Big Data in Environmental Research

Gavin Shaddick

University of Bath

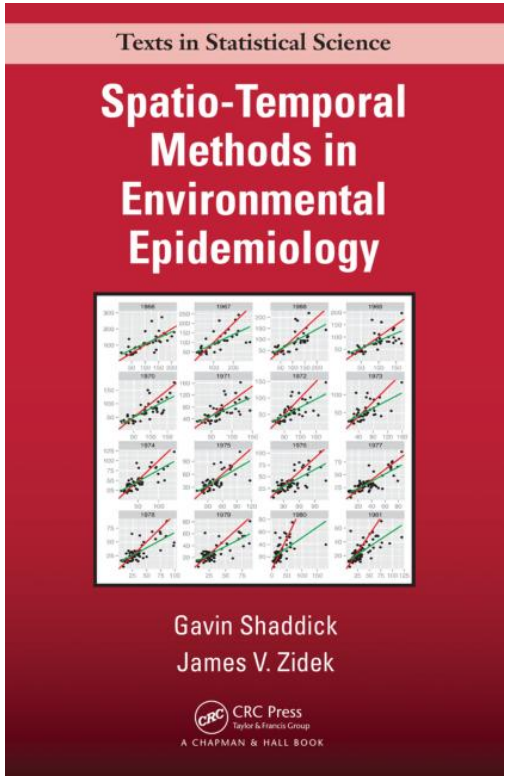
7^{th} July 2016

Outline

- *10:00 - 11:00* Introduction to Spatio-Temporal Modelling
- *14:30 - 15:30* Bayesian Inference
- *17:00 - 18:00* Examples of Spatial Modelling

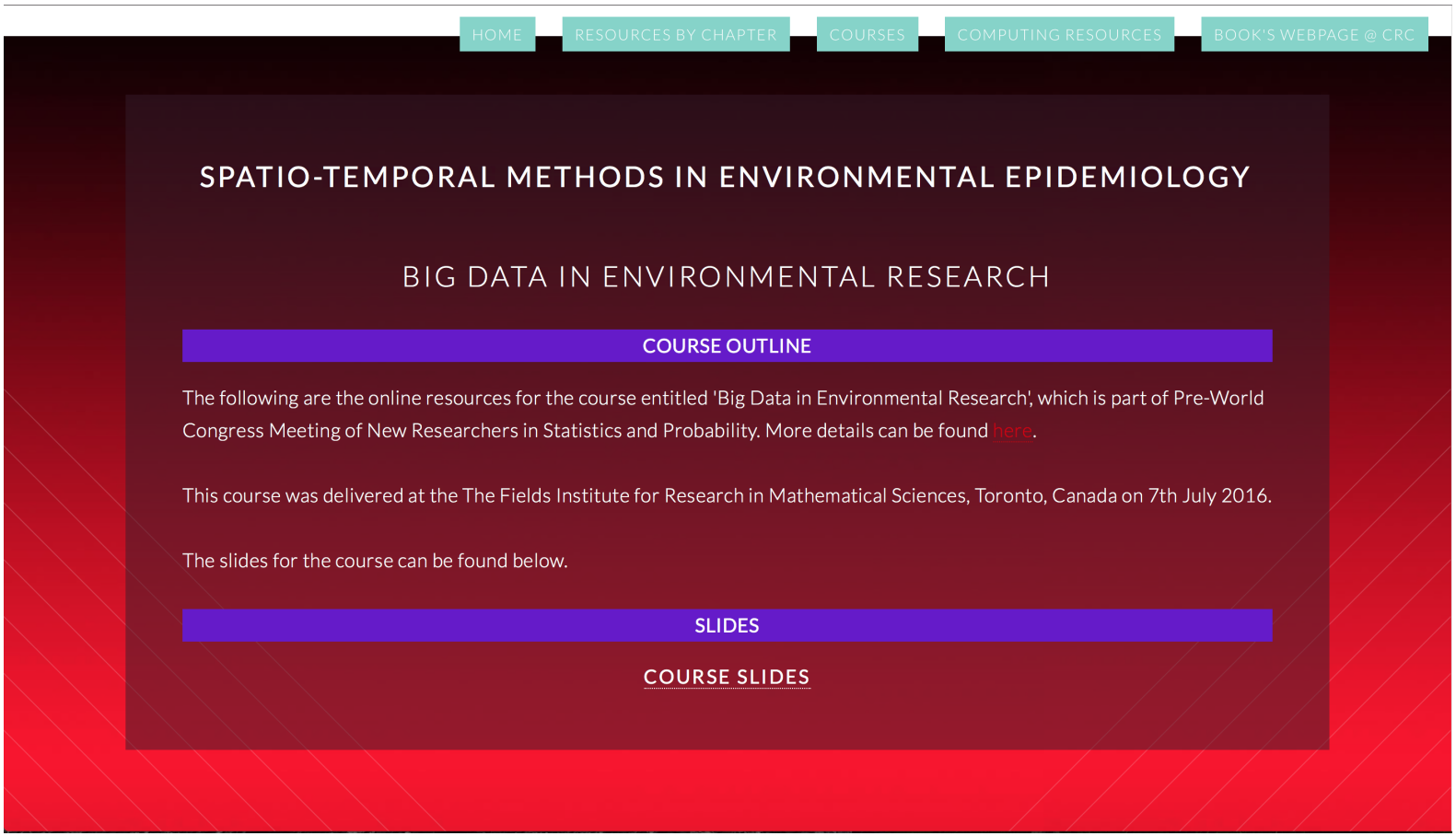
TEXTBOOK

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/`



WEBSITE

http://stat.ubc.ca/~gavin/STEPIBookNewStyle/course_toronto.html



Contact Information

Dr. Gavin Shaddick, University of Bath

- Email: G.Shaddick@bath.ac.uk
- Webpage: `http://people.bath.ac.uk/masgs/`

Session 1: Introduction to Spatio-Temporal Modelling

THE NEED FOR SPATIO-TEMPORAL MODELLING

- In recent years there has been an explosion of interest in spatio–temporal modelling.
- One major area where spatio-temporal is developing is environmental epidemiology, where interest is in the relationship between human health and spatio–temporal processes of exposures to harmful agents.

THE NEED FOR SPATIO-TEMPORAL MODELLING

- Example include the relationship between deaths and air pollution concentrations or future climate simulations, the latter of which may involve 1000's of monitoring sites that gather data about the underlying multivariate spatio–temporal field of precipitation and temperature.

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

- In order to assess and manage 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.

Dependencies over space and time

- Environmental epidemiologists commonly seek associations between a 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.

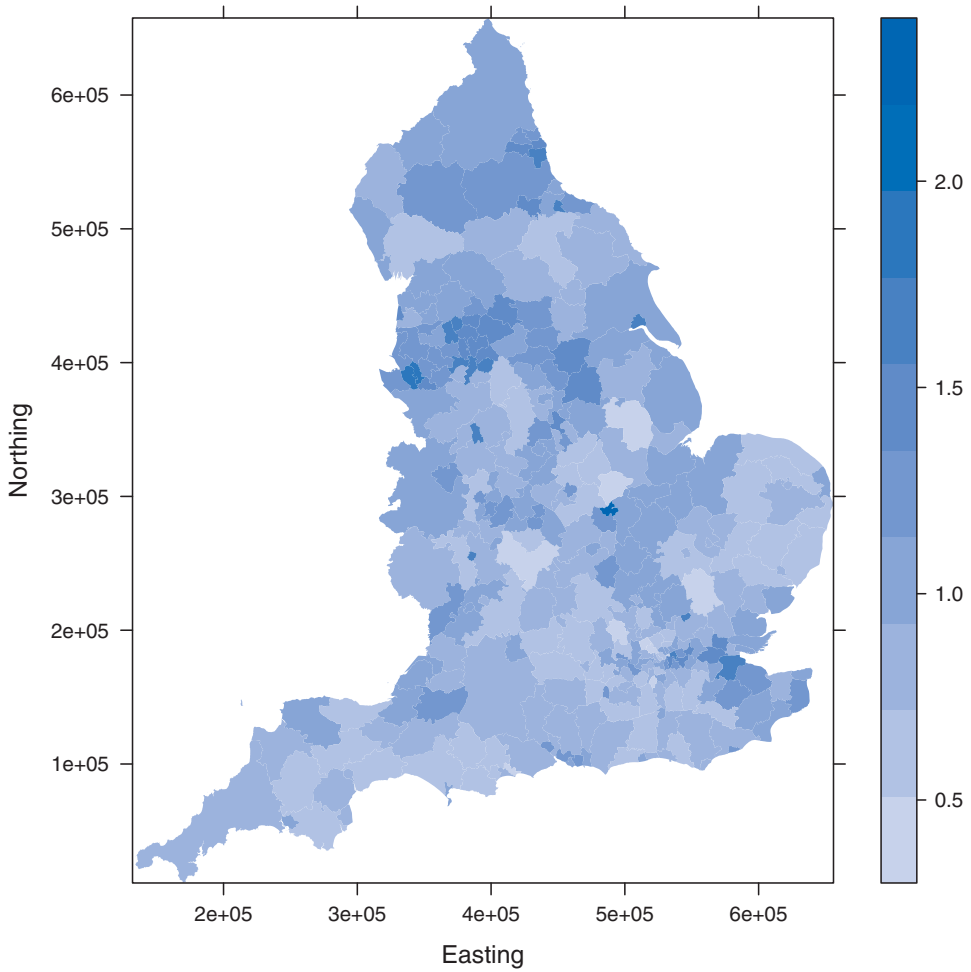


Figure: Map of the spatial distribution of risks of hospital admission for a respiratory condition, chronic obstructive pulmonary disease (COPD), in the UK for 2001. The shades of blue correspond to standardised admission rates, which are a measure of risk. Darker shades indicate higher rates of hospitalisation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

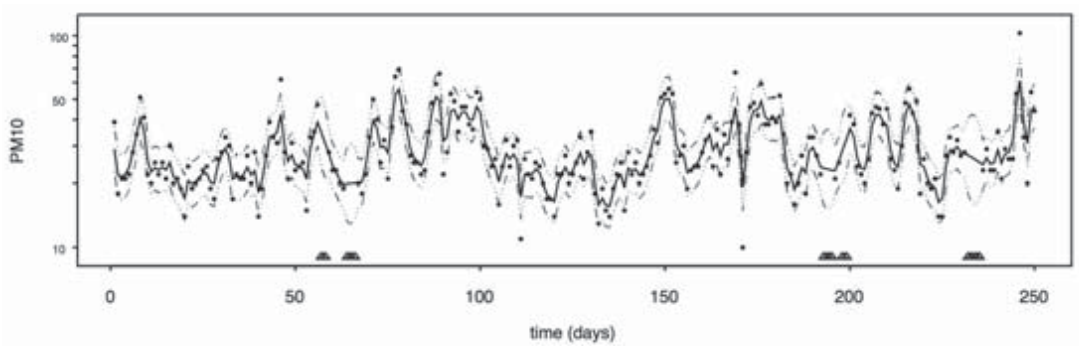


Figure: Time series of daily measurements of particulate matter (PM_{10}) for 250 days in 1997 in London. Measurements are made at the Bloomsbury monitoring site in central London. Missing values are shown by triangles. The solid black line is a smoothed estimate produced using a Bayesian temporal model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

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.

