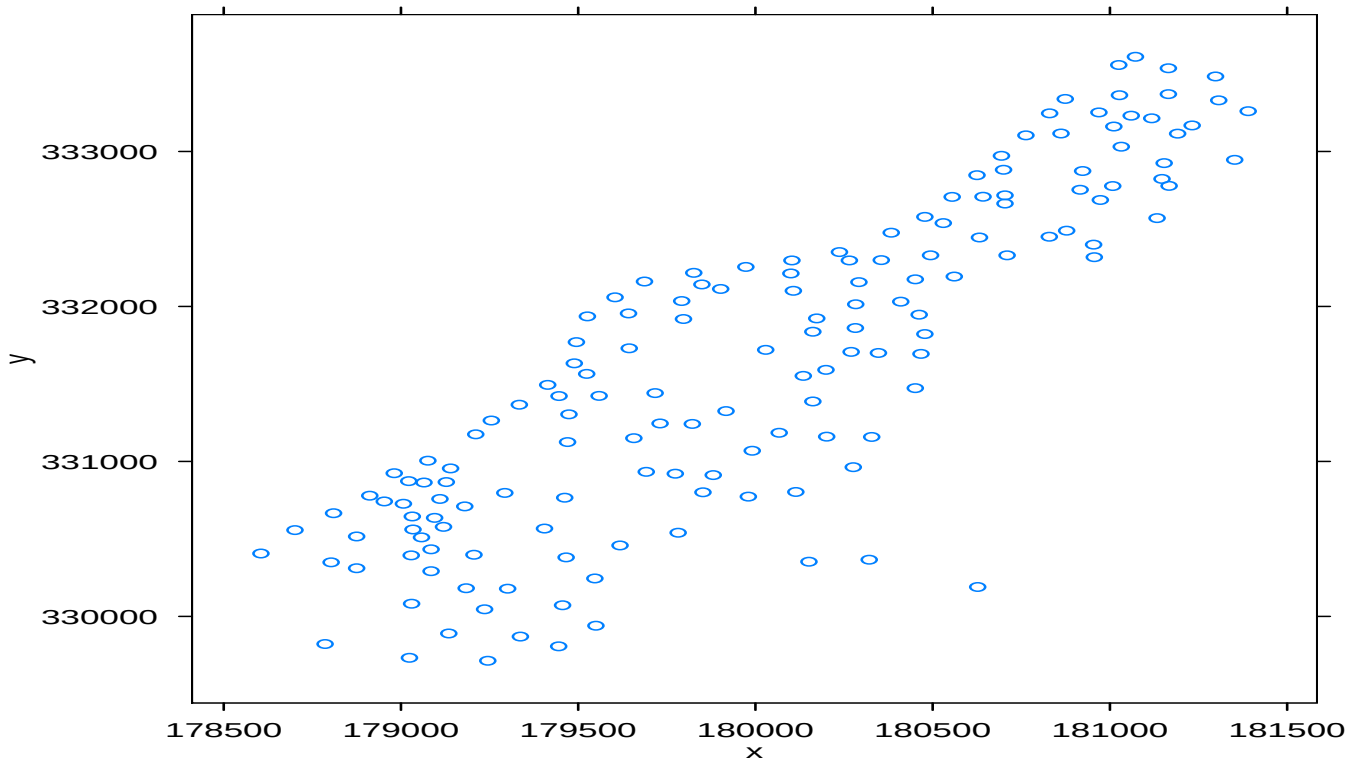
SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

```
'data.frame':   155 obs. of  13 variables:
 $ x       : num  181072 181025 181165 181298 181307 ...
 $ y       : num  333611 333558 333537 333484 333330 ...
 $ cadmium: num  11.7 8.6 6.5 2.6 2.8 3 3.2 2.8 2.4 1.6 ...
 $ copper : num  85 81 68 81 48 61 31 29 37 24 ...
 $ lead    : num  299 277 199 116 117 137 132 150 133 80 ...
 $ zinc    : num  1022 1141  640  257  269 ...
 $ elev    : num  7.91 6.98 7.80 7.66 7.48 ...
 $ dist    : num  50 30 150 270 380 470 240 120 240 420 ...
 $ om      : num  13.6 14 13 8 8.7 7.8 9.2 9.5 10.6 6.3 ...
 $ ffreq   : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ soil    : Factor w/ 3 levels "1","2","3": 1 1 1 2 2 2 2 1 1 2 ...
 $ lime    : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
 $ landuse: Factor w/ 15 levels "Aa","Ab","Ag",..: 4 4 4 11 4 11 4 2 2 15 ...$
```

SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

First of all: inspect sampling locations.

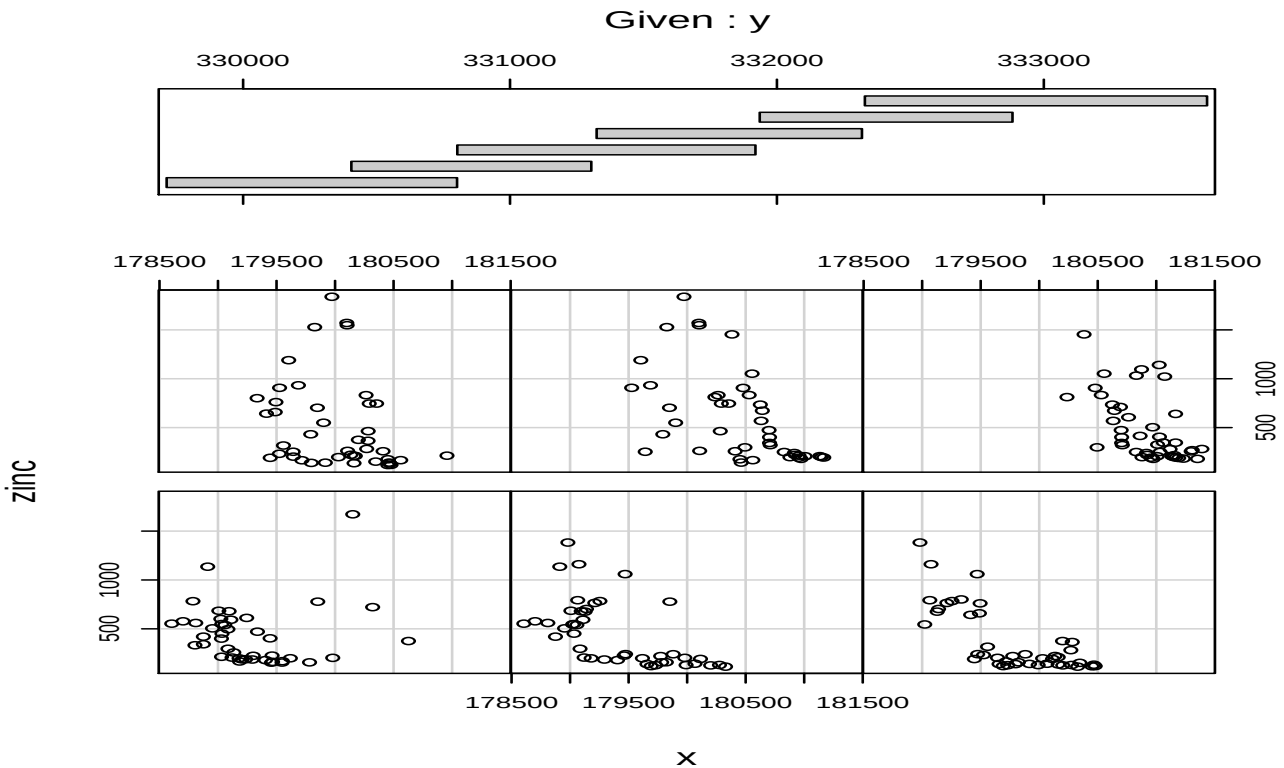
```
> print(xyplot(y ~ x, data = meuse))
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

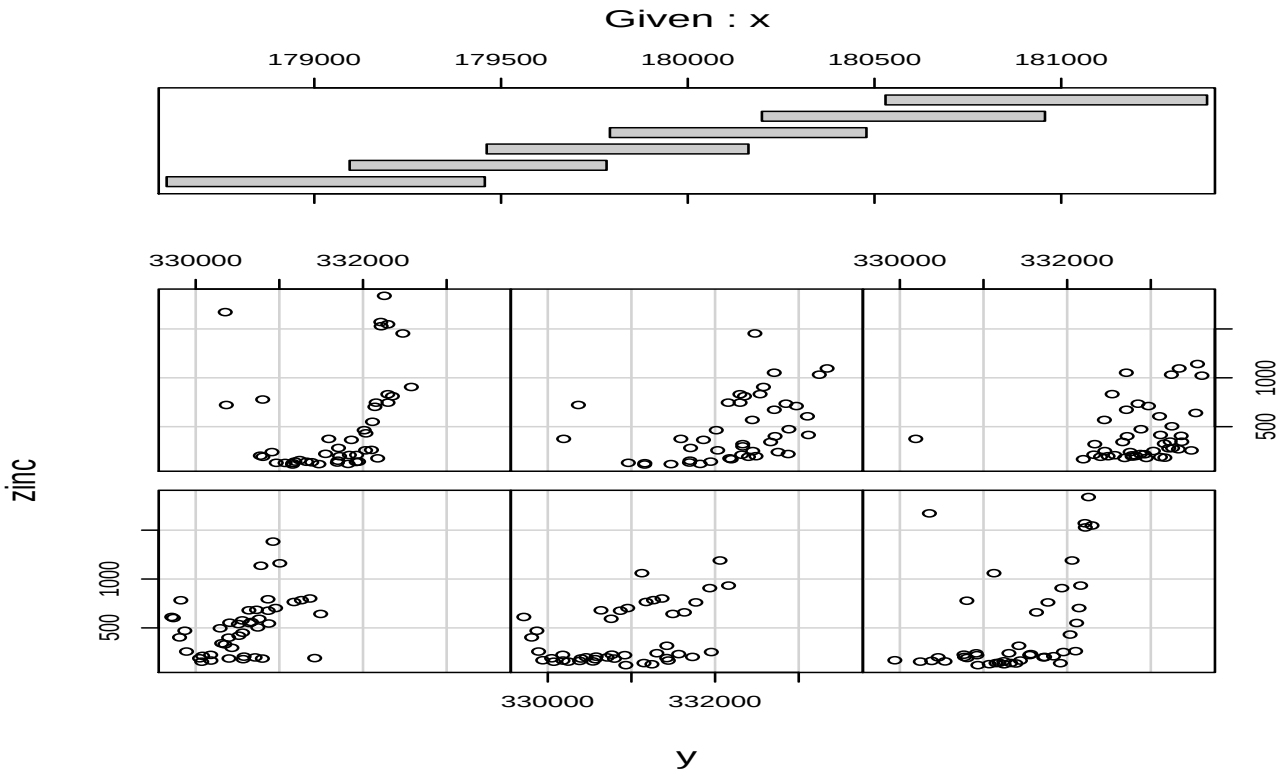
Ad hoc checks for non-stationarity can be done. Conditioning plots are one such approach.

```
> coplot(zinc ~ x | y, data = meuse)
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

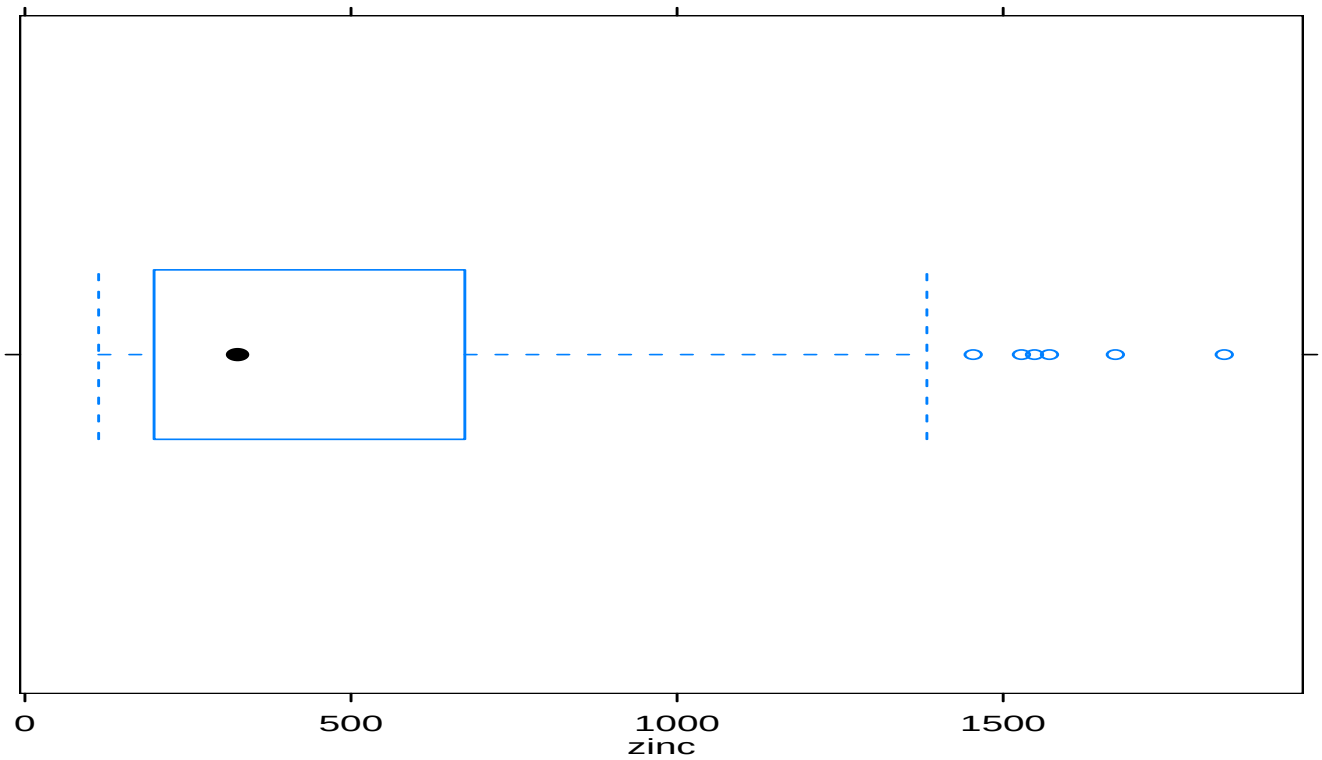
```
> coplot(zinc ~ y | x, data = meuse)
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

Next distributional checks, e.g. "Box and Whisker plot".

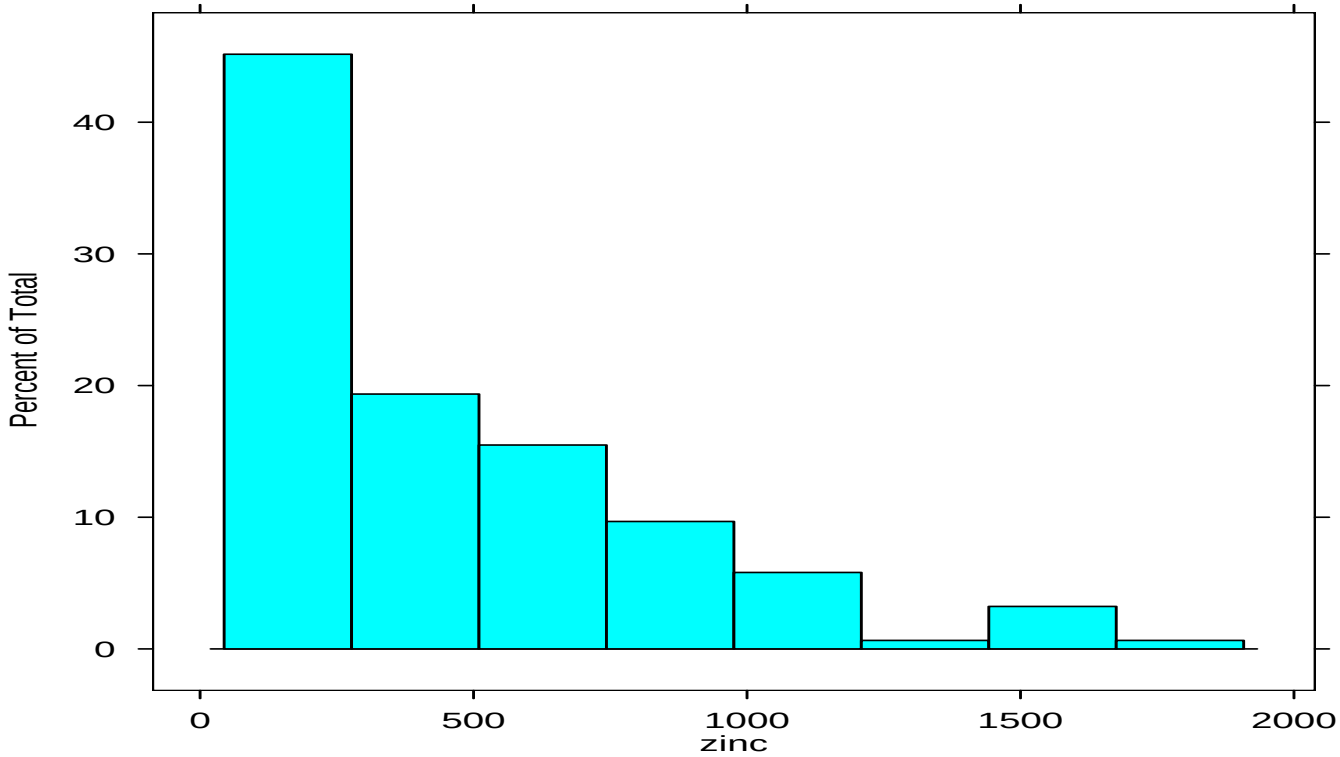
```
> print(bwplot(~zinc, data = meuse))
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

Or a histogram.

