CARBayes version 5.1.1: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors

Duncan Lee
University of Glasgow

Abstract

This is a vignette for the R package **CARBayes** version 5.1.1, and is an updated version of a paper in the Journal of Statistical Software in 2013 Volume 55 Issue 13 by the same author. The package implements univariate and multivariate spatial generalised linear mixed models for areal unit data, with inference in a Bayesian setting using Markov chain Monte Carlo (MCMC) simulation. The response variable can be binomial, Gaussian, multinomial, Poisson or zero-inflated Poisson (ZIP), and spatial autocorrelation is modelled by a set of random effects that are assigned a conditional autoregressive (CAR) prior distribution. A number of different models are available for univariate spatial data, including models with no random effects as well as random effects modelled by different types of CAR prior. Additionally, a multivariate CAR (MCAR) model for multivariate spatial data is available, as is a two-level hierarchical model for modelling data relating to individuals within areas. The initial creation of this package was supported by the Economic and Social Research Council (ESRC) grant RES-000-22-4256, and on-going development has been supported by the Engineering and Physical Science Research Council (EPSRC) grant EP/J017442/1, ESRC grant ES/K006460/1, Innovate UK / Natural Environment Research Council (NERC) grant NE/N007352/1, and the TB Alliance. Version 5.1.1 has a number of changes to version 5.0, including:

1. Multinomial and ZIP data models with either no random effects or random effects modelled by the Leroux CAR prior. For the multinomial model the latter is an MCAR model.
2. A ZIP model with the BYM CAR prior.
3. An MCAR model for a Gaussian data likelihood.
4. The use of data augmentation to account for missing values in the respone variable.
5. An updated vignette using the **leaflet** package for mapping.

Keywords: Bayesian inference, conditional autoregressive priors, R package **CARBayes**.

1. Introduction

Data relating to a set of non-overlapping spatial areal units are prevalent in many fields, including agriculture (Besag and Higdon (1999)), ecology (Brewer and Nolan (2007)), education (Wall (2004)), epidemiology (Lee (2011)) and image analysis (Gavin and Jennison (1997)). There are numerous motivations for modelling such data, including ecological regression (see

Wakefield (2007) and Lee *et al.* (2009)), disease mapping (see Green and Richardson (2002) and Lee (2011)) and Wombling (see Lu *et al.* (2007), Ma and Carlin (2007)). The set of areal units on which data are recorded can form a regular lattice or differ largely in both shape and size, with examples of the latter including the set of electoral wards or census tracts corresponding to a city or country. In either case such data typically exhibit spatial autocorrelation, with observations from areal units close together tending to have similar values. A proportion of this spatial autocorrelation may be modelled by known covariate risk factors in a regression model, but it is common for spatial structure to remain in the residuals after accounting for these covariate effects. This residual spatial autocorrelation can be induced by a number of factors, and violates the assumption of independence that is common in many regression models. One possible cause is unmeasured confounding, which occurs when an important spatially autocorrelated covariate is either unmeasured or unknown. The spatial structure in this covariate induces spatial autocorrelation into the response, which hence cannot be accounted for in a regression model. Other possible causes of residual spatial autocorrelation are neighbourhood effects, where subjects behaviour is influenced by that of neighbouring subjects, and grouping effects, where subjects choose to be close to similar subjects.

The most common remedy for this residual autocorrelation is to augment the linear predictor with a set of spatially autocorrelated random effects, as part of a Bayesian hierarchical model. These random effects are typically represented with a conditional autoregressive (CAR, Besag *et al.* (1991)) prior, which induces spatial autocorrelation through the adjacency structure of the areal units. A number of CAR priors have been proposed in the literature, including the intrinsic and Besag-York-Mollié (BYM) models (both Besag *et al.* (1991)), as well as alternatives developed by Leroux *et al.* (2000) and Stern and Cressie (1999).

However, the CAR priors listed above force the random effects to exhibit a single global level of spatial autocorrelation, ranging from independence through to strong spatial smoothness. Such a uniform level of spatial autocorrelation for the entire region maybe unrealistic for real data, which instead may exhibit sub-regions of spatial autocorrelation separated by discontinuities. Such localised spatial autocorrelation may occur where rich and poor communities live side-by-side, and in this context the response variable is likely to evolve smoothly within each community with a sudden change in its value at the border where the two communities meet. However, covariate data quantifying this localised structure may not be available, meaning that it has to be modelled by the random effects. A number of approaches have been proposed for extending the class of CAR priors to deal with localised spatial smoothing amongst the random effects, including papers by Lawson and Clark (2002), Brewer and Nolan (2007), Lu *et al.* (2007), Lee and Mitchell (2012), and Lee *et al.* (2014).

The models described above are typically implemented in a Bayesian setting, where inference is based on Markov chain Monte Carlo (MCMC) simulation. The most commonly used software to implement this class of models is the BUGS project (Lunn *et al.* (2009), WinBUGS and OpenBUGS), which has in-built functions `car.normal()` and `car.proper()` to implement the intrinsic, BYM and Stern and Cressie (1999) models. The intrinsic and BYM models can also be implemented in BayesX (Belitz, C and Brezger, A and Kneib, T and Lang, S (2009)), while the R software (R Core Team 2016) packages **hSDM** (Vieilledent *et al.* 2014), **spatcounts** (Schabenberger 2009) and **spdep** (Bivand 2013) can implement a restricted set of CAR models. CAR models can also be implemented in R using Integrated Nested Laplace Approximations (INLA, *http://www.r-inla.org/*), using the package **INLA** (Rue *et al.* 2009).

However, these software packages either only fit a limited set of CAR models or require a degree of programming to implement them, which was the original motivation for creating **CARBayes** (Lee 2013). Its main advantage is its ease of use because: (1) the spatial adjacency information is easy to specify as a neighbourhood (adjacency) matrix; and (2) given the neighbourhood matrix, models can be implemented by a single function call. **CARBayes** can implement a much wider class of spatial areal unit models than is possible using the R packages listed above, because the univariate or multivariate response data can follow binomial, Gaussian, multinomial, Poisson or zero-inflated Poisson (ZIP) distributions, while a range of CAR priors can be specified for the random effects. Additionally, a two-level hierarchical model is available for modelling data relating to individuals within areas. Spatio-temporal models for areal unit data using CAR type priors can be implemented using the sister package **CARBayesST** (Lee *et al.* 2018).

The aim of this vignette is to present the software **CARBayes**, by outlining the class of models that it can implement and illustrating its use by means of 3 worked examples. The remainder of this vignette is organised as follows. Section two outlines the general Bayesian hierarchical model that can be implemented in the **CARBayes** package, while Section three gives details about the software. Sections four to six give three worked examples of using the software, including how to create the neighbourhood matrix and produce spatial maps of the results. Finally, Section 7 contains a concluding discussion, and outlines areas for future development.

2. Spatial models for areal unit data

This section outlines the class of spatial generalised linear mixed models for areal unit data that can be implemented in **CARBayes**. Inference for all models is set in a Bayesian framework, and is based on MCMC simulation. The majority of the models in **CARBayes** relate to univariate spatial data and are described in Section 2.1, whiel models for multivariate spatial data and two-level data relating to individuals within areas are described in Sections 2.2 and 2.3.

2.1. Univariate spatial data models

The study region $\mathcal{S}$ is partitioned into K non-overlapping areal units $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$, which are linked to a corresponding set of responses $\mathbf{Y} = (Y_1, \ldots, Y_K)$, and a vector of known offsets $\mathbf{O} = (O_1, \ldots, O_K)$. Missing, `NA`, values are allowed in the response $\mathbf{Y}$ except for the `S.CARlocalised()` function, which does not allow them due to model complexity and corresponding poor predictive performance. These missing values are treated as additional unknown parameters, and are updated in the MCMC algorithm using a data augmentation approach Tanner and Wong (1987). The spatial variation in the response is modelled by a matrix of covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ and a spatial structure component $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)$, the latter of which is included to model any spatial autocorrelation that remains in the data after the covariate effects have been accounted for. The vector of covariates for areal unit $\mathcal{S}_k$ are denoted by $\mathbf{x}_k = (1, x_{k1}, \ldots, x_{kp})$, the first of which corresponds to an intercept term. The general spatial generalised linear mixed model is given by

$$Y_k|\mu_k \quad \sim \quad f(y_k|\mu_k, \nu^2) \qquad \text{for } k = 1, \ldots, K \tag{1}$$

$$\begin{aligned} g(\mu_k) &= \mathbf{x}_k^\top \boldsymbol{\beta} + O_k + \psi_k \\ \boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\ \nu^2 &\sim \text{Inverse-Gamma}(a, b). \end{aligned}$$

The expected value of Y_k is denoted by $\mathsf{E}(Y_k) = \mu_k$, while ν^2 is an additional scale parameter that is required if the Gaussian family is used. The latter is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\nu^2 \sim \text{Inverse-Gamma}(1, 0.01)$. The vector of regression parameters are denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, and non-linear covariate effects can be incorporated into the above model by including natural cubic spline or polynomial basis functions of the covariates in $\mathbf{X}$. A multivariate Gaussian prior is assumed for $\boldsymbol{\beta}$, and the mean $\boldsymbol{\mu}_\beta$ and diagonal variance matrix $\boldsymbol{\Sigma}_\beta$ can be chosen by the user. Default values specified by the software are a constant zero-mean vector and diagonal elements of $\boldsymbol{\Sigma}_\beta$ equal to 100,000. The expected values of the responses are related to the linear predictor via an invertible link function $g(.)$, and **CARBayes** can fit the following data likelihood models:

- **Binomial -** $Y_k \sim \text{Binomial}(n_k, \theta_k)$ and $\ln(\theta_k/(1-\theta_k)) = \mathbf{x}_k^\top \boldsymbol{\beta} + O_k + \psi_k$.
- **Gaussian -** $Y_k \sim \mathrm{N}(\mu_k, \nu^2)$ and $\mu_k = \mathbf{x}_k^\top \boldsymbol{\beta} + O_k + \psi_k$.
- **Poisson -** $Y_k \sim \text{Poisson}(\mu_k)$ and $\ln(\mu_k) = \mathbf{x}_k^\top \boldsymbol{\beta} + O_k + \psi_k$.
- **ZIP -** $Y_k \sim \text{ZIP}(\mu_k, \omega_k)$. The zero-inflated Poisson model is used to represent data containing an excess of zeros, and is a mixture of a point mass distribution based at zero and a Poisson distribution with mean μ_k. The probability that observation Y_k is in the point mass distribution based at zero (called a structural zero) is ω_k, and (μ_k, ω_k) are modelled by

  $$\ln(\mu_k) = \mathbf{x}_k^\top \boldsymbol{\beta} + O_k + \psi_k \qquad \ln\left(\frac{\omega_k}{1-\omega_k}\right) = \mathbf{v}_k^\top \boldsymbol{\delta} + O_k^{(2)}.$$

  Here $(\mathbf{v}_k, O_k^{(2)})$ are respectively covariates and an offset term that determine the probability that observation Y_k is in the point mass distribution, while $\boldsymbol{\delta}$ are the corresponding regression parameters. In implementing the model a binary random variable Z_k is sampled for each observation Y_k, where $Z_k = 1$ if Y_k comes from the point mass distribution, and $Z_k = 0$ if Y_k comes from the Poisson distribution. Further details about ZIP models are given by Ugarte *et al.* (2004).

