CARBayes: 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 4.1, and is an updated version of a paper in the Journal of Statistical Software in 2013 Volume 55 Issue 13 by the same author. This vignette describes the class of models that can be implemented in CARBayes, and gives 3 worked examples of spatial data analysis using the package.

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 county. 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.* (1999) 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 is 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 strucure 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) (combining the intrinsic model with a 'jump' component for discontinuities), Brewer and Nolan (2007) (variable smoothing via a spatially varying variance), Lu *et al.* (2007) (modelling the adjacency structure of the areal units using logistic regression), Reich and Hodges (2008) (varible smoothing via a spatially varying variance in a spatio-temporal setting), Lee and Mitchell (2012) (modelling the partial correlation between random effects in adjacent areal units as a function of their dissimilarity), and Lee *et al.* (2014) (localised smoothing in an ecological regression context).

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, as well as allowing users to write code to implement their own spatial random effects 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 (2013)) has packages **CARramps** (for Gaussian data), **hSDM** (for binomial data), **spatcounts** (for count data including Poisson and zero-inflated Poisson distributions) and **spdep** (for Gaussian data) that can also 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**.

However, each of these software packages either can only fit a limited set of CAR models or require a degree of programming to implement them, which is the motivation for creating the R package **CARBayes**. The main advantage of this package is its ease of use in fitting CAR models, because: (1) the spatial adjacency information is easy to specify as a binary neighbourhood matrix; and (2) given the neighbourhood matrix the models can be implemented by a single function call in R. In addition, **CARBayes** can implement a much wider class of spatial areal unit models than is possible using the R packages listed above, as the response data can follow binomial, Gaussian or Poisson distributions, while a range of CAR priors have been implemented for the random effects. We note that **CARBayes** is only designed to fit CAR models (for a full list of models see Sections 2 and 3), and is in no way a competitor to the general purpose BUGS software for Bayesian modelling.

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 paper 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 two 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 generalised linear mixed 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 the **CARBayes** software. Inference for all models is set in a Bayesian framework, and is based on Markov chain Monte Carlo (McMC) simulation.

2.1. Level 1 - data likelihood

The study region $\mathcal{S}$ is partitioned into n non-overlapping areal units $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$, which are linked to a corresponding set of responses $\mathbf{Y} = (Y_1, \ldots, Y_n)$, and a vector of known offsets $\mathbf{O} = (O_1, \ldots, O_n)$. The spatial pattern in the response is modelled by a matrix of covariates $X = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})$ and a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, the latter of which are 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 $\mathbf{x}_k^{\mathrm{T}} = (1, x_{k1}, \ldots, x_{kp})$, the first of which corresponds to an intercept term. The general model class that **CARBayes** can implement is a generalised linear mixed model for spatial areal unit data which is given by

$$
\begin{aligned}
Y_k | \mu_k &\sim f(y_k | \mu_k, \nu^2) \quad \text{for } k = 1, \ldots, n, \\
g(\mu_k) &= \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\beta} + \phi_k + O_k.
\end{aligned}
\tag{1}
$$

The responses Y_k come from an exponential family of distributions $f(y_k | \mu_k, \nu^2)$, and in **CARBayes** these can be the binomial, Gaussian or Poisson families. The expected value of Y_k is denoted by $\mathbb{E}[Y_k] = \mu_k$, while ν^2 is an additional scale parameter that is required if the Gaussian family is used. The expected values of the responses are related to the linear predictor via an invertible link function $g(.)$, which in this software is one of the logit (binomial family), identity (Gaussian family) and natural log (Poisson family) functions. The vector of regression parameters are denoted by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, 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 X.

2.2. Level 2 - prior distributions

An independent Gaussian prior is specified for each regression parameter β_j, that is $\beta_j \sim \mathrm{N}(m_j, v_j)$ for $j = 0, \ldots, p$, and the default values specified by the software are $(m_j = 0, v_j = 1000)$. The scale parameter ν^2 for the Gaussian likelihood is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\nu^2 \sim \mathrm{Inverse\text{-}Gamma}(0.001, 0.001)$.

CARBayes can implement a number of different random effects models, and they are summarised below.