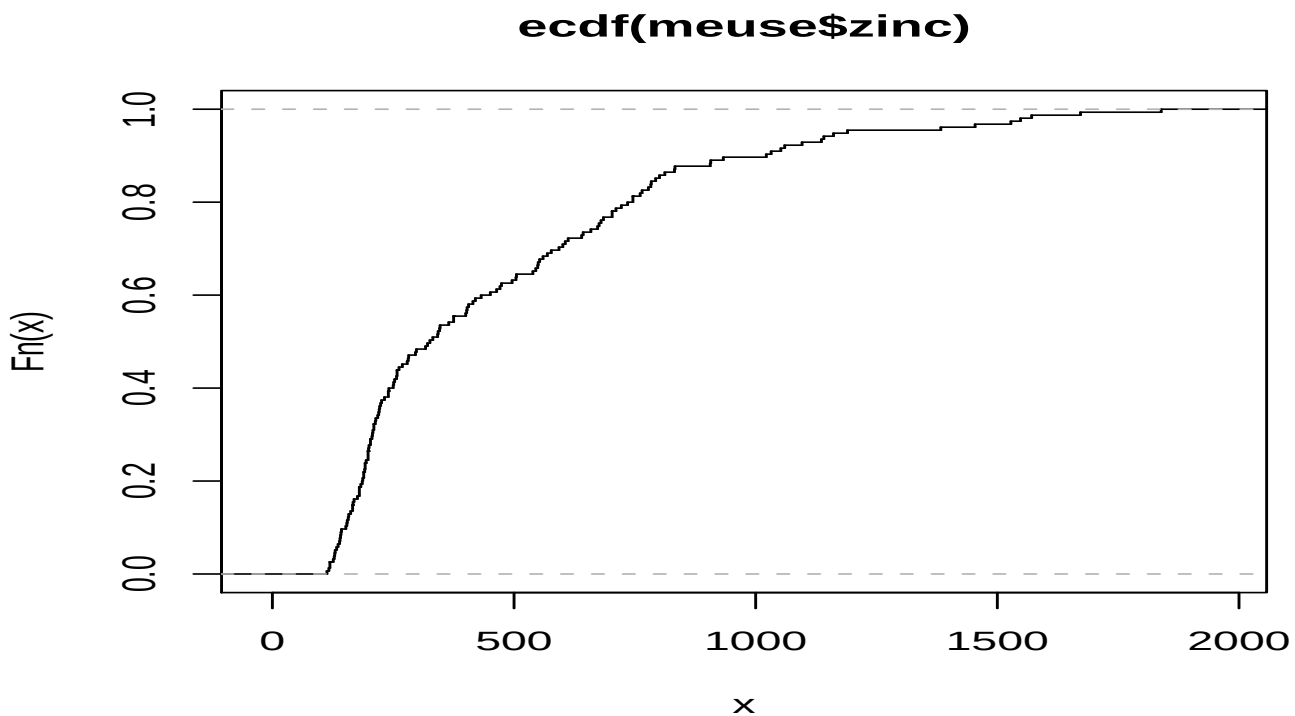
```
> print(histogram(~zinc, data = meuse))
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

The empirical cdf

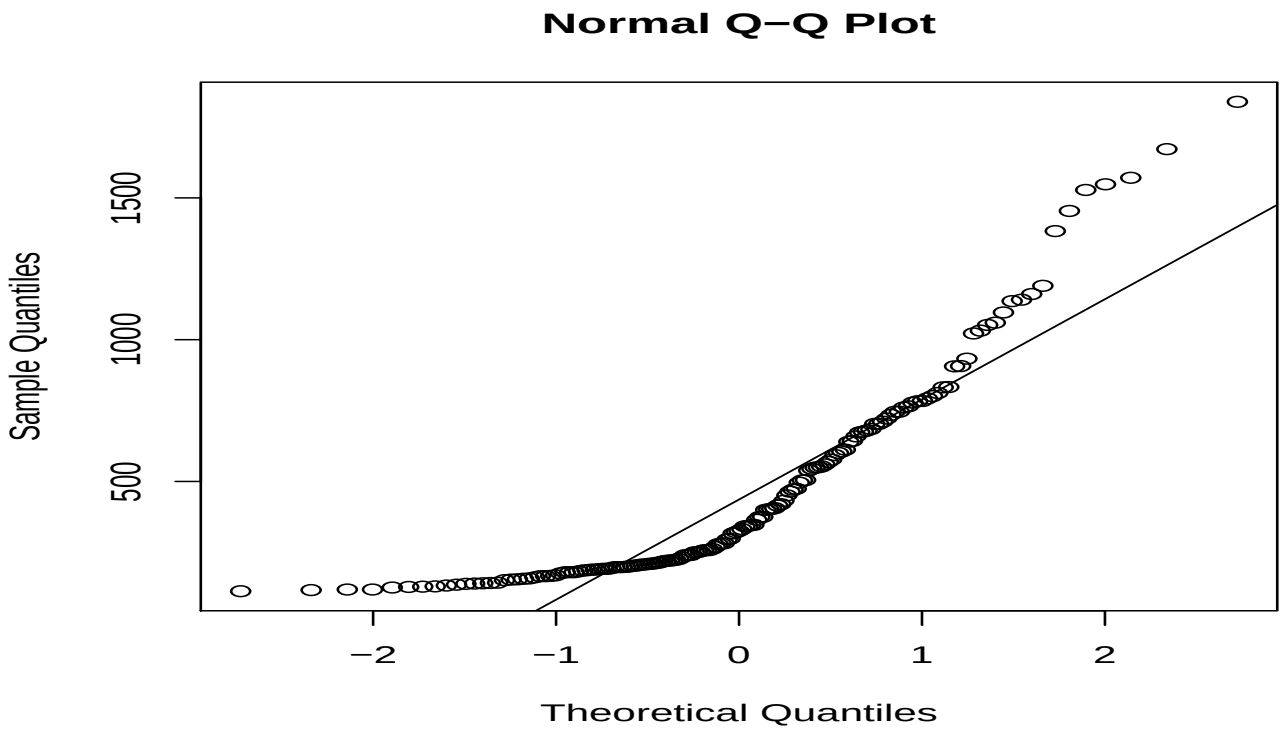
```
> library(stepfun)
> cdf.zinc <- ecdf(meuse$zinc)
> plot(cdf.zinc, verticals = T, do.points = F)
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

The q–q (quantile-quantile) plot

```
> qqnorm(meuse$zinc)
> qqline(meuse$zinc)
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

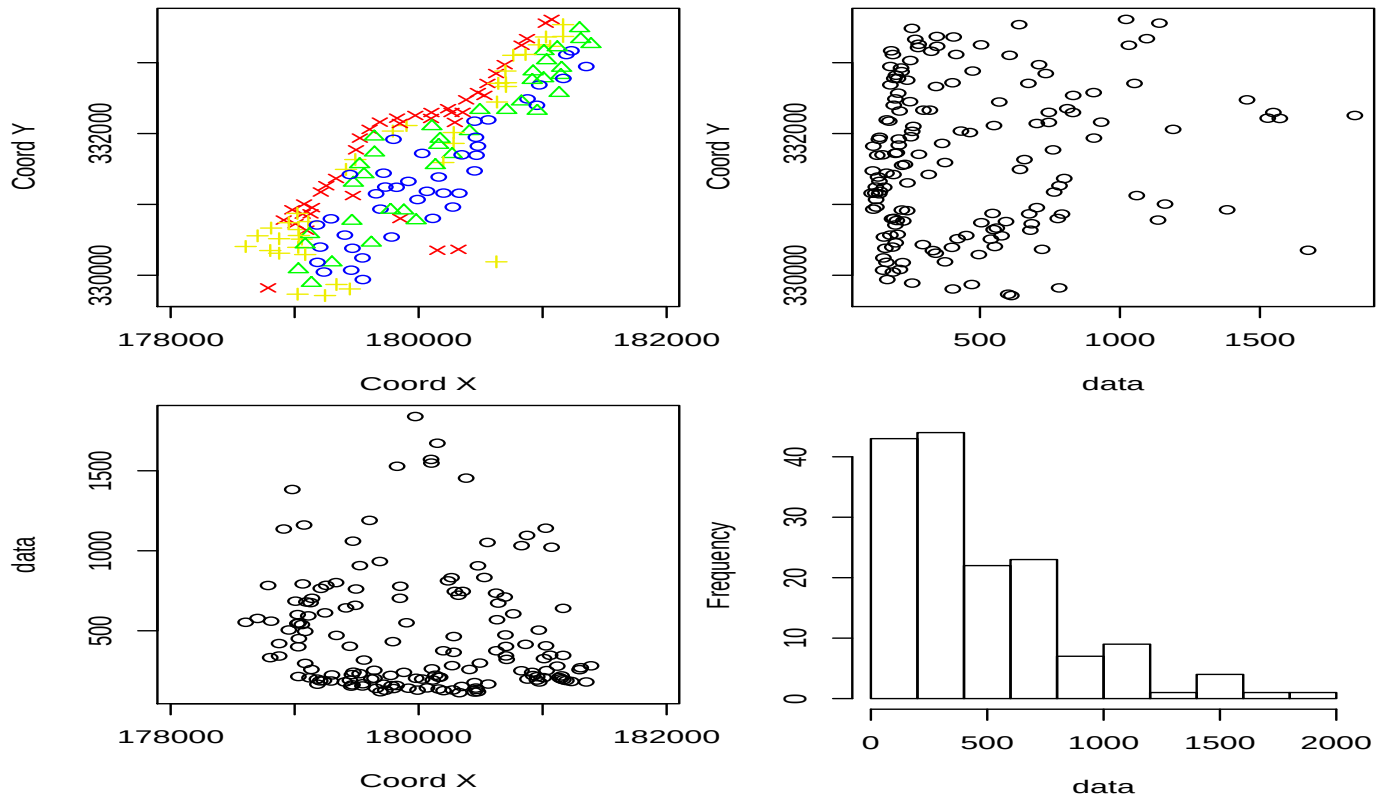
geoR alternative to gstat has nice plot function. But data must be converted to a geodata object from the meuse dataset, a dataframe object.

```
> library(geoR, warn = F)
> meuse.geo <- as.geodata(meuse, data.col = 6)

Loading required package: mva
```

SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

```
> plot(meuse.geo)
```



SPATIAL MODELLING: START WITH EDA! CASE STUDY 1

CONCLUSION: Log zinc fitted reasonably well with a student t distribution - very heavy tails.

SOME THEORY: MOMENTS

$Y \sim F$: random vector field. Fixed time t omitted in sequel. s and x commonly used for spatial coordinates, e.g. (lat, long). We use s. For locations $\{s_1, \ldots, s_g\}$ for any g

$$F_{s_1,\ldots,s_g}(y_1, \ldots, y_g) \equiv P\{Y(s_1) \leq y_1, \ldots, Y(s_g) \leq y_g\}.$$

$F_{s_1,\ldots,s_g}(y)$ is joint distribution distribution (DF)

- **Moment** of k^{th}*-order:*

$$E[Y(s)]^k \equiv \int y^k dF_s(y)$$

SOME THEORY

- **Expectation**: If exists, defined as the 1^{st}-order moment for any s

$$\mu(s) \equiv E[Y(s)]$$

- **Variance:**

$$Var[Y(s)] \equiv E[Y(s) - \mu(s)]^2.$$

- **Covariance** between locations s_1 & s_2,

$$C(s_1, s_2) \equiv E[(Y(s_1) - \mu(s_1))(Y(s_2) - \mu(s_2))]$$

NOTE: $C(s_1, s_1) \equiv Var[Y(s_1)]$

SOME THEORY: STATIONARITY

An important concept in characterizing the random field Y

- **Strict stationarity** Y *strictly stationary* if:

$$F_{s_1,\ldots,s_n}(\mathbf{y}) = F_{s_1+h,\ldots,s_n+h}(\mathbf{y})$$

  for any vector h & an arbitrary n

- **Second–order stationarity** Y is *second-order stationary* if:

$$\begin{array}{lcccl} \mu(s) & = & E[Y(s)] & = & \mu \\ C(s, s+h) & = & C(s+h-s) & = & C(h) \end{array}$$

  when $h = 0$: $Var[Y(s)] = C(s,s) = C(0)$
  ie. ***Mean, Variance do not depend on location***

SOME THEORY: STATIONARITY

- **Second–order stationarity - cont'd**
  $C(h)$: *covariogram* (or *autocovariance* in time series) implies ***Intrinsic Stationarity*** *(weaker)*

$$\begin{aligned} Var[Y(s) - Y(s+h)] &= Var[Y(s)] + Var[Y(s+h)] \\ &\quad -2Cov[Y(s), Y(s+h)] \\ &= C(0) + C(0) - 2C(h) \\ &= 2[C(0) - C(h)]. \end{aligned}$$

or equivalently semi-variogram

$$\gamma(h) = C(0) - C(h)$$

PROPERTIES OF $C(h)$

X second-order stationary process with covariance function $C(h)$.

- **Positive Definiteness (PD):** If $\Sigma = \{C(h_{ij})\}$ being covariance matrix of random vector $(Y(s_1), \ldots, Y(s_n))$ makes it PD implying for any vector a that:

$$\sum_i \sum_j a_i a_j C(h_{ij}) > 0$$

- **Anisotropy**: $C(h)$ - function of length & direction
- **Isotropy**: $C(h)$ - function only of length $|h|$

VARIOGRAMS

Matheron supposed that at least for small $| h |$

$$E[Y(s+h) - Y(s)] = 0$$

would be reasonable assumption. He then defined the

- **Variogram:**

$$\begin{aligned} 2\gamma(h) &\equiv var[Y(s+h) - Y(s)] \\ &= E[Y(s+h) - Y(s) - (\mu(s+h) - \mu(s))]^2. \\ &= E[Y(s+h) - Y(s)]^2. \end{aligned}$$

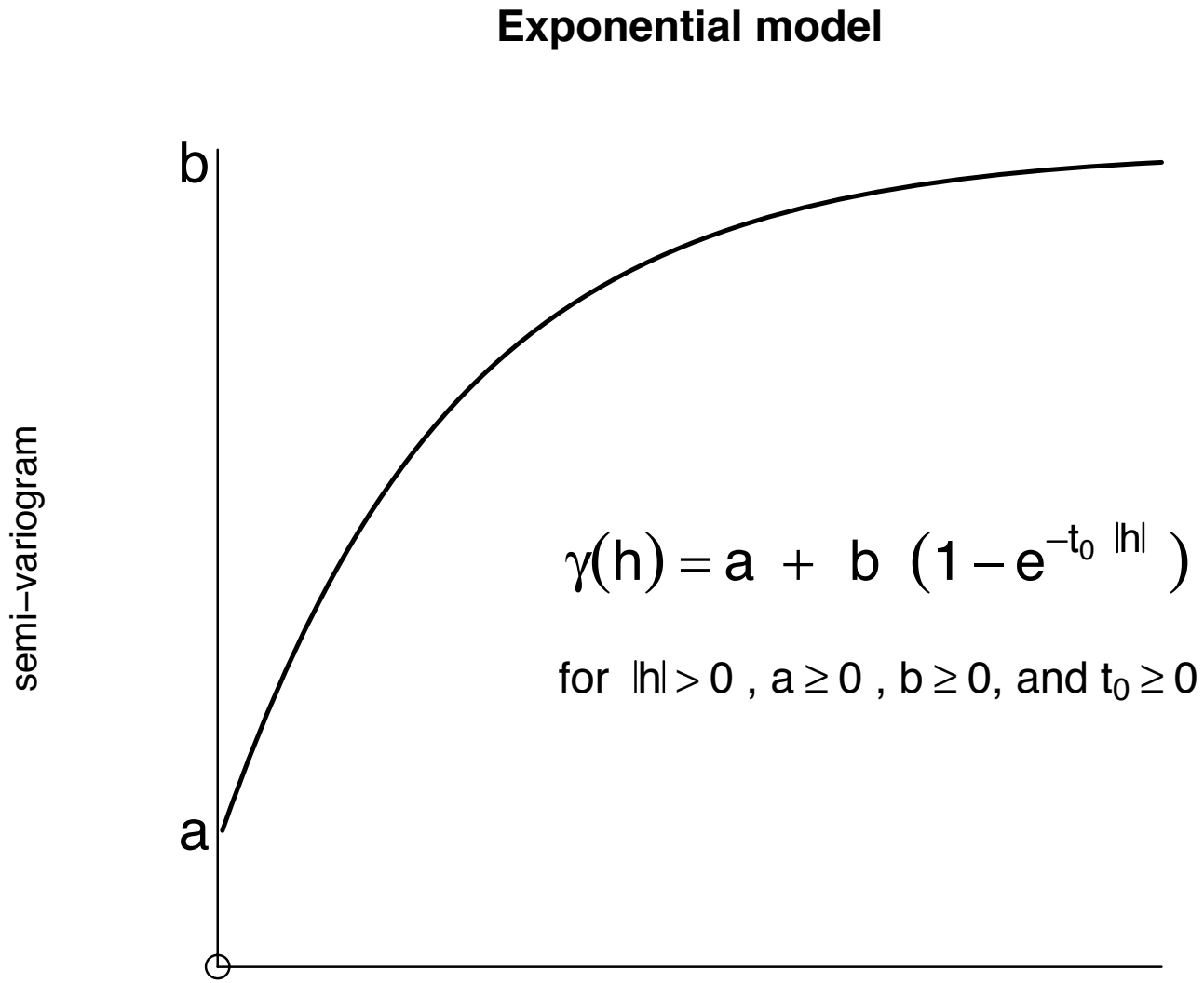
- $\gamma(h)$ is called *semi-variogram*.

Isotropic Semi-Variogram Models

Second order stationarity implies $\gamma(h) = C(0) - C(h) \to \gamma(0) = 0$

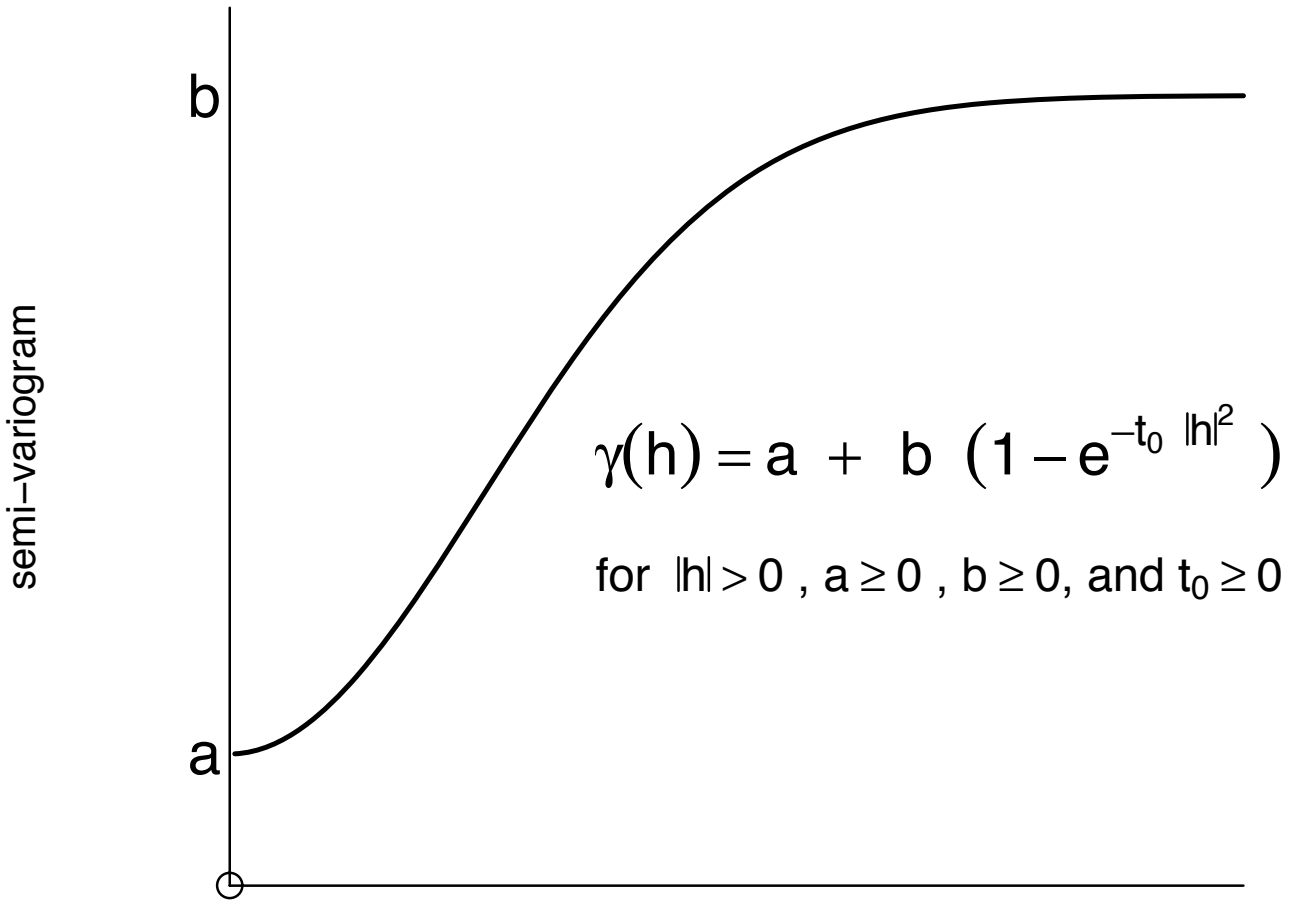
- But often $\lim_{h \to 0} \gamma(h) \neq 0$. Discontinuity called ***nugget effect***.