In the binomial model above n_k is the number of trials in the kth area, while θ_k is the probability of success in a single trial. **CARBayes** can implement a number of different spatial random effects models for $\boldsymbol{\psi}$, and they are summarised below.

- `S.glm()` - fits a model with no random effects and thus is a generalised linear model. This model can be implemented with binomial, Gaussian, Poisson and ZIP data likelihoods.
- `S.CARbym()` - fits the convolution or Besag-York-Mollie (BYM) CAR model outlined in Besag *et al.* (1991). This model can be implemented with binomial, Poisson and zip data likelihoods.

- `S.CARleroux()` - fits the CAR model proposed by Leroux *et al.* (2000). This model can also fit the intrinsic CAR model proposed by Besag *et al.* (1991), as well as a model with independent random effects. This model can be implemented with binomial, Gaussian, Poisson and ZIP data likelihoods.

- `S.CARdissimilarity()` - fits the localised spatial autocorrelation model proposed by Lee and Mitchell (2012). This model can be implemented with binomial, Gaussian and Poisson data likelihoods.

- `S.CARlocalised()` - fits the localised spatial autocorrelation model proposed by Lee and Sarran (2015). This model can be implemented with binomial and Poisson data likelihoods.

The spatial structure component $\boldsymbol{\psi}$ includes a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, which come from a conditional autoregressive model. These models are a special case of a Gaussian Markov Random Field (GMRF), and can be written in the general form $\boldsymbol{\phi} \sim \text{N}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W}, \rho)^{-1})$, where $\mathbf{Q}(\mathbf{W}, \rho)$ is a precision matrix that may be singular (intrinsic model). This matrix controls the spatial autocorrelation structure of the random effects, and is based on a non-negative symmetric $K \times K$ neighbourhood (or adjacency) matrix $\mathbf{W}$, and potentially a spatial dependence parameter ρ depending on the model chosen. The kjth element of the neighbourhood matrix w_{kj} represents the spatial closeness between areas $(\mathcal{S}_k, \mathcal{S}_j)$, with positive values denoting geographical closeness and zero values denoting non-closeness. Additionally, diagonal elements $w_{kk} = 0$.

A binary specification for $\mathbf{W}$ based on geographical contiguity is most commonly used, where $w_{kj} = 1$ if areal units $(\mathcal{S}_k, \mathcal{S}_j)$ share a common border (denoted $k \sim j$), and is zero otherwise. This specification forces (ϕ_k, ϕ_j) relating to geographically adjacent areas (that is where $w_{kj} = 1$) to be autocorrelated, whereas random effects relating to non-contiguous areal units are conditionally independent given the values of the remaining random effects. A binary specification is not necessary in **CARBayes** except for the function `S.CARdissimilarity()`, as the only requirement is that $\mathbf{W}$ is non-negative and symmetric. However, each area must have at least one positive element $\{w_{kj}\}$, meaning the row sums of $\mathbf{W}$ must be positive. CAR priors are commonly specified as a set of K univariate full conditional distributions $f(\phi_k|\boldsymbol{\phi}_{-k})$ for $k = 1, \ldots, K$ (where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_K)$), which is how they are presented below. We now outline the five models that **CARBayes** can fit.

A model with no random effects

`S.glm()`
The simplest model that **CARBayes** can implement is a generalised linear model, which is based on (1) with the simplification that $\psi_k = 0$ for all areas k.

Globally smooth CAR models

`S.CARbym()`
The convolution or Besag-York-Mollie (BYM) CAR model outlined in Besag *et al.* (1991) contains spatially autocorrelated and independent random effects andis given by

$$\begin{aligned} \psi_k &= \phi_k + \theta_k \\ \phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W}, \tau^2 &\sim \text{N}\left(\frac{\sum_{i=1}^K w_{ki}\phi_i}{\sum_{i=1}^K w_{ki}}, \frac{\tau^2}{\sum_{i=1}^K w_{ki}}\right) \\ \theta_k &\sim \text{N}(0, \sigma^2) \\ \tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a, b). \end{aligned} \tag{2}$$

Here $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ are independent with zero mean and a constant variance, while spatial autocorrelation is modelled via $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$. For the latter the conditional expectation is the average of the random effects in neighbouring areas, while the conditional variance is inversely proportional to the number of neighbours. This is appropriate because if the random effects are strongly spatially autocorrelated, then the more neighbours an area has the more information there is from its neighbours about the value of its random effect, hence the uncertainty reduces. In common with the other variance parameters the default prior specification for (τ^2, σ^2) has $(a = 1, b = 0.01)$. This model contains two random effects for each data point, and as only their sum is identifiable from the data only $\psi_k = \phi_k + \theta_k$ is returned to the user.

`S.CARleroux()`

Leroux *et al.* (2000) proposed the following alternative CAR prior for modelling varying strengths of spatial autocorrelation using only a single set of random effects.

$$\begin{aligned} \psi_k &= \phi_k \\ \phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W}, \tau^2, \rho &\sim \text{N}\left(\frac{\rho\sum_{i=1}^K w_{ki}\phi_i}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}\right) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho &\sim \text{Uniform}(0, 1). \end{aligned} \tag{3}$$

Here ρ is a spatial dependence parameter taking values in the unit interval, and can be fixed (using the argument `rho`) if required. specifically, $\rho = 1$ corresponds to the intrinsic CAR model, while $\rho = 0$ corresponds to independence.

Locally smooth CAR models

The CAR priors described above enforce a single global level of spatial smoothing for the set of random effects, which for model (3) is controlled by ρ. This is illustrated by the partial autocorrelation structure implied by that model, which for (ϕ_k, ϕ_j) is given by

$$\text{COR}(\phi_k, \phi_j|\boldsymbol{\phi}_{-kj}, \mathbf{W}, \rho) = \frac{\rho w_{kj}}{\sqrt{(\rho\sum_{i=1}^K w_{ki} + 1 - \rho)(\rho\sum_{i=1}^K w_{ji} + 1 - \rho)}}. \tag{4}$$

For non-neighbouring areal units (where $w_{kj} = 0$) the random effects are conditionally independent, while for neighbouring areal units (where $w_{kj} = 1$) their partial autocorrelation is

controlled by ρ. This representation of spatial smoothness is likely to be overly simplistic in practice, as the random effects surface is likely to include sub-regions of smooth evolution as well as boundaries where abrupt step changes occur. Therefore **CARBayes** can implement the localised spatial autocorrelation models proposed by Lee and Mitchell (2012) and Lee and Sarran (2015).

`S.CARdissimilarity()`

Lee and Mitchell (2012) proposed a method for capturing localised spatial autocorrelation and identifying boundaries in the random effects surface. The underlying idea is to model the elements of $\mathbf{W}$ corresponding to geographically adjacent areal units as random quantities, rather than assuming they are fixed at one. Conversely, if areal units $(\mathcal{S}_k, \mathcal{S}_j)$ are not adjacent as specified by $\mathbf{W}$, then w_{kj} is fixed at zero. From (4), it is straightforward to see that if w_{kj} is estimated as one then (ϕ_k, ϕ_j) are spatially autocorrelated and are smoothed over in the modelling process, whereas if w_{kj} is estimated as zero then no smoothing is imparted between (ϕ_k, ϕ_j) as they are modelled as conditionally independent. In this case a boundary is said to exist in the random effects surface between areal units $(\mathcal{S}_k, \mathcal{S}_j)$. We note that for this model $\mathbf{W}$ must be binary.

The model is based on (3) with ρ fixed at 0.99, which ensures that the random effects exhibit strong spatial smoothing globally, which can be altered locally by estimating $\{w_{kj}|k \sim j\}$. They model each w_{kj} as a function of the dissimilarity between areal units $(\mathcal{S}_k, \mathcal{S}_j)$, because large differences in the response are likely to occur where neighbouring populations are very different. This dissimilarity is captured by q non-negative dissimilarity metrics $\mathbf{z}_{kj} = (z_{kj1}, \ldots, z_{kjq})$, which could include social or physical factors, such as the absolute difference in smoking rates, or the proportion of the shared border that is blocked by a physical barrier (such as a river or railway line) and cannot be crossed. Using these measures of dissimilarity two distinct models are proposed for $\{w_{kj}|k \sim j\}$.

Binary model

$$\begin{aligned} w_{kj}(\boldsymbol{\alpha}) &= \begin{cases} 1 & \text{if } \exp(-\sum_{i=1}^{q} z_{kji}\alpha_i) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases} \\ \alpha_i &\sim \text{Uniform}(0, M_i) \qquad \text{for } i = 1, \ldots, q. \end{aligned} \tag{5}$$

Non-binary model

$$\begin{aligned} w_{kj}(\boldsymbol{\alpha}) &= \exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) \\ \alpha_i &\sim \text{Uniform}(0, 50) \qquad \text{for } i = 1, \ldots, q. \end{aligned} \tag{6}$$

The q regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ determine the effects of the dissimilarity metrics on $\{w_{kj}|k \sim j\}$, and for the binary model if $\alpha_i < -\ln(0.5)/\max\{z_{kji}\}$, then the ith dissimilarity metric has not solely identified any boundaries because $\exp(-\alpha_i z_{kji}) > 0.5$ for all $k \sim j$. The upper limits M_i for the priors for α_i in the binary model depend on the distribution of z_{kji}, and are chosen to be weakly informative and fixed in the software. Users can choose between (5) and (6) by the logical argument `W.binary`, where `TRUE` corresponds to (5), while `FALSE` corresponds to (6).

