Session 2: Modelling spatial data

Duncan Lee

Overview

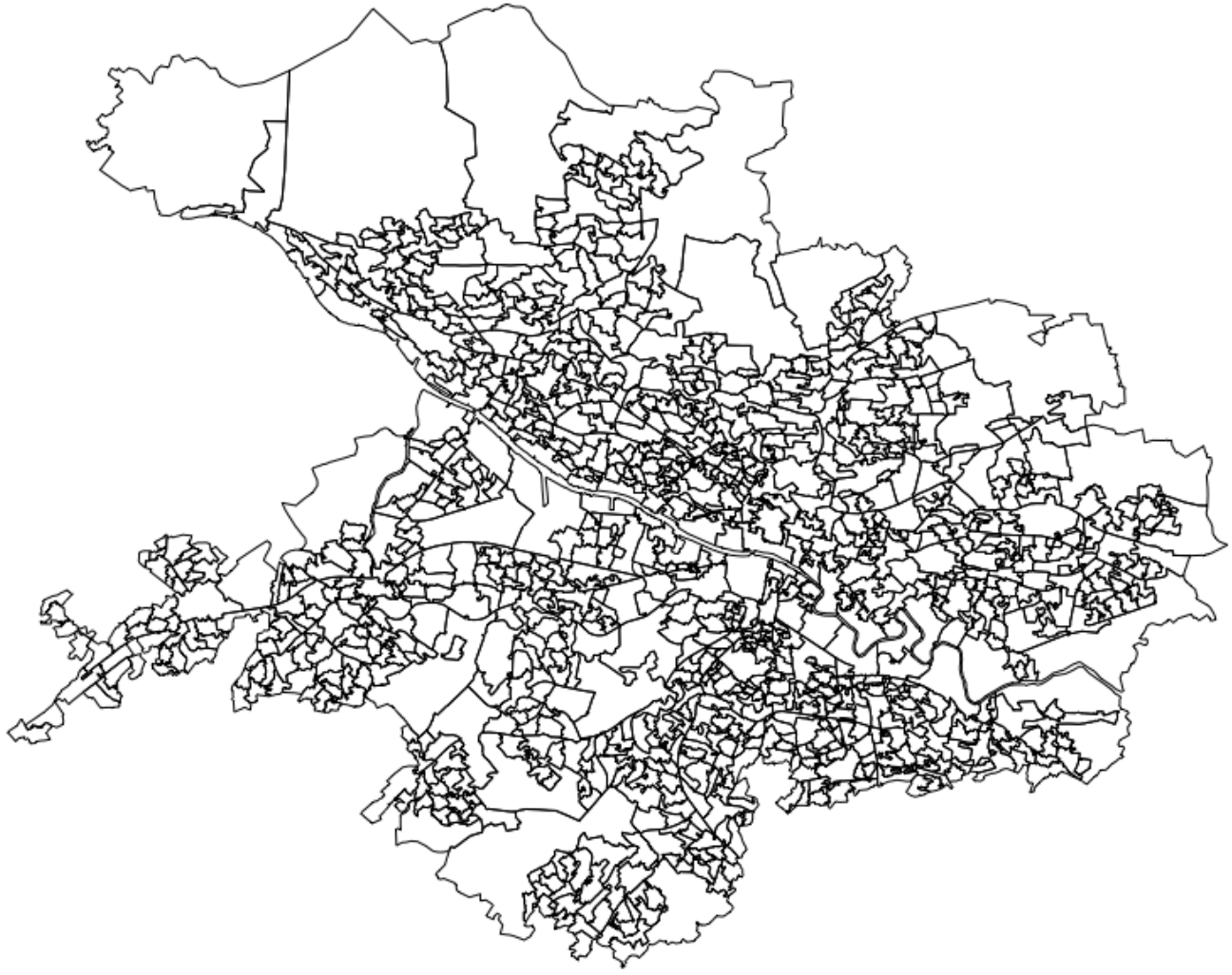
This session will cover the following:

- Issues when modelling spatial data.
- Modelling spatial data.
- Bayesian estimation and inference.