- When $\gamma(h) \to B$ as $h \to \infty$, B called a ***sill***
- **Note**: Few functions satisfy positive definiteness condition - only certain ones (eg. variogram)

COMMON ISOTROPIC MODELS

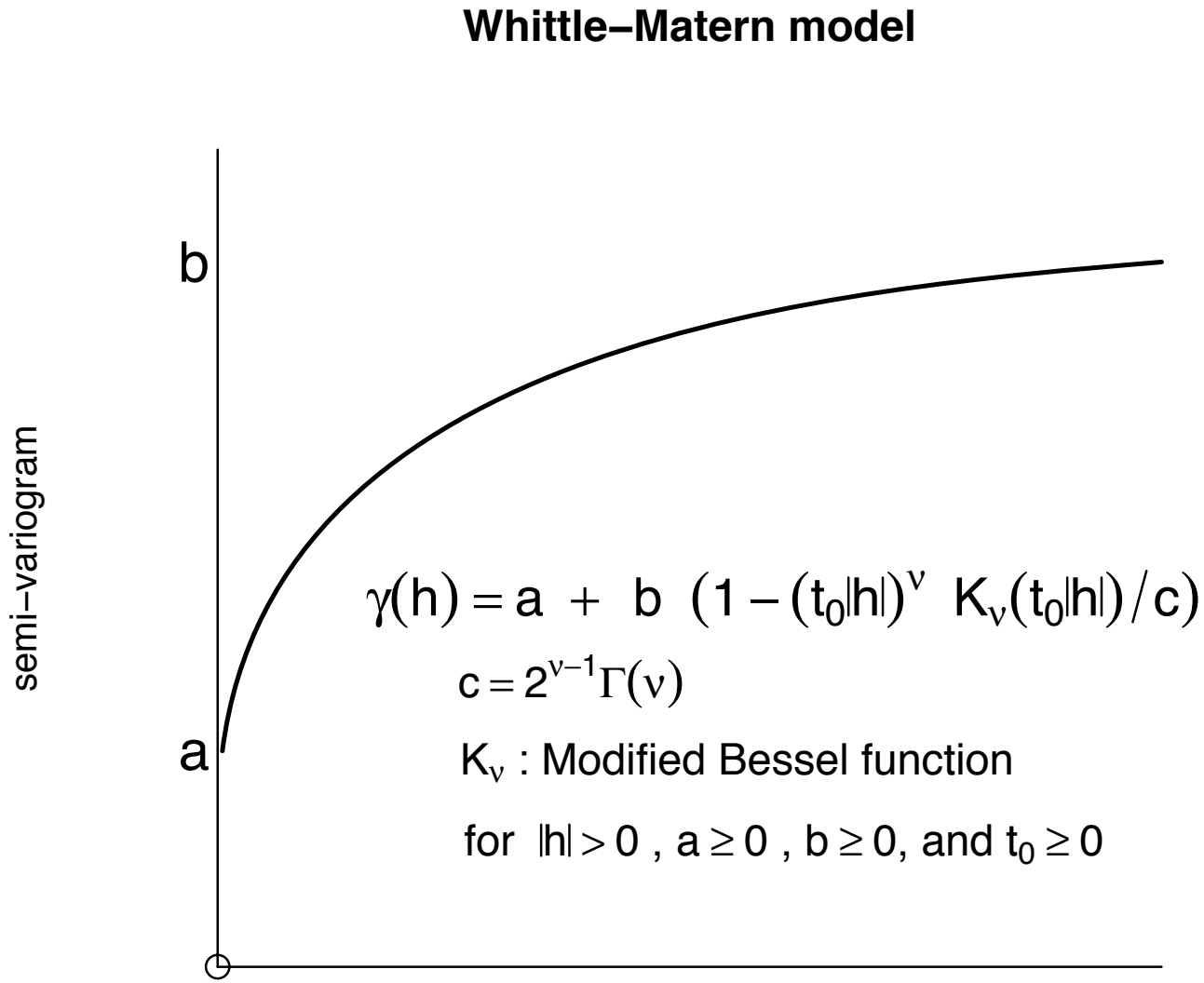


COMMON ISOTROPIC MODELS

Gaussian model



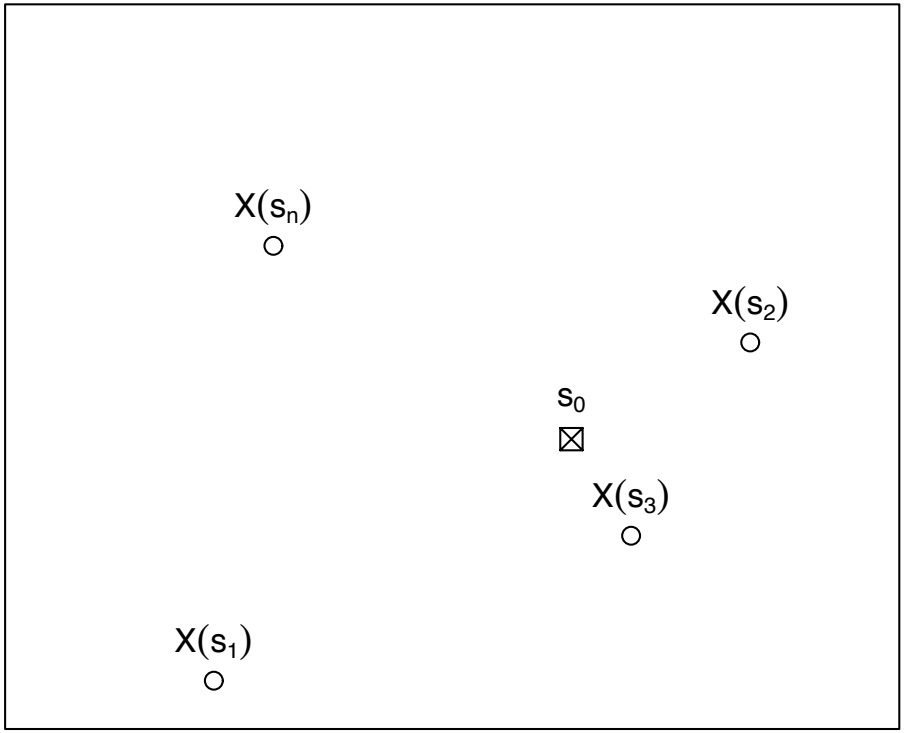
COMMON ISOTROPIC MODELS



SPATIAL PREDICTION

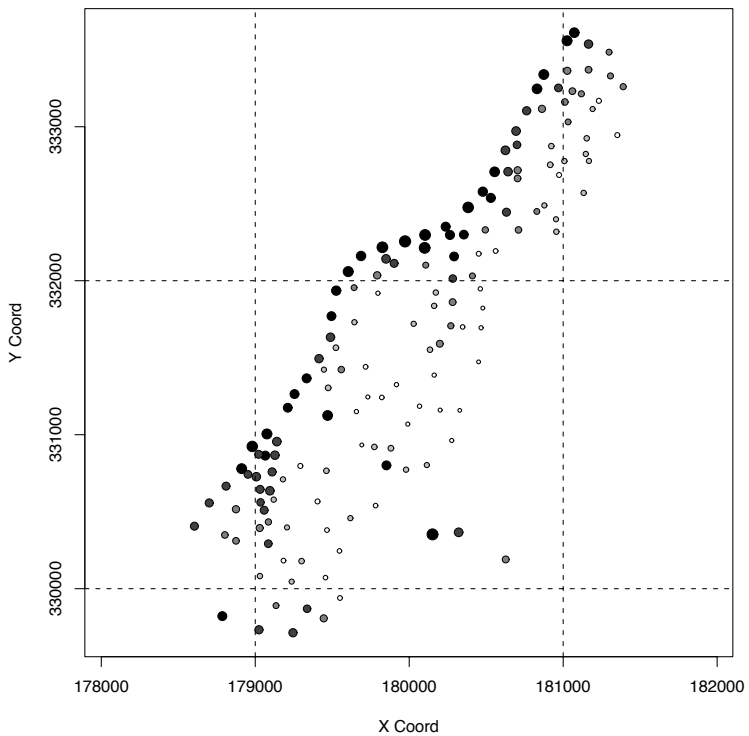
Typo: Change X to Y

Problem: Estimate at location s_0 given observed levels $X(s_i)$?



Case study 1: Zinc levels in the Netherlands

Values of log zinc at sampling locations. Mapping the basin would mean predicting unmeasured responses at other sites without measurements.



Ordinary Kriging

Goal:. Ignoring measurement error for simplicity predict $Y(s_0)$ given observations $y_1, \ldots, y_n$ at locations $s_1, \ldots, s_n$. Assumption

- Covariance structure known
- $Y(s) = \mu + W(s)$ & intrinsic stationary, ie.

$$\begin{aligned} E[Y(s)] &= \mu \\ Var[Y(s) - Y(s+h)] &= 2\gamma(|h|) \end{aligned}$$

- *Linear predictors:*

$$Y^*(s_0) = \sum_{i=1}^{n} \alpha_i Y(s_i)$$

Ordinary Kriging

Reaching the goal: choose $\{\alpha\}$ to get unbiasedness & minimal prediction error

$$\sigma^2_{s_0} \equiv E\left[Y^*(s_0) - Y(s_0)\right]^2$$

Result: Kriging predictor = best linear unbiased predictor (BLUP)

Ordinary Kriging System

- $E[Y^*(s_0)] = E\left[\sum_{i=1}^n \alpha_i Y s_i)\right] = \mu \sum_{i=1}^n \alpha_i \qquad (1)$
  implies $\sum_{i=1}^n \alpha_i = 1.$
- **Prediction error (Kriging variance).**

$$\begin{aligned}
\sigma^2_{s_0} &\equiv E\left[Y^*(s_0) - Y(s_0)\right]^2 = E\left[\sum_{i=1}^n \alpha_i [Y(s_i) - Y(s_0)]\right]^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j E[Y(s_i) - Y(s_j)]^2/2 \\
&\quad - \sum_{i=1}^n \alpha_i E[Y(s_i) - Y(s_0)]^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \gamma(|h_{ij}|) - 2\sum_{i=1}^n \alpha_i \gamma(|h_{i0}|) \qquad (2)
\end{aligned}$$

α's chosen to minimize (2) & satisfy (1)

IMPLEMENTATION IN SUMMARY

- Select good semi-variogram model. Estimate $\hat{\gamma}(.)$ since it will not be known as assumed.
- Solve the *Kriging system* to obtain $\hat{\alpha}$'s

Resulting Kriging predictor & estimated **Kriging variance**

$$\begin{aligned} \hat{Y}^*(s_0) &= \sum_{i=1}^{n} \hat{\alpha}_i y_i \\ \hat{\sigma}^2_{s_0} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\alpha}_i \hat{\alpha}_j \hat{\gamma}(|h_{ij}|) - \sum_{i=1}^{n} \hat{\alpha}_i \hat{\gamma}(|h_{i0}|) \end{aligned}$$

REMARKS

- $Y \sim$ Gaussian implies 95% prediction interval:

$$[Y^*(s_0) - 1.96\sigma_{s_0}, Y^*(s_0) + 1.96\sigma_{s_0}]$$

- Kriging predictor is **exact interpolator**;
(interpolator = observed value at that location)
- $\sigma^2_{s_0}$ is

$$\sigma^2_{s_0} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j C(s_i, s_j) - 2\sum_{i=1}^{n} \alpha_i C(s_i, s_0) + Var(Y(s_0))$$

- Stationarity required only because cannot otherwise estimate the covariance.

Universal Kriging

Random fields with non-constant means. Let

$$Y(s) = \mu(s) + W(s)$$

- Here $W(s)$ is 2^{nd}-order stationary with mean $E[W(s)] = 0$
- $\mu(s) = \sum_{l=1}^{k} a_l f_l(s)$ $\{f_l(s), l = 1, \ldots, k\}$: known functions with parameters and $\{a_l\}$. Can be dummy variables.

Universal Kriging Estimator:

$$Y^*(s_0) = \sum_{i=1}^{n} \alpha_i Y(s_i)$$

Weights α's chosen to get unbiased estimate with smallest prediction error.

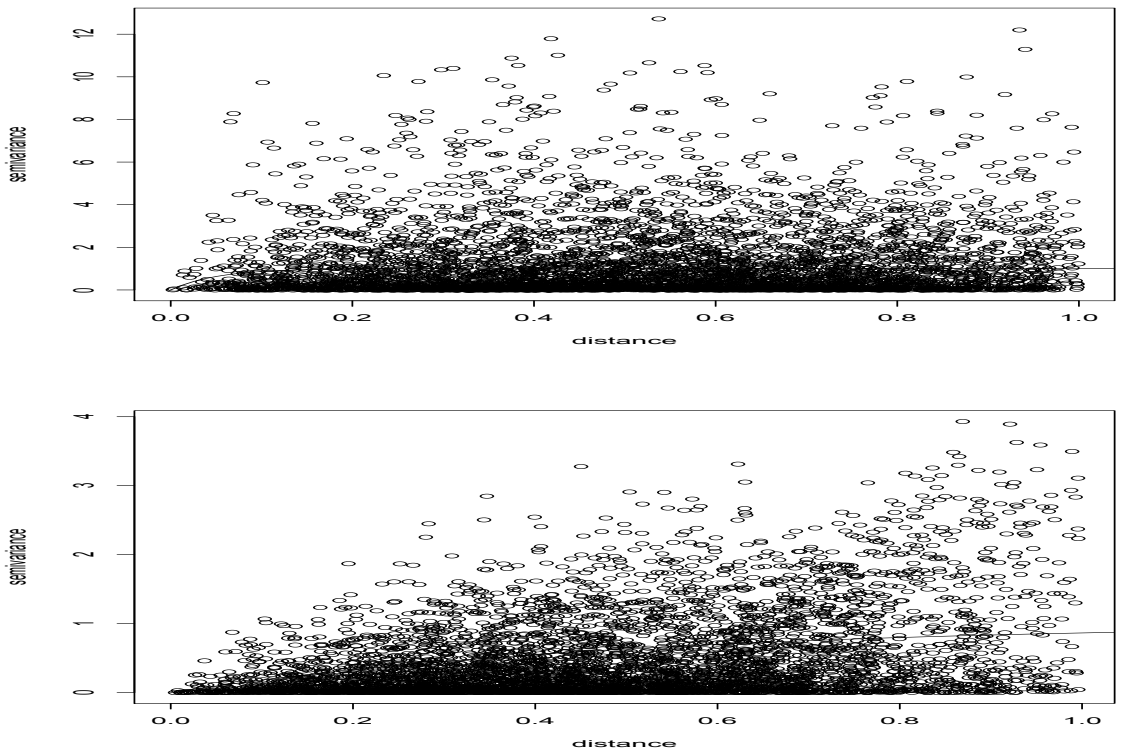
SIMULATION STUDY

geoR provides a random field simulation function. Notice that we have used the Matern covariance function to generate the data with $\kappa = 0.5$ so it gives an exponential variogram. The range is $\phi = 0.05$ but this varies in the simulation study.

grf(n, grid = "irreg", nx, ny, xlims = c(0, 1),
ylims = c(0, 1), borders, nsim = 1,
cov.model = "matern", cov.pars = c(1,0.04) kappa = 0.5, nugget =
0, lambda = 1, aniso.pars,
mean = 0, method, RF=TRUE, messages)

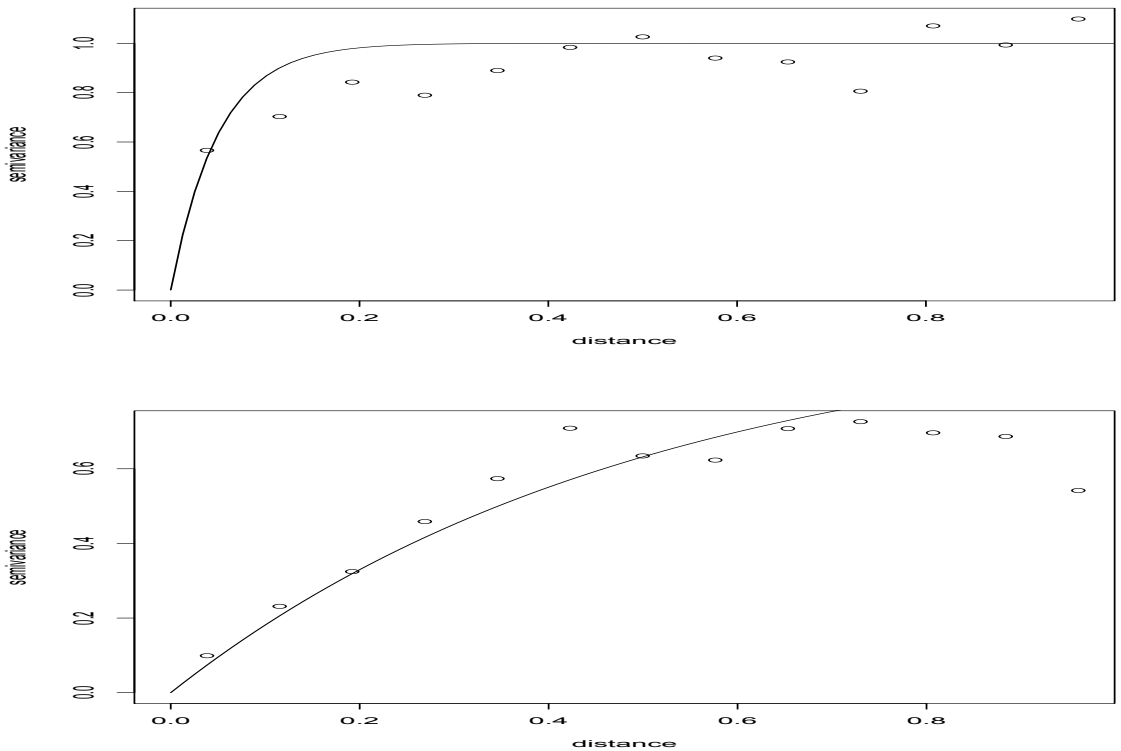
SIMULATION STUDY

We begin with the variogram clouds for $\phi = 0.05, 0.50$.



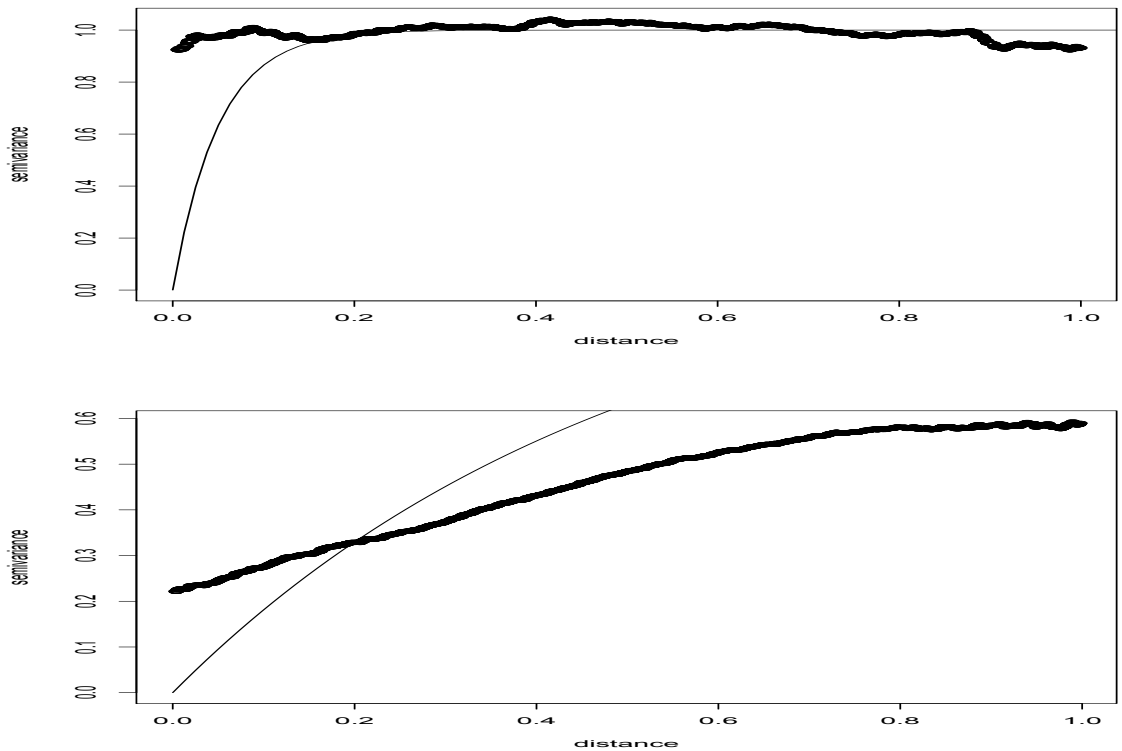
SIMULATION STUDY

We turn to bins or $\phi = 0.05, 0.50$.



