Modelling spatial data in R with CARBayes

Part 2: Spatial modelling with CARBayes

Duncan Lee and Eilidh Jack

GEOMED 2017

1. Introduction

This session will cover the following:

- Modelling spatial data allowing for spatial correlation.
- Using `CARBayes` for modelling.
- Inference from spatial models.

Beyond exploratory measures of risk

Exploratory measures of risk such as the standardised morbidity (mortality) ratio provide a good initial estimate for visualising the data. However, they are deficient for a number of reasons:

- They are unstable estimates as they are often simple ratios, especially if the population size is small.
- They do not borrow strength in the estimation of risk using the spatial correlation in the data.
- You cannot estimate the effects of covariates on disease risk.

Therefore a new modelling approach is needed.

2. Modelling spatial data

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

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

- **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. Even if it is not correlated one can use models to better estimate risks.

Notation

Suppose we have the following data on K areas (e.g. intermediate zones, etc):

- $\mathbf{Y} = (Y_1, \ldots, Y_K)$ is a vector of disease response data to be modelled, where Y_k is the value for area k.
- $\mathbf{E} = (E_1, \ldots, E_K)$ is a vector of population sizes or indirectly standardised expected numbers of disease cases, where E_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 $\mathbf{x}_k = (1, x_{k2}, \ldots, x_{kp})$.

`CARBayes` can fit the following 3 spatial data models.

(i) Continuous data model

For continuous data, the **Gaussian spatial linear model** is given by.

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

- Here Y_k is assumed to be normally distributed with fitted (mean) value μ_k.
- The fitted values depend on covariates with regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and spatially correlated random effects ϕ_k.
- σ^2 represents the amount of unexplained random variation (error).

(ii) Count data model 1

For count data where the disease in question is not that rare, the **Binomial spatial logistic model** is given by:

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

- Y_k is the number of people in area k with the event and E_k is the total population size.
- The key parameter θ_k is the probability of having the event in area k.
- Again the spatial variation is modelled by covariates and unexplained spatial variation.

(iii) Count data model 2

For count data where the disease in question is rare, the **Poisson spatial log-linear model** is given by:

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

- Y_k is the number of people in area k with the event (e.g. hospital admissions) and E_k is the indirectly standardised expected number of events based on population demographics.
- The key parameter θ_k is the relative (to E_k) risk of having the event in area k, and is on the same scale as the SMR.
- Again the spatial variation is modelled by covariates and unexplained spatial variation.

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 estimated risk or fitted value 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 disease 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.

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 expectation and variance.

- The expectation 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.

Two common models

However, the ICAR model does not have a spatial dependence parameter, which led *Besag et al. (1991)* to propose the BYM or convolution model.

$$\begin{aligned} \phi_k &= \phi_k^{(1)} + \phi_k^{(2)} \\ \phi_k^{(1)} | \boldsymbol{\phi}_{-k}^{(1)}, \mathbf{W} &\sim \mathsf{N}\left(\frac{\sum_{j=1}^{K} w_{kj} \phi_j^{(1)}}{\sum_{j=1}^{K} w_{kj}}, \frac{\tau^2}{\sum_{j=1}^{K} w_{kj}} \right) \\ \phi_k^{(2)} &\sim \mathsf{N}(0, \sigma^2). \end{aligned}$$

So here the spatial structure is split into:

- $\phi_k^{(1)}$ - strongly spatially correlated variation.
- $\phi_k^{(2)}$ - independent spatial variation.

This model can be fitted using the function `S.CARbym()` in `CARBayes`.

The second commonly used model was proposed by *Leroux et al. (2000)* and is given by.

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

So here:

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

This model can be fitted using the function `S.CARleroux()` in `CARBayes`.

3. Fitting the model in `CARBayes`

- 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 called the Markov Chain Monte Carlo (MCMC) algorithm.
- Here we illustrate the `CARBayes R` package, which is available on CRAN and comes with a vignette with fully worked examples.