Spatial data analysis

- Recall that the data relate to K non-overlapping areal units, such as datazones.
- Each data point is a summary measure relating to the population living in that area, such as a total or average.
- Spatial data typically display spatial correlation, with data points close together in space tending to have more similar values than data points further apart in space.
- This spatial correlation violates the assumption of independence in simple statistical models, requiring the need for more complex models.

Example - datazones in Glasgow

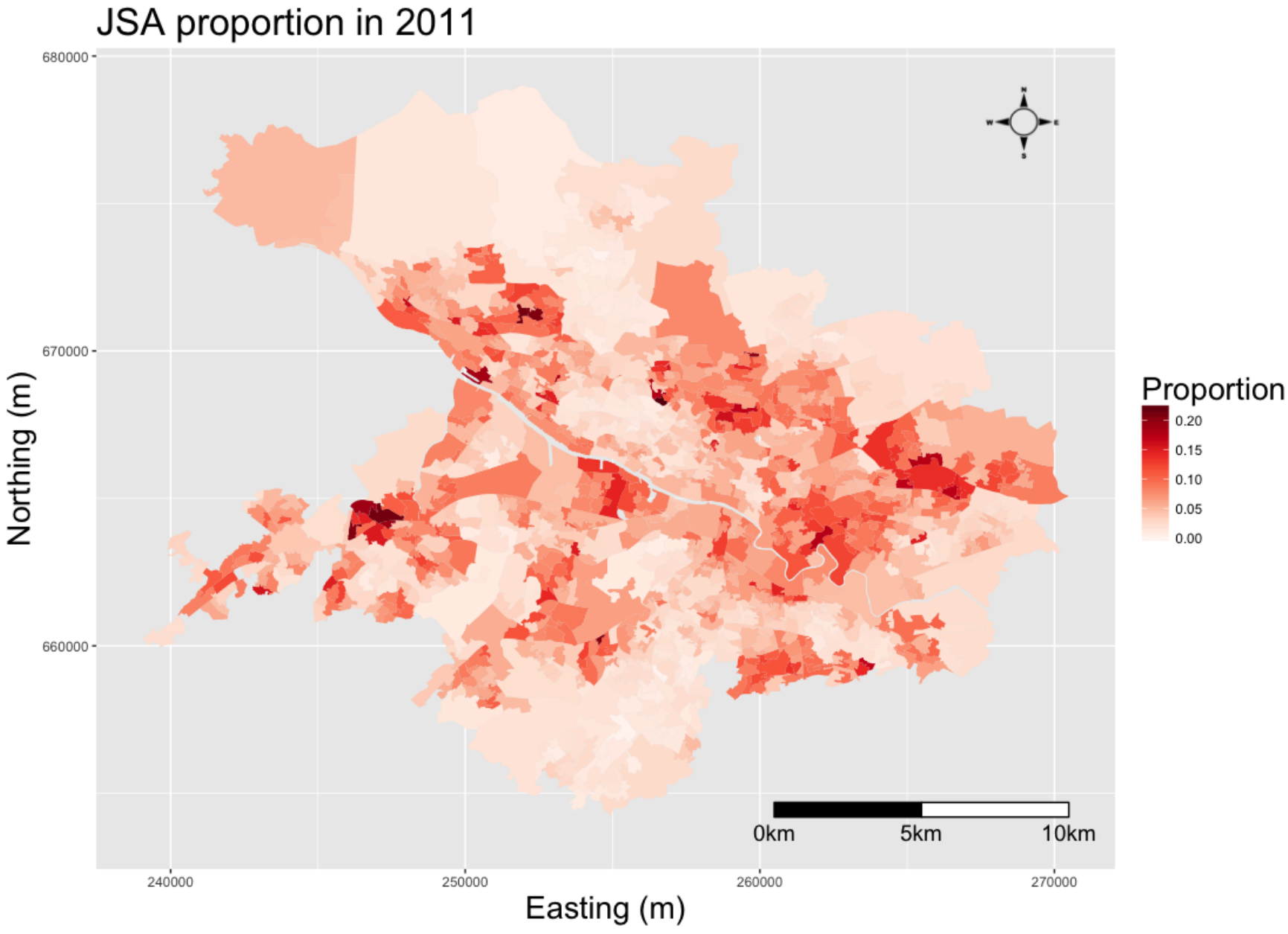


Causes of spatial correlation

There are many possible causes of spatial correlation, including:

- **Unmeasured confounding** - where an unmeasured or unknown spatially correlated risk factor induces spatial correlation in the data being modelled.
- **Grouping effects** - where subjects choose to be close to similar subjects.
- **Neighbourhood effects** - where subjects' behaviour is influenced by that of neighbouring subjects.

Example - JSA in 2011



General modelling idea

When modelling data one is attempting to represent the data as:

$$\textbf{Data} = \textbf{Fit} + \textbf{Error}$$

where

- **Fit** is the variation in the data explained by the model, such as by covariates, etc.
- **Error** is the remaining variation in the data unexplained by the model.

Simple models typically assume the latter is independent, but in most cases with spatial data it will be spatially correlated.

What happens if I ignore spatial correlation

If one is fitting a regression model then ignoring spatial correlation will not bias the regression parameter estimates. However, it will result in 95% uncertainty intervals that are too narrow, which may lead to non-significant effects appearing to be significant.

Modelling stages

Therefore the approach to modelling depends on whether one has any covariates available to include in the model.

- **No covariates** - Check the data for the presence of spatial correlation using Moran's I statistic, and if it is present fit a spatial correlation model to the data.
- **Covariates** - Initially fit a model for the covariates ignoring the spatial correlation and assess the residuals from that model for the presence of spatial correlation. If no correlation is present then the modelling is finished, otherwise if correlation is present re-fit the same covariate model but allow for spatial correlation.

Common data types

The type of model to fit will depend on the type of data being modelled.

- **Continuous variable -** If the response variable is continuous then a **normal (Gaussian)** distribution is usually appropriate (possibly after transformation).
- **Discrete variable -** If the response variable is the number of people that have the event of interest in each area, then a normal distribution is not appropriate. Instead:
  - A **binomial** model is used if the event is not that rare.
  - A **Poisson** model is used if the event is rare, as in this case the binomial probabilities would be hard to model being so close to zero.

Notation

Suppose we have the following data on K areas (e.g. datazones):

- $\mathbf{Y} = (Y_1, \ldots, Y_K)$ is a vector of response data to be modelled, where Y_k is the value for area k.
- $\mathbf{N} = (N_1, \ldots, N_K)$ is a vector of population sizes, where N_k is the value for area k.
- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ is a matrix of p covariates including the intercept term, where for area k the covariates are denoted by $\mathbf{x}_k = (1, x_{k2}, \ldots, x_{kp})$.

Continuous data models

A spatial model for continuous data is the **normal linear spatial regression model**, which has the form

$$\begin{aligned} Y_k &\sim \mathsf{N}(\theta_k, \sigma^2), \\ \theta_k &= \beta_1 + \beta_2 x_{k2} + \ldots + \beta_p x_{kp} + \phi_k. \end{aligned}$$

The fitted values θ_k depend on:

- known covariates with regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. If x_{kj} increases by 1 then the fitted value increases by β_j.
- random effects (spatial errors) ϕ_k that control for spatial correlation.

Spatial binomial logistic model

The spatial binomial logistic regression model has the general form

$$\begin{aligned} Y_k &\sim \text{Binomial}(N_k, \theta_k), \\ \ln\left(\frac{\theta_k}{1-\theta_k}\right) &= \beta_1 + \beta_2 x_{k2} + \ldots + \beta_p x_{kp} + \phi_k. \end{aligned}$$

The probability of success in area k is denoted by θ_k and is modelled by:

- Covariates $(x_{k1}, \ldots, x_{kp})$ with regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.
- random effects (spatial errors) ϕ_k that control for spatial correlation.