Dependencies over space and time

- These measures of uncertainty should reflect the inherent uncertainties that will be present at each of the stages in the modelling process.
- 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.

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

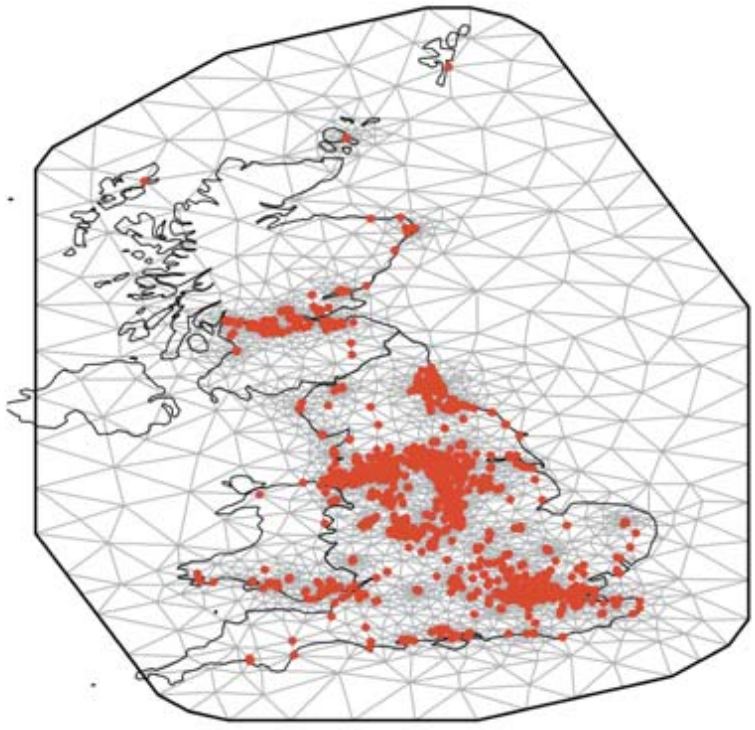


Figure: Triangulation for the locations of black smoke monitoring sites within the UK for use with the SPDE approach to modelling point-referenced spatial data with INLA. The mesh comprises 3799 edges and was constructed using triangles that have minimum angles of 26 and a maximum edge length of 100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

Example: Global Modelling of $PM_{2.5}$ using Multiple Data Sources

- Air pollution is an important determinant of health and poses a significant threat globally.
- It is known to trigger cardiovascular and respiratory diseases in addition to some cancers.
- Particulate Matter ($PM_{2.5}$) is estimated to be
  - 4th highest health risk factor in the world
  - attributable to 5.5 million premature deaths
- There is convincing evidence for the need to model air pollution effectively.

Example: Global Modelling of $PM_{2.5}$ using Multiple Data Sources

- WHO and other partners plan to strengthen air pollution monitoring globally.
- This will produce accurate and convincing evidence of risks posed.
- Allow data integration from different sources.
- This will allow borrowing from each methods respective strengths.
- Currently, three methods are considered:
  - Ground Monitoring,
  - Satellite Remote Sensing and
  - Atmospheric Modelling

Example: Global Modelling of $PM_{2.5}$ using Multiple Data Sources

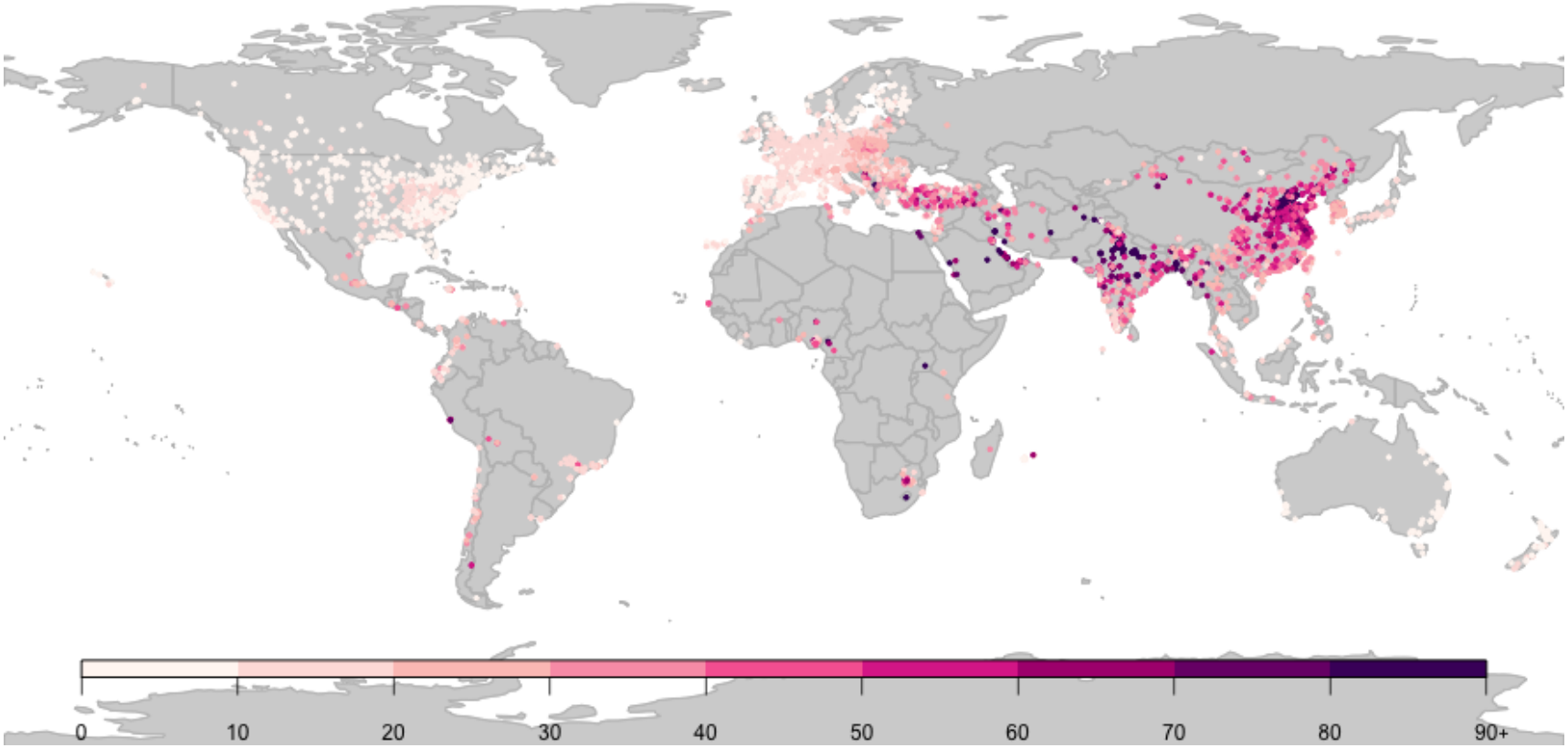


Figure: World map with ground monitor locations, coloured by the estimated level of $PM_{2.5}$