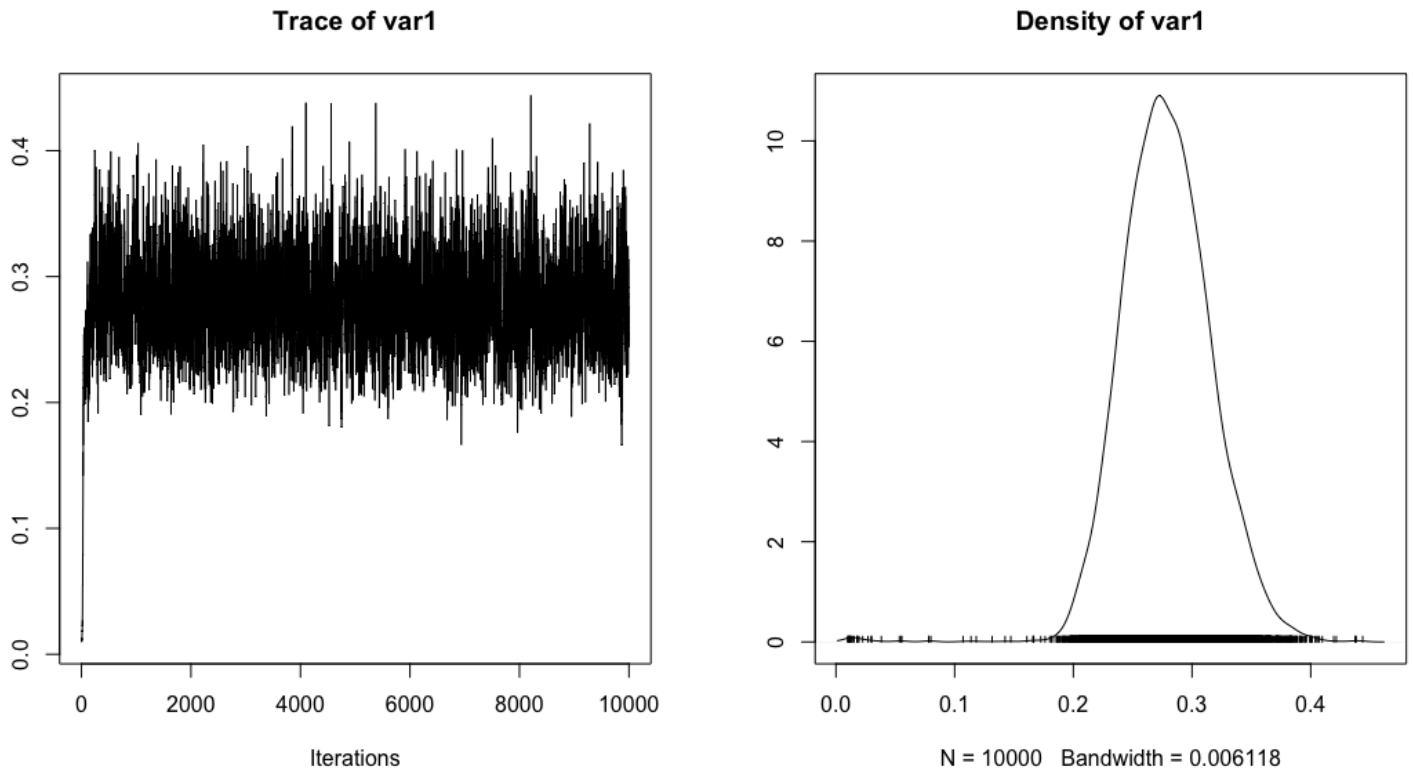
Markov Chain Monte Carlo simulation

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.
- `CARBayes` 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 `CARBayes` 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

Non-Bayesian model comparison statistics such as AIC are not commonly used in a Bayesian setting, and instead `CARBayes` produces the following model fit criteria.

- The Deviance Information Criterion (DIC) and the Watanabe-Akaike Information Criterion (WAIC), which have similar interpretations to AIC.
- The percentage of the deviance (variation) explained by the model.
- The log marginal predictive likelihood (LMPL), which is a predictive measure of model fit.

Fitting the model in `CARBayes`

To fit the model in the `CARBayes` software you need to specify the following arguments.

- `formula` - The response and covariates to include in the model.
- `family` - Which data likelihood model to fit.
- `W` - The spatial neighbourhood matrix $\mathbf{W}$.
- `burnin` - The number of MCMC samples to remove as the burnin period.
- `n.sample` - The total number of MCMC samples to generate.

4. Summarising the results

The main quantities of interest you may want to compute from the fitted model are:

- The effects of the covariates on disease risk.
- The estimated risks in each of the K areas.
- The probability in each area that the risk is elevated above the average. These are called **Posterior exceedence probabilities**.

We illustrate how to compute each of these in the practical session that follows.

Other models `CARBayes` will fit

`CARBayes` will fit the following other models:

- Locally smooth models such as `S.CARdissimilarity()` and `S.CARlocalised()`.
- A simple generalised linear model `S.glm()` with no correlation for comparison purposes.
- A two-level model for individuals within areas using `S.CARmultilevel()`.
- A simple multivariate model `MVS.CARleroux()` for multiple disease data for each spatial unit.