SIMULATION STUDY

We finish with smoothers $\phi = 0.05, 0.50$.



VARIOGRAM FITTING STRATEGIES

First choose a parametric variogram family.

Then use:

Least squares: We use four LS methods below, all of which fit to the binned variogram:

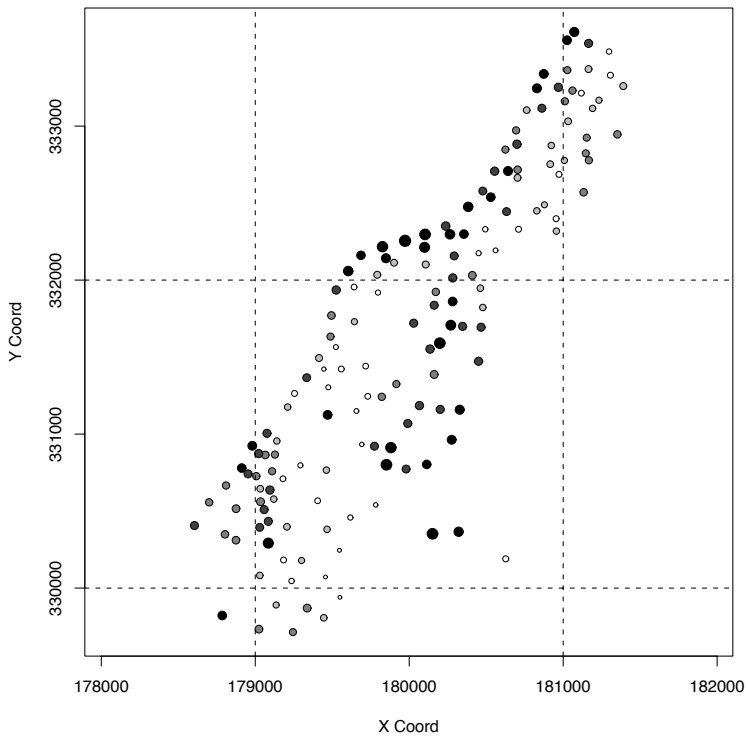
1. ordinary least squares
2. weighted least squared- bin counts; variances; Cressie weights.

Maximum likelihood: Needs to have a specified sampling distribution.

Bayes: Distributions put on the parameters.

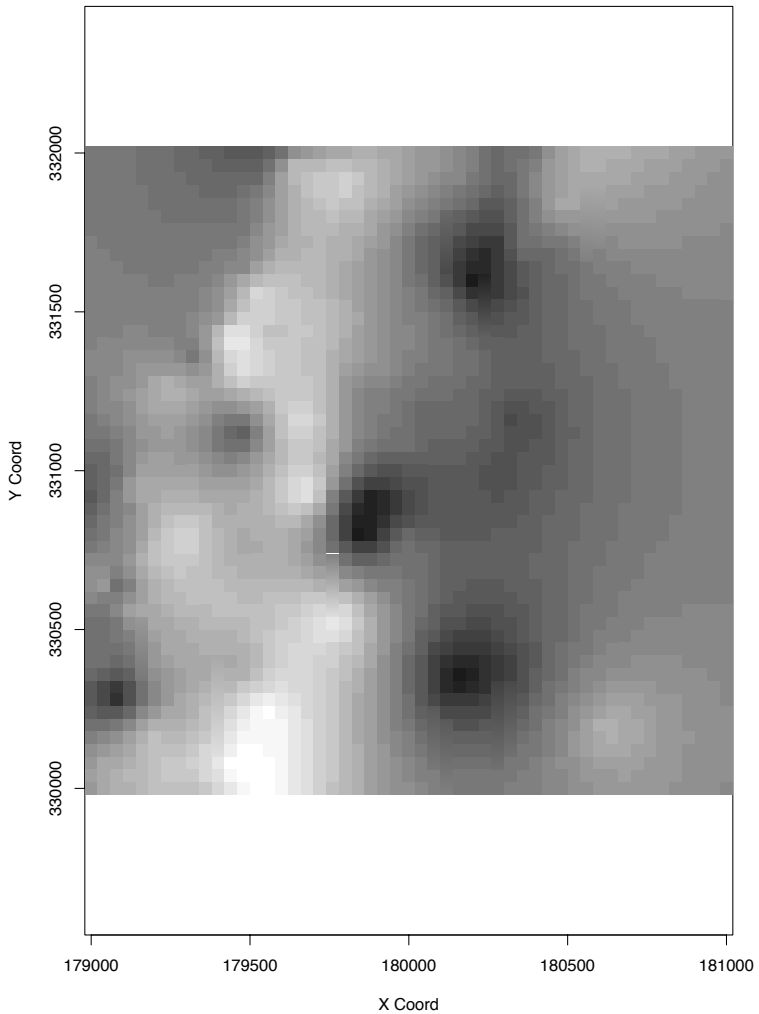
Case study 1 (continued)

Values of log residuals, after detrending the data by removing effect of "distance from river" and "elevation" through universal kriging.



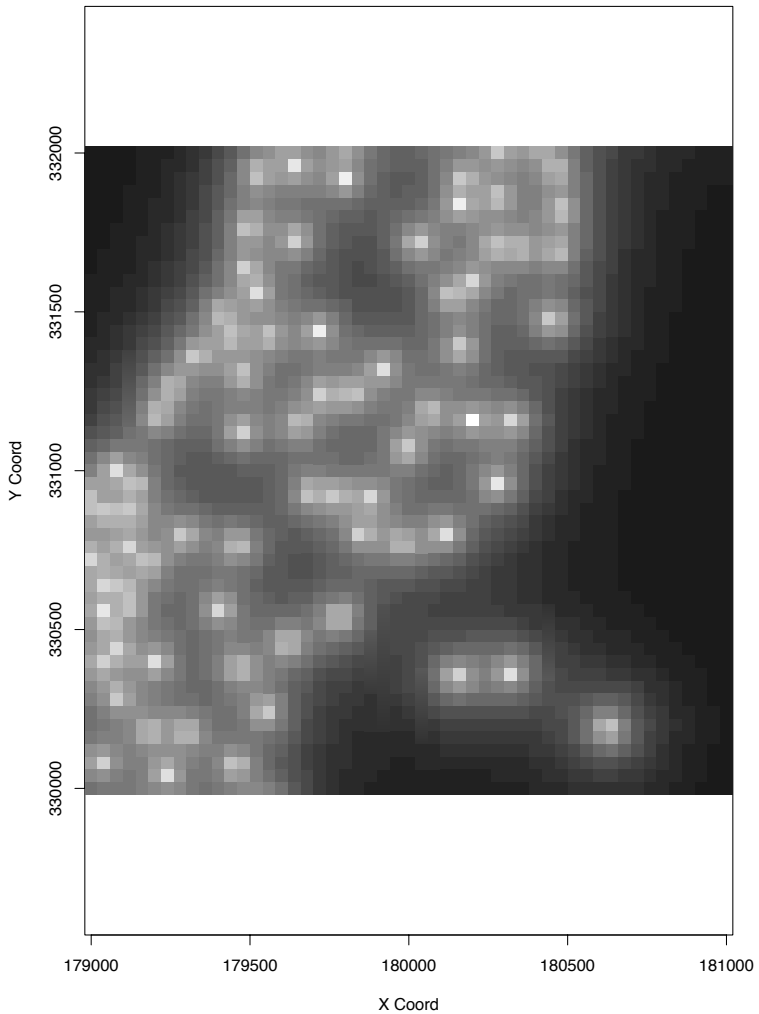
Case study 1 (continued)

Predicted spatial residual surface.



Case study 1 (continued)

Standard error of prediction of residual.



Least Square Estimators

Ordinary least squares: Choose θ to minimize

$$(\hat{\gamma} - \gamma_\theta)'(\hat{\gamma} - \gamma_\theta).$$

Ordinary LS immediately implementable by a nonlinear least squares procedure. But estimates $\hat{\gamma}(h)$ may vary a lot so assigning equal weights to all $\hat{\gamma}(h)$ unsatisfactory.

Least Square Estimators

Number weighted least squares: Modification of equal weights scheme uses weights given by number of pairs in each bin as in second method above. Choose θ to minimize

$$(\hat{\gamma} - \gamma_\theta)' M (\hat{\gamma} - \gamma_\theta),$$

where M is a diagonal matrix of the number of pairs of points in each bin.

Least Square Estimators

Weighted least squares: Choose θ to minimize

$$(\hat{\gamma} - \gamma_\theta)' W_\theta (\hat{\gamma} - \gamma_\theta),$$

where W_θ is a diagonal matrix of the variances of the entries of γ_θ.

Least Square Estimators

Generalized least squares: Choose θ to minimize

$$(\hat{\gamma} - \gamma_\theta)' V_\theta (\hat{\gamma} - \gamma_\theta),$$

where V_θ denotes the covariance matrix of γ_θ.

NOTES:

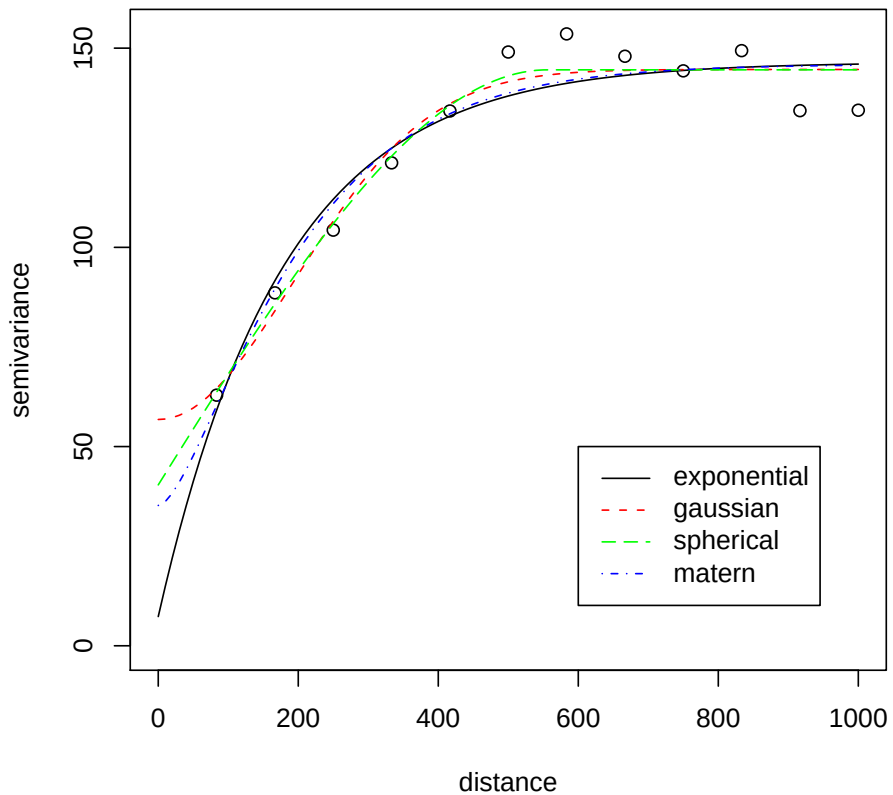
- The weighted and generalized least squares method require specification of the matrices W_θ and V_θ.
- Generalized LS is possible in principle, but complicated to implement.

Case study 2: Soil Calcium in Brazil

These data consist of calcium content in soil from a region in Brazil. They are in the geoR library. For a description use > ?ca20 on the command line in R.

Case study 2: Soil Calcium in Brazil

Fitting variograms by ordinary least squares.



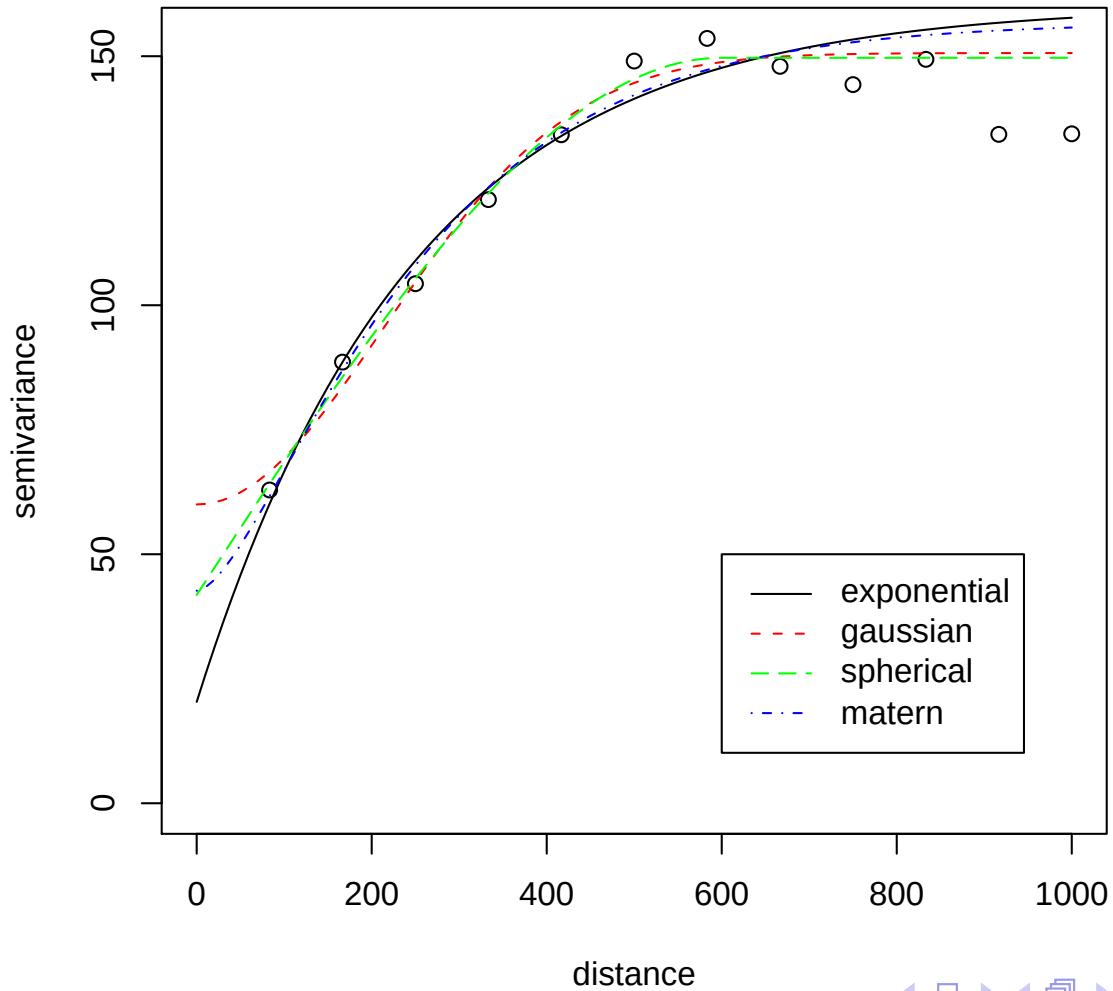
Case study 2: Soil Calcium in Brazil

Estimated parameters.

model	sill	range	nugget	RSS
exponential	139.1881	179.0273	7.3292	678.683
gaussian	87.8525	273.3551	56.7929	443.8803
spherical	104.1764	555.6048	40.3781	375.0742
matern	110.7297	135.1417	35.213	601.5659

Case study 2: Soil Calcium in Brazil

Number weighted least square fitting.



CASE STUDY 2: SOIL CALCIUM IN BRAZIL

Estimated parameters.

model	sill	range	nugget	RSS
exponential	139.901	249.1264	20.3517	430123.7
gaussian	90.6371	303.4283	59.9994	251816.7
spherical	107.8644	599.9593	41.8199	149820.6
matern	114.1321	170.633	42.6684	355798.4

Case study 2: Soil Calcium in Brazil

Estimated parameters –spherical model – different fitting methods.

method	sill	range	nugget	RSS
ordinary	104.1764	555.6048	40.3781	375.0742
number	107.8644	599.9593	41.8199	149820.6
cressie	108.1	598.2109	41.6151	8.8353

LIKELIHOOD APPROACH

The likelihood approach is viewed as best method since points in the empirical variogram are highly correlated. Makes LS inefficient and misleading. geoR has that option.

ENSURING SIMPLE MODELS

Adding parameters can always reduce residal sums of squares. But also need to minimize # of parameters. Distributional assumptions & Akaike Information Criterion (AIC) can do this:

$$\text{AIC} = -2\log(\text{maximized likelihood}) + 2(\text{number of parameters}),$$

AIC's variable part is estimated by

$$n\log(\text{RSS}) + 2p.$$

Here n = # of points, p= #of model parameter and RSS = residual sum of squares.

CROSS-VALIDATION WITH KRIGING

Spatial prediction important goal of kriging. So choose model that does this best. How? By leave–one–out cross–validation.

1. Estimate variogram using sample data & fitted plausible models.
2. For each model, predict excluded Y's using kriging value there. Calculate kriging variance as well.

CROSS-VALIDATION WITH KRIGING

Diagnostics from results: *mean–deviation (ME)*; *mean–squared–deviation (MSE)*; *mean–squared–deviation–ratio (MSDR)* found from squared–errors & kriging variances, $\hat{\sigma}^2(s_i)$:

$$\text{ME} = \sum_{i=1}^{N} \mid y(s_i) - \hat{y}(s_i) \mid /N$$

$$\text{MSE} = \sum_{i=1}^{N} \mid y(s_i) - \hat{y}(s_i) \mid^2 /N$$

$$\text{MSDR} = \sum_{i=1}^{N} \frac{(y(s_i) - \hat{y}(x_i))^2}{\hat{\sigma}^2(s_i)} /N.$$

CROSS-VALIDATION WITH KRIGING

NOTES:

- ME should be close to 0, since kriging is an unbiased prediction method.
- MSE should be as small as possible.
- If the model is accurate then the MSDR should be close to 1.

CASE STUDY 2 (CONT'D).