Example: Global Modelling of $PM_{2.5}$ using Multiple Data Sources

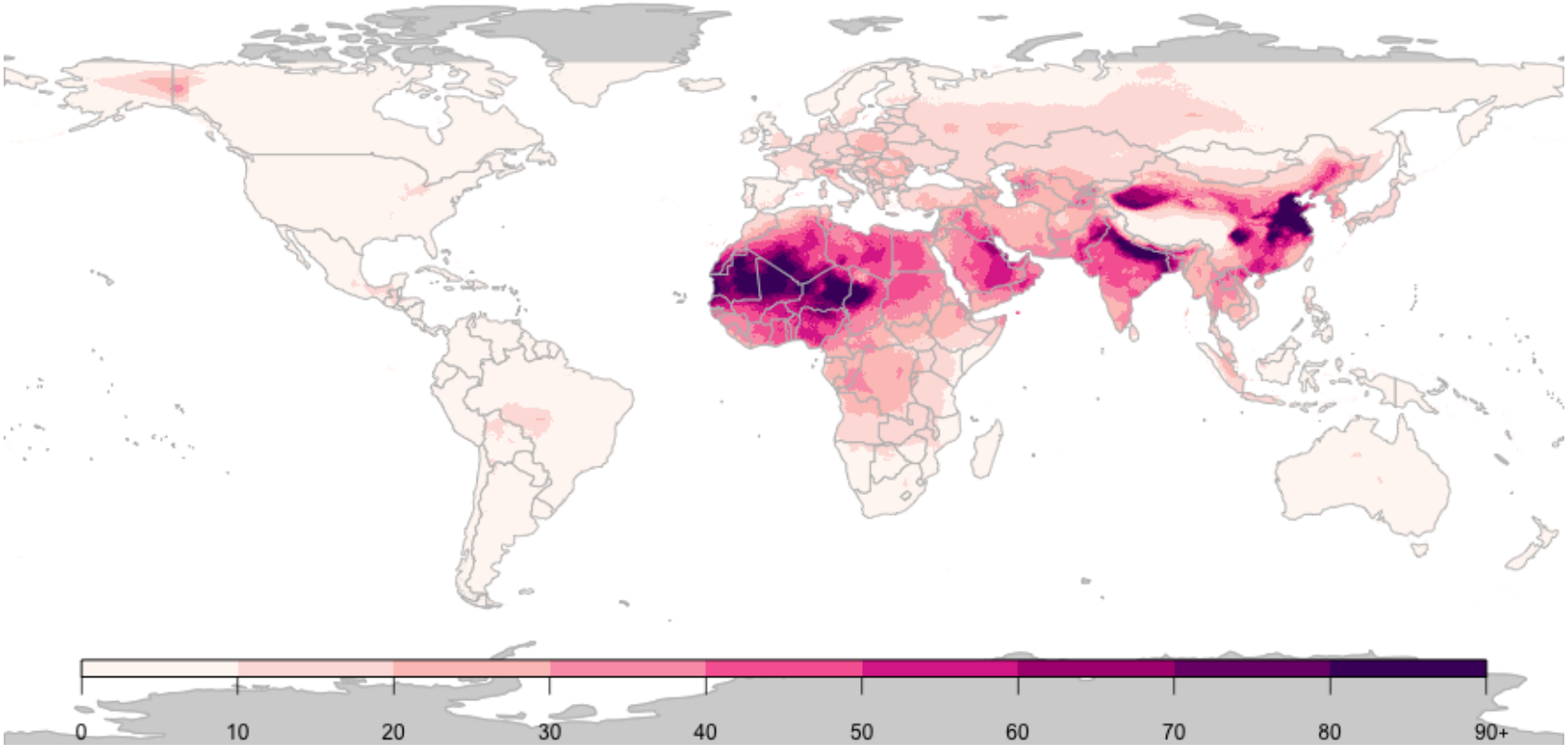


Figure: Global satellite remote sensing estimates of $PM_{2.5}$ for 2014.

Example: Global Modelling of $PM_{2.5}$ using Multiple Data Sources

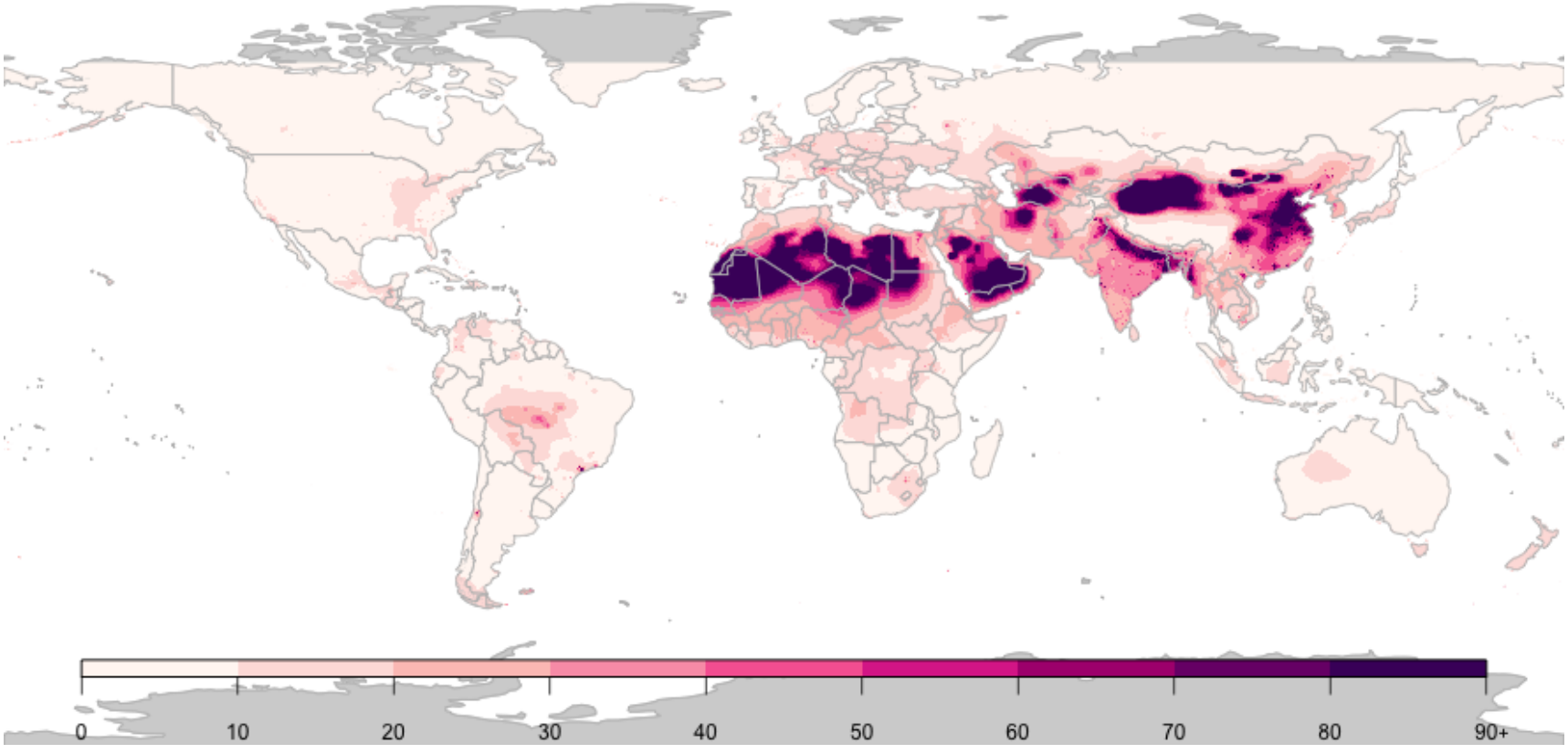


Figure: Global chemical transport model estimates of $PM_{2.5}$ for 2014.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- There are many ways in which space and time can be incorporated into a statistical model and we now consider a selection. One must first choose the model's spatio-temporal domain.
- Is it to be a continuum in both space and time?
- Or a discrete space with a finite number of locations at which measurements may be made?

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- Time is obviously different than space.
- For one thing, it is directed, whereas any approach to adding direction in space is bound to be artificial.
- A major challenge in the development of spatio–temporal theory has been combining these fundamentally different fields in a single modelling framework.
- Much progress has been made in this area over the last three or four decades to meet the growing need in applications of societal importance, including those in epidemiology.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- There are competing advantages to using finite (discrete) and continuous domains.
- Indeed a theory may be easier to formulate over a continuous domain, but practical use may entail projecting them onto a discrete domain.
- Time is regarded as discrete because measurements are made at specified, commonly equally spaced, time points.
- The precise methodology will be determined by the nature of the data that is available over space, for example is it point-referenced or collected on a lattice?

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

Some general approaches to incorporating time are as follows:

Approach 1: Treat continuous time as another spatial dimension,

- For example, spatio–temporal Kriging
- There is extra complexity in constructing covariance models compared to purely spatial process modelling and possible reductions in the complexity based on time having a natural ordering (unlike space) are not realised.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

Approach 2: Represent the spatial fields represented as vectors $\mathbf{Z}_t : N_S \times 1$, and combine them across time to get a multivariate time series.

Approach 3: Represent the time series as vectors, $\mathbf{Z}_s : 1 \times N_T$, and use multivariate spatial methods

- For example, co-kriging

Approach 4: Build a statistical framework based on deterministic models that describe the evolution of processes over space and time.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- Approach 1 may appeal to people used to working in a geostatistical framework.
- Approach 2 may be best where temporal forecasting is the inferential objective while Approach 3 may be best for spatial prediction of unmeasured responses.
- Approach 4 is an important new direction that has promise because it includes background knowledge through numerical computer models.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- If the primary aim is spatial prediction then you would want to preserve the structure of the spatial field.
- However if the primary interest is in forecasting this would lead to an emphasis in building time series models at each spatial location.
- The exact strategy for constructing a spatio–temporal model will also depend on the purpose of the analysis.

STRATEGIES FOR SPATIO–TEMPORAL MODELLING

- Interest may lie in forecasting an ambient measurement twenty-four hours ahead of time. Or to spatially predict such levels at unmonitored sites to get a better idea of the exposure of susceptible school children in a school far from the nearest ambient monitor.
- In deciding how to expand or contract an existing network of monitoring sites in order to improve prediction accuracy or to save resources, a spatio–temporal model will be required together with a criterion on which to evaluate the changes you recommend.

Spatio–temporal Processes

We can represent the spatio-temporal random field Z_{st} in terms of a hierarchical model for the measurement and process models

$$\begin{aligned} Y_{st} &= Z_{st} + v_{st} \\ Z_{st} &= \mu_{st} + \omega_{st} \end{aligned}$$

where

- v_{st} represents independent random measurement error.
- μ_{st} is a spatio–temporal mean field (trend) that is often represented by a model of the form $\mu_{st} = x_{st}\beta_{st}$.
- ω_{st} is the underlying spatio–temporal process

Spatio–temporal Processes

- For many processes the mean term μ_{st} represents the largest source of variation in the responses.
- Over a broad scale it might be considered as deterministic if it can be accurately estimated,
  - An average of the process over a very broad geographical area.
- However where there is error in modelling μ_{st} the residuals ω_{st} play a vital role in capturing the spatial and temporal dependence of the process.