Notes

- The logistic link function $\ln\left(\frac{\theta_k}{1-\theta_k}\right)$ is used to ensure that θ_k is between 0 and 1 as it is a probability.
- Covariate effects β_j are summarised as odds ratios, which is the ratio of the odds of the event of interest with the jth covariate increased by one divided by the odds of the event of interest with the jth covariate fixed.
- An odds ratio for x_{kj} is calculated as $\exp(\beta_j)$, and interpretation is that an odds ratio of 2 means that if the jth covariate increases by 1 then the odds have doubled.
- You may not have covariates in your data, in which case you just have ϕ_k in the model.

Spatial Poisson log-linear model

The spatial Poisson log-linear regression model has the general form

$$\begin{aligned} Y_k &\sim \text{Poisson}(N_k\theta_k), \\ \ln(\theta_k) &= \beta_1 + \beta_2 x_{k2} + \ldots + \beta_p x_{kp} + \phi_k. \end{aligned}$$

The rate of the event in area k is denoted by θ_k and is modelled by:

- Covariates $(x_{k1}, \ldots, x_{kp})$ with regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.
- random effects (spatial errors) ϕ_k that control for spatial correlation.

What is $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$?

- $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ are called **random effects**, and there is one for each areal unit, i.e. K in total.
- The value ϕ_k provides an adjustment to the fitted values or rates θ_k in area k, and is simply added to the regression component $\beta_1 + \beta_2 x_{k2} + \ldots + \beta_p x_{kp}$.
- The set of random effects are constrained to be spatially correlated, which essentially means that when mapped they produce a spatially smooth surface.
- They allow for the unmeasured spatial correlation and variation in the data resulting from factors such as unmeasured confounding.

How is ϕ modelled?

- There are a number of possible approaches to ensuring that ϕ is modelled as spatially autocorrelated.
- The most commonly used model class is **Conditional AutoRegressive (CAR)** models, which are known as **CAR** models for short.
- They are a spatial analogue of autoregressive models in time series modelling.
- They use the neighbourhood matrix $\mathbf{W}$ to induce spatial correlation into ϕ.

Measuring spatial closeness

- Recall that spatial closeness is captured by the neighbourhood (adjacency) matrix, that is commonly denoted by $\mathbf{W}$.
- As our data have K areas $\mathbf{W}$ is a $K \times K$ matrix, where w_{kj} determines whether areas (k, j) are spatially close.
- Most often a binary specification is taken where if $w_{kj} = 0$ then areas (k, j) are not spatially close but if $w_{kj} = 1$ then they are spatially close. Always $w_{kk} = 0$.
- The most common specification is that two areas are spatially close if they share a common border.

CAR models

CAR models are commonly specified as a set of K univariate conditional distributions for ϕ_k given the remaining elements $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_K)$. The simplest CAR model is called the **Intrinsic CAR model (ICAR)** and is given by:

$$\phi_k | \boldsymbol{\phi}_{-k}, \mathbf{W} \sim \mathsf{N}\left(\frac{\sum_{j=1}^{K} w_{kj}\phi_j}{\sum_{j=1}^{K} w_{kj}}, \frac{\tau^2}{\sum_{j=1}^{K} w_{kj}} \right).$$

So each ϕ_k is modelled as normally distributed, with a mean and a variance that depend on the neighbourhood (spatial closeness) information $\mathbf{W}$.

Why does this represent spatial correlation?

This model captures spatial correlation through its mean and variance.

- The mean of ϕ_k is $\frac{\sum_{j=1}^{K} w_{kj}\phi_j}{\sum_{j=1}^{K} w_{kj}}$, which is the mean of the random effects in neighbouring areas (areas for whom $w_{kj} = 1$).
- The variance of ϕ_k is $\frac{\tau^2}{\sum_{j=1}^{K} w_{kj}}$, which is inversely proportional to the number of neighbouring areas. Thus the more areas that are close to area k and have similar values to ϕ_k, the more information there is about ϕ_k and hence its uncertainty goes down.