Independence priors

The simplest prior for the random effects that can be specified is the independence prior:

$$
\begin{aligned}
\theta_k &\sim \mathrm{N}(0, \sigma^2), \\
\sigma^2 &\sim \text{Inverse-Gamma}(a, b),
\end{aligned}
\tag{2}
$$

where θ_k replaces ϕ_k in the data likelihood (1). The variance parameter is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\sigma^2 \sim \text{Inverse-Gamma}(0.001, 0.001)$. This random effects prior is appropriate if the covariates included in model (1) have removed all of the spatial structure in the response, leaving the random effects to account for the possible effects of over-dispersion (for binomial and Poisson models). However, for most data sets there is likely to be residual spatial autocorrelation, in which case one of the global or local CAR priors described below is more appropriate.

Global smoothing CAR priors

Three different globally smooth conditional autoregressive priors for modelling spatial autocorrelation are available in the **CARBayes** package, the intrinsic and BYM models (both Besag *et al.* (1991)), as well as the alternative developed by Leroux *et al.* (1999). Each model is a special case of a Gaussian Markov Random Field (GMRF), and can be written in the general form $\boldsymbol{\phi} \sim \mathrm{N}(\mathbf{0}, \tau^2 Q^{-1})$, where Q 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 $n \times n$ neighbourhood or weight matrix $\mathbf{W}$. A binary specification 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**, as the only requirement is that $\mathbf{W}$ is non-negative and symmetric.

CAR priors are commonly specified as a set of n univariate full conditional distributions $f(\phi_k | \boldsymbol{\phi}_{-k})$ for $k = 1, \ldots, n$ (where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$), rather than via the multivariate specification described above. The first CAR prior to be proposed was the intrinsic model (Besag *et al.* (1991)), which is given by

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

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. The latter

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. In common with the other variance parameters the default prior specification for τ^2 has $a = b = 0.001$. The limitation with this model is that it can only represent strong spatial autocorrelation, and is well known to produce random effects that are overly smooth. Therefore, the same authors proposed an extension to allow for both weak and strong spatial autocorrelation, by replacing ϕ_k in (1) with $\theta_k + \phi_k$, which are respectively defined by (2) and (3). This model is known as the BYM or convolution model, and is the most commonly used conditional autoregressive model in practice. However, it requires two random effects to be estimated for each data point, whereas only their sum is identifiable from the data. Therefore in **CARBayes** only McMC samples for the sum $re_k = \phi_k + \theta_k$ are returned to the user.

As a result of this limitation Leroux *et al.* (1999) proposed an alternative CAR prior for modelling varying strengths of spatial autocorrelation, using only a single set of random effects. Their model is given by

$$
\begin{aligned}
\phi_k | \boldsymbol{\phi}_{-k}, \mathbf{W}, \tau^2, \rho & \sim \text{N}\left(\frac{\rho \sum_{i=1}^{n} w_{ki} \phi_i}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho} \right), \\
\tau^2 & \sim \text{Inverse-Gamma}(a, b), \\
\rho & \sim \text{Uniform}(0, 1),
\end{aligned}
\tag{4}
$$

where ρ is a spatial autocorrelation parameter, with $\rho = 0$ corresponding to independence, while $\rho = 1$ corresponds to strong spatial autocorrelation. A uniform prior on the unit interval is specified for ρ while the usual inverse gamma prior is adopted for τ^2. For (4) when $\rho = 1$ the intrinsic model give by (3) is obtained, while when $\rho = 0$ the random effects are independent and have the prior given by (2). These global CAR models were compared in a review by Lee (2011), who concluded that the model proposed by Leroux *et al.* (1999) was the most appealing from both theoretical and practical standpoints.

Localised smoothing CAR priors

The CAR priors described above enforce a single global level of spatial smoothing for the set of random effects, which for model (4) 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

$$
\text{COR}(\phi_k, \phi_j | \boldsymbol{\phi}_{-kj}, \mathbf{W}, \rho) = \frac{\rho w_{kj}}{\sqrt{(\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho)(\rho \sum_{i=1}^{n} w_{ji} + 1 - \rho)}}.
\tag{5}
$$

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. However, 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. A number of approaches have been proposed for modelling this localised spatial autocorrelation, and **CARBayes** can implement the proposals of Lee and Mitchell (2012) and Lee and Sarran (2014) which are

both summarised below.