Spatio–temporal Processes

- The spatio–temporal process modelled by ω can be broken down into separate components representing space, m, time, γ and the interaction between the two, κ.

$$\omega_{st} = m_s + \gamma_t + \kappa_{st}$$

- Here, **m** would be a collection of zero mean, site-specific deviations (spatial random effects) from the overall mean, μ_{st} that are common to all times.
- For time, γ would be a set of zero mean time-specific deviations (temporal random effects) common to all sites.
- The third term κ_{st} represents the stochastic interaction between space and time.

Spatio–temporal Processes

- For example, the effect of latitude on temperature depends on the time of year.
- The mean term, μ_{st} may constitute a function of both time and space but the interaction between the two would also be manifest in κ_{st}.
- This would capture the varying intensity of the stochastic variation in the temperature field over sites which might also vary over time. In a place such as California the temperature field might be quite flat in summer but there will be great variation in winter.
- It is likely that there will be interaction acting both through the mean and covariances of the model.

SEPARABLE MODELS

- In most applications, modelling the entire spatial–temporal structure will be impractical because of high dimensionality.
- A number of approaches have been suggested to deal with this directly and we now discuss the most common of these, that of assuming that space and time are **separable**.
- This is in contrast to cases where the spatio-temporal structure is modelled jointly which are known as **non-separable** models.

SEPARABLE MODELS

- Separable models impose a particular type of independence between space and time components. It is assumed the correlation between Z_{st} and $Z_{s't}$ is $\rho_{ss'}$ at every time point t while the correlation between Z_{st} and $Z_{t's}$ is $\rho_{tt'}$ at all spatial time points s.
- The covariance for a separable process is therefore defined as

$$Cov(Z_{st}, Z_{s't'}) = \sigma^2 \rho_{ss'} \rho_{tt'}$$

for all $(s,t), (s',t') \in \mathcal{S} \times \mathcal{T}$.

NON-SEPARABLE MODELS

- The complexity of non-separable spatio–temporal processes often combined with computational issues has resulted in the development of a number of different approaches to modelling them.
- We now provide a brief description of a selection of the available approaches.

NON-SEPARABLE MODELS

- A spatio–temporal model for hourly ozone measurements was developed by Carroll *et al.* (1997).
- The model,

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

combines a trend term incorporating temperature and hourly/monthly effects,

$$\mu_t = \alpha_{hour} + \beta_{month} + \beta_1 temp_t + \beta_2 temp_t^2,$$

which is constant over space, and an error model in which the correlation in the residuals was a nonlinear function of time and space.

NON-SEPARABLE MODELS

- In particular the spatial structure was a function of the lag between observations,

$$COV(v_{st}, v_{s't'}) = \sigma^2 \rho(d, v),$$

where d is the distance between sites and $v = |t' - t'|$ is the time difference, with the correlation being given by

$$\rho(d, v) = \begin{cases} 1 & d = v = 0 \\ \phi_v^d \psi_v & d \text{ otherwise} \end{cases}$$

where

$$\log(\psi_v) = a_0 + a_1 v + a_2 v^2 \text{ and } \log(\phi_v) = b_0 + b_1 v + b_2 v^2$$

NON-SEPARABLE MODELS

- The correlation of the random field is thus a product of two factors, the first, ψ_v^d depends on both the time and space, the second only on the time difference.
- Unfortunately, as Carroll *et al.* (1997) pointed out, this correlation function is not positive definite.
- Using results from the model, there were occasions when

$$Cov(Z_{st}, Z_{s't}) > Cov(Z_{st}, Z_{st}).$$

- This highlights a genuine lack of a rich set of functions that can be used as spatio-temporal correlation functions.

SUMMARY

In this section we have seen the many ways in which the time can be added to space in order to characterise random exposure fields. In particular we have looked at the following topics:

- Additional power that can be gained in an epidemiological study by combining the contrasts in the process over both time and space while characterising the stochastic dependencies across both space and time for inferential analysis.
- Criteria that good approaches to spatio–temporal modelling should satisfy.
- General strategies for developing such approaches.
- Separability and non-separability in spatio–temporal models, and how these could be characterised using the Kronecker product of correlation matrices.
- Examples of the use of spatio–temporal models in modelling environmental exposures.

Session 2: Bayesian Inference

Bayes' Theorem

- Bayes' theorem:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

  - For some it is just a theorem,
  - For others, it is a way of life!

Bayesian Inference

- It allows us to specify a model for some data Y in terms of some parameters θ in a *Likelihood* function:

$$p(Y|\theta)$$

and any a-priori knowledge about the model parameters in a *prior* probability distribution

$$p(\theta)$$

- Combining these, we can obtain the posterior distribution

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

Bayesian Inference

- The denominator $p(Y)$ is the marginal distribution of the observation Y
- This a normalisation constant so we often take proportionality with respect to θ

$$Posterior \propto Likelihood \times Prior$$
$$p(\theta|Y) \propto p(Y|\theta) \times p(\theta)$$

- $p(Y)$ will be of the form

$$p(Y) = \int p(Y|\theta)p(\theta)\, d\theta$$

- This integral is often analytically intractable and thus we must use other techniques to be able to find the posterior

Latent Gaussian Models

- Suppose we have observation vector Y that arises from some distribution.
- We are often interested in estimating the mean $\boldsymbol{\mu}$ which is related to the linear predictor,

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ji} + \sum_{k=1}^{q} f_k(U_{ki}) + \epsilon_i, \qquad i = 1, \ldots, n$$

where

- β_0 is an intercept term
- β_j is the linear effect of covariates X_{ji}
- ϵ_i is the iid noise term (i.e. $\epsilon_i \sim N(0, \sigma_\epsilon^2)$)
- $f_k(\cdot)$ is non-linear function of covariate U_{ji}. We often represent this function as $f_k(s) = \sum_m \gamma_{km} \psi_{km}(s)$ where $\psi_{km}(\cdot)$ are the basis functions and γ_{km} are the weights.

Latent Gaussian Models

- By letting,

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T \quad Z = (\beta_0, \ldots, \beta_p, \{\gamma_{km}\})^T$$

  we can write this as a linear system

$$\boldsymbol{\eta} = AZ$$

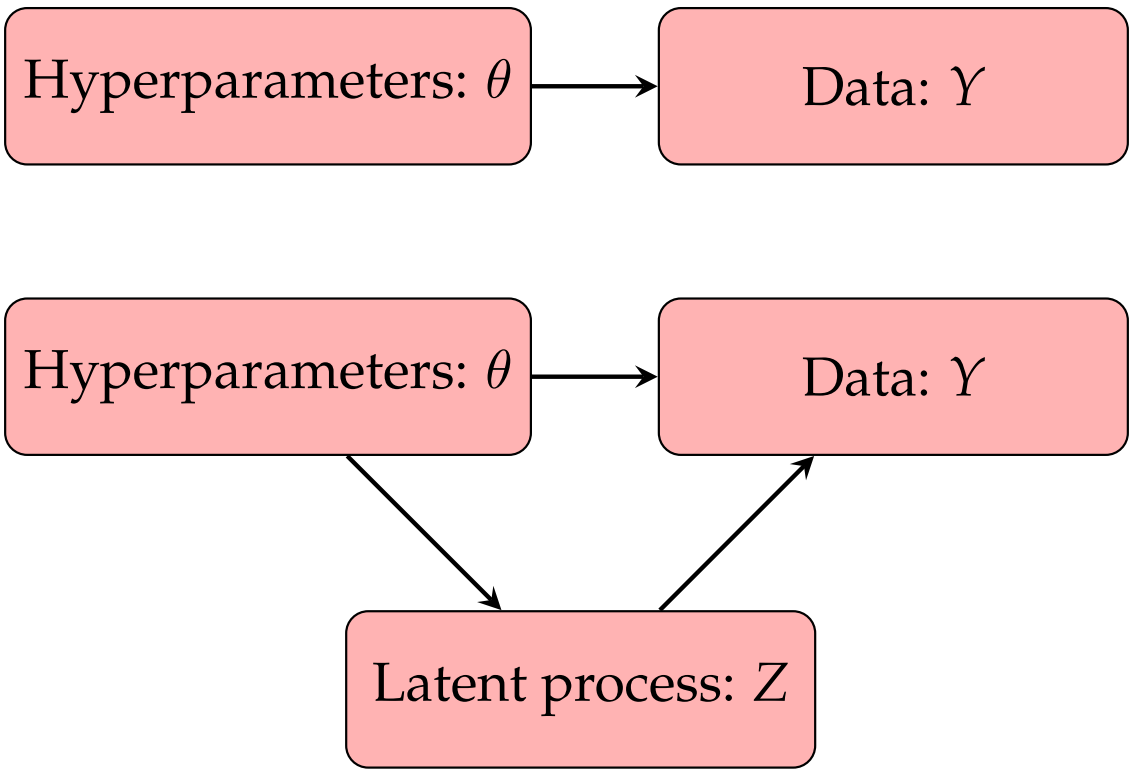
- A model is then classed as a latent Gaussian model if we assign a Gaussian distribution to the vector Z i.e.

$$Z \sim N(\boldsymbol{\mu}, \Sigma)$$

  where $\boldsymbol{\mu}$ is the mean vector and Σ is a positive-definite covariance matrix.

- We then define hyperparameters θ to account for scale of dependency and variability.

- This offers a very flexible framework so that we can work with a large range of models.



Bayesian Hierarchical Models

- The observation level $Y|Z, \theta$ - Data Y, are assumed to arise from the underlying latent (Gaussian) process Z, which is unobservable, although may be measured with error. For example consider,

$$Y|Z, \theta \sim N(AZ, \sigma_\epsilon^2 I)$$

- The underlying process level $Z|\theta$ - The latent process Z assumed to drive the observable data and represents the true value of the quantity of interest. For example consider,

$$Z|\theta \sim N(\boldsymbol{\mu}, \sigma_z^2 \Sigma)$$

Bayesian Hierarchical Models

- The prior level θ - This level describes known prior information about the model parameters θ and controls the scale and variability of the data and the latent process. For example consider,

$$\theta = (\sigma_\epsilon, \sigma_z)^T \sim p(\theta)$$

- Inference on all model parameters in a hierarchical model such as these can be done as follows.