Note that $\sum_{j=1}^{K} w_{kj}$ corresponds to the number of spatially neighbouring areas for area k.

A better model

However, the ICAR model does not have a spatial dependence parameter, which led *Leroux et al. (2000)* to propose the following extended model.

$$\phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W} \sim \mathsf{N}\left(\frac{\rho\sum_{j=1}^{K} w_{kj}\phi_j}{\rho\sum_{j=1}^{K} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{j=1}^{K} w_{kj} + 1 - \rho}\right).$$

So here:

- $\rho = 1$ corresponds to strong spatial dependence and is the ICAR model.
- $\rho = 0$ corresponds to independence as then $\phi_k \sim \mathsf{N}(0, \tau^2)$.

Fitting the model

- This model is most often fitted in a Bayesian setting instead of using maximum likelihood, as it allows the correct propagation of uncertainty through the model.
- This means that any 95% uncertainty intervals that are computed are called **95% credible intervals** and not 95% confidence intervals.
- Parameter estimation is typically done via a simulation based approach, using a Markov Chain Monte Carlo (MCMC) algorithm.
- The upside of this is that it is very flexible, the downside is that the model takes longer to fit.

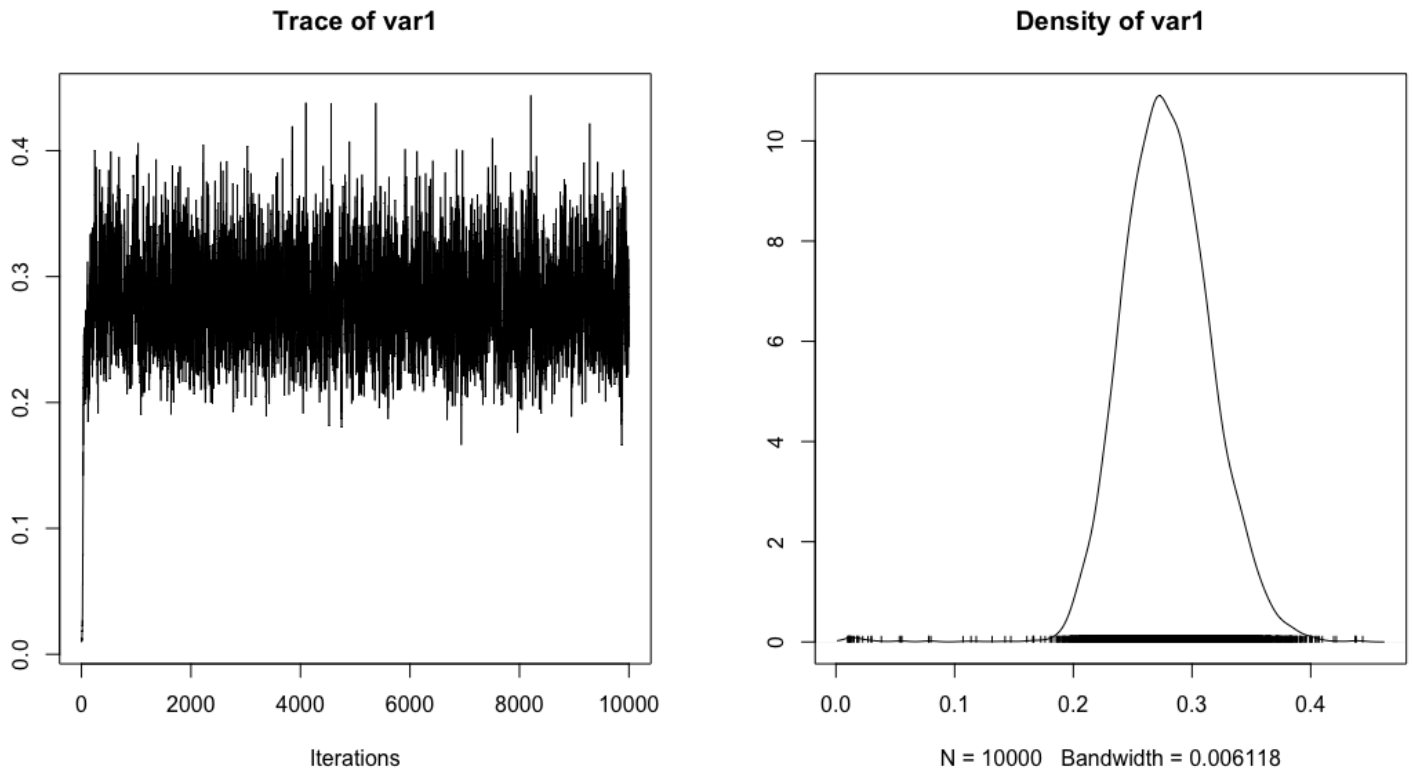
Markov Chain Monte Carlo simulation

The underlying idea with MCMC simulation is as follows.

- Randomly generate starting values for each parameter (e.g. for $(\boldsymbol{\beta}, \boldsymbol{\phi}, \tau^2, \rho)$).
- Repeat the following M (say $M = 100,000$) times. For each parameter simulate a new value using the data and the current values of the other parameters.
- As the starting values were randomly generated, the first part of the M simulated values for each parameter will not be good estimates. This is known as the **burnin** period and these simulated values are removed.
- After the burnin period the simulated values of the parameters are said to have **converged**, and will be good estimates of the parameters.

Simulated parameter values

The simulated parameter values are shown in the left plot, while the right one is a density estimate of all values.



The simulated values (samples) appear to have converged from around 500 iterations onwards. So the first 500 samples should be removed here as the burnin period.

Checking convergence of the parameter samples

There are two easy ways to check convergence of a set of parameter samples.