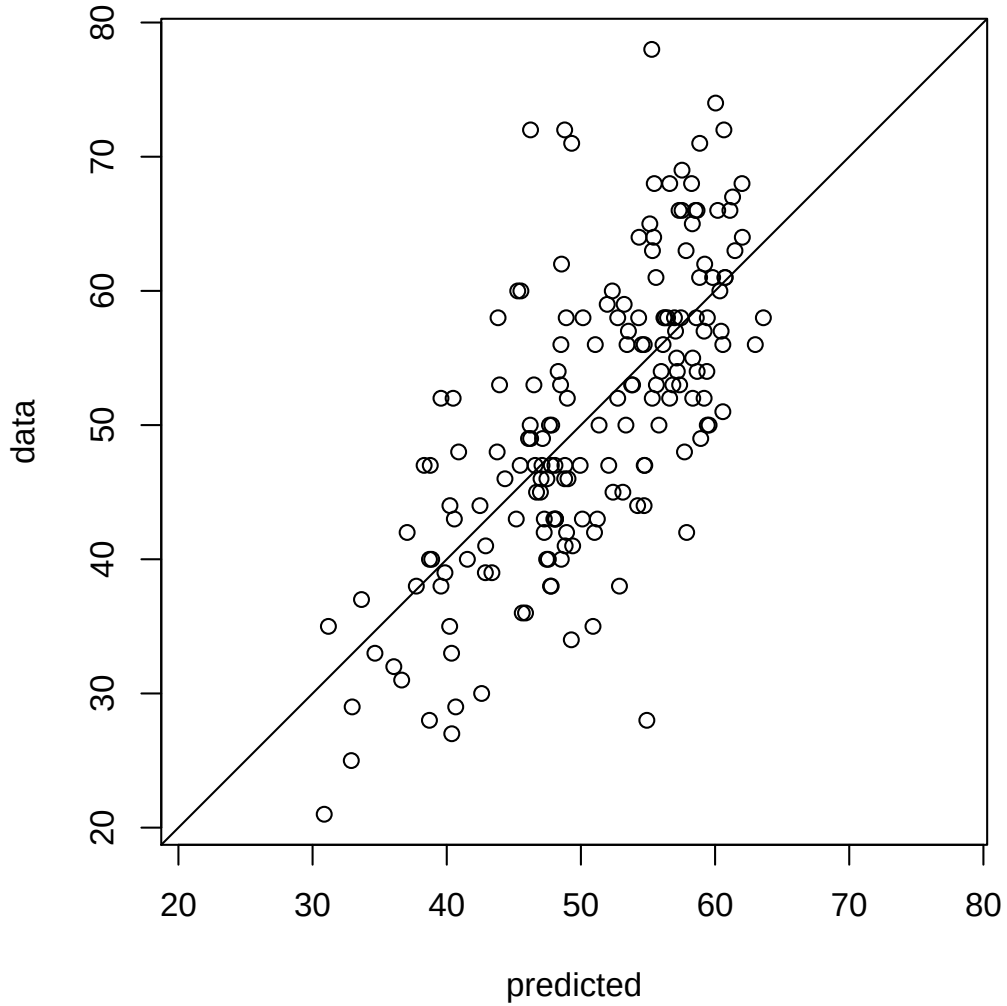
Do cross-validation for the four variogram models "exponential", "gaussian", "spherical", and "matern" on the ca20 data from the geoR package. Then calculate diagnostic indices.

model	ME	MSE	MSDR
exponential	-0.008028705	60.94539	1.103823
gaussian	-0.007405837	69.02756	1.064712
spherical	-0.008975785	62.69338	1.022848
matern	-0.00870061	62.96571	1.057020

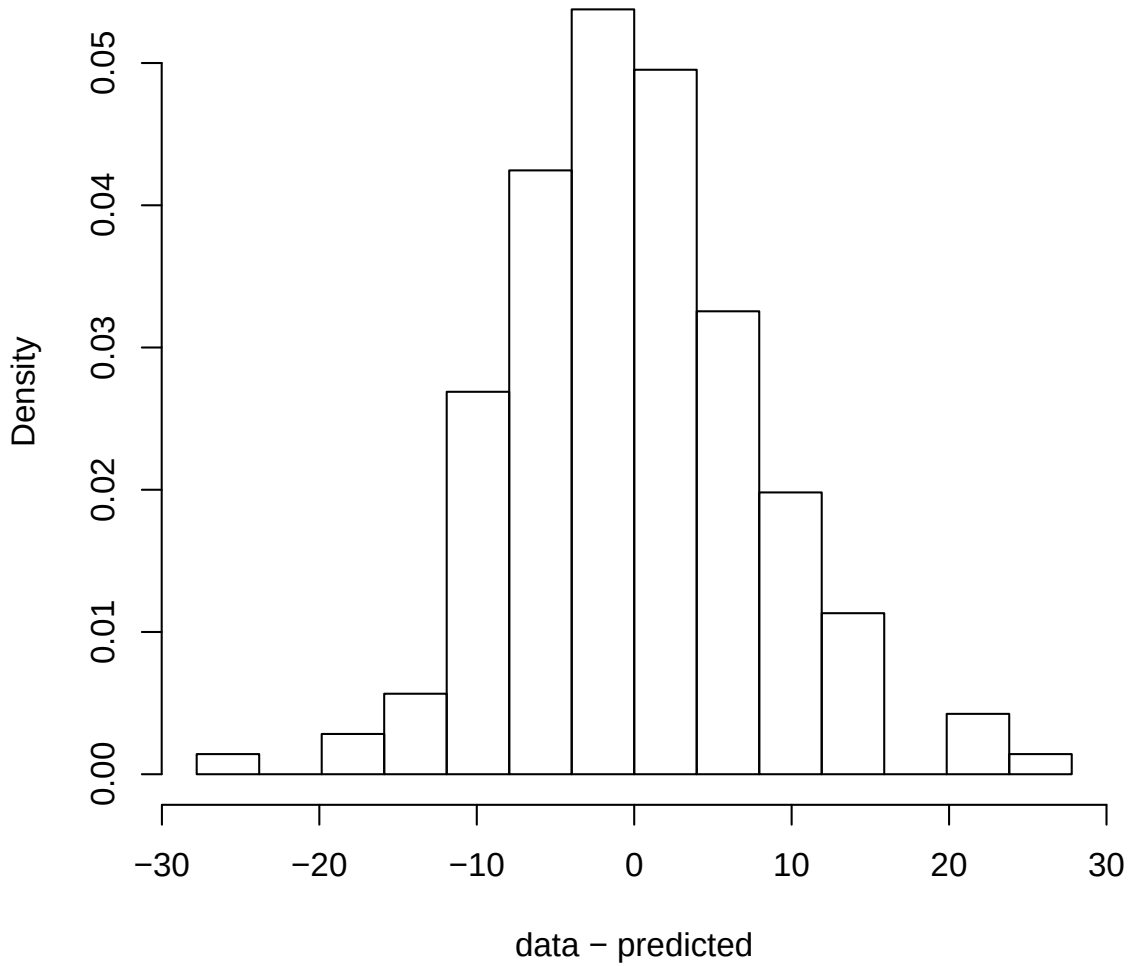
Some diagnostic plots

The spherical model seems to win also in the cross-validation competition. But diagnostic plots seen in the slides that follow can also be useful.

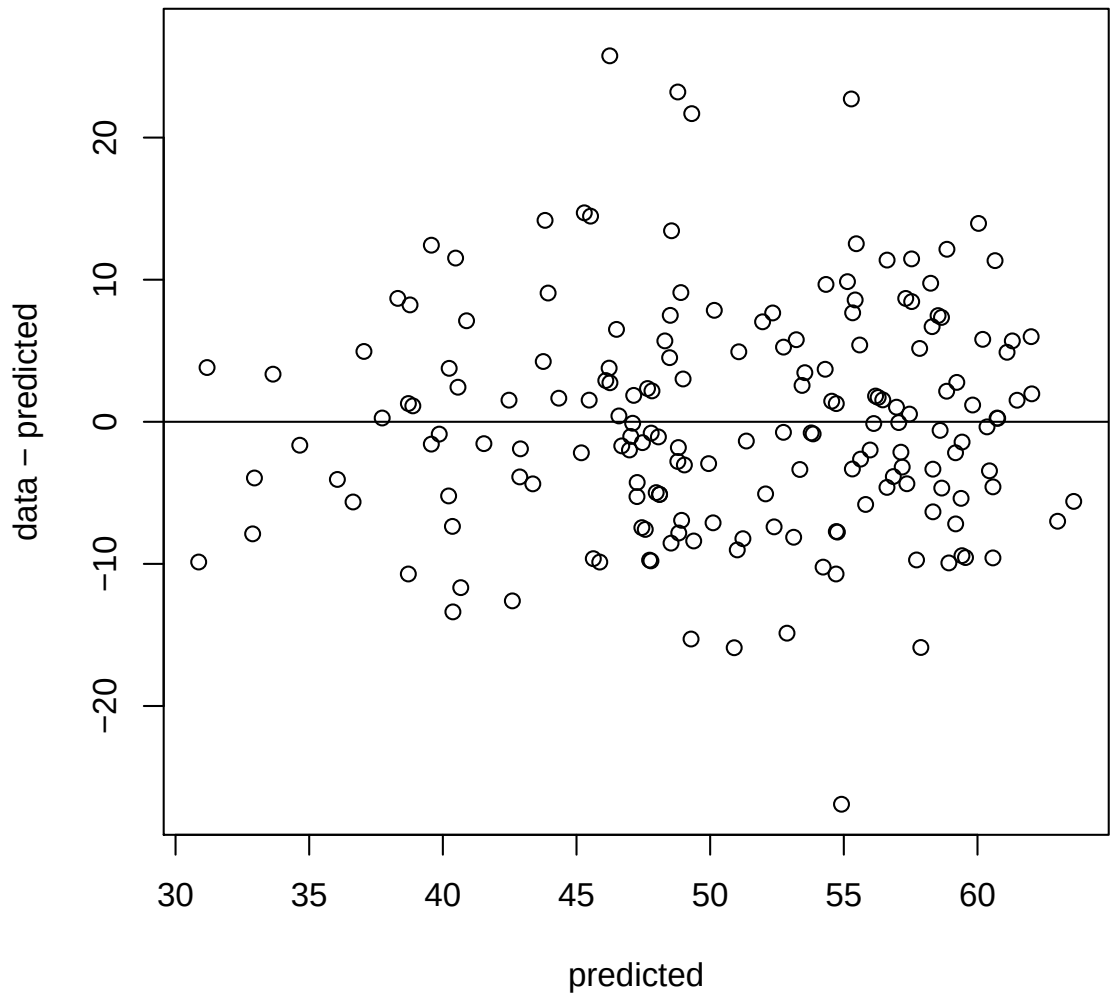
Some diagnostic plots



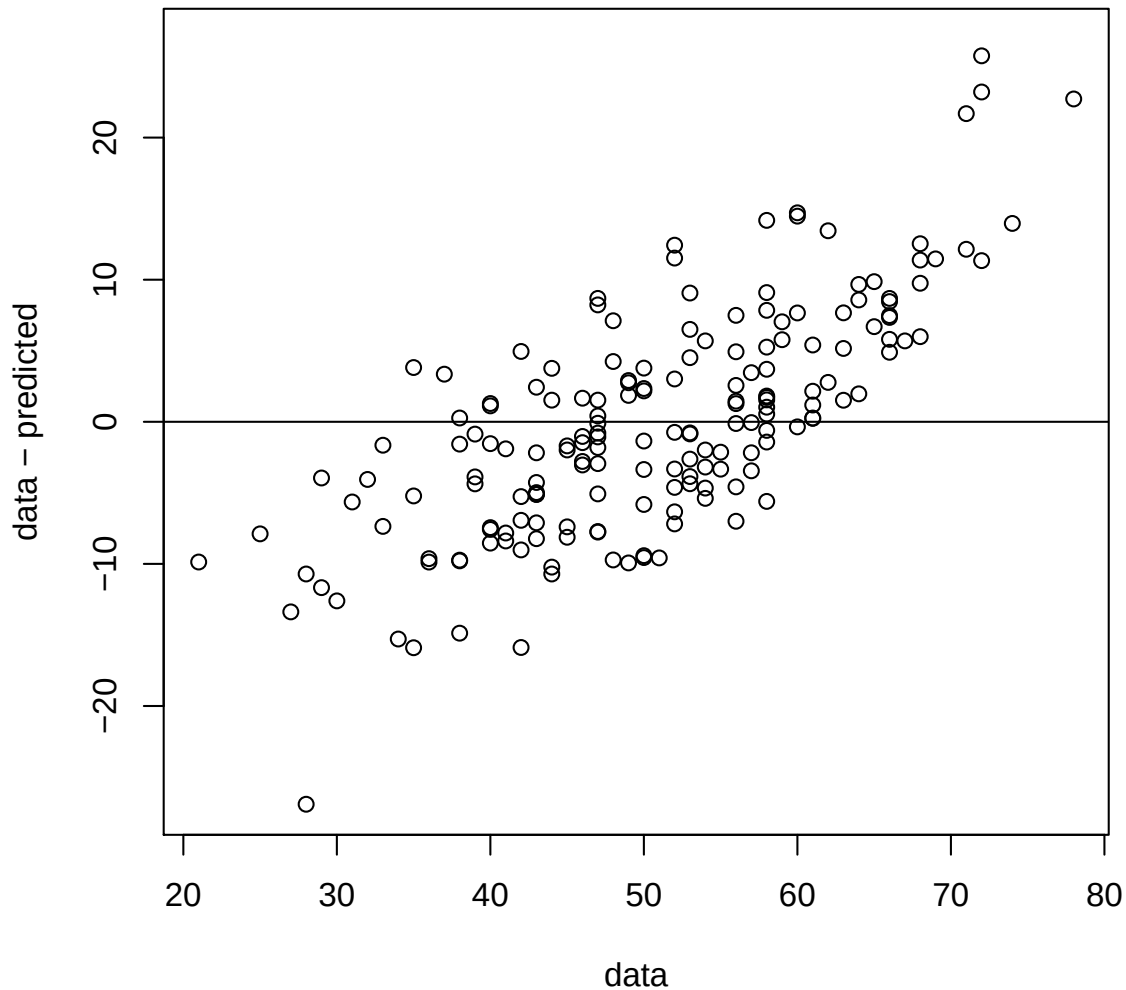
Some diagnostic plots



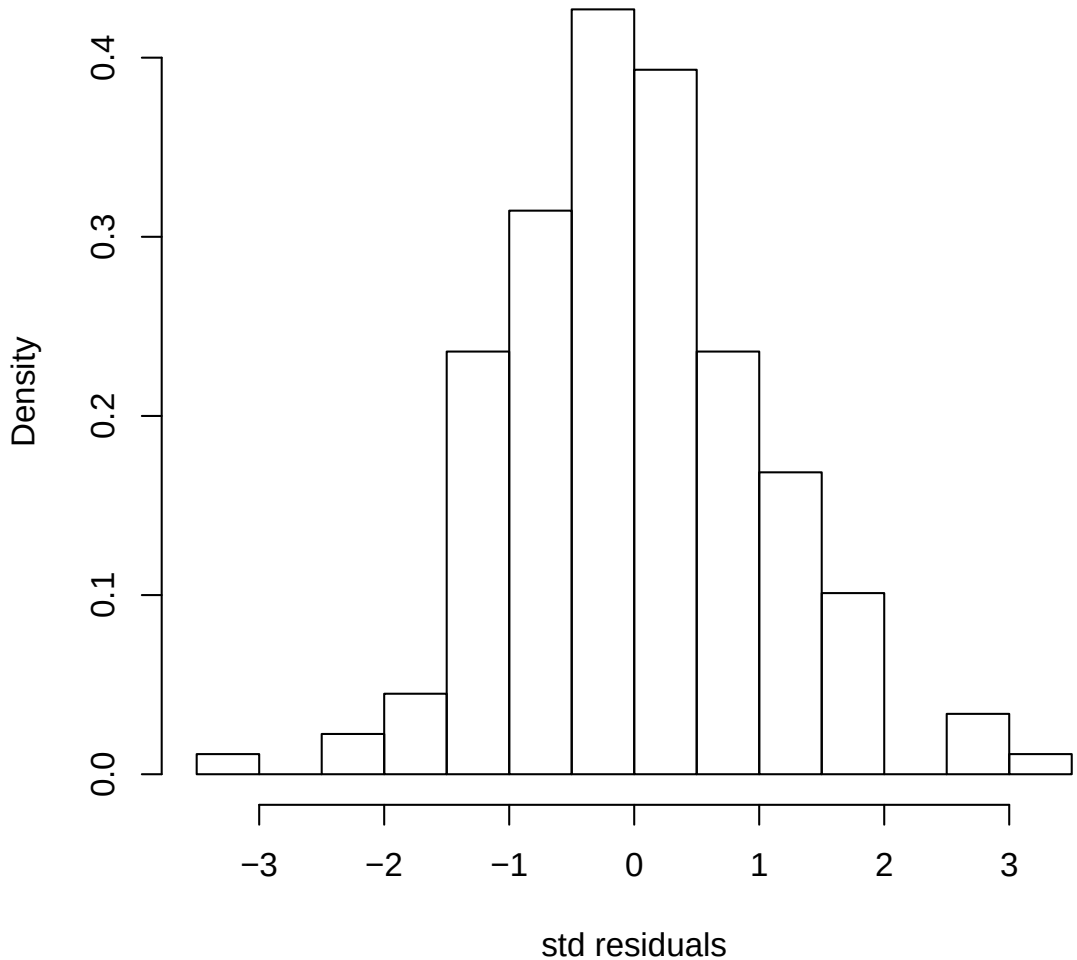
Some diagnostic plots



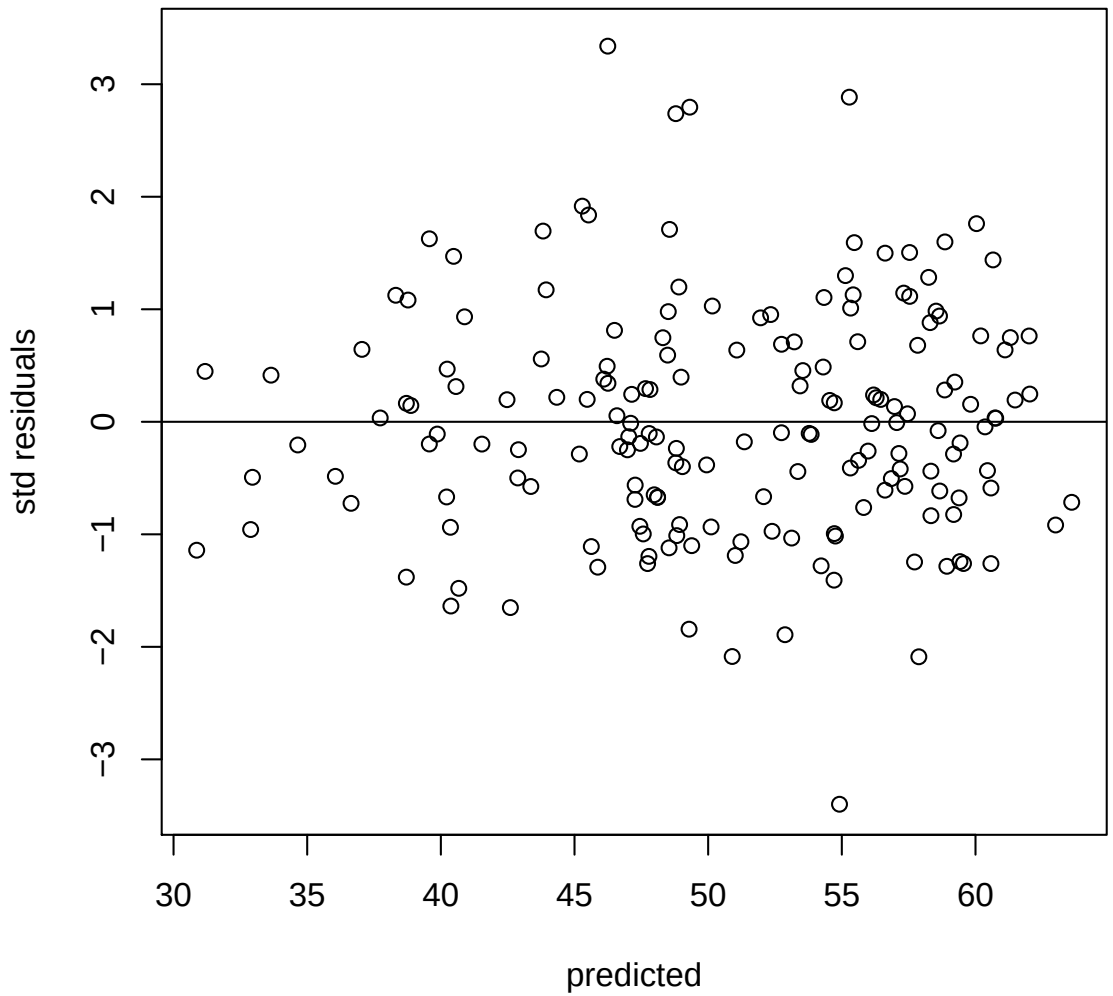
SOME DIAGNOSTIC PLOTS



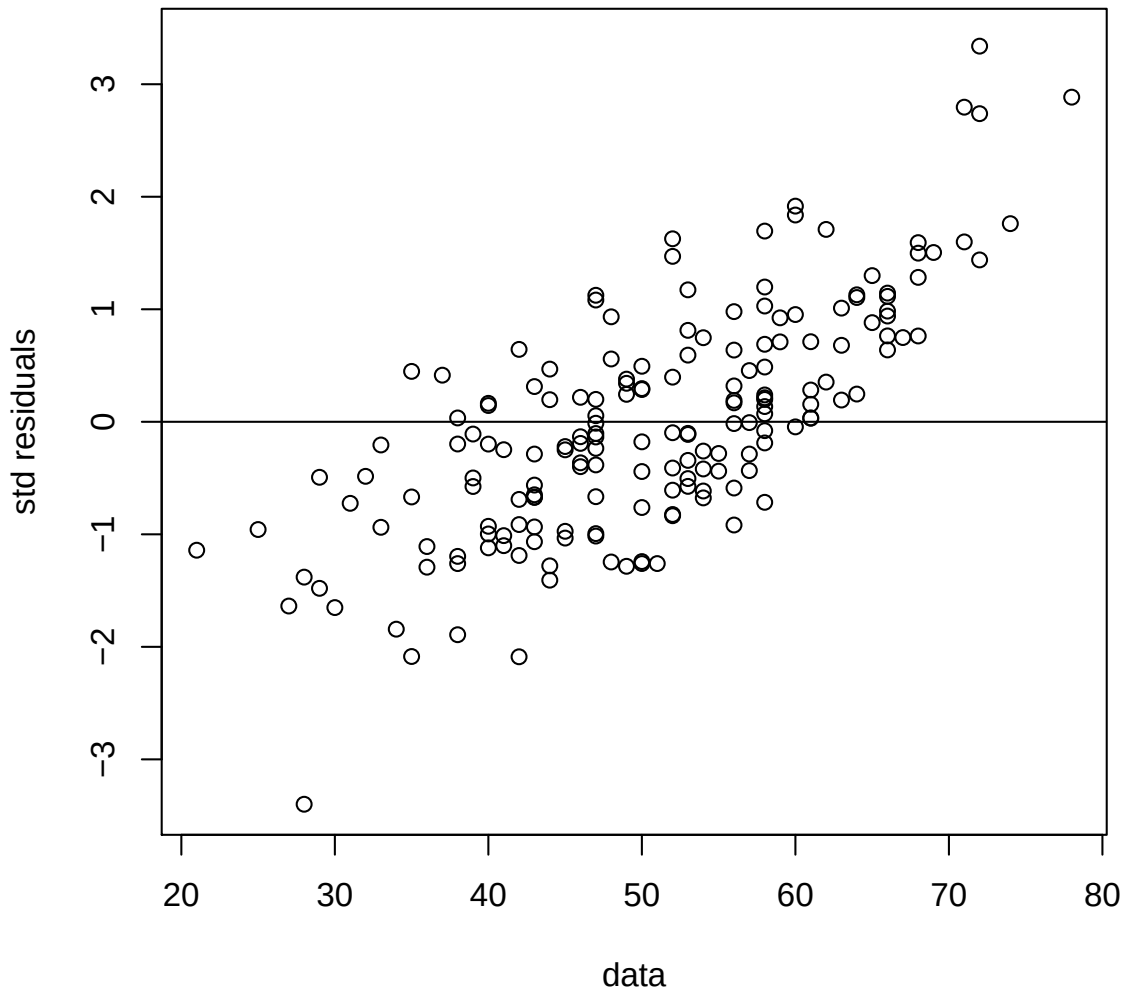
Some diagnostic plots



Some diagnostic plots



SOME DIAGNOSTIC PLOTS



WHAT IF PROCESS SEEMS NON-STATIONARY?