INFERENCE

- We can write the posterior of the model parameters in a similar way as before

$$p(\theta, Z|Y) \propto p(Y|Z,\theta)p(Z,\theta) = p(Y|Z,\theta)p(Z|\theta)p(\theta)$$

- We are interested in the marginal effects of all the latent process parameters and the hyperparameters

$$p(\theta_i|Y) = \int\int p(\theta, Z|Y)\, dZ\, d\theta_{-i}, \quad p(z_i|Y) = \int\int p(\theta, Z|Y)\, dZ_{-i}\, d\theta$$

- Typically $dim(Z) = 10^2 - 10^6$ and $dim(\theta) \leq 10$ so these are, high-dimensional integrals, so we simplify

INFERENCE

- Using the fact that $p(\theta, Z|Y) = p(Z|\theta, Y)p(\theta|Y)$ then we see that

$$p(\theta_i|Y) = \iint p(\theta, Z|Y)\, dZ\, d\theta_{-i} = \int p(\theta|Y)\, d\theta_{-i}$$

$$p(Z_i|Y) = \iint p(\theta, Z|Y)\, dZ_{-i}\, d\theta = \int p(Z_i|\theta, Y)p(\theta|Y)d\theta$$

- So instead of having to find $p(Z, \theta|Y)$ and do very high dimensional integrals we just need distributions $p(\theta|Y)$ and $p(Z_i|\theta, Y)$ (lower dimensional numerical integration).

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) methods are based on sampling, and are extensively used in Bayesian inference. We aim to sample from the posterior

$$p(\theta|Y) \propto p(Y|\theta) \times p(\theta)$$

  to estimate such as mean and variance.
- Some advantages and disadvantages:
  - Very flexible with well-known algorithms
  - Software available (JAGS, WinBUGS, etc.)
  - May be computationally infeasible in large-scale problems
- Techniques for performing Bayesian inference can be extremely useful.

Integrated Nested Laplace Approximations

- Integrated Nested Laplace Approximations (INLA) is a recent development in approximate Bayesian Inference.
- It was introduced as an alternative to methods such as MCMC for a general set of statistical models called latent Gaussian models.
- Posterior distributions are approximated using a series of Laplace approximations meaning we do not need to sample from the posterior.
- It has been shown to be accurate in all but extreme cases and can substantially reduce computational burden compared to MCMC in many cases.
- Software suite called `R-INLA` suite allows implementation in `R`.

Laplace Approximation

- Laplace approximation is an technique that can be used to perform inference by approximating a posterior distribution by a Gaussian distribution.
- To make a Laplace approximation we take a Taylor expansion of the density $\log p(\theta|Y)$ around its mode $\hat{\theta}$, i.e.

$$\log p(\theta|Y) = \log p(\hat{\theta}|Y) + (\theta - \hat{\theta})^T \frac{\partial}{\partial \theta} \log p(\theta|Y)\Big|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial}{\partial \theta \partial \theta^T} \log p(\theta|Y)\Big|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + h.o.t.$$

- As the mode is the maximum of $p(\theta|Y)$ then the second term will be zero, i.e.

$$\frac{\partial}{\partial \theta} \log p(\theta|Y)\Big|_{\theta=\hat{\theta}} = 0$$

Laplace approximation

- Letting

$$H(\hat{\theta}|Y) = -\left.\frac{\partial}{\partial\theta\partial\theta^T}\log p(\theta|Y)\right|_{\theta=\hat{\theta}}$$

and omit higher order terms then we see that

$$p(\theta|Y) \propto \exp\left(-\frac{1}{2}(\theta-\hat{\theta})^T H(\hat{\theta}|Y)(\theta-\hat{\theta})\right)$$

which is the kernel of a Gaussian distribution and therefore,

$$\theta|Y \sim N(\hat{\theta}, H(\hat{\theta}|Y)^{-1})$$

- In general, the mode will have to found numerically (often performed using Newton optimisation)

Gaussian Random Fields

- A random vector $Z = (z_1, \ldots, z_n)^T$ is called a Gaussian Markov Field if it has mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ, with density

$$p(Z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(Z-\boldsymbol{\mu})^T\Sigma^{-1}(Z-\boldsymbol{\mu})\right)$$

- There are some issues with working with this parameterisation in practice, especially when n is large.
  - Covariance matrix has $\mathcal{O}(n^2)$ elements.
  - Computation often rises $\mathcal{O}(n^3)$ (determinants, inverse, etc.)

Gaussian Random Fields

- In some cases it may be advantegeous to parameterise the distribution in terms of the precision matrix $Q = \Sigma^{-1}$ (inverse of the covariance),

$$p(Z) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}(Z-\boldsymbol{\mu})^T Q(Z-\boldsymbol{\mu})\right)$$

- This often reduces computation.
- The non-zero pattern in the precision matrix tells us a lot about the conditional distributional structure.
- The Markov property states that if $Q_{ij} = 0$ if and only if z_i and z_j are conditionally independent given all other elements Z_{-ij}

Gaussian Markov Random Field

- A random vector $Z = (z_1, \ldots, z_n)^T$ is called a Gaussian Markov Random Field (GMRF) with mean $\boldsymbol{\mu}$ and positive definite precision matrix Q, if

$$p(Z) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}(Z-\boldsymbol{\mu})^T Q(Z-\boldsymbol{\mu})\right)$$

$$Q_{ij} = 0 \iff (i,j) \notin \mathcal{E}$$

- To reduce computation, we assume that latent variables follow a GMRF.
- The conditional independence allows for computing with *sparse* matrices.

Example: AR(1) Process

- Suppose we have a Autoregressive process of order 1,

$$Z_0 \sim N(0, (1-\alpha)^{-1}\sigma^2)$$
$$\{Z_t | Z_s, s < t\} \sim N(\alpha Z_{t-1}, \sigma^2); t = 1, \ldots, T$$

  where $\alpha \in (0, 1)$ and $\sigma > 0$.

- The covariance and precision matrices of the joint distribution Z has the form

$$\Sigma_{ij} = \frac{\alpha^{|i-j|}}{1-\alpha}\sigma^2, \quad Q = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -\alpha & & & \\ -\alpha & 1+\alpha^2 & -\alpha & & \\ & \ddots & \ddots & \ddots & \\ & & -\alpha & 1+\alpha^2 & -\alpha \\ & & & -\alpha & 1 \end{bmatrix}$$

  so Σ is dense and Q is sparse.

Integrated Nested Laplace Approximation

- As mentioned earlier, we have to find

$$p(\theta|Y) \text{ and } p(Z_i|\theta, Y)$$

to be able to find the marginal posterior distributions

$$p(\theta_i|Y) = \int p(\theta|Y)\, d\theta_{-i}, \quad p(Z_i|Y) = \int p(Z_i|\theta, Y)p(\theta|Y)d\theta$$

- This is done in three steps.
  1. Find an approximation to the distribution $p(\theta|Y)$
  2. Find an approximation to the marginal distributions $p(Z_i|\theta, Y)$
  3. Numerically integrate to get marginal distributions

R-INLA

- The R–INLA package provides a practical implementation of Integrated Nested Laplace Approximations (INLA).
- It can be used with hierarchical GMRF models
- The class of models that can be expressed in this form and thus can be used with R–INLA is very large and includes, amongst others, the following:
  - Dynamic linear models.
  - Stochastic volatility models.
  - Generalised linear (mixed) models.
  - Generalised additive (mixed) models.
  - Spline smoothing.
  - Semi–parametric regression.
  - Disease mapping.
  - Log–Gaussian Cox–processes.
  - Model–based geostatistics.
  - Spatio–temporal models.
  - Survival analysis.

THE SYNTAX OF R–INLA

- There are three main parts to fitting a model using R–INLA:
  1. The data.
  2. Defining the model formula.
  3. The call to the INLA program.
- The basic syntax of running models in R–INLA is very similar in appearance to that of `glm` in R and takes the general form `formula, data, family` but with the addition of the specification of the nature of the random effects, `f()`.
- For the latter component, common examples include `f(i, model="iid")` (independent), `f(i, model="rw")` (random walk of order one) and `f(i, model="ar")` (autoregressive of order p).

Fitting a Poisson regression model in R–INLA

- An extension of the standard Poisson model to include log–normal random effects in the linear predictor
- $$\log \mu_l = \beta_0 + \beta_{0i} + \beta_1 X_l + \beta_d X_l + \epsilon_l \quad (1)$$

  where β_l represents the effect of exposure, β_d is the effect of an area-level covariate and β_{0i} denotes the random effect for area i.
- The syntax of the R–INLA code to fit this model is very similar to that of a standard `glm` in R.

Fitting a Poisson regression model in R–INLA

```
> formula = Y ~ X1+X2 + f(i, model="iid")
> model   = inla(formula, family="poisson", data=data)

Call:
"inla(formula = formula, family = "poisson", data = data)"

Time used:
 Pre-processing    Running inla Post-processing         Total
       0.278389        0.286911        0.125699      0.690999

Integration Strategy: Central Composite Design

Model contains 1 hyperparameters
The model contains 3 fixed effect (including a possible
    intercept)

Likelihood model: poisson

The model has 1 random effects:
1.'i' is a IID model
```

Fitting a Poisson regression model in R–INLA

```
> summary(model)

Call:
"inla(formula = formula, family = "poisson", data = data)"

Time used:
 Pre-processing     Running inla Post-processing         Total
         0.2784           0.2869           0.1257        0.6910

Fixed effects:
               mean    sd 0.025quant 0.5quant 0.975quant
```

Fitting a Poisson regression model in R–INLA

```
(Intercept) 2.4960 0.0713    2.3553    2.4962      2.6355
X1          0.1187 0.0310    0.0578    0.1186      0.1796
X2          0.0578 0.0074    0.0433    0.0578      0.0722

Random effects:
Name       Model
 i    IID model

Model hyperparameters:
                  mean      sd 0.025quant 0.5quant 0.975quant
Precision for i 3.784 0.3548       3.131    3.769       4.525

Expected number of effective parameters(std dev):
    321.42(3.926)
Number of equivalent replicates : 1.223

Marginal Likelihood:  -1513.92
```

FITTING MODELS IN R–INLA

- Future details on R–INLA, including the latent process models that can be
  accommodated, can be found on the R–INLA webpage:
  `http://www.R-INLA.org`.

