



# Bayesian Hierarchical Models

Gavin Shaddick,  
Millie Green, Matthew Thomas  
University of Bath

6<sup>th</sup> - 9<sup>th</sup> December 2016

# COURSE OVERVIEW

- ▶ *Day 1* - An introduction to Bayesian Hierarchical Models
- ▶ *Day 2* - Implementing Bayesian models using R-INLA (Practical)
- ▶ *Day 3* - Applications of Bayesian Hierarchical Models
- ▶ *Day 4* - Bayesian disease mapping (Practical)

# TEXTBOOK

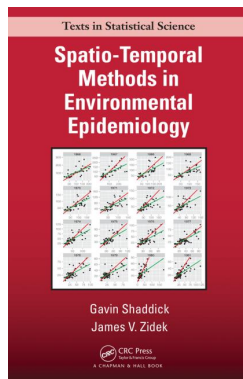
**Title:** Spatio-Temporal Methods in Environmental Epidemiology

**Authors:** Gavin Shaddick and Jim Zidek

**Publisher:** CRC Press

**Resource Website:**

<http://www.stat.ubc.ca/~gavin/STEPIDBookNewStyle/>



# WEBSITE

`http://stat.ubc.ca/~gavin/STEP1BookNewStyle/course_clapem.html`

The screenshot shows a website with a dark red background. At the top, there is a navigation bar with five teal buttons: HOME, RESOURCES BY CHAPTER, COURSES, COMPUTING RESOURCES, and BOOK'S WEBPAGE @ CRC. Below the navigation bar, the main heading reads "BAYESIAN HIERARCHICAL MODELS". Underneath this, a purple bar contains the text "COURSE OUTLINE". The main content area contains two paragraphs of text. The first paragraph describes the course's aim to provide an interactive experience for students and researchers, covering modelling relationships in space and time, with a focus on fitting complex models to big data. The second paragraph states that the course will be delivered at the Latin American Congress of Probability and Mathematical Statistics (CLAPEM) in Universidad de Costa Rica between 6th-9th December 2016, presented by Gavin Shaddick, Amelia Green, and Matthew Thomas.

# CONTACT INFORMATION

Dr. Gavin Shaddick, University of Bath

- ▶ Email: [gavinshaddick@bath.edu](mailto:gavinshaddick@bath.edu)
- ▶ Webpage: <http://people.bath.ac.uk/masgs/>

# AN INTRODUCTION TO BAYESIAN HIERARCHICAL MODELS

# OUTLINE

Spatio-temporal modelling

Bayesian hierarchical models

Dealing with 'big' data

Bayesian inference

# Spatio-temporal modelling



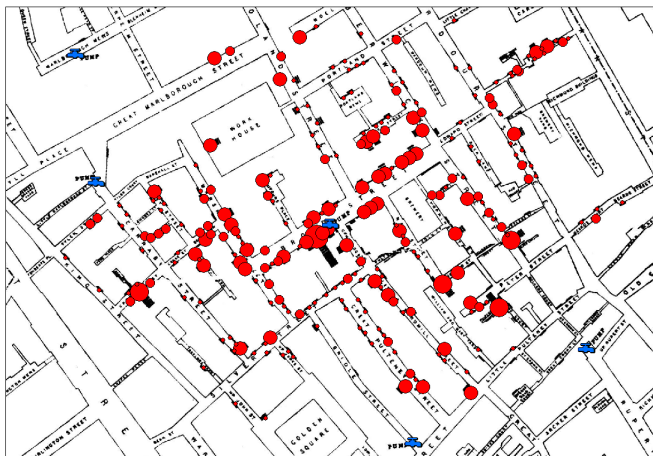
# THE NEED FOR SPATIO-TEMPORAL MODELLING

- ▶ 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

- ▶ 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.

# 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.

# 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.

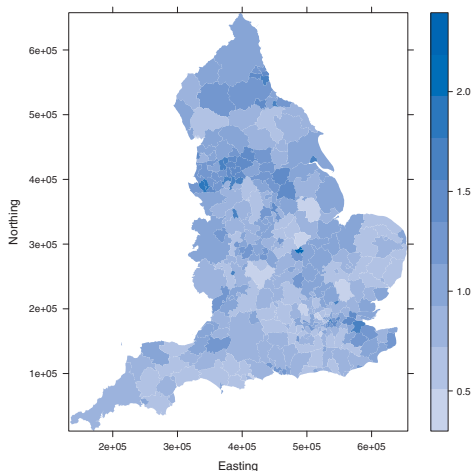
# DEPENDENCIES OVER SPACE AND TIME

- ▶ 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.