`S.CARlocalised()`
An alternative to the above is to augment the set of spatially smooth random effects with a piecewise constant intercept or cluster model, thus allowing large jumps in the mean surface between adjacent areal units in different clusters. Lee and Sarran (2015) proposed a model that partitions the K areal units into a maximum of G clusters each with their own intercept term $(\lambda_1, \ldots, \lambda_G)$. The model is given by

$$\begin{aligned} \psi_k &= \phi_k + \lambda_{Z_k} \\ \phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W}, \tau^2 &\sim \text{N}\left(\frac{\sum_{i=1}^K w_{ki}\phi_i}{\sum_{i=1}^K w_{ki}}, \frac{\tau^2}{\sum_{i=1}^K w_{ki}}\right) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \lambda_i &\sim \text{Uniform}(\lambda_{i-1}, \lambda_{i+1}) \qquad \text{for } i = 1, \ldots, G \\ f(Z_k) &= \frac{\exp(-\delta(Z_k - G^*)^2)}{\sum_{r=1}^G \exp(-\delta(r - G^*)^2)} \\ \delta &\sim \text{Uniform}(1, M). \end{aligned} \tag{7}$$

The cluster means $(\lambda_1, \ldots, \lambda_G)$ are ordered so that $\lambda_1 < \lambda_2 < \ldots < \lambda_G$, which prevents the label switching problem common in mixture models, and $\lambda_0 = -\infty$ and $\lambda_{G+1} = \infty$. Area k is assigned to one of the G intercepts by $Z_k \in \{1, \ldots, G\}$, and G is the maximum number of different intercept terms. Here we penalise Z_k towards the middle intercept value, so that the extreme intercept classes (e.g. 1 or G) may be empty. This is achieved by the penalty term $\delta(Z_k - G^*)^2$ in the prior for Z_k, where $G^* = (G+1)/2$ if G is odd and $G^* = G/2$ if G is even, and is the middle group. A weakly informative uniform prior is specified for the penalty parameter $\delta \sim \text{Uniform}(1, M)$ (by default $M = 10$), so that the data play the dominant role in estimating its value. Note, a Gaussian likelihood is not allowed with this model because of a lack of identifiability among the parameters, and missing values are not allowed in the response for the same reasons.

2.2. Multivariate spatial data models

The study region $\mathcal{S}$ is again partitioned into K non-overlapping areal units $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$, and each unit contain J responses $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kJ})$ and J offsets $\mathbf{O}_k = (O_{k1}, \ldots, O_{kJ})$. The model therefore has to represent both spatial autocorrelation and between variable correlation, and the general multivariate spatial mixed model is given by

$$\begin{aligned} Y_{kj}|\mu_{kj} &\sim f(y_{kj}|\mu_{kj}, \nu^2) \quad \text{for } k = 1, \ldots, K, \quad j = 1, \ldots, J \\ g(\mu_{kj}) &= \mathbf{x}_k^\top \boldsymbol{\beta}_j + O_{kj} + \phi_{kj} \\ \boldsymbol{\beta}_j &\sim \text{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \end{aligned} \tag{8}$$

In common with the univariate models $\mathbf{x}_k^\top$ is a vector of p covariates, and the same covariates are used for each of the J categories of response variable. The regression coefficients $\boldsymbol{\beta}_j$ vary by category j allowing for category specific effects, and Gaussian priors are assumed for the regression parameters $\boldsymbol{\beta}_j$ as before. The following data likelihood models are allowed:

- **Binomial -** $Y_{kj} \sim \text{Binomial}(n_{kj}, \theta_{kj})$ and $\ln(\theta_{kj}/(1-\theta_{kj})) = \mathbf{x}_k^\top \boldsymbol{\beta}_j + O_{kj} + \phi_{kj}$.
- **Gaussian -** $Y_{kj} \sim \text{N}(\mu_{kj}, \nu^2)$ and $\mu_{kj} = \mathbf{x}_k^\top \boldsymbol{\beta}_j + O_{kj} + \phi_{kj}$.

  The scale parameter ν^2 for the Gaussian likelihood is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\nu^2 \sim \text{Inverse-Gamma}(1, 0.01)$.
- **Multinomial -** $\mathbf{Y}_k \sim \text{Multinomial}(n_k, \theta_{k1}, \ldots, \theta_{kJ})$ and $\ln(\theta_{kj}/\theta_{k1}) = \mathbf{x}_k^\top \boldsymbol{\beta}_j + O_{kj} + \phi_{kj}$, where $n_k = \sum_{j=1}^{J} Y_{kj}$.

  The above holds for categories $j = 2, \ldots, J$, and thus category $j = 1$ is a baseline and has no regression parameters or random effects or offset terms (they are all zero). Here θ_{kj} is the probability of a single outcome in area k being in category j, and hence $\sum_{j=1}^{J} \theta_{kj} = 1$.
- **Poisson -** $Y_{kj} \sim \text{Poisson}(\mu_{kj})$ and $\ln(\mu_{kj}) = \mathbf{x}_k^\top \boldsymbol{\beta}_j + O_{kj} + \phi_{kj}$.

When fitting this model the response variable and offset should be $K \times J$ matrices, while each covariate should be a $K \times 1$ vector. As the multinomial model models the first category as a baseline there will be $J-1$ different regression parameter sets and random effect surfaces, where as for the other data likelihood models there will be J regression parameter sets and random effect surfaces. The set of random effects are denoted by $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$, where $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kJ})$ are the set of J values ($J-1$ for the multinomial model where $\phi_{k1} = 0$) for area k. The random effects need to model both spatial autocorrelation and between variable correlation, and this is achieved using a multivariate conditional autoregressive (MCAR) model, for details see Gelfand and Vounatsou (2003). **CARBayes** can fit the following multivariate data models.

`S.glm()`

The `S.glm()` function discussed earlier can also be applied to multinomial data, where in the above equation $\phi_{kj} = 0$ for all (k, j).

`MVS.CARleroux()`

This model can be implemented with binomial, Gaussian, multinomial and Poisson data likelihoods.The random effects $\boldsymbol{\phi}$ are modelled using the approach outlined in Kavanagh *et al.* (2016) given by:

$$\boldsymbol{\phi} \sim \text{N}\left(\mathbf{0}, \left[\mathbf{Q}(\mathbf{W}, \rho) \otimes \boldsymbol{\Sigma}^{-1}\right]^{-1}\right). \tag{9}$$

Here $\mathbf{Q}(\mathbf{W}, \rho) = \rho[\text{diag}(\mathbf{W1}) - \mathbf{W}] + (1-\rho)\mathbf{I}$ is the precision matrix for the joint distribution corresponding to the CAR prior proposed by Leroux *et al.* (2000) and described above, while $\boldsymbol{\Sigma}_{J \times J}$ is a cross variable covariance matrix. In common with the univariate models, the correlation structure imposed by (9) is more easily seen by its full conditional form, that is:

$$\boldsymbol{\phi}_k | \boldsymbol{\phi}_{-k}, \mathbf{W}, \boldsymbol{\Sigma}, \rho \sim \text{N}\left(\frac{\rho \sum_{i=1}^{K} w_{ki} \boldsymbol{\phi}_i}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho}, \frac{\boldsymbol{\Sigma}}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho}\right)$$

$$\begin{aligned} \boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}(df, \boldsymbol{\Omega}) \\ \rho &\sim \text{Uniform}(0, 1), \end{aligned}$$

where $\boldsymbol{\phi}_{-k}$ denotes the vector of random effects except those relating to the kth areal unit. Here df is the degrees of freedom for the Inverse-Wishart prior for $\boldsymbol{\Sigma}$ and the default value is $df = J + 1$. Similarly, $\boldsymbol{\Omega}$ is the $J \times J$ scale matrix, with the default value being the identity matrix. In common with the univariate model `S.CARleroux()`, the spatial autocorrelation parameter ρ can be fixed to any value in the unit interval using the argument `rho`.

2.3. Two-level spatial data models

The study region $\mathcal{S}$ is again partitioned into K non-overlapping areal units $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$, and data are available on m_k individuals within area k. Thus for areal unit $\mathcal{S}_k$ there are m_k different response variables being modelled, leading to both spatial variation and individual-level variation. The general likelihood model allowed for these data is given by

$$\begin{aligned} Y_{kj}|\mu_{kj} &\sim f(y_{kj}|\mu_{kj}, \nu^2) \quad \text{for } k = 1, \ldots, K, \quad j = 1, \ldots, m_k, \\ g(\mu_{kj}) &= \mathbf{x}_{kj}^\top \boldsymbol{\beta} + O_{kj} + \psi_{kj}, \\ \boldsymbol{\beta} &\sim \text{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \end{aligned} \tag{10}$$

In common with the univariate models $(\mathbf{x}_{kj}^\top, O_{kj})$ are respectively a vector of p covariates and an offset for individual j within area k. For this model the response and each covariate vector is of length $m = \sum_{k=1}^K m_k$. Gaussian priors are again assumed for the regression parameters $\boldsymbol{\beta}$. Binomial, Gaussian and Poisson data likelihood models are allowed, that is:

- **Binomial -** $Y_{kj} \sim \text{Binomial}(n_{kj}, \theta_{kj})$ and $\ln(\theta_{kj}/(1-\theta_{kj})) = \mathbf{x}_{kj}^\top \boldsymbol{\beta} + O_{kj} + \psi_{kj}$.
- **Gaussian -** $Y_{kj} \sim \text{N}(\mu_{kj}, \nu^2)$ and $\mu_{kj} = \mathbf{x}_{kj}^\top \boldsymbol{\beta} + O_{kj} + \psi_{kj}$.

  The scale parameter ν^2 for the Gaussian likelihood is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\nu^2 \sim \text{Inverse-Gamma}(1, 0.01)$.
- **Poisson -** $Y_{kj} \sim \text{Poisson}(\mu_{kj})$ and $\ln(\mu_{kj}) = \mathbf{x}_{kj}^\top \boldsymbol{\beta} + O_{kj} + \psi_{kj}$.

CARBayes can only fit the following model for ψ_{kj}.

`S.CARmultilevel()`
This model can be implemented with binomial, Gaussian and Poisson data likelihoods. The spatial and individual-level variation are modelled by the decomposition:

$$\begin{aligned} \psi_{kj} &= \phi_k + \zeta_{\lambda(k,j)}, \\ \phi_k|\boldsymbol{\phi}_{-k} &\sim \text{N}\left(\frac{\rho \sum_{j=1}^K w_{kj}\phi_j}{\rho \sum_{j=1}^K w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^K w_{kj} + 1 - \rho} \right), \end{aligned} \tag{11}$$

$$\begin{aligned} \zeta_r &\sim \text{N}(0, \sigma^2) \quad \text{for all } r, \\ \tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a, b). \\ \rho &\sim \text{Uniform}(0, 1). \end{aligned}$$

The spatial variation is modelled by $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, which is common to all individuals within each area and is modelled by the CAR prior proposed by Leroux *et al.* (2000). Again ρ can be fixed to any value in the unit interval using the argument `rho`. The ordering of the response and covariate data vectors are not constrained to have all individuals in area 1 followed by all individuals in area 2, etc. Instead, the `S.CARmultilevel()` function requires the `ind.area` argument to be specified, which is a vector of length m. Each element in that vector must be an integer between 1 and K (where K is the number of areas), and denotes which area an individual belongs to as ordered by the $\mathbf{W}$ matrix. For example, if the r*th* element of `ind.area` is 5, then the r*th* element in each response and covariate data vector refers to an individual in area 5, that is the area represented by the 5th row of the neighbourhood matrix $\mathbf{W}$.