Session 3: Applications of Spatial Modelling

Spatial Data

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

(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.

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

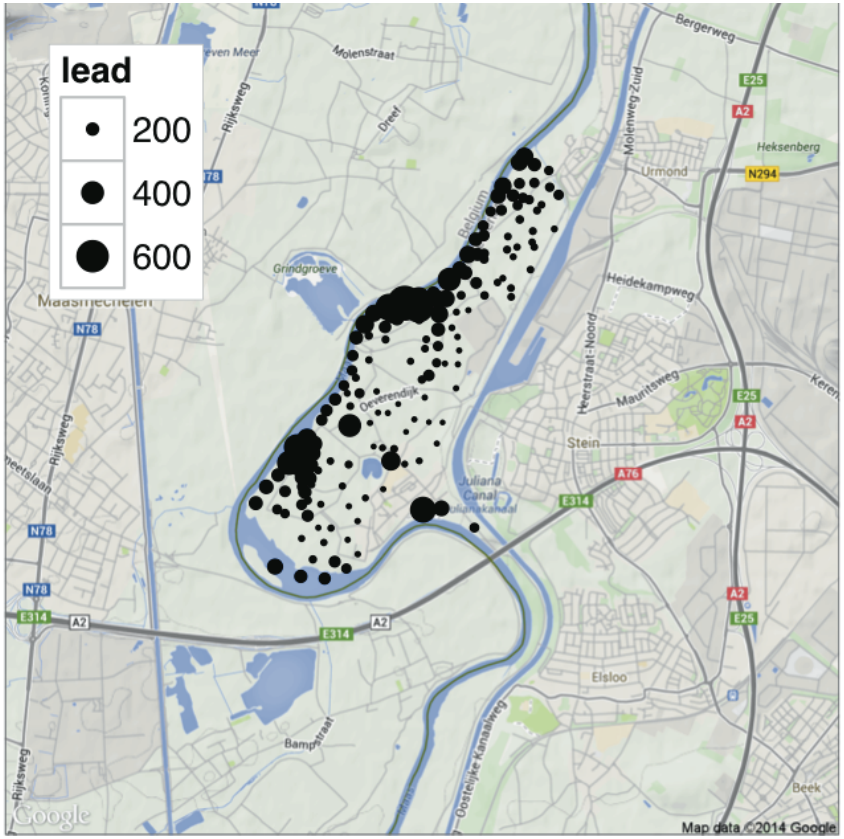
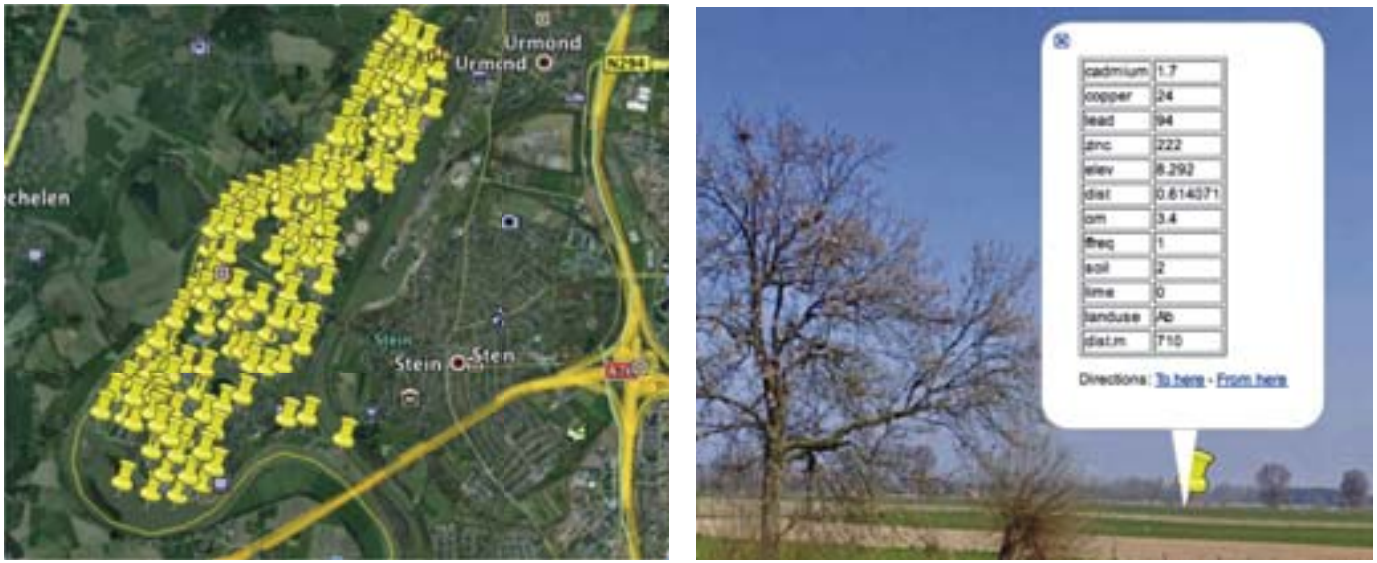


Figure: Bubble plot showing the size of lead concentrations measured in samples taken at 155 locations in the Meuse River flood plain.

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: 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.

VARIOGRAMS

- The covariance function and the semi–variogram are both functions that summarise the strength of association as a function of distance and, in the case of anisotropy, direction.
- When dealing with a purely spatial process where there are no independent realisations, patterns in correlation and variances from different parts of the overall region of study are used as if they were replications of the underlying process.
- Under the assumption of stationarity, a common covariance function for all parts of the regions can then be estimated.

VARIOGRAMS

- The semi–variogram will be zero at a distance of zero as the value at a single spot is constant and has no variance.
- It may then rise and reach a plateau, indicating that past a certain distance, the correlation between two units is zero.
- This plateau will occur when the semi–variogram reaches the variance.

Variogram Models

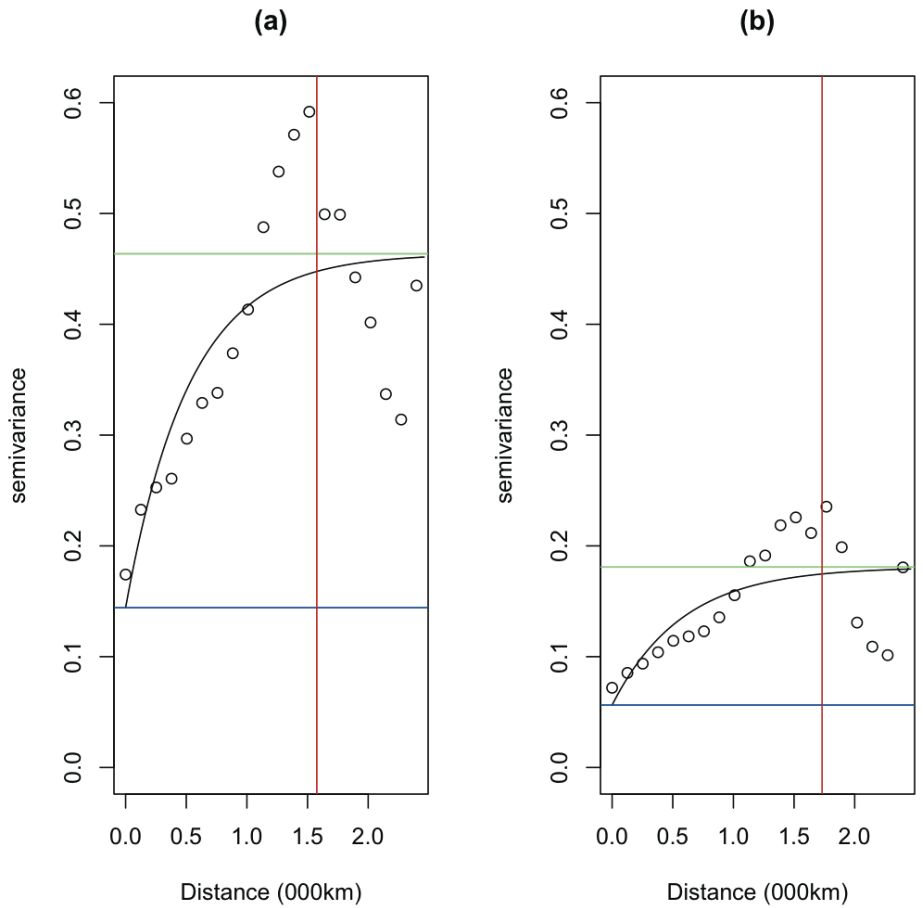


Figure: Variograms for (a) log values of Nitrogen Dioxide (NO_2) measured at monitoring sites throughout Europe in 2001; (b) residuals after fitting a model with land–use covariates.

Matern Class

- A common class of models used for variogram models is the Matern class: *Matern class of models*

$$\gamma(h|\theta) = \begin{cases} 0 & h = 0 \\ \theta_1 \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right)^{\theta_2} K_{\theta_2}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right) & h > 0 \end{cases} \quad (2)$$

- Where $\theta_1 > 0$ is a scalable parameter controlling the range of the spatial correlation, and $\theta_2 > 0$ is the smoothness parameter.
- K_{θ_2} is a the modified Bessel function of order θ_2.

Exponential Model

- The exponential model is a special case of this, with $\theta_2 = 1/2$.

$$\gamma(d|\theta) = \begin{cases} 0 & d = 0 \\ \theta_1 \exp(-\theta_2 d) & d > 0 \end{cases}$$

The limiting case of the Matern class of models, when $\theta_2 \to \infty$, is the Gaussian model.