- Look at trace plots such as that on the previous slide. When it shows no trend or pattern then it has converged.
- The software we use presents the Geweke diagnostic for a number of the parameters, which is a Z-score, so that values between (-1.96, 1.96) are indicative of convergence.

In practice both methods can be used together on a subset of the parameters. Checking each one would take too long!

Correlation in MCMC samples

Here we use the spatial dependence parameter ρ as an example. Once convergence has been checked and the samples in the burnin period have been removed, one is left with G sample values $\{\rho^{(1)}, \ldots, \rho^{(G)}\}$ that represent the true parameter ρ.

- These G samples $\{\rho^{(1)}, \ldots, \rho^{(G)}\}$ are not independent estimates of ρ, but are instead correlated by design.
- This correlation can be checked using the `acf()` function in `R`.
- These G correlated samples provide much less information about ρ than G independent samples would, so the software we use computes the effective number of independent samples.

Summarising MCMC samples

Again using the spatial dependence parameter ρ as an example, we have G sample values $\{\rho^{(1)}, \ldots, \rho^{(G)}\}$ that represent the true parameter ρ. They can be summarised as follows:

- A point estimate can be obtained by computing the sample mean or median of the G values.
- A 95% credible interval (the Bayesian equivalent of a 95% confidence interval) can be constructed by calculating the (2.5, 97.5) percentile points of the G values, and the model thinks there is a 95% chance the true value lies in that interval.

Model comparison

- Model comparison statistics such as adjusted R^2 or AIC are not commonly used in a Bayesian setting, and instead other measures are used.
- The most popular is the Deviance Information Criterion (DIC), which is very similar to the AIC.
- The best fitting model is the one that minimises the DIC.

Software

- These spatial models can be fitted using the `R` package `CARBayes`, which can fit a range of spatial models.
- The package is on the R website (`https://cran.r-project.org`) and contains a comprehensive vignette covering modelling details and worked examples.
- The CAR model proposed by *Leroux et al. (2000)* is implemented by the function `S.CARleroux()`, and we illustrate this in the next session.
- The function has many of the arguments of the `glm()` function, such as formula and family arguments.