The second term $\zeta_{\lambda(k,j)}$ is a random effect allowing for individual-level variation, which is given an independent and identically distributed zero-mean Gaussian prior with a constant variance σ^2. It can incorporate correlation between individuals (if desired) by allowing different individuals to have the same random effect value. For example, if individual r in area $\mathcal{S}_k$ had the same random effect value as individual s in area $\mathcal{S}_t$, then $\lambda(k, r) = \lambda(t, s)$. Operationally, this random effect structure is achieved by specifying the `ind.re` argument in the `S.CARmultilevel()` function as a factor variable, which is the same length as each response and covariate data vector, namely m. Two data points with the same level of this factor variable will have the same random effect value. This individual-level variation term $\zeta_{\lambda(k,j)}$ can be omitted from the model by omitting the `ind.re` argument from the `S.CARmultilevel()` function call.

2.4. Inference

All models in this package are fitted in a Bayesian setting using MCMC simulation, via a combination of Gibbs sampling (when the appropriate full conditional distributions are proportional to standard distributions) and Metropolis type steps. The Metropolis steps for the random effects and the regression parameters use the Metropolis adjusted Langevin algorithm (MALA, Roberts and Rosenthal 1998), although for the random effects there is the option of using simple random walk Metropolis steps by setting `MALA=FALSE` in the function call. Note also that simple random walk Metropolis updates are used in the multinomial models. The overall functions that implement the MCMC algorithms are written in R, while the computationally intensive updating steps are written as computationally efficient C++ routines using the R package **Rcpp** (Eddelbuettel and Francois 2011). Additionally, the sparsity of the neighbourhood matrix $\mathbf{W}$ is utilised via its triplet form when updating the random effects within the algorithms, which increases the computational efficiency of the software. Additionally, matrix identities and Kronecker product forms are used to speed up the computation where possible. Missing values are allowed in the response variable $\mathbf{Y}$ for most models (not the `S.CARlocalised()` model), and are treated as additional parameters to be updated in the MCMC algorithm using a data augmentation approach (Tanner and Wong 1987).

3. Loading and using the software

3.1. Loading the software

CARBayes is an add-on package to the statistical software R, and is freely available to download from the Comprehensive R Archive Network (CRAN, http://cran.r-project.org/) for Windows, Linux and Apple platforms. The package requires R ($\geq$ 3.0.0) and depends on packages **MASS** (Venables and Ripley 2002), and **Rcpp** ($\geq$ 0.11.5). Additionally, it imports functionality from the following other packages: **CARBayesdata** (Lee 2016), **coda** (Plummer *et al.* 2006), **leaflet** (Cheng *et al.* 2018), **matrixcalc** (Novomestky 2012), **MCMCpack** (Martin *et al.* 2011), **spam** (Furrer and Sain 2010), **sp** (Bivand *et al.* 2013), **spdep**, **stats**, **truncnorm** (Trautmann *et al.* 2014) and **utils**. Once installed it can be loaded using the command `library(CARBayes)`.

Note, certain functionality from the packages listed in the previous paragraph are automatically loaded upon loading **CARBayes**, but only for use within the package. However, a complete spatial analysis will typically also include the creation of the neighbourhood matrix **W** from a shapefile, the production of spatial maps of the fitted values and residuals, and tests for the presence of spatial autocorrelation. To achieve these tasks the following packages should be loaded separately into R: **leaflet**, **maptools**, **rgdal** (Bivand *et al.* 2018), **shapefiles**, **sp** and **spdep**.

3.2. Using the software

The software can fit seven models: `S.glm()`, `S.CARbym()`, `S.CARleroux()`, `S.CARdissimilarity()` and `S.CARlocalised()` for univariate spatial data, `MVS.CARleroux()` for multivariate spatial data, and `S.CARmultilevel()` for two-level data relating to individuals within areas. Full details of the arguments required for each model are given in the helpfiles. However, the main common arguments that are required for a baseline analysis (for example using default priors) are as follows.

- `formula` - A formula for the covariate part of the model using the same syntax as the `lm()` function. Offsets can be included here using the `offset()` function.
- `family` - The likelihood model which must be one of `"binomial"`, `"gaussian"`, `"multinomial"`, `"poisson"` or `"zip"`.
- `trials` - This is only needed if `family="binomial"` or `family="multinomial"`, and is a vector the same length as the response containing the total number of trials for each area.
- `W` - A $K \times K$ symmetric and non-negative neighbourhood matrix, whose row sums must all be positive.
- `burnin` - The number of MCMC samples to discard as the burn-in period.
- `n.sample` - The number of MCMC samples to generate.

When a model has run **CARBayes** has the following summary extractor functions:

- `coef()` - returns the estimated (posterior median) regression coefficients.
- `fitted()` - returns the fitted values based on posterior medians.
- `logLik()` - returns the estimated loglikelihood.
- `model.matrix()` - returns the design matrix of covariates.
- `print()` - prints a summary of the fitted model to the screen, including both parameter summaries and convergence diagnostics for the MCMC run.
- `residuals()` - returns either the "response" (raw) or "pearson", residuals from the model (based on posterior means).

Additionally, **CARBayes** has functions `summarise.samples()` and `summarise.lincomb()` to summarise the results, and both functions are illustrated in the examples that follow. The software updates the user on its progress to the R console, which allows the user to monitor the function's progress. However, using the `verbose=FALSE` option will disable this feature. Once run, each model returns a list object with the following components.

- `summary.results` - A summary table of selected parameters that is presented when using the `print()` function. The table includes the posterior median (`Median`) and 95% credible interval (`2.5%, 97.5%`), the number of samples generated (`n.sample`), the acceptance rate for the Markov chain (`% accept`), the effective number of independent samples using the `effectiveSize()` function from the **coda** package (`n.effective`), and the convergence diagnostic proposed by Geweke (1992) and implemented in the **coda** package (`Geweke.diag`). This diagnostic takes the form of a Z-score, so that convergence is suggested by the statistic being within the range (-1.96, 1.96).
- `samples` - A list containing the MCMC samples generated from the model, where each element in the list is a matrix. The names of these matrix objects correspond to the parameters defined in Section 2 of this paper, and each column of a matrix contains the set of samples for a single parameter. This list includes samples from the posterior distribution of the fitted values for each data point (`fitted`). Additionally, if the response variable $\mathbf{Y}$ contains missing values, then samples from its posterior predictive distribution obtained via data augmentation are available (`Y`).
- `fitted.values` - The fitted values based on posterior medians from the model. For the univariate data models this is a vector, while for the multivariate data models this is a matrix.
- `residuals` - For the univariate data models this is a matrix with 2 columns, where each column is a type of residual and each row relates to a single data point. The types are *response* (raw) and *pearson*. For the multivariate data models this is a list with 2 $K \times J$ matrix elements, where each matrix element is a type of residual (*response* or *pearson*).
- `modelfit` - Model fit criteria including the Deviance Information Criterion (DIC, Spiegelhalter *et al.* (2002)) and its corresponding estimated effective number of parameters (p.d), the Watanabe-Akaike Information Criterion (WAIC, Watanabe (2010)) and its

corresponding estimated number of effective parameters (p.w), the Log Marginal Predictive Likelihood (LMPL, Congdon (2005)), and the loglikelihood. The best fitting model is one that minimises the DIC and WAIC but maximises the LMPL. If the response data contains missing data, the DIC is computed based on only the observed data (see Celeux *et al.* (2006)).

- `accept` The acceptance probabilities for the parameters.
- `localised.structure` - This element is `NULL` except for the models `S.CARdissimilarity()` and `S.CARlocalised()`. For `S.CARdissimilarity` it is a list containing two matrices, `W.posterior` and `W.border.prob`. `W.posterior` contains posterior medians for each element w_{kj} of the $K \times K$ neighbourhood matrix $\mathbf{W}$, while `W.border.prob` contains posterior probabilities that each w_{kj} element equals zero, which corresponds to the posterior probability of a boundary in the random effects surface. The latter is only present if `W.binary=TRUE`, otherwise it is missing (`NA`). In all cases elements of $\mathbf{W}$ that correspond to non-neighbouring areas as determined by the original $\mathbf{W}$ matrix have `NA` values. For `S.CARlocalised()` this element is a vector of length K, and gives the posterior median class (Z_k value) that each data point is assigned to.
- `formula` - The formula (as a text string) for the response, covariate and offset part of the model.
- `model`- A text string describing the model that has been fitted.
- `X` - The design matrix of covariates inherited from the `formula` argument.

The remainder of this vignette illustrates the **CARBayes** software via 3 worked examples.

4. Example 1 - Scottish lip cancer data

The first example is the famous Scottish lip cancer data set, which is included purely to illustrate how to combine a data frame and shapefile together into a `SpatialPolygonsDataFrame` object. The creation of this object allows spatial maps to be produced of variables of interest, as well as allowing the neighbourhood matrix $\mathbf{W}$ to be created for use in the models implemented in **CARBayes**. The Scottish lip cancer data are contained in the **CARBayesdata** package and can be loaded into R using the following commands:

```
R> library(CARBayesdata)
R> library(shapefiles)
R> library(sp)
R> data(lipdata)
R> data(lipdbf)
R> data(lipshp)
```

To create a `SpatialPolygonsDataFrame` object you essentially need two types of data. The first is a `data.frame` containing the data you wish to model. If this is a comma separated variable (csv) file then it can be read into R using the command `read.csv()`. The second

data type is a shapefile, which comprises many separate components with different file extensions. Here we need two of these components: `shapefile.shp` containing the polygons, and `shapefile.dbf` containing a unique identifier linking each row in the `data.frame` to a polygon in the `shapefile.shp` file. The shapefiles can be read in to R using the `read.shp()` and `read.dbf()` functions. In the above example the code `data(lipdata)` loads the `data.frame` object, `data(lipdbf)` loads the *.dbf* component of the shapefile, while `data(lipshp)` loads the *.shp* component of the shapefile. These three data sets can be combined together to create a `SpatialPolygonsDataFrame` object using the `combine.data.shapefile()` function.

```
R> library(CARBayes)
R> lipdbf$dbf <- lipdbf$dbf[ ,c(2,1)]
R> data.combined <- combine.data.shapefile(data=lipdata, shp=lipshp, dbf=lipdbf)
```

For this function to work the row-names of the data-frame (`lipdata`) must be contained in the first column of the *.dbf* (`lipdbf$dbf`) object, which is the reason for re-ordering the columns in the second line of the above code. The `data.combined` object is a `SpatialPolygonsDataFrame` object, which is what is created in **R** if shapefiles are read in using the **rgdal** package (Bivand *et al.* 2018).