INLA AND MODELLING IN A CONTINOUS DOMAIN

- The methods presented here are for use with point referenced data, particularly cases where there is a Gaussian field (GRF) with responses measured with error.
- A GRF doesn't have a natural Markov structure and so INLA, as originally developed, does not apply directly as it does with GMRFs as would be the case when using areal data.
- It is possible to use a bridge between a GF and a GMRF, to which INLA can be applied.

INLA AND MODELLING IN A CONTINOUS DOMAIN

- We now describe the SPDE–GRMF approximation following Lindgren (2011).
- INLA assumes the GF $Z_s, s \in \mathcal{S}$ has a Matern spatial covariance that is the solution of the SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} Z_s = v_s, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \tag{3}$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo–difference operator, Δ is the Laplacian and v is spatial white noise with unit variance.

- The marginal variance of the process is given by

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}} \tag{4}$$

Representing the process in this way is key to the developments that follow; it provides the bridge over which we can cross from the GF to the GMRF via an approximate solution to the SPDE.

INLA AND MODELLING IN A CONTINOUS DOMAIN

- An infinite dimensional solution, Z_s, of the SPDE over its domain, $\mathcal{S}$, is characterised by the requirement that for all members of an appropriate class of test functions, ϕ,

$$\int \phi_{js}(\kappa^2 - \Delta)^{\alpha/2} Z_s dZ = \int \phi_{js} v_s ds \tag{5}$$

- However in practice, only approximate solutions are available and Lindgren (2011) use the conventional finite element approach, which uses a Delauney triangulation (DT) over $\mathcal{S}$.
- Initially the triangles are formed with vertices at the points of the sparse network where observations are available with additional triangles added until $\mathcal{S}$ is covered, leading to an irregular array of locations (vertices).

INLA AND MODELLING IN A CONTINOUS DOMAIN

- The result is a GF model for the process but with an associated GMRF that can be used (by INLA) for performing the computations that would be computationally prohibitive using the GF directly.
- The resulting algorithm is implemented in R-INLA (www.r-inla.org).
- To best illustrate the SPDE approach in practice, we consider an example.

EXAMPLE: BLACK SMOKE IN THE UK

- The Great Smog was a severe air pollution event in London during December 1952
- A large amount of snow in London led to more coal being burnt.
- An area of high pressure and light winds trapped the resulting smog in London.
- This event led to over 4000 excess deaths in the following weeks

EXAMPLE: BLACK SMOKE IN THE UK

EXAMPLE: BLACK SMOKE IN THE UK

- Management of air pollution began in the middle of the 20th century when serious concern arose about the possible effects of air pollution on health.
- The Great Smog and other events led to the Clean Air Act in 1956 which established a large scale monitoring network and the National Survey.

EXAMPLE: BLACK SMOKE IN THE UK

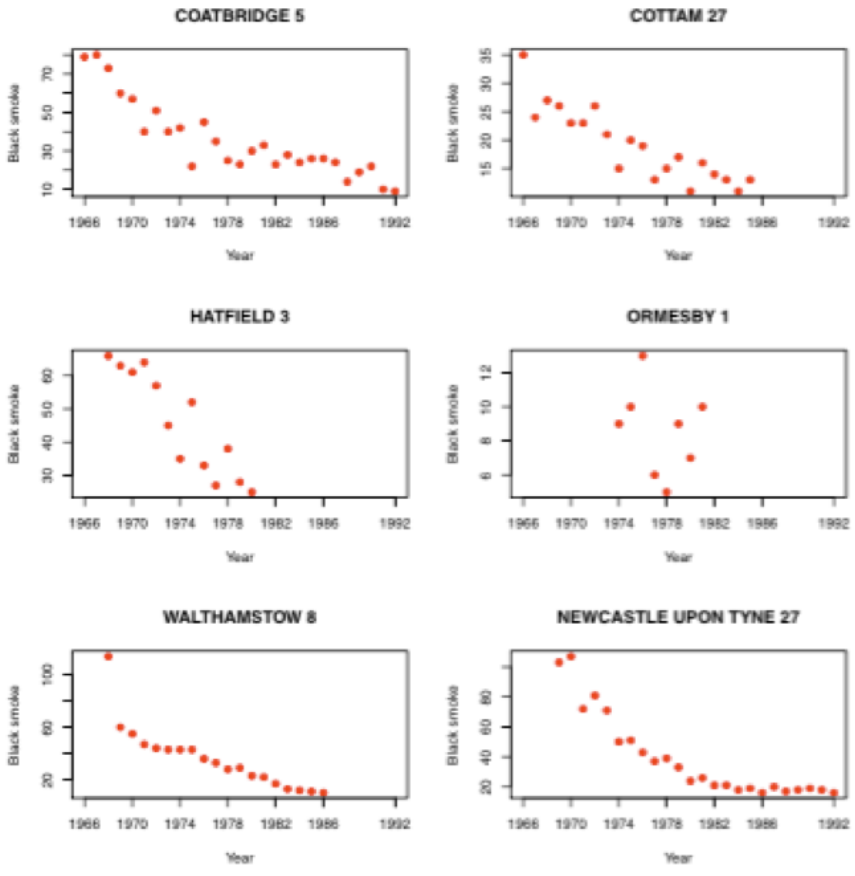
- The National Survey measured black smoke and sulphur dioxide.
  - In mid-1960s 1000+ sites
  - In mid-1990s 200 sites
- This was done to examine the changes over time and variations in space.
- Also interested in effects of reduction in network over time.

EXAMPLE: BLACK SMOKE IN THE UK

- Black smoke consists of fine particulate matter.
- Mainly emitted from fuel combustion.
- Following the large reductions in domestic coal use the main source is diesel-engined vehicles.
- We can measure black smoke by its blackening effects on filters.

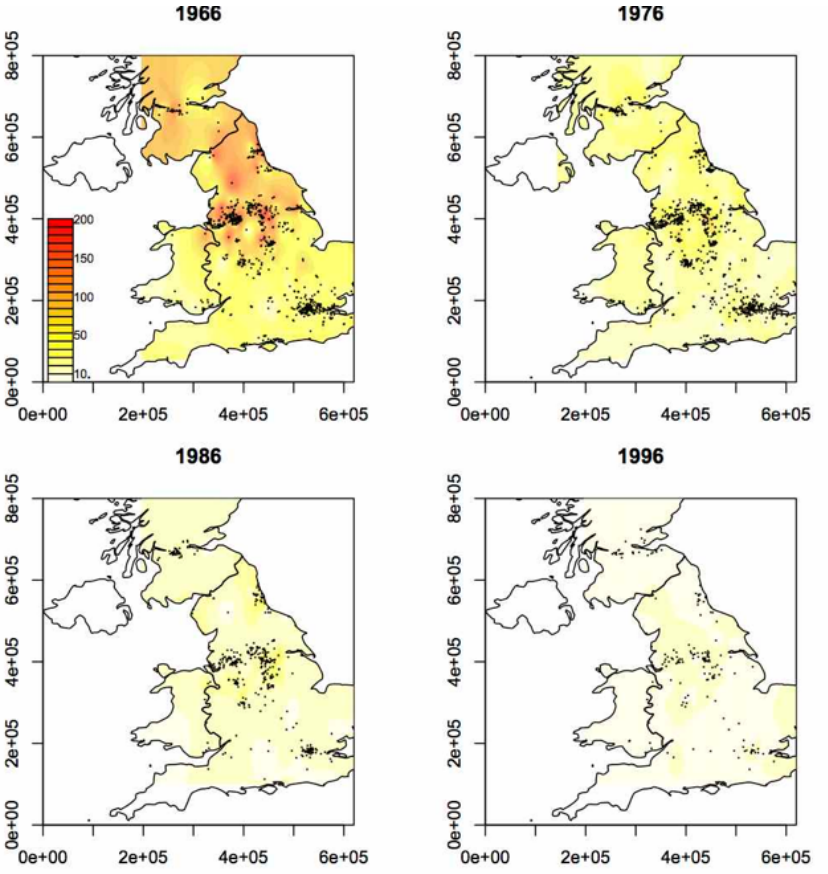
EXAMPLE: BLACK SMOKE IN THE UK

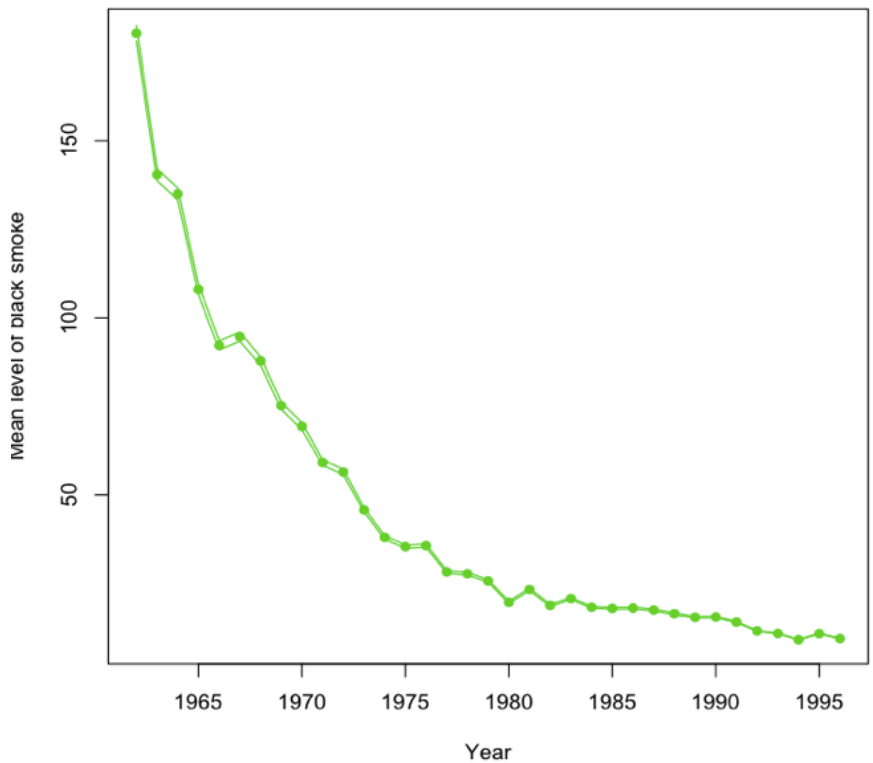
Decrease in concentrations over time by site



EXAMPLE: BLACK SMOKE IN THE UK

Overall decrease in concentrations over time.