## 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.

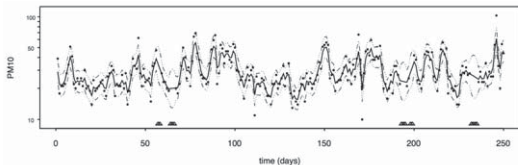




**Figure:** Map of the spatial distribution of risks of hospital admission for a respiratory condition, chronic obstructive pulmonary disease (COPD), in the UK for 2001. The shades of blue correspond to standardised admission rates, which are a measure of risk. Darker shades indicate higher rates of 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<sub>10</sub>) 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.

# 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.

# Bayesian hierarchical models

# BAYESIAN HIERARCHICAL MODELS

Bayesian hierarchical models are an extremely useful and flexible framework in which to model complex relationships and dependencies in data.

In the hierarchy we consider, there are three levels;

- (1) The observation, or measurement, level
- (2) The underlying process level
- (3) The parameter level.

# THE OBSERVATION LEVEL

- ▶ 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  $\theta_1$  is the set of parameters for this model and may include, for example, regression coefficients and error variances.

# THE UNDERLYING PROCESS LEVEL

- ▶ 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  $\theta_2$  is the set of parameters for this level of the model.

# THE PARAMETER LEVEL

- ▶ 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  $T = \{t_0, t_1, \dots, t_{N_T}\}$ .
- ▶ 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,  $S = \{s_0, s_1, \dots, s_{N_S}\}$ .

# A HIERARCHICAL APPROACH TO MODELLING SPATIO-TEMPORAL DATA

- ▶ There are three levels to the hierarchy that we consider.
- ▶ The observed data,  $Y_{st}, s = 1, \dots, N_S, t = 1, \dots, N_T$ , 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,  $\mu_{st}$  and (ii) a random process,  $\omega_{st}$ .

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

# A HIERARCHICAL APPROACH TO MODELLING SPATIO-TEMPORAL DATA

- ▶ The trend, or mean term,  $\mu_{st}$  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,  $\omega_{st}$ , 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

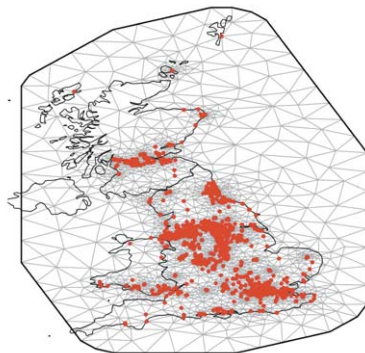
# DEALING WITH 'BIG' DATA

- ▶ 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 required, it may be computationally impractical to perform Bayesian analysis using MCMC or packages such as WinBUGS in any straightforward fashion.
- ▶ This can be due to both the requirement to manipulate large matrices within each simulation of the MCMC.
- ▶ During this course, we will show examples of recently developed techniques that perform 'approximate' Bayesian inference.
- ▶ These are based on integrated nested Laplace approximations (INLA).
- ▶ 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.

# 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

- ▶ Likelihood function
  - ▶ A model for  $Y$  in terms of parameters  $\theta$ ,  $p(Y|\theta)$
- ▶ Prior distribution
  - ▶ A-priori knowledge about  $\theta$ ,  $p(\theta)$
- ▶ Combining these, we can obtain the posterior distribution

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

- ▶ The denominator  $p(Y)$  a normalisation constant.

# BAYESIAN INFERENCE

- ▶  $p(Y)$  is the marginal distribution

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

- ▶ This integral is often analytically intractable
- ▶ Use proportionality with respect to  $\theta$

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}$$

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

# MARKOV CHAIN MONTE CARLO

- ▶ No need to explicitly specify posterior
- ▶ Specify the prior and likelihood separately
- ▶ Create Markov chain to allow samples to be drawn from posterior distributions
  - ▶ Generate candidate sample from proposal distribution, based on previous value in chain.
  - ▶ Either accept or reject based on acceptance function
- ▶ Gibbs sampling: proposal distributions are the full conditionals, therefore acceptance probabilities are one.
- ▶ Metropolis-Hastings: very flexible
- ▶ Software available, e.g. WinBUGS, JAGS, STAN
- ▶ May be computationally infeasible in large-scale problems

# INTEGRATED NESTED LAPLACE APPROXIMATIONS

- ▶ For latent Gaussian models.
- ▶ Does not rely on sampling.
- ▶ Posterior distributions are approximated using a series of Laplace approximations.
- ▶ Uses numerical integration.
- ▶ It has been shown to be accurate in all but extreme cases.
- ▶ It can substantially reduce computational burden compared to MCMC.
- ▶ Implementation in R using R-INLA.