Some options follow:

1. Change spatial mean: $\mu(s)$ will inevitably be misspecified as $\mu^*(s)$ so the residual is misspecified as $W^*(s) = Y(s) - \mu^*(s)$. Thus the calculated variogram will be non-stationary

$$E[W^*(s^1) - W^*(s^2)]^2 = E[W(s^1) - W(s^2)]^2 + [\{\mu^*(s^1) - \mu^*(s^1)\} - \{\mu^*(s^2) - \mu^*(s^2)\}]^2$$

WHAT IF PROCESS SEEMS NONSTATIONARY?

2. Adopt non-stationary modelling approach, convolution approach: Represent the residual as

$$W(s) = \int K(s - s')W^*(s')ds'$$

where W^* is stationary. **NOTE:** Allows only modest degree of nonstationarity.

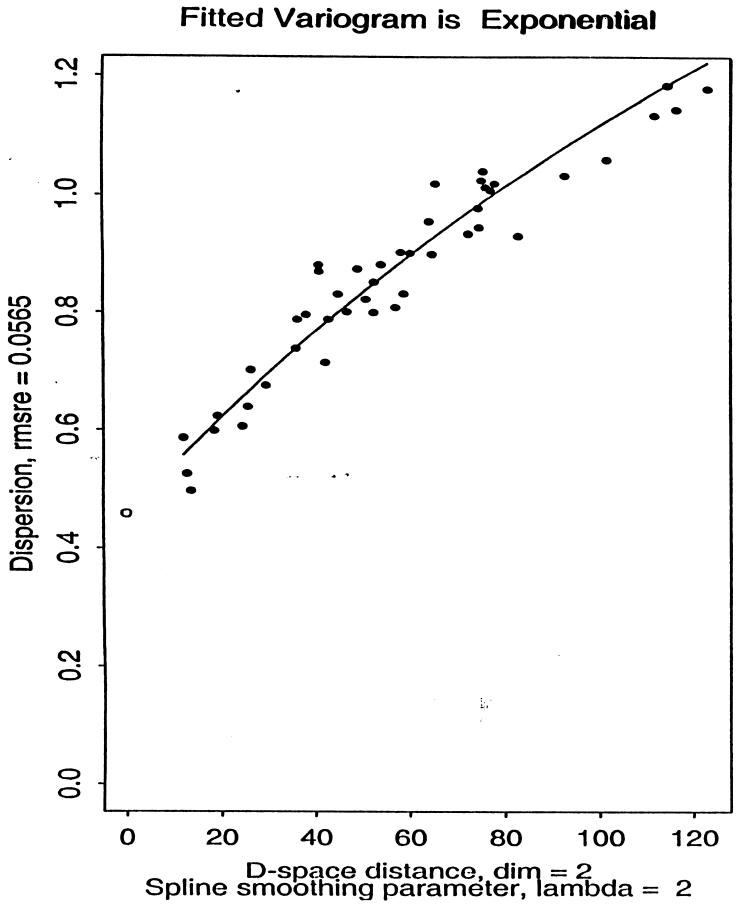
WHAT IF PROCESS SEEMS NONSTATIONARY?

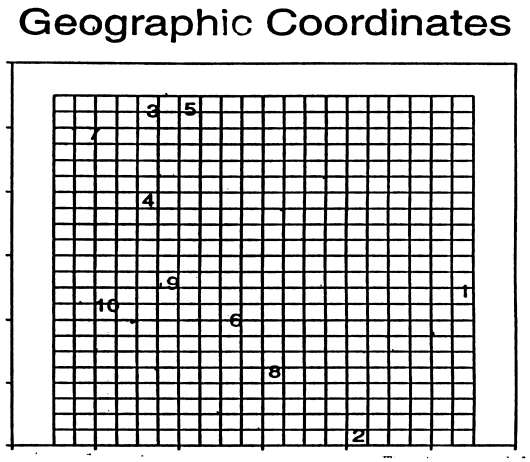
3. Warping: The famous Sampson–Guttorp approach warps the geographic space into dispersion space so that strongly correlated sites are moved closet together, uncorrelated ones further apart.
4. Dimension expansion: Keep the geographic space as is but add additional dimensions.

EXAMPLE: PARTICULATE MATTER IN VANCOUVER

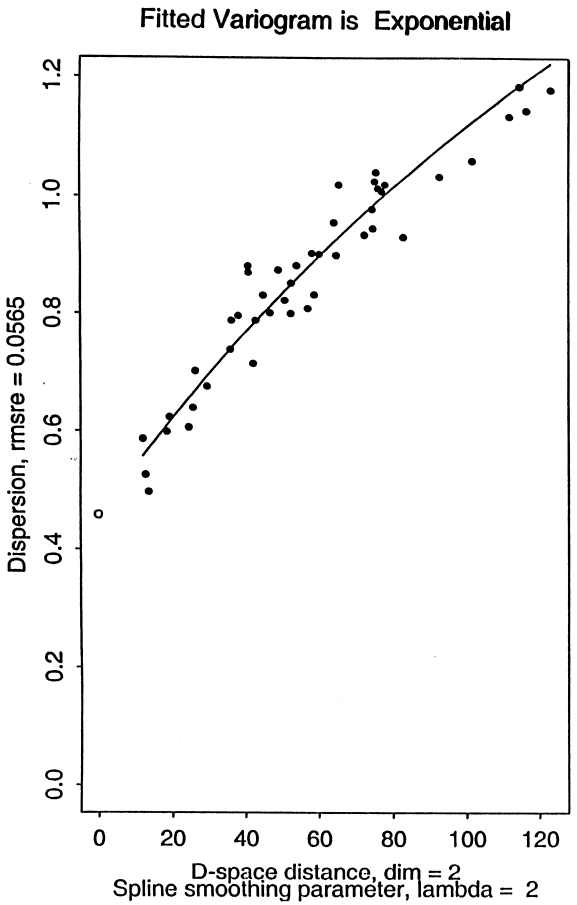
- Small particulates, the size of those in cigarette smoke are nasty.
- They get deep into the lung to the gas exchange membrane where they can generate antiinflammatory mediators.
- These in turn affect the cardio–vascular system and cause heart problems.
- PM10 are all up to 10 microns in size. PM2.5 is the fraction with the smallest sizes and are now of primary concern.
- However the spatial field can be quite nonstationary since these particulates come from mobile and local sources.

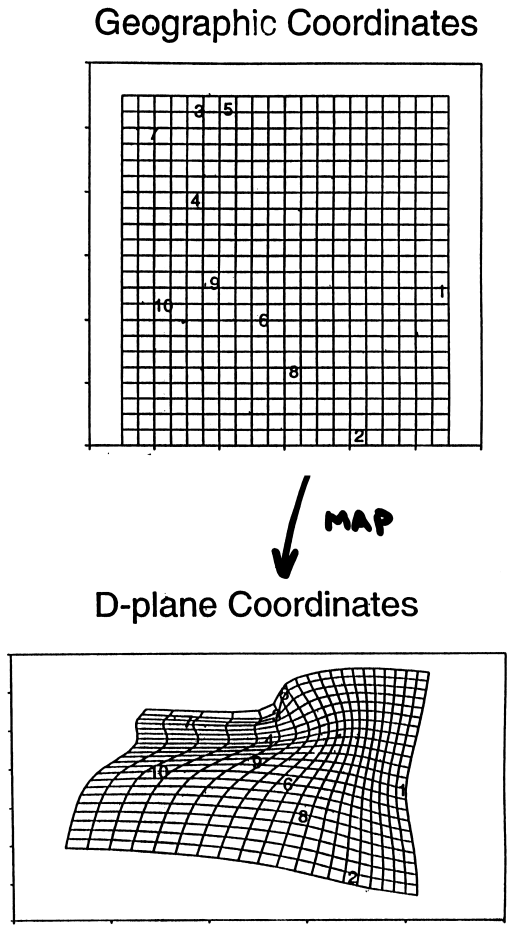
HOURLY PM_{10} IN VANCOUVER -1994-1999





HOURLY PM_{10} IN VANCOUVER -1994-1999





Spatio-Temporal Processes

Bayesian Hierarchical Models

Bayesian hierarchical models are an extremely useful and flexible framework in which to model complex relationships and dependencies in data and they are used extensively throughout the book. In the hierarchy we consider, there are three levels;

(1) The observation, or measurement, level; $Y|Z, X_1, \theta_1$.
Data, Y, are assumed to arise from an underlying process, Z, which is unobservable but from which measurements can be taken, possibly with error, at locations in space and time. Measurements may also be available for covariates, X_1. Here θ_1 is the set of parameters for this model and may include, for example, regression coefficients and error variances.

Bayesian Hierarchical Models

(2) The underlying process level; $Z|X_2, \theta_2$.
The process Z drives the measurements seen at the observation level and represents the true underlying level of the outcome. It may be, for example, a spatio–temporal process representing an environmental hazard. Measurements may also be available for covariates at this level, X_2. Here θ_2 is the set of parameters for this level of the model.

(3) The parameter level; $\theta = (\theta_1, \theta_2)$.
This contains models for all of the parameters in the observation and process level and may control things such as the variability and strength of any spatio–temporal relationships.

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- A spatial–temporal random field, $Z_{st}, s \in \mathcal{S}, t \in \mathcal{T}$, is a stochastic process over a region and time period.
- This underlying process is not directly measurable, but realisations of it can be obtained by taking measurements, possibly with error.
- Monitoring will only report results at N_T discrete points in time, $T \in \mathcal{T}$ where these points are labelled $T = \{t_0, t_1, \ldots, t_{N_T}\}$.
- The same will be true over space, since where air quality monitors can actually be placed may be restricted to a relatively small number of locations, for example on public land, leading to a discrete set of N_S locations $S \in \mathcal{S}$ with corresponding labelling, $S = \{s_0, s_1, \ldots, s_{N_T}\}$.

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- There are three levels to the hierarchy that we consider.
- The observed data, $Y_{st}, s = 1, ..., N_S, t = 1, ..., N_T$, at the first level of the model are considered conditionally independent given a realisation of the underlying process, Z_{st}.

$$Y_{st} = Z_{st} + v_{st}$$

  where v_{st} is an independent random, or measurement, error term
- The second level describes the true underlying process as a combination of two terms: (i) an overall trend, μ_{st} and (ii) a random process, ω_{st}.

$$Z_{st} = \mu_{st} + \omega_{st}$$

A HIERARCHICAL APPROACH TO MODELLING SPATIO–TEMPORAL DATA

- The trend, or mean term, μ_{st} represents broad scale changes over space and time which may be due to changes in covariates that will vary over space and time.
- The random process, ω_{st} has spatial–temporal structure in its covariance.
- In a Bayesian analysis, the third level of the model assigns prior distributions to the hyperparameters from the previous levels.

SPATIO–TEMPORAL MODELLING

Handling time.

- Depends on random response paradigm: point referenced; lattice; point process.
- Active area of current development

General approaches to incorporating time

- **Approach 1: Treat continuous time as like another spatial dimension** with stationarity assumptions. Eg. Spatio–temporal Kriging. **NOTE:** Constructing covariance models is more involved
- **Approach 2: Integrate spatial fields over time.** Eg. Given a spatial lattice let **X(t)** : $m \times 1$ be vectors of spatial responses at lattice points. Eg. use multivariate autoregression.
- **Approach 3: Integrate times series across space.** For a temporal lattice let **X(s)** : $1 \times T$ be vector of temporal responses at - use multivariate spatial methods. Eg.co–Kriging; BSP.

Specialized approaches

- **Approach 4: Build a statistical framework on physical models that describe the evolution of physical processes over time**

EXAMPLE: THE DLM

Combine dynamic linear models across space to get spatial predictor & temporal forecastor. **Result**: model for hourly $\sqrt{(O_3)}$

field over Mexico City - data from 19 monitors in Sep 1997.
Measurement model:

$$X(s,t) = \beta(t) + S'(t)\alpha(s,t) + Z(s,t)\gamma(t) + \epsilon(s,t)$$

where

- $S_t : 2 \times 1$ has sin's and cos's;
- α has their amplitudes, Z temperature covariate
- $\epsilon(s,t)$: un-autocorrelated error with isotropic exponential spatial covariance.

Specialized approaches: Eg DLM

Process model:

$$\begin{aligned}\beta(t) &= \beta(t-1) + \omega^{\beta}(t)\\ \alpha(s,t) &= \alpha(s,t-1) + \omega^{\alpha}(s,t)\\ \gamma(t) &= \gamma(t-1) + \omega^{\gamma}(t)\end{aligned}$$

SPECIALIZED APPROACHES: EG DLM

PROS:

- intuitive, flexible
- allows incorporation of physical/prior knowledge

CONS:

- computationally intensive - maximum of 10 measurement sites
- non - unique model specification - finding good one can be difficult
- unrealistic covariance
- empirical tests suggest simpler multivariate BSP works better for spatial prediction and temporal forecasting but much less computationally demanding, Eg. 300 measurement sites

PHYSICAL STATISTICAL MODELLING

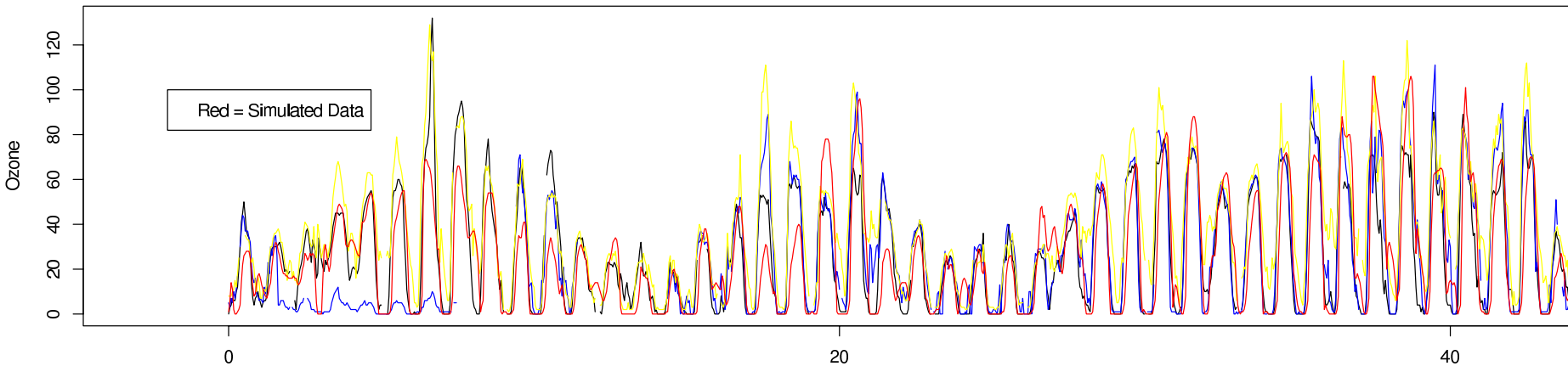
- physical models needed for background
  - prior knowledge often expressed by differential equations (de's)
  - can lead to big computer models
  - yield deterministic response predictions
  - can encounter difficulties:
    - butterfly effect
    - nonlinear dynamics
    - lack of relevant background knowledge
    - lack of sufficient computing power

PHYSICAL STATISTICAL MODELLING

- statistical models also desirable
  - prior knowledge expressed by statistical models
  - often lead to big computer models
  - yield predictive distributions
  - can encounter difficulty:
    - off-the-shelf-models too simplistic
    - lack of relevant background knowledge
    - lack of sufficient computing power

EXAMPLE: CHEMICAL TRANSPORTATION MODELS

MAQIP hourly ozone concentration prediction model outputs version data. A CMAQ prototype. Red is from the model. Blue are the data.



PHYSICAL STATISTICAL MODELLING

May be strength in unity but:

- big gulf between two cultures
- communication between camps difficult
- approaches different
- route to reconciliation unclear

PHYSICAL STATISTICAL MODELLING

Approach to reconciliation - depends on: purpose; context; # of (differential) equations; etc.

With many equations (e.g. 100):

- build a better predictive response density for [field response | deterministic model outputs]
  eg. input model value as prior mean
- view model output as response and create joint density for
  $[\text{field response}, \text{model output}] =$
  $\int [\text{field response}|\lambda][\text{model output}|\lambda] \times \pi(\lambda|\text{data})d\lambda$

PHYSICAL STATISTICAL MODELLING

With a few differential equations (de's)

Example: $dX(t)/dt = \lambda X(t)$.