5. Example 2 - property prices in Greater Glasgow

The **CARBayes** software is illustrated by modelling the spatial pattern in average property prices across Greater Glasgow, Scotland, in 2008. This is an ecological regression analysis, whose aim is to identify the factors that affect property prices and quantify their effects.

5.1. Data and exploratory analysis

The data come from the Scottish Statistics database (*http://statistics.gov.scot*), but are also included in the **CARBayesdata** R package. The study region is the Greater Glasgow and Clyde health board (GGHB), which is split into 271 Intermediate Geographies (IG). These IGs are also known as Intermediate zones (IG), but hereafter we refer to them as Intermediate Geographies. These IGs are small areas that have a median area of 124 hectares and a median population of 4,239. These data can be loaded into R using the code shown below:

```
R> library(CARBayesdata)
R> library(sp)
R> data(GGHB.IG)
R> data(pricedata)
```

The `GGHB.IG` object is a `SpatialPolygonsDataFrame` containing the spatial information for the GGHB, which is used to plot the data, construct the neighbourhood matrix **W**, and conduct a test for spatial autocorrelation. The `pricedata` object is a `data.frame` containing the property price data for 270 of the 271 IGs in GGHB, because one area had outlying values and was hence removed. These two data sets can be merged using the code below

```
R> propertydata.spatial <- merge(x=GGHB.IG, y=pricedata, by="IG", all.x=FALSE)
```

Variable	**Percentiles**				
	0%	25%	50%	75%	100%
Property price (in thousands)	50.0	95.0	121.8	159.2	372.8
Crime rate (per 10,000)	85.0	303.2	517.0	728.0	1994.0
Number of rooms (median)	3	3	4	4	6
Property sales	4	46	58	85	266
Drive time to a shop (minutes)	0.3	0.8	1.2	1.9	8.5

Table 1: Summary of the distribution of the data.

The variables in `pricedata` are summarised in Table 1, which displays the percentiles of their distribution (with the exception of the categorical variable `type`). The response variable in this study is the median price (in thousands, `price`) of all properties sold in 2008 in each IG. The table shows large variation in this variable, with average prices ranging between £50,000 and £372,800 across the study region. The first covariate in this study is the crime rate (`crime`) in each IG, because areas with higher crime rates are likely to be less desirable to live in. Crime rate is measured as the total number of recorded crimes in each IG per 10,000 people that live there, and the values range between 85 and 1994. Other covariates included in this study are the median number of rooms in a property (`rooms`), the number of properties that sold in a year (`sales`), and the average time taken to drive to the nearest shopping centre (`driveshop`). The latter is a proxy measure of access to services which may affect property prices. Finally, a categorical variable measuring the most prevalent property type in each area is available (`type`), with levels; 'flat' (68% of areas), 'terraced' (7%), 'semi-detached' (13%) and 'detached' (12%).

A spatial map of the `price` response variable can be overlaid on a OpenStreetMap using the functionality of the **leaflet** package. However, first the `propertydata.spatial` object needs to have its coordinate reference system changed to longitude and latitude as this is what the **leaflet** package requires, which can be done using the following R code.

```
R> library(rgdal)
R> propertydata.spatial <- spTransform(propertydata.spatial,
+                                      CRS("+proj=longlat +datum=WGS84 +no_defs"))
```

Then a map of `price` can be drawn using the following code.

```
R> library(leaflet)
R> colours <- colorNumeric(palette = "BuPu", domain = propertydata.spatial@data$price)
R> map1 <- leaflet(data=propertydata.spatial) %>%
+     addTiles() %>%
+     addPolygons(fillColor = ~colours(price), color="red", weight=1,
+                 fillOpacity = 0.7) %>%
+     addLegend(pal = colours, values = propertydata.spatial@data$price, opacity = 1,
+                 title="Price") %>%
+     addScaleBar(position="bottomleft")
R> map1
```

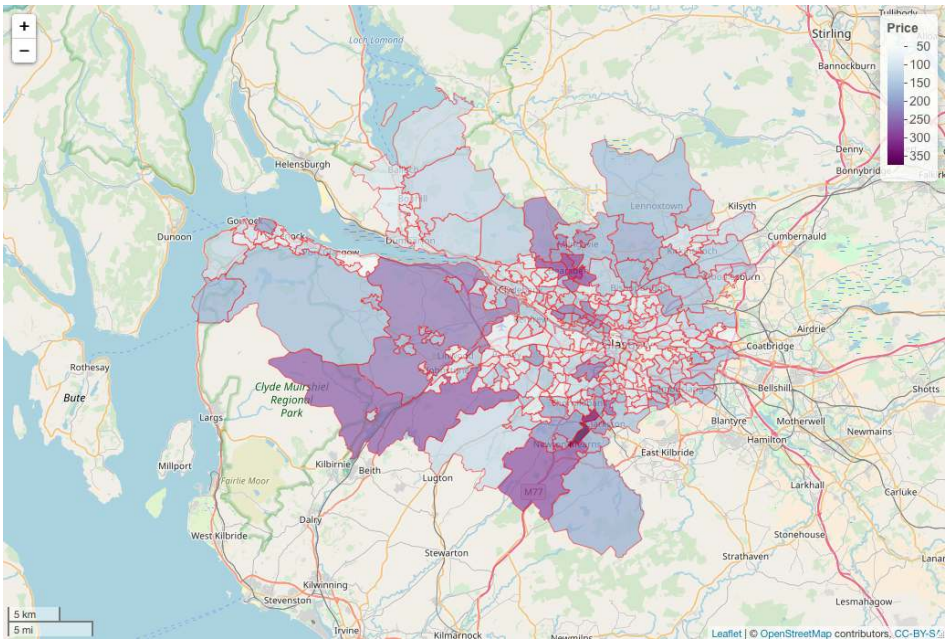

Figure 1: Map showing the average property price in each IG (in thousands).

The map is shown in Figure 1 and suggests that Glasgow has a number of property submarkets, whose prices are not related to those in neighbouring areas. An example of this is the two groups of higher priced regions north of the river Clyde, which are the highly sought after Westerton / Bearsden (northerly cluster) and Dowanhill / Hyndland (central cluster) districts.

5.2. Non-spatial modelling

The natural log of the median property price is treated as the response and assumed to be Gaussian, and an initial covariate only model is built in a frequentist framework using linear models. Initial plots of the data using the `pairs()` command suggest that the natural log of drive time to a shopping centre (`driveshop`) is linearly related to the response, and that crime rate (`crime`) has a non-linear relationship to the response. The natural log transformations of `price` and `driveshop` are created using the following commands.

```
R> propertydata.spatial@data$logprice <- log(propertydata.spatial@data$price)
R> propertydata.spatial@data$logdriveshop <- log(propertydata.spatial@data$driveshop)
```

A model with all the covariates is fitted to the data, where the crime rate variable is modelled as non-linear using a natural cubic spline with 3 degrees of freedom.

```
R> library(splines)
R> form <- logprice~ns(crime,3)+rooms+sales+factor(type) + logdriveshop
R> model <- lm(formula=form, data=propertydata.spatial@data)
```

From fitting this model all of the numeric covariates are significantly related to the response at the 5% level, suggesting they all play an important role in explaining the spatial pattern in median property price. The predominant property type variable also appears to be important, with areas where the level is 'detached' (the baseline level) having significantly higher property prices than the other three levels.

To quantify the presence of spatial autocorrelation in the residuals from this model we can compute Moran's I statistic (Moran 1950) and conduct a permutation test. The permutation test has the null hypothesis of no spatial autocorrelation and an alternative hypothesis of positive spatial autocorrelation, and is conducted using the `moran.mc()` function from the **spdep** package. The test can be implemented using the code below. Lines 2 and 3 turn `propertydata.spatial` into a neighbourhood (`nb`) object and then into a `listw` object, which is required by the `moran.mc()` function.

```
R> library(spdep)
R> W.nb <- poly2nb(propertydata.spatial, row.names = rownames(propertydata.spatial@data))
R> W.list <- nb2listw(W.nb, style="B")
R> resid.model <- residuals(model)
R> moran.mc(x=resid.model, listw=W.list, nsim=1000)

        Monte-Carlo simulation of Moran I

data:  resid.model
weights: W.list
number of simulations + 1: 1001

statistic = 0.2733, observed rank = 1001, p-value = 0.000999
alternative hypothesis: greater
```

The Moran's I test has a p-value much less than 0.05, which suggests that the residuals contain substantial positive spatial autocorrelation.