The paper by Lee and Mitchell (2012) proposes a method for capturing such localised spatial autocorrelation, including the identification of boundaries in the random effects surface. The underlying idea is to model the elements of $\mathbf{W}$ corresponding to geographically adjacent areal units as binary random quantities, rather than assuming they are fixed at one. Conversely, if areal units $(\mathcal{S}_k, \mathcal{S}_j)$ do not share a common border then w_{kj} is fixed at zero. From (5), 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. In contrast, 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 if covariates are excluded from (1) then any boundaries identified also relate to the mean surface $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ in the absence of an offset term, because it has the same spatial structure as the random effects as $g(\mu_k) = \beta_0 + \phi_k$.

The model proposed by Lee and Mitchell (2012) is based on the Poisson log-linear specification of (1) and the CAR prior (4), with the restriction that ρ is fixed at 0.99 (**CARBayes** can also fit binomial and Gaussian specifications). This restriction on ρ was made by Lee and Mitchell (2012) to ensure 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, $\{w_{kj}|k \sim j\}$ are collectively modelled as

$$
\begin{aligned}
w_{kj}(\boldsymbol{\alpha}) &= \begin{cases} 1 & \text{if } \exp(-\sum_{i=1}^{q} z_{kji}\alpha_i) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases}, \\
\alpha_i &\sim \text{Uniform}(0, M_i) \qquad \text{for } i = 1, \ldots, q.
\end{aligned} \tag{6}
$$

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 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 aim of Lee and Mitchell (2012) was to identify the locations of any boundaries (abrupt step changes) in disease risk surfaces, so the available covariates were used to construct dissimilarity metrics rather than being incorporated into the linear predictor. In contrast, if the aim of the analysis was to explain the spatial pattern in the response, then covariates would be included in (1), and only metrics directly describing the dissimilarity between two areas, such as the existence of a physical boundary or the distance between the areas centroids, would be included in (6).

The above approch allows for localised spatial autocorrelation by modelling the elements in $\mathbf{W}$, which can reduce the partial autocorrelations between certain pairs of adjacent random effects. An alternative is to augment a 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. One such model was proposed by Lee and Sarran (2014), and partitions the n areal units into a maximum of G clusters each with their own intercept term $(\lambda_1, \ldots, \lambda_G)$. The mean function for this model is given by

$$
\begin{aligned}
g(\mu_k) &= \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta} + \phi_k + \lambda_{Z_k} + O_k, &\qquad(7)\\
\phi_k | \boldsymbol{\phi}_{-k}, \mathbf{W}, \tau^2, \rho &\sim \mathrm{N}\left(\frac{\rho \sum_{i=1}^n w_{ki}\phi_i}{\rho \sum_{i=1}^n w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^n w_{ki} + 1 - \rho}\right), \\
\tau^2 &\sim \mathrm{Inverse\text{-}Gamma}(0.001, 0.001), \\
\rho &\sim \mathrm{Uniform}(0, 1), \\
\lambda_i &\sim \mathrm{Uniform}(\lambda_{i-1}, \lambda_{i+1}) &\qquad \text{for } i = 1, \ldots, G, \\
f(Z_k) &= \frac{\exp(-\delta(Z_k - G^*)^2)}{\sum_{r=1}^G \exp(-\delta(r - G^*)^2)}, \\
\delta &\sim \mathrm{Uniform}(0, M = 100).
\end{aligned}
$$
$$\qquad(8)$$

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. Thus in the above $\lambda_0 = -\infty$ 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 $G^* = (G+1)/2$ is the middle of the intercept terms. A weakly informative uniform prior with a large range is specified for the penalty parameter δ, so that the data play the dominant role in estimating its value. We note that we do not put any spatial smoothing constraints on the indicator vector $\mathbf{Z}$, because similar random effect values can occur at opposite ends of the study region. The clustering is thus inherently non-spatial, and the spatial autocorrelation is accounted for by the CAR prior for ϕ. The **CARBayes** package can fit binomial, Gaussian and Poisson variants of this model.