- **Option 1:** solve it and make known or unknown constants uncertain (i.e. random):
  $X(t) = \beta_1 \exp \lambda t + \beta_0$
- **Option 2:** discretize the de and add noise to get a state space model: $X(t+1) = (1+\lambda)X(t) + \epsilon(t)$
- **Option 3:** use functional data analytic approach - incorporate de through a penalty term as in splines
  $\sum_t (Y_t - X_t)^2 + (\textit{smoothing parameter}) \int (DX - \lambda X)^2 dt$

DOWNSCALING PHYSICAL MODELS

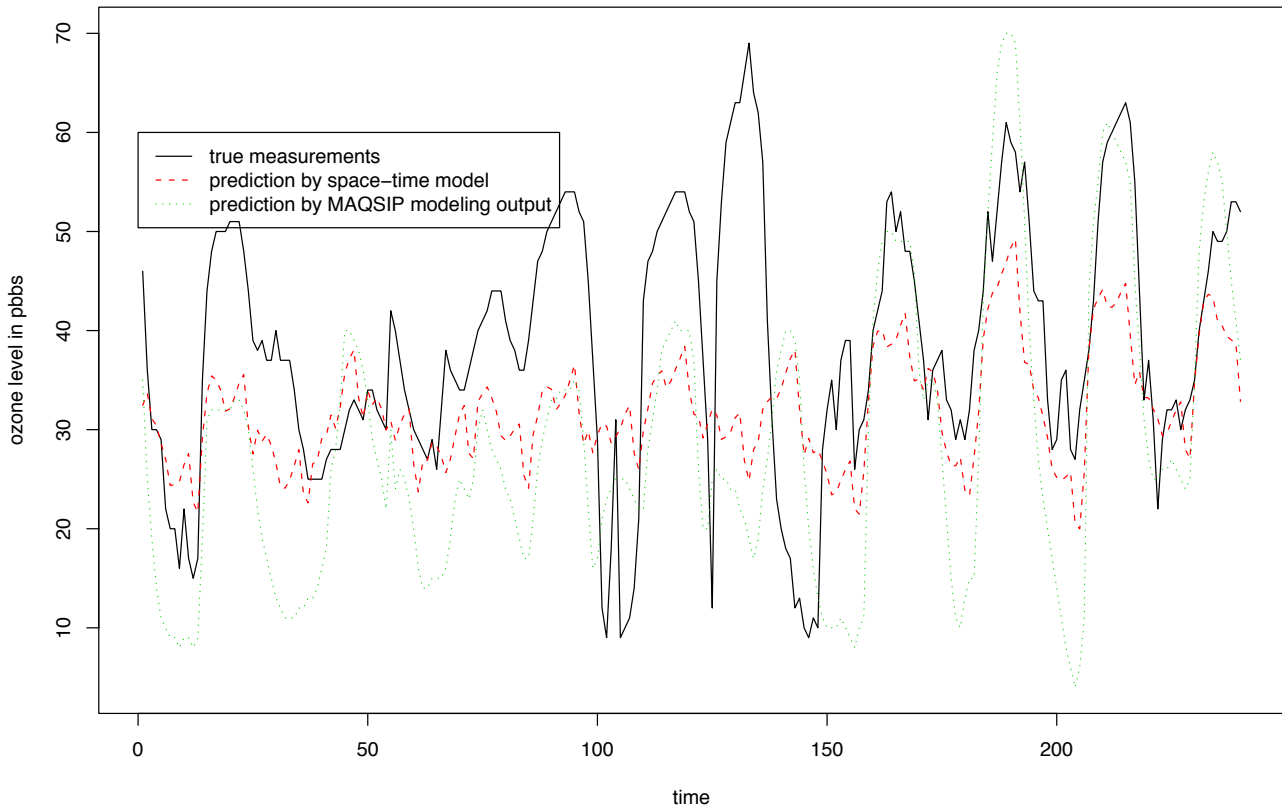
Regression – like approaches may be used:

$$X(s,t) = \alpha_{st} + \beta_{Mst} M(S,T) + \beta_{st} Z^{\text{covariates}}(s,t)\delta(s,t)$$

where M is physical model output, $s \in S^{\text{grid cell}}$ & $t \in T^{\text{Time Interval}}$.

EXAMPLE: MAQSIP REVISITED

MAQIP hourly ozone concentration prediction model outputs version the downscaling model above.



Modelling multiple pollutants at multiple sites: a case study in Bayesian hierarchical modelling using MCMC

Aims

- Investigate the spatial-temporal modelling of pollutants.
- Assess the contribution of different components of variability; spatial, temporal and random variability.
- Develop methodology to provide:
  - exposures (and measures of uncertainty) for use in mapping of environmental factors
  - studies investigating the health effects of pollution.
- Fit models and perform analyses in MCMC.

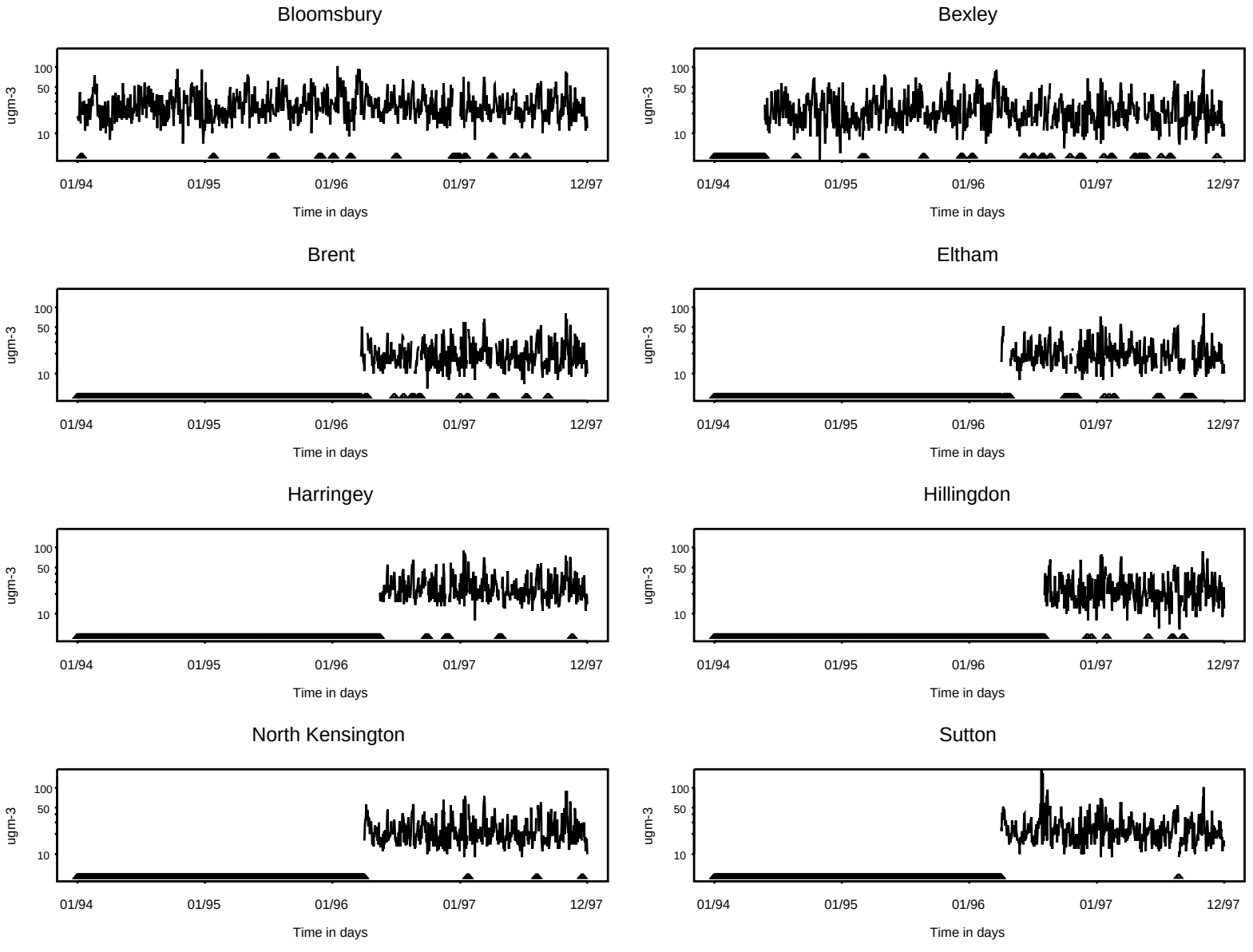
Background

- Daily measurements often available for different pollutants from a number of sites
- May be subject to measurement error
- Contain missing values
  - Pollutants not measured at all sites
  - Monitor being moved by design, e.g. six-day monitoring schedule
  - Unreliable or faulty monitors

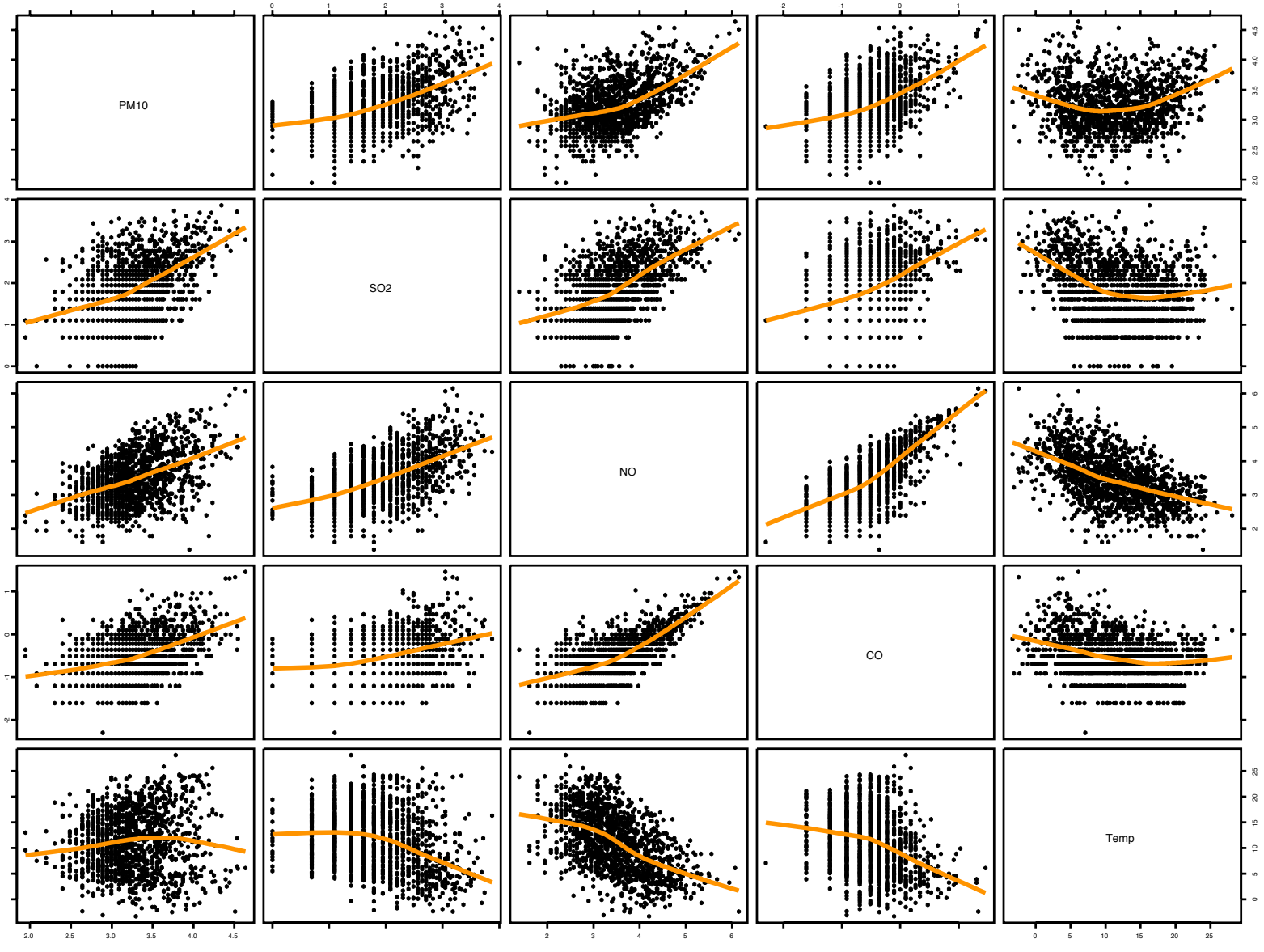
DATA

- Eight sites within London, 1997-94
- PM_{10}, SO_2, NO and CO.
- All pollutants only measured at only 4 sites.
- Periods of operation between 1 and 4 years.
- Percentage of missing values as great as 37%.

Time series plots of (logged) values of PM_{10}



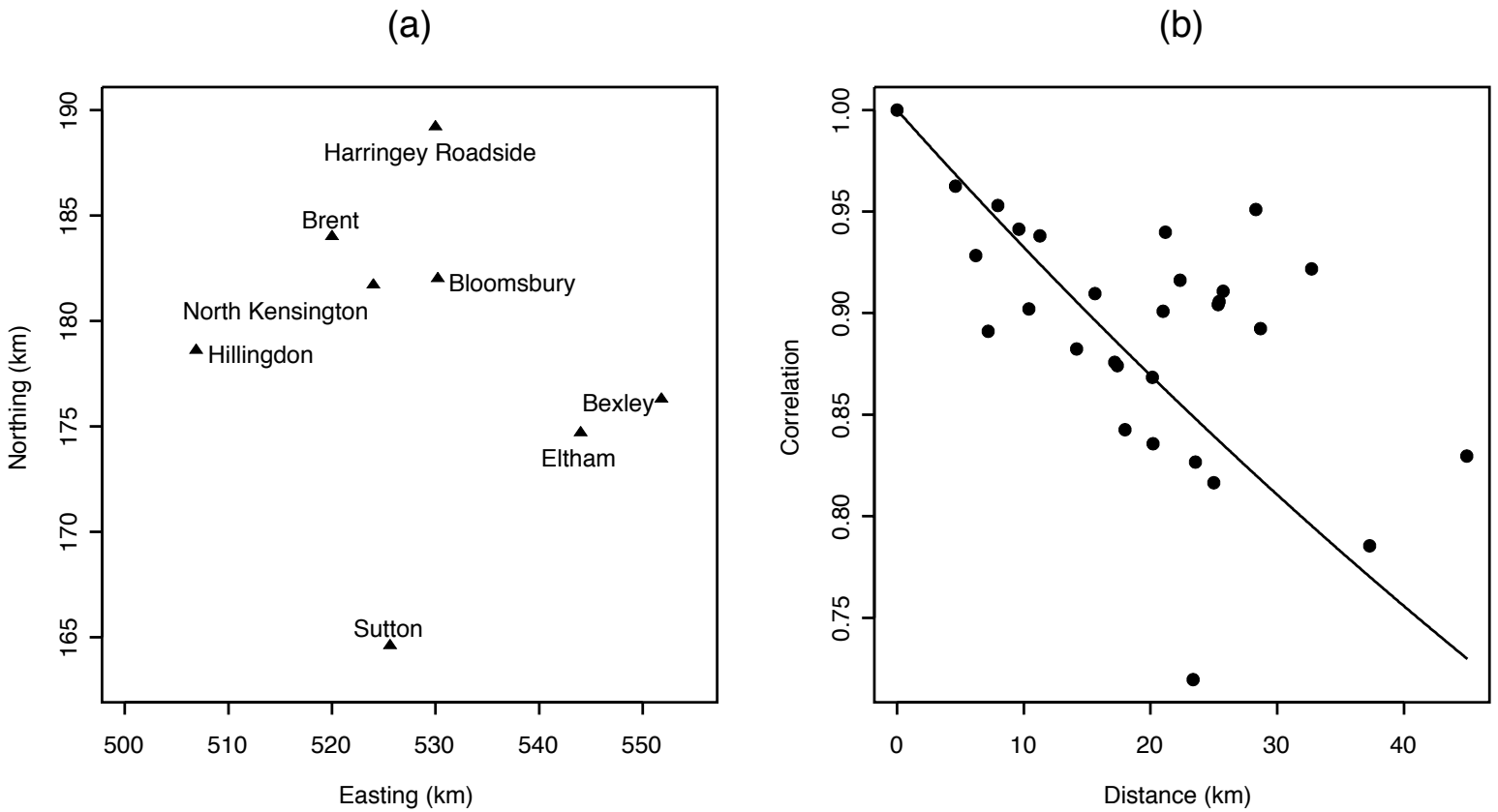
Correlations between pollutants and temperature



Data dependencies

- There are dependencies, both temporally and spatially, between daily measurements of different pollutants.
  - Pollutant dependence - common processes by which they are formed and the relationship with meteorological conditions.
  - Temporal dependence - atmospheric lifetimes and relationship with meteorological conditions.
  - Spatial dependencies - distance between sites and site type.

Locations of monitoring sites and correlations with distance



Model framework

- Bayesian hierarchical model.
- Pollutants modelled as a function of the true underlying level with measurement error.
- Incorporate covariate information, e.g. temperature.
- Underlying level is a function of the previous day's level.
- Missing values treated as unknown parameters within the Bayesian framework and can be estimated.

SINGLE POLLUTANT, SINGLE MONITORING SITE

- **Stage One, Observed Data Model:**

$$Y_t = X_t^T \beta_1 + \theta_t + v_t,$$

v_t is referred to as *measurement error*, and assumed to be are independent and identically distributed (i.i.d.) as $N(0, \sigma_v^2)$

- **Stage Two, Temporal Model:**

Autoregressive first order model

$$\theta_t = \rho \theta_{t-1} + w_t$$

w_t i.i.d. as $N(0, \sigma_w^2)$.

SINGLE POLLUTANT, SINGLE MONITORING SITE

- **Stage Three, Hyperprior:**
  Normal prior $N(c, C)$ for β_1, where c is a $q_1 \times 1$ vector and C a $q_1 \times q_1$ variance-covariance matrix.
  $\sigma_v^{-2} \sim Ga(a_v, b_v)$ and $\sigma_w^{-2} \sim Ga(a_w, b_w)$.