5.3. Spatial modelling with CARBayes

The residual spatial autocorrelation can be accounted for by adding a set of random effects to the model, and we apply model (1) and (3) to the data as shown below. Note, line 2 creates the binary neighbourhood matrix `W` from the `W.nb` object.

```
R> library(CARBayes)
R> W <- nb2mat(W.nb, style="B")
R> model.spatial <- S.CARleroux(formula=form, data=propertydata.spatial@data,
+   family="gaussian", W=W, burnin=100000, n.sample=300000, thin=20)
```

In the above code the covariate and offset component defined by `formula` is the same as for the simple linear model fitted earlier, and the neighbourhood matrix `W` is binary and defined by whether or not two areas share a common border. Inference for this model is based on 10,000 MCMC samples, which were obtained following a burn-in period of 100,000 and thinning the remaining 200,000 samples by 20 to reduced their autocorrelation. A summary of the fitted model can be obtained using the following code.

```
R> print(model.spatial)

#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - logprice ~ ns(crime, 3) + rooms + sales + factor(type) + logdrivesho
Number of missing observations - 0

#############
#### Results
#############
Posterior quantities and DIC

                     Median    2.5%   97.5% n.sample % accept n.effective Geweke.diag
(Intercept)          4.2423  3.9625  4.5213    10000    100.0     10000.0        -0.4
ns(crime, 3)1       -0.2475 -0.4000 -0.0945    10000    100.0      9398.8        -0.8
ns(crime, 3)2       -0.4081 -0.7055 -0.1105    10000    100.0      9268.4         0.8
ns(crime, 3)3       -0.2021 -0.4052  0.0045    10000    100.0     10000.0         0.4
rooms                0.2205  0.1682  0.2714    10000    100.0     10000.0        -0.3
sales                0.0022  0.0016  0.0029    10000    100.0     10000.0         0.5
factor(type)flat    -0.2474 -0.3660 -0.1279    10000    100.0      9483.0         0.6
factor(type)semi    -0.1616 -0.2606 -0.0598    10000    100.0     10516.5        -0.5
factor(type)terrace -0.2918 -0.4206 -0.1641    10000    100.0     10000.0         0.1
logdriveshop        -0.0055 -0.0621  0.0521    10000    100.0      8321.1         0.5
nu2                  0.0246  0.0146  0.0340    10000    100.0      4364.2         0.3
tau2                 0.0423  0.0196  0.0831    10000    100.0      3626.0        -0.1
rho                  0.9344  0.7499  0.9922    10000     45.5      6323.1         0.2

DIC =  -145.598       p.d =  93.88278       LMPL =  56.89
```

The Summary is presented in two parts, the first of which describes the model that has been fit. The second summarises the results, and includes the posterior median (`Median`) and 95% credible intervals (`2.5%, 97.5%`) for selected parameters (not the random effects), the convergence diagnostic proposed by Geweke (1992) (`Geweke.diag`) as a Z-score, and the effective number of independent samples (`n.effective`). Also displayed are the actual number of samples kept from the MCMC run (`n.sample`), as well as the acceptance rate for each parameter (`% accept`). Note, parameters that have an acceptance rate of 100% have been Gibbs sampled. Finally, the DIC and LMPL model fit criteria are displayed. In addition to producing the summary above, the model returns a list object with the following components:

```
R> summary(model.spatial)

Length Class   Mode
summary.results       91   -none-  numeric
```

```
samples                7   -none-  list
fitted.values        270   -none-  numeric
residuals              3   data.frame list
modelfit               7   -none-  numeric
accept                 5   -none-  numeric
localised.structure    0   -none-  NULL
formula                3   formula call
model                  2   -none-  character
X                   2700   -none-  numeric
```

The first element is the summary results table used by the `print()` function. The next element is a list containing matrices of the thinned and post burn-in MCMC samples for each set of parameters. For example, `model.spatial$samples$beta` is a matrix containing the MCMC samples for all the regression parameters. The next two elements in the list `fitted.values` and `residuals` are vectors of fitted values and residuals from the model, while `modelfit` gives a selection of model fit criteria. These criteria include the Deviance Information Criterion (DIC), the log Marginal Predictive Likelihood (LMPL), the Watanabe-Akaike Information Criterion (WAIC), and the log likelihood. For further details about Bayesian modelling and model fit criteria see Gelman *et al.* (2003). The item `accept` contains the acceptance rates for the model, while `localised.structure` is `NULL` for this model and is used for compatibility with the other functions in the package. Finally, the `formula` and `model` elements are text strings describing the formula used and the model fit, while `X` gives the design matrix corresponding to the `formula` object.

5.4. Inference

The summary table above gives posterior medians and 95% credible intervals for a selection of model parameters, but these can be recreated (or similar summarise created for other parameters) using the function

```
R> summarise.samples(model.spatial$samples$beta, quantiles=c(0.5, 0.025, 0.975))

$quantiles
0.5        0.025       0.975
[1,]  4.241896323  3.962957346  4.51787943
[2,] -0.245877968 -0.396698998 -0.09679782
[3,] -0.401030811 -0.704894560 -0.10461861
[4,] -0.200704600 -0.407327946  0.01018534
[5,]  0.219841199  0.169342794  0.27108763
[6,]  0.002237906  0.001611683  0.00287336
[7,] -0.248767056 -0.368024984 -0.12990882
[8,] -0.162194165 -0.263154252 -0.06109645
[9,] -0.294254065 -0.422182354 -0.16841391
[10,] -0.005002887 -0.061142928  0.04995092

$exceedences
NULL
```

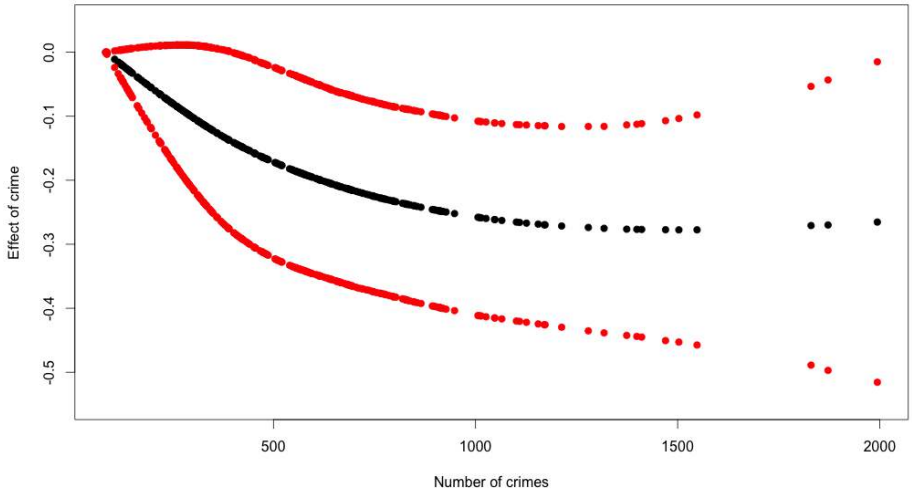

Figure 2: Plot showing the estimated non-linear relationship between crime rate and log-price.

which here has summarised posterior medians and 95% credible intervals for the covariate effects. However, for the crime variable its relationship is non-linear and summarised by the results for all 3 basis functions `ns(crime, 3)1`, `ns(crime, 3)2`, `ns(crime, 3)3`. Therefore to summarise the entire non-linear relationship we can use the `summarise.lincomb()` function, which allows us to compute the posterior distribution and quantiles of a linear combination of the covariates. This can be achieved and then plotted using the code:

```
R> crime.effect <- summarise.lincomb(model=model.spatial, columns=c(2,3,4),
+    quantiles=c(0.5, 0.025, 0.975), distribution=FALSE)
R> plot(propertydata.spatial@data$crime, crime.effect$quantiles[ ,1], pch=19,
+    ylim=c(-0.55,0.05), xlab="Number of crimes", ylab="Effect of crime")
R> points(propertydata.spatial@data$crime, crime.effect$quantiles[ ,2], pch=19,
+   col="red")
R> points(propertydata.spatial@data$crime, crime.effect$quantiles[ ,3], pch=19,
+   col="red")
```

The first line creates the crime effect, while the remaining lines plot the graph which is shown in Figure 2 and shows a decreasing relationship as crime rate increases as expected.

6. Example 3 - identifying high-risk disease clusters

The third example illustrates the utility of the localised spatial autocorrelation model proposed by Lee and Mitchell (2012), which can identify boundaries that represent step changes in the (random effects) response surface between geographically adjacent areal units. The aim of this analysis is to identify boundaries in the risk surface of respiratory disease in Greater Glasgow, Scotland, in 2010, so that the spatial extent of high-risk clusters can be identified.

The identification of boundaries in spatial data is affectionately known as *Wombling*, after the seminal paper by Womble (1951).

6.1. Data and exploratory analysis

The data again relate to the Greater Glasgow and Clyde health board, and are also freely available to download from Scottish Statistics (*http://statistics.gov.scot*). However, the river Clyde partitions the study region into a northern and a southern sub-region, and no areal units on opposite banks of the river border each other. This means that boundaries could not be identified across the river, and therefore here we only consider those areal units that are on the northern side of the study region. This leaves 134 areal units in the new smaller study region, and data on respiratory disease for this region are included with the **CARBayesdata** package and can be loaded with the command:

```
R> library(CARBayesdata)
R> library(sp)
R> data(GGHB.IG)
R> data(respiratorydata)
```

The `GGHB.IG` `spatialPolygonsDataFrame` object can then be subset to just include IGs in the `respiratorydata` data.frame using the following code.

```
R> respiratorydata.spatial <- merge(x=GGHB.IG, y=respiratorydata, by="IG", all.x=FALSE)
```

The first 6 rows of the data can be viewed using the code below.

```
R> head(respiratorydata.spatial@data)
```

```
         IG                          name  easting northing observed  expected
1 S02000260                   Auchinairn 261624.5 669657.4      107 106.45661
2 S02000261                Woodhill East 262927.1 670027.8       23  50.97354
3 S02000262                Woodhill West 262142.9 670428.0       53 104.49236
4 S02000263               Westerton East 254570.5 670593.8       40  90.35747
5 S02000264 Bishopbriggs West and Cadder 261248.4 670928.0       60 140.16546
6 S02000265               Westerton West 253764.4 670982.6       25  63.93549
  incomedep       SMR
1        22 1.0051044
2         7 0.4512145
3         6 0.5072141
4         5 0.4426861
5         7 0.4280655
6         6 0.3910191
```

In addition to the unique identifier IG codes (`IG`), the name of each IG (`name`), and the geographical coordinates of each area's centroid (`easting, northing`), the data contain 4 variables. `observed` is the number of hospital admissions in 2010 in each IG due to respiratory

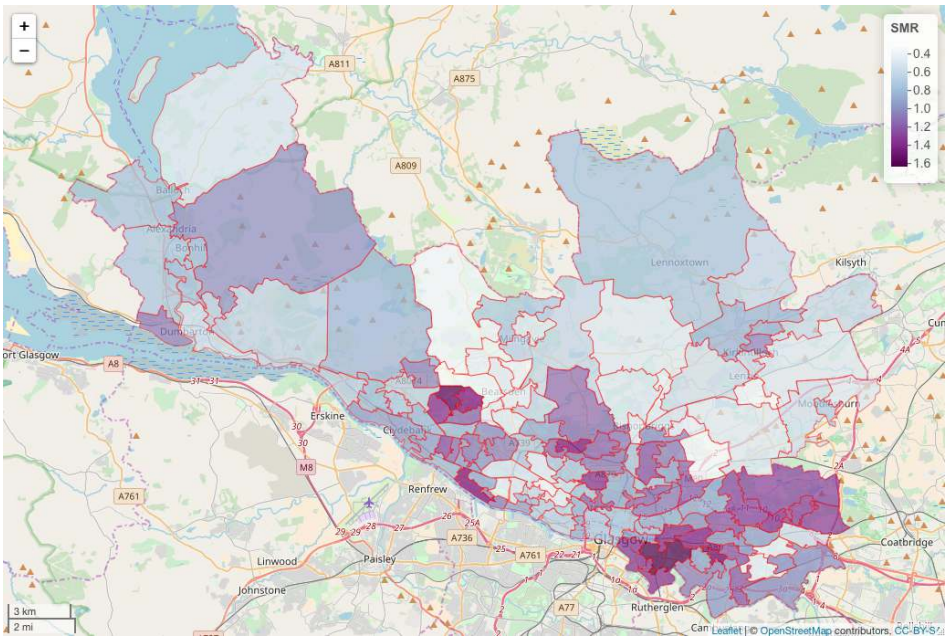

Figure 3: Map displaying the SMR for each area.