The Poisson variant of this model has the option to allow for ecological bias, by changing the form of the mean function. Assume that one of the covariates (now called the exposure) is spatially misaligned with the response data (counts), as the former is available at more than one (point) location within each areal unit. Therefore, areal unit k has q_k different values (exposures) $\mathbf{w}_k = (w_{k1}, \ldots, w_{kq_k})$, and hence there is variation within an areal unit for which only a single response variable (count) exists. Furthermore, the population may be unevenly distributed within each areal unit, and the proportions of the population corresponding to each of the q_k exposures within areal unit k are denoted by $\mathbf{p}_k = (p_{k1}, \ldots, p_{kq_k})$, where $\sum_{i=1}^{q_k} p_{ki} = 1$. Then an appropriate model is given by:

$$
\mu_k = \exp(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta} + \phi_k + \lambda_{Z_k} + O_k) \sum_{i=1}^{q_k} p_{ki} \exp(w_{ki}\alpha), \qquad(9)
$$

and further details can be found in Lee and Sarran (2014).

3. CARBayes

3.1. Obtaining the software

CARBayes is an add-on package to the statistical software R ($\geq$ 2.10.0), and is freely available to download from CRAN (http://cran.r-project.org/). The package can be downloaded for Windows, Linux and Apple platforms, and in addition to the base implementation of R, it requires the **coda**, **MASS**, **Rcpp**, **sp**, **spam** and **truncdist** packages. Once R and the required packages have been installed, **CARBayes** can be loaded using the following code.

```
> library(CARBayes)
```

Note, the packages listed in the previous paragraph are automatically loaded upon loading **CARBayes**, as they are the only ones required for **CARBayes** to implement the Bayesian spatial models described in the previous section. 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 additional packages are also required, which need to be loaded into R using the code:

```
> library(shapefiles)
> library(sp)
> library(maptools)
> library(spdep)
```

Note, that the **sp** package is only loaded for use with **CARBayes** functions, and needs to be loaded again for general use in R, such as drawing maps. Finally, to run the examples in the next three sections of this vignette the data package **CARBayesdata** is required, which can be loaded using the command:

```
> library(CARBayesdata)
```

3.2. Functionality

CARBayes can fit the general exponential family Bayesian hierarchical model outlined in the previous section, where the response data can be binomial, Gaussian or Poisson. The models listed below can be implemented by the software.

1. `S.independent()` - the independent random effects model given by (2), with the likelihood model (1).

2. `S.CARiar()` - the intrinsic autoregressive model proposed by Besag *et al.* (1991) and given by (3), with the likelihood model (1).

3. `S.CARbym()` - the BYM model proposed by Besag *et al.* (1991) and given by a linear combination of (2) and (3), with the likelihood model (1).

4. `S.CARleroux()` - the CAR prior proposed by Leroux *et al.* (1999) and given by (4), with the likelihood model (1).

5. `S.CARdissimilarity()` - the local spatial smoothing model proposed by Lee and Mitchell (2012) and given by (4) and (6), with the likelihood model (1).

6. `S.CARcluster()` - the CAR model with a cluster component given by (7).

The linear predictor for each of the Bayesian hierarchical models is specified as an R `formula` object, in common with the `glm()` and `gam()` functions. To fit non-linear functions the function `ns()` can be used in the `formula` object, but the **splines** library needs to be loaded. The spatial neighbourhood information required to run the CAR models needs to be provided as an $n \times n$ neighbourhood matrix $\mathbf{W}$, which is simpler to construct than the series of list objects required by the BUGS software. A full list of arguments for each function can be found in the manual accompanying the package. **CARBayes** contains two further functions `combine.data.shapefile()` and `highlight.borders()`, which aid in plotting spatial maps of the data, and their use is illustrated below. Additionally, two further functions allow the posterior samples (`summarise.samples()`) and linear combinations of the covariate component (`summarise.lincomb()`), for example to summarise non-linear relationships, to be summarised.

3.3. Inference

Inference for all of the Bayesian hierarchical models is based on McMC simulation, using a combination of Gibbs sampling and Metropolis steps. The variance parameters are Gibbs sampled from their full conditional truncated inverse gamma distributions, while the remaining parameters are updated using Metropolis steps with univariate or multivariate random walk proposal distributions. The exception to this is for Gaussian response data, where the covariate regression parameters and the random effects can also be Gibbs sampled. 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 a model has been fitted the `print()` function can be applied to the model object to print out a summary results table to the console, which includes details of the model fitted, parameter estimates and uncertainty intervals.