EXAMPLE: BLACK SMOKE IN THE UK

- We model the pollution field using a Bayesian hierarchical model.
- The annual average (log) for each site is modelled as a function of space and time

$$Y_{it} = (\beta_0 + \beta_{0i}) + (\beta_x + \beta_{x_i})t + (\beta_{x2} + \beta_{x2_i})t^2 + \beta_u U_i + \epsilon_{it}$$

where i is a the location of the ground monitor, t is time and U_i is a urban/rural indicator.

- We use a linear and quadratic effect of time.
- Site random effects are assumed to be multivariate normal

$$\boldsymbol{\beta}_s \sim N(\mathbf{0}, \Sigma)$$
$$\boldsymbol{\beta}_s \sim N(\mathbf{0}, Q^{-1})$$

Example: Black Smoke in the UK

- The next figure shows the mesh that was constructed using Delauney triangulation for the locations of black smoke monitors in the UK.

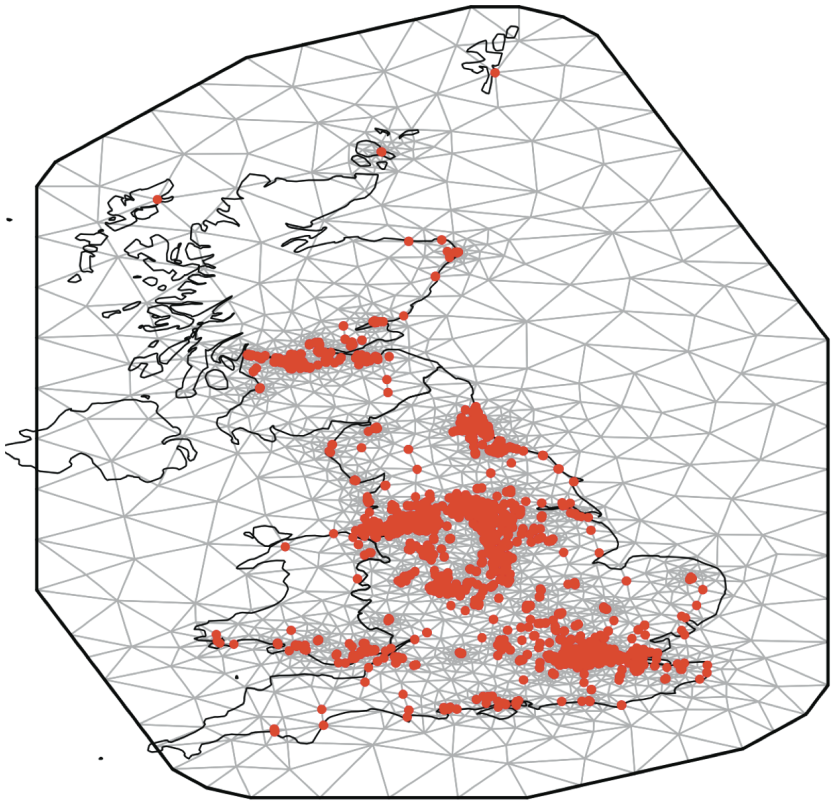


Figure: Triangulation for the black smoke network in the UK. The red dots show the locations of black smoke monitoring sites.

EXAMPLE: BLACK SMOKE IN THE UK

```
    mesh = inla.mesh.create(locations[,1:2],
extend=list( offset=-0.1), cutoff=1,
# Refined triangulation,
# minimal angles >=26 degrees,
# interior maximal edge lengths 0.08,
# exterior maximal edge lengths 0.2:
refine=(list(min.angle=26,
max.edge.data = 100,
max.edge.extra=200))
)
```

EXAMPLE: BLACK SMOKE IN THE UK

```
ukmap <- readShapeLines("uk_BNG.shp")
plot(mesh, col="gray", main="")
lines(ukmap)
points(locations, col="red", pch=20,bg="red")
```

EXAMPLE: BLACK SMOKE IN THE UK

- Here we use the shapefile for the UK using the British National Grid projection that provides the outline of the UK coastline and are included in the online resources.
- In this case, the distance between the points is expressed in metres and so the distances used in creating the mesh to ensure the plots will overlay should also be in metres.
- In this case, there are 3799 edges and the mesh was constructed using triangles that have minimum angles of 26 and a maximum edge length of 100km.
- There are 1466 monitoring locations being considered over the period of study and these are highlighted in red.

EXAMPLE: BLACK SMOKE IN THE UK

- This lattice underlies the GMRF and gives a finite element representation of the solution of the SPDE

$$Z_s = \sum_{k=1}^{n} \psi_{ks} w_k \tag{6}$$

where n is the number of vertices of the DT, $\{w_k\}$ are Gaussian weights and ψ_{ks} are piecewise linear in each triangle (1 at vertex k and 0 at all other vertices).

EXAMPLE: BLACK SMOKE IN THE UK

- Using the mesh set up above we now create the INLA SPDE object that will be used as the model.
- The model is then fit by defining a formula and then running the `inla` command using the SPDE object in a random effects term.

EXAMPLE: BLACK SMOKE IN THE UK

```
# Field std.dev. for theta=0
 sigma0 = 1
 # find the range of the location data
 size = min(c(diff(range(mesh$loc[,1])),
              diff(range(mesh$loc[,2]))))
# A fifth of the approximate domain width.
 range0 = size/5
 kappa0 = sqrt(8)/range0
 tau0 = 1/(sqrt(4*pi)*kappa0*sigma0)
 spde = inla.spde2.matern(mesh,
   B.tau=cbind(log(tau0), -1, +1),
   B.kappa=cbind(log(kappa0), 0, -1),
   theta.prior.mean=c(0,0),
constr=TRUE)
```

EXAMPLE: BLACK SMOKE IN THE UK

```
formula = logbs ~ 1+ urban.rural +f(site, model=spde)

model = inla(formula, family="gaussian", data = BSdata,
control.predictor = list(compute=TRUE),
control.compute = list(dic = TRUE, config=TRUE))
```

EXAMPLE: BLACK SMOKE IN THE UK

```
model$fixed.effects

               mean     sd 0.025quant 0.5quant 0.975quant
(Intercept)   0.042 0.707      -1.346    0.042       1.430
ur           -0.066 0.013      -0.092   -0.066      -0.040

  mode     kld
  0.042    0
-0.066     0
```

EXAMPLE: BLACK SMOKE IN THE UK

- We are interested in the posterior marginals of the latent field,

$$\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)d\theta \tag{7}$$

$$\pi(\theta_j|y) = \int \pi(\theta|y)\pi(\theta_{-j})d\theta \tag{8}$$

  where $i = 1, ..., N_S + P$ where N_S is the number of monitoring locations and P the number of predictions to be made and $j = 1, \ldots J$ is the number of parameters in the model.
- With regards to spatial prediction, the SPDE–INLA algorithm provides the posterior conditional distribution of the random effects terms at all the vertices of the triangulation.

EXAMPLE: BLACK SMOKE IN THE UK

- Given these, there is a mapping to the response variable which allows samples of predictions to be obtained.
- In order to produce a map displaying the spatial predictions of the model we first need to define a lattice projection starting from the mesh object back to the grid on which the data lies and will be plotted.
- Then the posterior mean and standard deviations can be extracted and then projected from the latent field space to the grid, and then plotted as a map.

Example: Black Smoke in the UK

- The example of the result can be seen below.

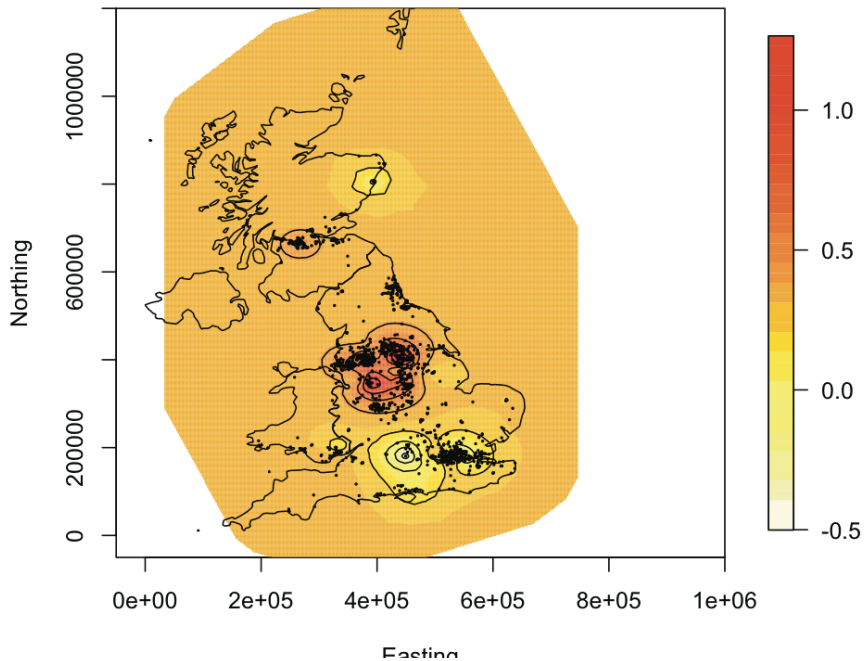


Figure: Map of predicted values of black smoke in the UK. Values are medians of posterior predicted distributions on the logarithmic scale from an SPDE–INLA model.

- There are a number of ways of converting the output from SPDE/R-INLA to a form that can produce maps and the package `geostatsp` provides many routines for doing this.

THE END

THANK YOU