disease (International Classification of Disease tenth revision codes J00-J99). These observed numbers will depend on the size and demographic structure of the populations living in each IG, and these factors need to be adjusted for before estimating disease risk. This is typically achieved by computing the expected numbers of hospital admissions in each IG based on this demographic information, using indirect standardisation. These expected numbers are stored in the `expected` column, and the simplest measure of disease risk is the Standardised Morbidity Ratio (SMR), which is the ratio of the observed to the expected numbers of hospital admissions and is stored in the `SMR` column. Finally, the percentage of people defined to be income deprived (in receipt of means tested benefits) is stored in the `incomedep` column. A map showing the SMR is displayed in Figure 3, and is created using the code below, where in common with the previous example the coordinate reference system is changed to longitude and latitude.

```
R> respiratorydata.spatial <- spTransform(respiratorydata.spatial,
+                                         CRS("+proj=longlat +datum=WGS84 +no_defs"))
R> library(leaflet)
R> colours <- colorNumeric(palette = "BuPu", domain = respiratorydata.spatial@data$SMR)
R> map2 <- leaflet(data=respiratorydata.spatial) %>%
+     addTiles() %>%
+     addPolygons(fillColor = ~colours(SMR), color="red", weight=1,
+                 fillOpacity = 0.7) %>%
+     addLegend(pal = colours, values = respiratorydata.spatial@data$SMR, opacity = 1,
+                 title="SMR") %>%
+     addScaleBar(position="bottomleft")
R> map2
```

Values of the SMR above one relate to areas exhibiting above average risks, while values below one correspond to below average risks. For example, an SMR of 1.2 corresponds to a 20% increased risk relative to the expected numbers of respiratory disease cases. The figure shows evidence of localised spatial structure, with numerous different locations where high and low risk areas border each other. This in turn suggests that boundaries are likely to be present, and their identification is the goal of this analysis. The method proposed by Lee and Mitchell (2012) identifies these boundaries using dissimilarity metrics, which are non-negative measures of the dissimilarity between all pairs of adjacent areas. In this example we use the absolute difference in the percentage of people in each IG who are defined to be income deprived (`incomedep`), because it is well known that socio-economic deprivation plays a large role in determining people's health. However, before fitting the model the spatial neighbourhood matrix $\mathbf{W}$ based on sharing a common border is computed using the following code.

```
R> W.nb <- poly2nb(respiratorydata.spatial, row.names =
+                  rownames(respiratorydata.spatial@data))
R> W <- nb2mat(W.nb, style="B")
```

6.2. Spatial modelling with CARBayes

Let the observed and expected numbers of hospital admissions be denoted by $\mathbf{Y} = (Y_1, \ldots, Y_K)$ and $\mathbf{E} = (E_1, \ldots, E_K)$ respectively. Then as the observed numbers of hospital admissions are counts, a Poisson likelihood model given by $Y_k \sim \text{Poisson}(E_k R_k)$ is appropriate, where R_k represents disease risk in areal unit $\mathcal{S}_k$. A log-linear model is specified for R_k, that is, $\ln(R_k) = \beta_0 + \phi_k$, and for a general review of disease mapping see Wakefield (2007). We note that in fitting this model in **CARBayes**, the offset is specified on the linear predictor scale rather than the expected value scale, so in this analysis the offset is $\log(\mathbf{E})$ rather than $\mathbf{E}$. The dissimilarity metric used here is the absolute difference in the level of income deprivation between each pair of areal units, which can be created from the vector of area level income deprivation scores using the following code.

```
R> income <- respiratorydata.spatial@data$incomedep
R> Z.incomedep <- as.matrix(dist(income, diag=TRUE, upper=TRUE))
```

The first line pulls out the income deprivation covariate while the second line computes the $K \times K$ matrix of absolute differences in income deprivation levels between each pair of areal units, that is $\texttt{Z.incomedep}_{kj} = |\texttt{income}_k - \texttt{income}_j|$. The function to implement the localised CAR model is called `S.CARdissimilarity()`, and it takes the same arguments as the other CAR models except that it additionally requires the dissimilarity metrics. These are required in the form of a list of $K \times K$ matrices, and for this example we only have a single dissimilarity metric. Additionally, we add the argument `W.binary=TRUE`, so that the estimated elements in $\mathbf{W}$ are ones or zeros (corresponding to (5)), the latter corresponding to boundaries. The model is run using the following code.

```
R> formula <- observed ~ offset(log(expected))
R> model.dissimilarity <- S.CARdissimilarity(formula=formula,
```

```
+    data=respiratorydata.spatial@data, family="poisson", W=W,
+    Z=list(Z.incomedep=Z.incomedep), W.binary=TRUE, burnin=100000,
+    n.sample=300000, thin=20)
R> print(model.dissimilarity)

#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Binary dissimilarity CAR
Dissimilarity metrics -  Z.incomedep
Regression equation - observed ~ offset(log(expected))
Number of missing observations - 0

#############
#### Results
#############
Posterior quantities and DIC

             Median    2.5%   97.5% n.sample % accept n.effective Geweke.diag alpha.min
(Intercept) -0.2197 -0.2414 -0.1987    10000     35.2     10000.0        -0.5        NA
tau2         0.1373  0.0980  0.1927    10000    100.0      9089.8        -0.5        NA
Z.incomedep  0.0500  0.0464  0.0513    10000     45.4     10000.0         0.3    0.0139

DIC =  1070.334       p.d =  105.4371       LMPL =  -615.38

The number of stepchanges identified in the random effect surface
     no stepchange stepchange
[1,]           261         99
```

The first line of the above code specifies the formula with an offset (the natural log of the expected numbers of cases) but no covariates, the latter being required so that boundaries identified in the random effects surface can also be interpreted as boundaries in the risk surface (that is boundaries in $\mathbf{R} = (R_1, \ldots, R_n)$). The above model bases inference on 10,000 post burn-in and thinned MCMC samples. When the model has been fit the `print()` function produces the summary output above, which is similar to that produced for the property price data in the previous example. The main difference between this and the corresponding output from the property price analysis is the addition of a column in the parameter summary table headed `alpha.min`. This column only applies to the dissimilarity metrics, which is why it is `NA` for the remaining parameters. The value of `alpha.min` is the threshold value for the regression parameter α, below which the dissimilarity metric has no effect in identifying boundaries in the response (random effects) surface. A brief description is given in Section 2.1, while full details are given in Lee and Mitchell (2012). For these data the posterior median and 95% credible interval lie completely above this threshold, suggesting that the income deprivation dissimilarity metric has identified a number of boundaries.

The number and locations of these boundaries are summarised in the element of the output list called `model.dissimilarity$localised.structure$W.posterior`, which is a

$K \times K$ symmetric matrix containing the posterior median for the set $\{w_{kj}|k \sim j\}$. Values equal to zero represent a boundary, values equal to one correspond to no boundary, while `NA` values correspond to non-adjacent areas. The locations of these boundaries can be overlaid on a map of the estimated disease risk (that is the posterior median of $\mathbf{R}$). This is done in two steps, the first being the creation of a `SpatialPoints` object using the following code.

```
R> border.locations <- model.dissimilarity$localised.structure$W.posterior
R> respiratorydata.spatial@data$risk <- model.dissimilarity$fitted.values /
+   respiratorydata.spatial@data$expected
R> boundary.final <- highlight.borders(border.locations=border.locations,
+   spdata=respiratorydata.spatial)
```

The first line saves the matrix of border locations, while the second adds the estimated risk values to the `respiratorydata.spatial` object. The next line identifies the boundary points (using the **CARBayes** function `highlight.borders()`) and formats them to enable plotting. Then plotting can be done using the code below, and the result is presented in Figure 4.

```
colours <- colorNumeric(palette = "BuPu", domain = respiratorydata.spatial@data$risk)
map3 <- leaflet(data=respiratorydata.spatial) %>%
    addTiles() %>%
    addPolygons(fillColor = ~colours(risk), color="red", weight=1,
                fillOpacity = 0.7) %>%
    addLegend(pal = colours, values = respiratorydata.spatial@data$risk, opacity = 1,
                title="Risk") %>%
    addCircles(lng = ~boundary.final$X, lat = ~boundary.final$Y, weight = 1,
                radius = 2) %>%
    addScaleBar(position="bottomleft")
map3
```

The figure shows the fitted risk surface and the locations of the boundaries (denoted by blue circles). The model has identified 99 boundaries in the risk surface. The majority of these visually seem to correspond to sizeable changes in the risk surface, suggesting that the model has the power to distinguish between boundaries and non-boundaries. The notable boundaries are the demarcation between the low risk (shaded green) city centre / west end of Glasgow in the middle of the region and the deprived neighbouring areas on both sides, which include Easterhouse / Parkhead in the east and Knightswood / Drumchapel in the west. The other interesting feature of this map is that the boundaries are not closed, suggesting that the spatial pattern in risk is more complex than being partitioned into groups of non-overlapping areas of similar risk.