Posterior distribution The posterior distribution is given by

$$p(\theta, \beta_1, \sigma_v^2, \sigma_w^2|y) = p(y)^{-1} \left\{ \prod_{t=1}^{T} p(y_t|\theta_t, \beta_1, \sigma_v^2) \right\} \times$$
$$\left\{ \prod_{t=2}^{T} p(\theta_t|\theta_{t-1}, \sigma_w^2) \right\} \times$$
$$p(\theta_1)p(\beta_1)p(\sigma_v^2)p(\sigma_w^2)$$

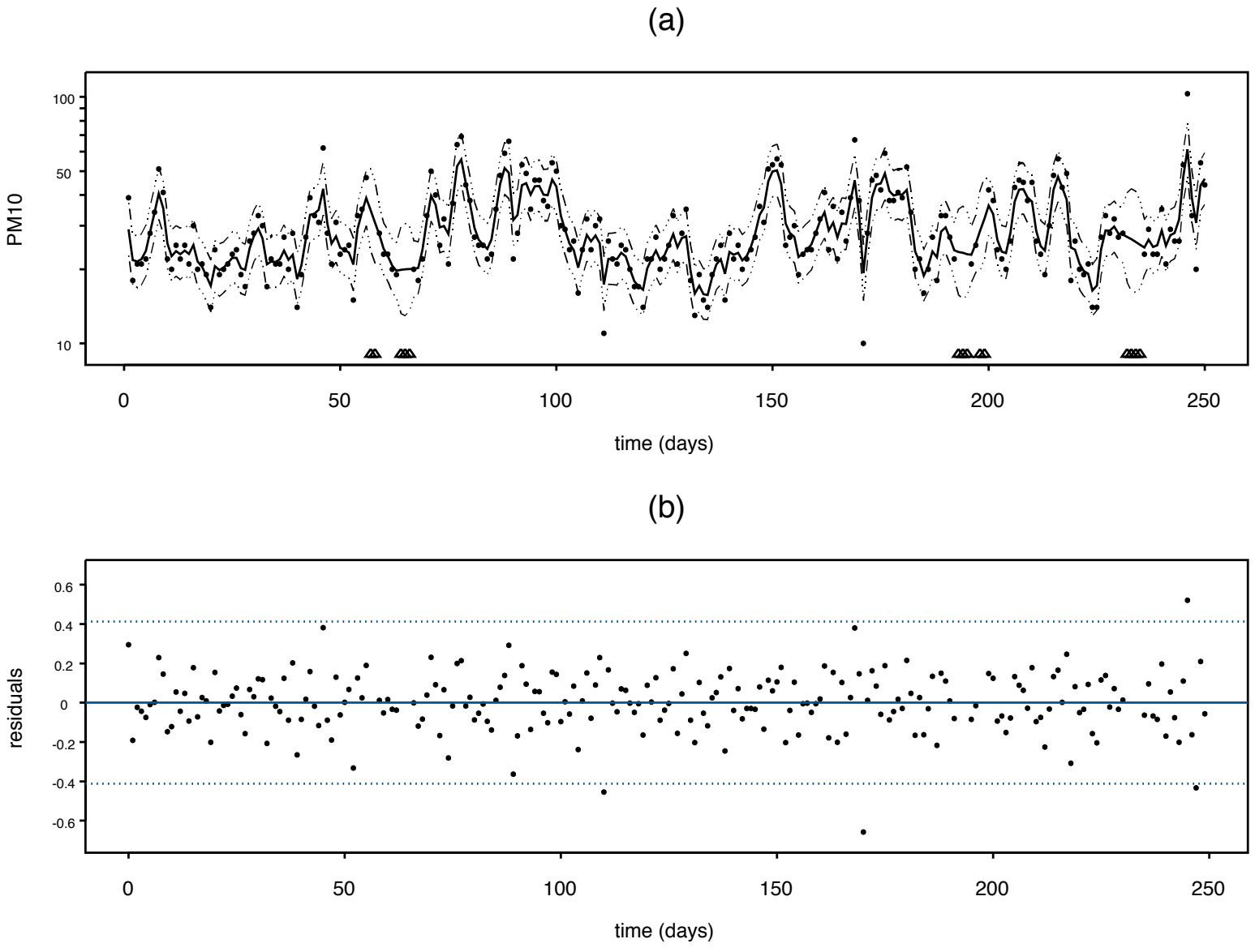
- Samples may be generated in a straightforward fashion using Markov chain Monte Carlo
- Dealing with the cyclical graph that arises at stage two, requires some of the conditional distributions to be explicitly specified

- Missing values are treated as parameters and the posterior obtained over these values and the model parameters. Samples can be generated from the distribution of missing values

$$p(y_m|y_o) = \int p(y_m|\lambda)p(\lambda|y_o)d\lambda$$

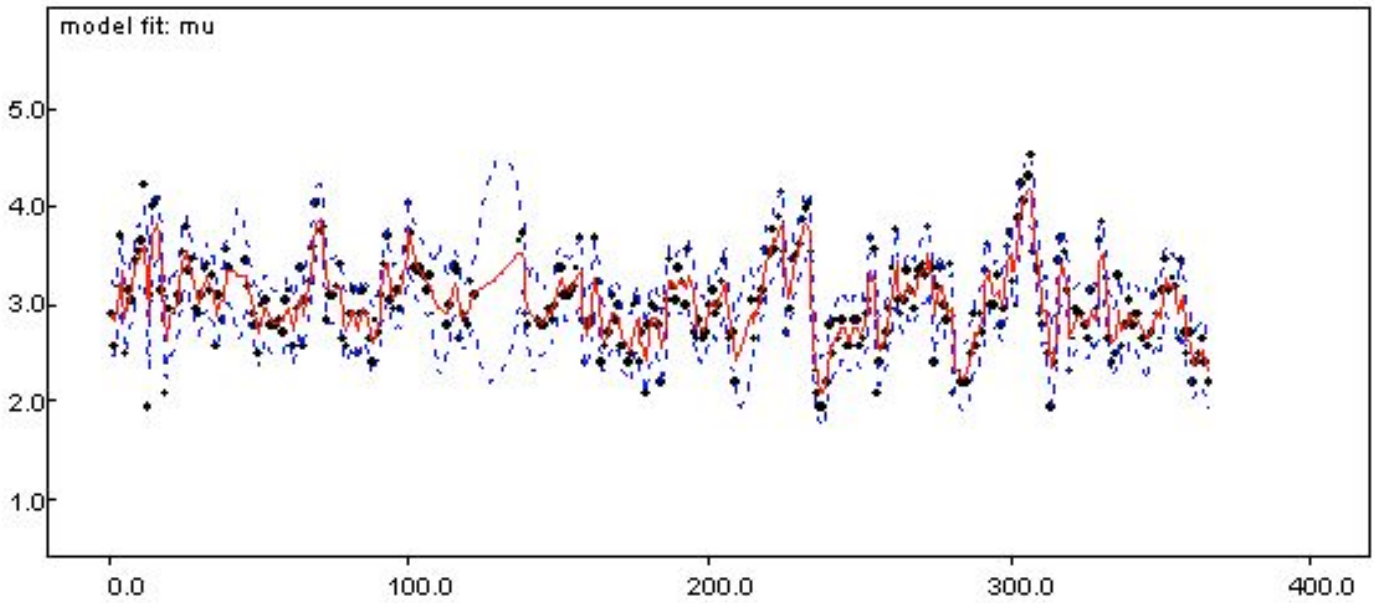
where $\lambda = (\theta, \beta_1, \sigma_v^2, \sigma_w^2)'$

Time series of 250 days of observed and estimated levels (together with their differences) of PM_{10} at Bloomsbury

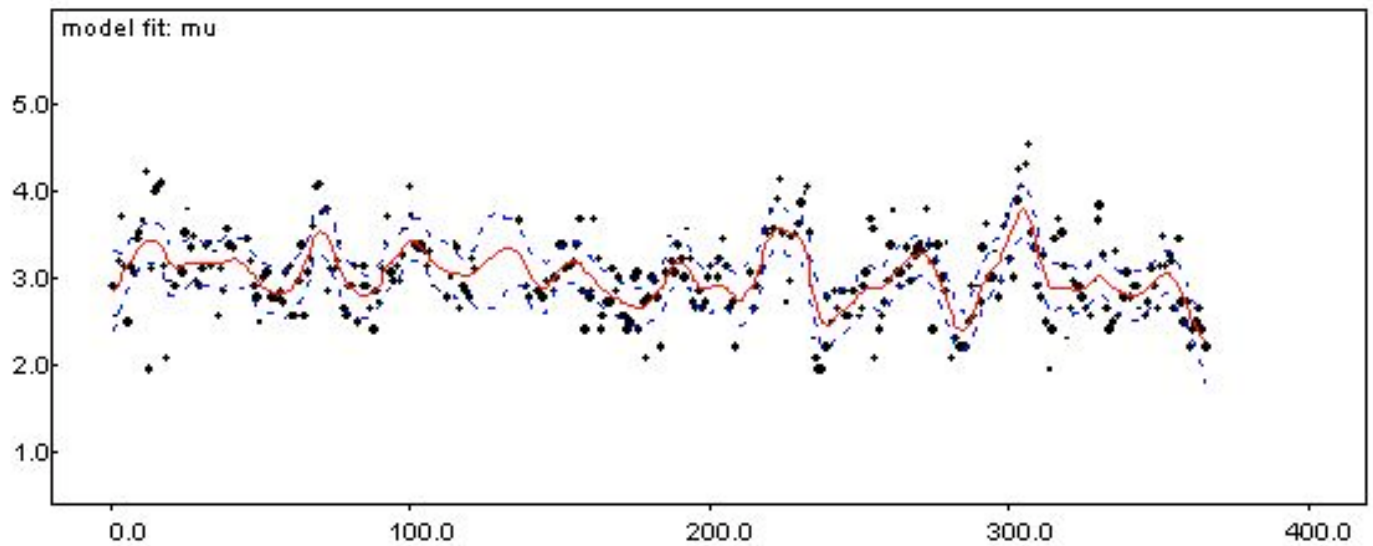


- The following is a plot of posterior median (red line) and posterior 95% intervals (dashed blue lines) for `mu[t]` (the underlying mean daily pollutant concentration), with observed concentrations shown as black dots.
- This plot was produced by selecting the model fit option from the Compare menu (available from the Inference menu), with `mu` specified as the node, day as the axis and y as other).
- Note that the dashed blue line shows the posterior 95% interval for the estimated mean daily concentration, and is not a predictive interval - hence we would not necessarily expect all of the observed data points to lie within the interval.

Using RW(1) model



Equivalent plot assuming an RW(2) prior. Note the greater amount of smoothing imposed by this prior



SINGLE POLLUTANT, MULTIPLE MONITORING SITE

- S monitoring sites measuring a single pollutant.
- The underlying autoregressive structure remains constant across sites with a constant adjustment in the mean level for site s by an amount m_s, $s = 1, ..., S$.
- **Stage One, Observed Data Model:**

$$Y_{st} = X'_{st}\beta_1 + X'_s\beta_2 + m_s + \theta_t + v_{st}$$

with v_{st} i.i.d. as $N(0, \sigma^2_{vs})$ and β_1, β_2, $q_1 \times 1$ and $q_2 \times 1$ vectors of site/day and site only regression coefficients.

- **Stage Two (a), Temporal Model:**

$$\theta_t = \rho\theta_{t-1} + w_t$$

with w_t i.i.d. as $N(0, \sigma^2_w)$.

- **Stage Two (b), Spatial Model:** The random effects $m = (m_1, ..., m_S)'$ arise from the multivariate normal distribution

$$m \sim MVN(0_S, \sigma_m^2 \Sigma_m),$$

where 0_S is an $S \times 1$ vector of zeros,
σ_m^2 the between-site variance and
Σ_m is the $S \times S$ correlation matrix, in which element (s, s') represents the correlation between sites s and s'.

- This model is stationary and assumes an isotropic covariance model in which the correlation between sites s and s' is assumed to be a function of the distance between them

$$f(d_{ss'}, \phi) = \exp(-\phi d_{ss'})$$

where $\phi > 0$ describes the strength of the correlation

- A simpler model assumes that the site-specific levels are (conditionally) independent

$$m_s \sim \text{ i.i.d } N(0, \sigma_m^2),$$

- **Stage Three, Hyperpriors:**
  - Unless there is specific information to the contrary, i.e. that a monitor with different characteristics is used at a particular site, we will assume $\sigma_{vs}^{-2} \sim Ga(a_v, b_v)$.
  - The between site precision has prior $\sigma_m^{-2} \sim Ga(a_m, b_m)$.
  - A uniform prior is used for ϕ, with the limits being based on beliefs about the relationship between correlation and distance.
  - The distance, d, at which the correlation, ρ, between two sites might be expected to fall to a particular level would be $d = -\log(\rho)/\phi$.

ESTIMATING LEVELS AT UNMEASURED LOCATIONS

- Based on the posterior estimates of the site effects, m_s and the variance-covariance matrix $\sigma_m^2 \Sigma_m$, it is possible to estimate the site effects, and thus pollution levels, at locations where there is no monitoring site.
- For a site at a new location, m_{S+1}, $(m_1, ..., m_S, m_{S+1})$ follows a multivariate normal distribution with zero mean and $(S+1) \times (S+1)$ variance-covariance matrix.
- Letting $m = (m_1, .., m_S)'$, the conditional distribution of $m_{S+1}|m$ is, normal with mean and variance given by

$$E[m_{S+1}|m] = \sigma_m^{-2} \Omega' \Sigma_m^{-1} m,$$

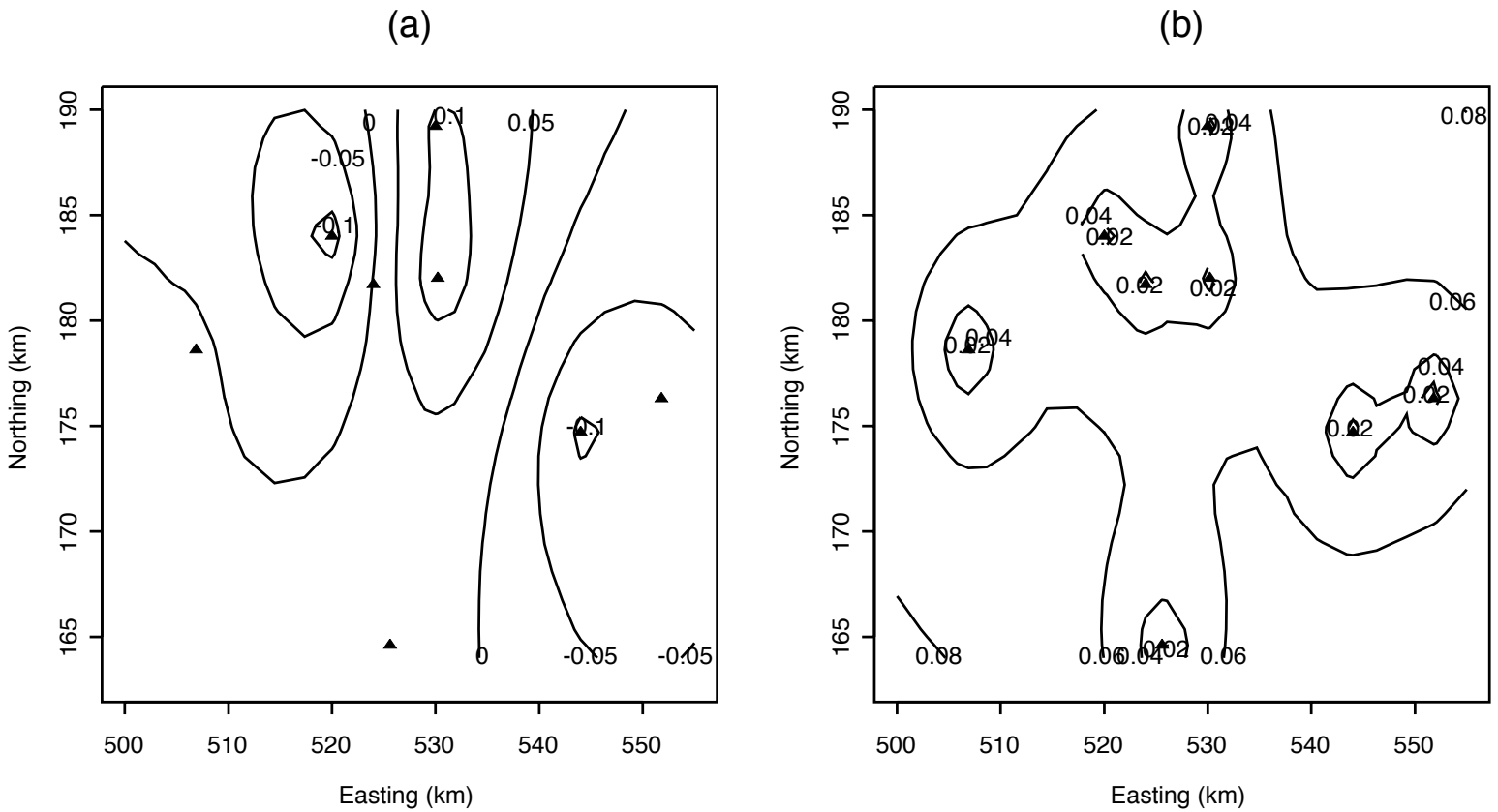
$$var(m_{S+1}|m) = \sigma_m^2 (1 - \Omega' \Sigma_m^{-1} \Omega),$$

- For exploratory purposes, the posterior medians may be substituted into these expressions (although this will ignore the inherent uncertainty in the estimates).

SITE EFFECTS

	Median	2.5%	97.5%
Bexley	-0.0696	-0.0785	-0.0607
Bloomsbury	0.1341	0.1257	0.1426
Brent	-0.1210	-0.1294	-0.1125
Eltham	-0.1105	-0.1205	-0.1005
Harringey	0.1098	0.0999	0.1195
Hillingdon	0.0132	-0.0032	0.0300
North Kensington	0.0030	-0.0031	0.0090
Sutton	0.0410	0.0250	0.0572
σ_m	0.1019	0.0668	0.1794
ϕ	0.05675	0.02158	0.09778

Contour plot of site effects based on a 20x20 grid of locations without a pollution monitor with corresponding standard deviations



MULTIPLE POLLUTANTS, SINGLE MONITORING SITE

- **Stage One, Observed Data Model:**

$$Y_{pt} = X_t'\beta_1 + \theta_{pt} + v_{pt}$$

with v_{pt} i.i.d. as $N(0, \sigma^2_{vp})$ and β_1 a $q_1 \times 1$ vector of regression coefficients.

- **Stage Two, Temporal and Pollutant Model:**

$$\theta_{pt} = \theta_{p,t-1} + w_{pt}$$

$w_t = (w_{1t}, ..., w_{Pt})'$ are i.i.d. multivariate normal random variables with zero mean and variance-covariance matrix Σ_P.

- **Stage Three, Hyperpriors:**
$\sigma^{-2}_{vp} \sim Ga(a_v, b_v)$, $p = 1, ..., P$. $\Sigma_P^{-1} \sim W_P(D, d)$, a $P-$dimensional Wishart distribution with mean D and precision parameter d.

- Model was applied to data from four pollutants (PM_{10}, $S0_2$, NO and CO) from the Bloomsbury site.
- Priors $\sigma_{vp}^{-2} \sim Ga(1, 0.01)$, $p = 1, .., P$, and $\beta_1 \sim N(0, 1000)$.
- For the parameters of the Wishart distribution, d was chosen to be equal to four, the dimension of Σ_P;
  D was then chosen so that the diagonals of the expected value (D/d) represent a 10% coefficient of variation. The off-diagonals were taken to be zero.
- Posterior correlations

	PM_{10}	SO_2	NO	CO
PM_{10}	1.0000	0.8806	0.8192	0.8134
SO_2	0.8806	1.0000	0.8472	0.9202
NO	0.8192	0.8472	1.0000	0.9146
CO	0.8134	0.9202	0.9146	1.0000

- Strong correlations mean that inference on missing values can be made on the values of pollutants

MULTIPLE POLLUTANTS, MULTIPLE MONITORING SITES

- **Stage One, Observed Data Model:**

$$Y_{spt} = X'_{pt}\beta_1 + X'_{st}\beta_2 + \theta_{pt} + m_s + v_{spt},$$

where v_{spt} are i.i.d. $N(0, \sigma^2_{sp})$, β_1 a $q_1 \times 1$ vector of pollutant regression coefficients, and β_2 a $q_2 \times 1$ vector of spatial regression coefficients.

- **Stage Two, Spatial, Temporal and Pollutant Model:**
The $(p \times 1)$ vector of daily pollution measurements, $(\theta_1, ..., \theta_P)'$, as a function of the previous days values with possible correlation between the values of the different pollutants.
An alternative approach would be to allow the spatial effects to be pollutant specific
- **Stage Three, Hyperprior:** In the absence of additional information, we assume that $\sigma^{-2}_{vsp} \sim Ga(a_v, b_v)$.

Components of variability

- Model 1 (Single pollutant, single site)
  - Temporal 70%
  - Measurement error 30%
- Model 2 (Single pollutant, multiple sites)
  - Temporal 80%
  - Spatial 10%
  - Measurement error 10%
- Model 3 (Multiple pollutants, single site)
  - Temporal 77%
  - Measurement error 23%
- Model 4 (Multiple pollutants, multiple sites)
  - Temporal 75%
  - Spatial 15%
  - Measurement error 10%

SUMMARY

- Examine the contribution of spatial, temporal and random variability.
- Allows levels to be estimated at non-measured locations.
- Calculate underlying levels of pollution for use in health studies.
- Estimates of missing values.