7. Discussion

This vignette has illustrated the R package **CARBayes**, which can fit a number of commonly used conditional autoregressive models to spatial areal unit data, as well as the localised spatial smoothing models proposed by Lee and Mitchell (2012) and Lee and Sarran (2015). The response data can be binomial, Gaussian, multinomial, Poisson or ZIP, with link functions

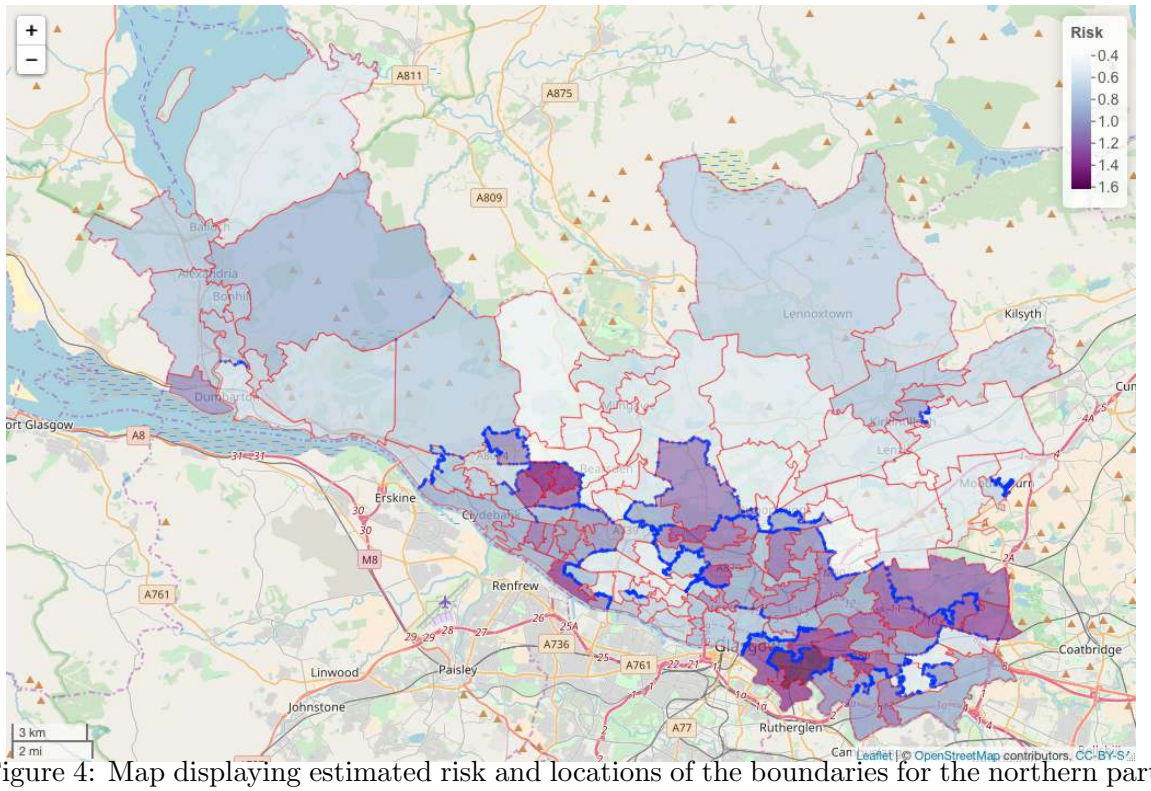

Figure 4: Map displaying estimated risk and locations of the boundaries for the northern part of Greater Glasgow.

logit, identity, logit, natural log and (natural log / logit) respectively. The availability of areal unit data has grown dramatically in recent times, due to the launch of freely available on-line databases. This increased availability of spatial data has fuelled a growth of modelling in this area, leading to the need for user friendly software such as **CARBayes** for use by both statisticians and non-statisticians alike. Finally, this software now has a sister spatio-temporal modelling package called **CARBayesST**, which can fit a range of spatio-temporal areal unit models based on CAR priors. These models include similar models to those proposed by Bernardinelli *et al.* (1995) and Knorr-Held (2000), as well as other alternatives.

Acknowledgements

The data and shapefiles used in sections 5 and 6 of this vignette were provided by the Scottish Government.

References

Belitz, C and Brezger, A and Kneib, T and Lang, S (2009). *BayesX - Software for Bayesian Inference in Structured Additive Regression Models.*

Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M (1995). "Bayesian Analysis of Space-Time Variation in Disease Risk." *Statistics in Medicine*, **14**, 2433–2443.

Besag J, Higdon D (1999). "Bayesian Analysis of Agricultural Field Experiments." *Journal of the Royal Statistical Society Series B*, **61**, 691–746.

Besag J, York J, Mollié A (1991). "Bayesian Image Restoration with Two Applications in Spatial Statistics." *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59.

Bivand R (2013). "spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. R package version 0.5-56, http://CRAN.R-project.org/package=spdep."

Bivand R, Keitt T, Rowlingson B (2018). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.3-4, URL https://CRAN.R-project.org/package=rgdal.

Bivand R, Pebesma E, Gomez-Rubio V (2013). *Applied spatial data analysis with R*. 2nd edition. Springer, NY.

Brewer M, Nolan A (2007). "Variable Smoothing in Bayesian Intrinsic Autoregressions." *Environmetrics*, **18**, 841–857.

Celeux G, Forbes F, Robert C, Titterington D (2006). "Deviance Information Criteria for Missing Data Models." *Bayesian Analysis*, **1**, 651–674.

Cheng J, Karambelkar B, Xie Y (2018). *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.0.1, URL https://CRAN.R-project.org/package=leaflet.

Congdon P (2005). *Bayesian models for categorical data*. 1st edition. John Wiley and Sons.

Eddelbuettel D, Francois R (2011). "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**, 8.

Furrer R, Sain SR (2010). "spam: A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields." *Journal of Statistical Software*, **36**(10), 1–25. URL http://www.jstatsoft.org/v36/i10/.

Gavin J, Jennison C (1997). "A subpixel Image Restoration Algorithm." *Journal of Computational and Graphical and Statistics*, **6**, 182–201.

Gelfand A, Vounatsou P (2003). "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, **4**, 11–25.

Gelman A, Carlin J, Stern H, Rubin D (2003). *Bayesian Data Analysis*. 2nd edition. Chapman and Hall/CRC, London.

Geweke J (1992). "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In *IN BAYESIAN STATISTICS*, pp. 169–193. University Press.

Green P, Richardson S (2002). "Hidden Markov Models and Disease Mapping." *Journal of the American Statistical Association*, **97**, 1055–1070.

Kavanagh L, Lee D, Pryce G (2016). "Is Poverty Decentralising? Quantifying Uncertainty in the Decentralisation of Urban Poverty." *Annals of the American Association of Geographers*, **106**, 1286–1298.

Knorr-Held L (2000). "Bayesian modelling of Inseparable Space-Time Variation in Disease Risk." *Statistics in Medicine*, **19**, 2555–2567.

Lawson A, Clark A (2002). "Spatial Mixture Relative Risk Models Applied to Disease Mapping." *Statistics in Medicine*, **21**, 359–370.

Lee D (2011). "A Comparison of Conditional Autoregressive Models Used in Bayesian Disease Mapping." *Spatial and Spatio-temporal Epidemiology*, **2**, 79–89.

Lee D (2013). "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software*, **55**, 13.

Lee D (2016). *CARBayesdata: Data Sets Used in the Vignette Accompanying the CARBayes Package*. R package version 2.0, URL https://CRAN.R-project.org/package=CARBayesdata.

Lee D, Ferguson C, Mitchell R (2009). "Air Pollution and Health in Scotland: A Multicity Study." *Biostatistics*, **10**, 409–423.

Lee D, Mitchell R (2012). "Boundary Detection in Disease Mapping Studies." *Biostatistics*, **13**, 415–426.

Lee D, Rushworth A, Napier G (2018). "Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package." *Journal of Statistical Software, Articles*, **84**(9), 1–39. doi:10.18637/jss.v084.i09.

Lee D, Rushworth A, Sahu S (2014). "A Bayesian Localized Conditional Autoregressive Model for Estimating the Health Effects of Air Pollution." *Biometrics*, **70**, 419–429.

Lee D, Sarran C (2015). "Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies." *Environmetrics*, **26**, 477–487.

Leroux B, Lei X, Breslow N (2000). "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In M Halloran, D Berry (eds.), *Statistical Models in Epidemiology, the Environment and Clinical Trials*, pp. 179–191. Springer-Verlag, New York.

Lu H, Reilly C, Banerjee S, Carlin B (2007). "Bayesian Areal Wombling Via Adjacency Modelling." *Environmental and Ecological Statistics*, **14**, 433–452.

Lunn D, Spiegelhalter D, Thomas A, Best N (2009). "The BUGS Project: Evolution, Critique and Future Directions ." *Statistics in Medicine*, **28**, 3049–3082.

Ma H, Carlin B (2007). "Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis." *Bayesian Analysis*, **2**, 281–302.

Martin AD, Quinn KM, Park JH (2011). "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software*, **42**(9), 22. URL http://www.jstatsoft.org/v42/i09/.

Moran P (1950). "Notes on continuous stochastic phenomena." *Biometrika*, **37**, 17–23, DOI:10.1093/biomet/37.1–2.17.

Novomestky F (2012). *matrixcalc: Collection of functions for matrix calculations*. R package version 1.0-3, URL https://CRAN.R-project.org/package=matrixcalc.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL http://CRAN.R-project.org/doc/Rnews/.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Roberts G, Rosenthal J (1998). "Optimal scaling of discrete approximations to the Langevin diffusions." *Journal of the Royal Statistical Society Series B*, **60**, 255–268.

Rue H, Martino S, Chopin N (2009). "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)." *Journal of the Royal Statistical Society Series B*, **71**, 319–392.

Schabenberger H (2009). *spatcounts: Spatial count regression*. R package version 1.1, URL http://CRAN.R-project.org/package=spatcounts.

Spiegelhalter D, Best N, Carlin B, Van der Linde A (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society series B*, **64**, 583–639.

Stern H, Cressie N (1999). *Disease Mapping and Risk Assessment for Public Health. Lawson, A and Biggeri, D and Boehning, E and Lesaffre, E and Viel, J and Bertollini, R (eds)*, chapter Inference for Extremes in Disease Mapping. Wiley.

Tanner M, Wong W (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, **82**, 528–540.

Trautmann H, Steuer D, Mersmann O, Bornkamp B (2014). *truncnorm: Truncated normal distribution*. R package version 1.0-7, URL https://CRAN.R-project.org/package=truncnorm.

Ugarte M, Ibanez B, Militino A (2004). "Testing for Poisson Zero Inflation in Disese Mapping." *Biometrical Journal*, **46**, 526–539.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

Vieilledent G, Merow C, GuÃl'lat J, Latimer AM, KÃl'ry M, Gelfand AE, Wilson AM, Mortier F, Silander Jr JA (2014). *hSDM: hierarchical Bayesian species distribution models*. R package version 1.4, URL https://CRAN.R-project.org/package=hSDM.

Wakefield J (2007). "Disease Mapping and Spatial Regression with Count Data." *Biostatistics*, **8**, 158–183.

Wall M (2004). "A Close Look at the Spatial Structure Implied by the CAR and SAR Models." *Journal of Statistical Planning and Inference*, **121**, 311–324.

Watanabe S (2010). "Asymptotic equivalence of the Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, **11**, 3571–3594.

Womble W (1951). "Differential Systematics." *Science*, **114**, 315–322.

Affiliation:

Duncan Lee
School of Mathematics and Statistics
University Place
University of Glasgow
Glasgow
G12 8SQ, Scotland
E-mail: Duncan.Lee@glasgow.ac.uk
URL: http://www.gla.ac.uk/schools/mathematicsstatistics/staff/duncanlee/