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 (assuming the **CARBayesdata**, **shapefiles** and **sp** packages have been loaded) using the commands:

```
> data(lipdata)
> data(lipdbf)
> data(lipshp)
```

For other data sets not inlcuded in an R package, you essentially need two types of data. The first is a dataframe, 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 two components: `shapefile.shp` containing the polygons, and `shapefile.dbf` containing a unique identifier linking each row in the dataframe to a polygon in the `shapefile.shp` file. These can be read in to R using the `read.shp()` and `read.dbf()` functions. These three data sets can be combined together to create a `"SpatialPolygonsDataFrame"` object using the following function:

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

For this function to work the rownames of the dataframe (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 first line of the above code. The `data.combined` object is a `"SpatialPolygonsDataFrame"` object, and the data it contains can be mapped using the `spplot()` function. A binary neighbourhood matrix **W** can be constructed for the data from this object using the code:

```
> W.nb <- poly2nb(data.combined, row.names = rownames(lipdata))
> W.mat <- nb2mat(W.nb, style="B")
```

This example was included purely to illustrate this process, and in the next two examples full spatial analyses using the **CARBayes** software will be undertaken.

5. Example 2 - property prices in Greater Glasgow

The utility of 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 Neighbourhood Statistics (SNS) database (*http://www.sns.gov.uk/*), but are also included with the **CARBayesdata** software. The study region is the Greater Glasgow and Clyde health board, which is split into 270 intermediate geographies (IG). These IGs are small areas that have a median area of 124 hectares and a median population of 4,239. Assuming the **CARBayesdata** and **sp** packages have been loaded these data can be loaded into R using the `data()` function as shown below:

```
> data(propertydata.spatial)
> propertydata <- propertydata.spatial@data
```

This data set is a SpatialPolygonsDataFrame object (from the sp package), which is a merger of a dataframe and spatial polygon data. The dataframe is extracted into the object `propertydata` using the second line of code above. These data are summarised in Table

Variable	Percentiles				
	0%	25%	50%	75%	100%
Property price (in thousands)	50.0	95.0	122.0	158.4	372.8
Crime rate (per 10,000)	85.0	303.5	519.0	733.0	8009.0
Number of rooms (median)	3.0	3.0	4.0	4.0	6.0
Property sales (%)	0.2	2.3	3.1	4.1	10.6
Drive time to a shop (minutes)	0.3	0.9	1.3	1.9	8.5

Table 1: Summary of the distribution of the data.

1, which displays the percentiles of their distribution. The response variable in this study is the median price (in thousands) of all properties sold in 2008 in each IG, with that year being chosen because covariate data for later years are not available. 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 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, the percentage of properties that sold in a year, and the average time taken to drive to the nearest shopping centre. Finally, a categorical variable measuring the most prevalent property type in each area is available, with levels; 'flat' (68% of areas), 'terraced' (7%), 'semi-detached' (13%) and 'detached' (12%).

A spatial map of this response variable can be plotted using the functionality of the **sp** package, using the following R code.

```
> northarrow <- list("SpatialPolygonsRescale", layout.north.arrow(),
offset = c(220000,647000), scale = 4000)
> scalebar <- list("SpatialPolygonsRescale", layout.scale.bar(),
offset = c(225000,647000), scale = 10000, fill=c("transparent","black"))
> text1 <- list("sp.text", c(225000,649000), "0")
> text2 <- list("sp.text", c(230000,649000), "5000 m")
> spplot(spatialhousedata, c("price"), sp.layout=list(northarrow, scalebar, text1, text2),
scales=list(draw = TRUE), at=seq(min(housedata$price)-1, max(housedata$price)+1,
length.out=8), col.regions=hsv(0,seq(0.05,1,length.out=7),1), col="transparent")
```

The plotting is achieved by the `spplot()` function, with the preceding lines adding a north arrow, a scale bar and accompanying text. The resulting plot is shown in Figure 1, which suggests that Glasgow has a number of property sub-markets, whose prices are not related to those in neighbouring areas. An example of this is the two groups of darker red regions (more expensive properties) north of the river Clyde (the thin white line running south east), 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 variable is treated as the response and assumed to be Gaussian, and an initial covariate only model is built in a frequentist framework using
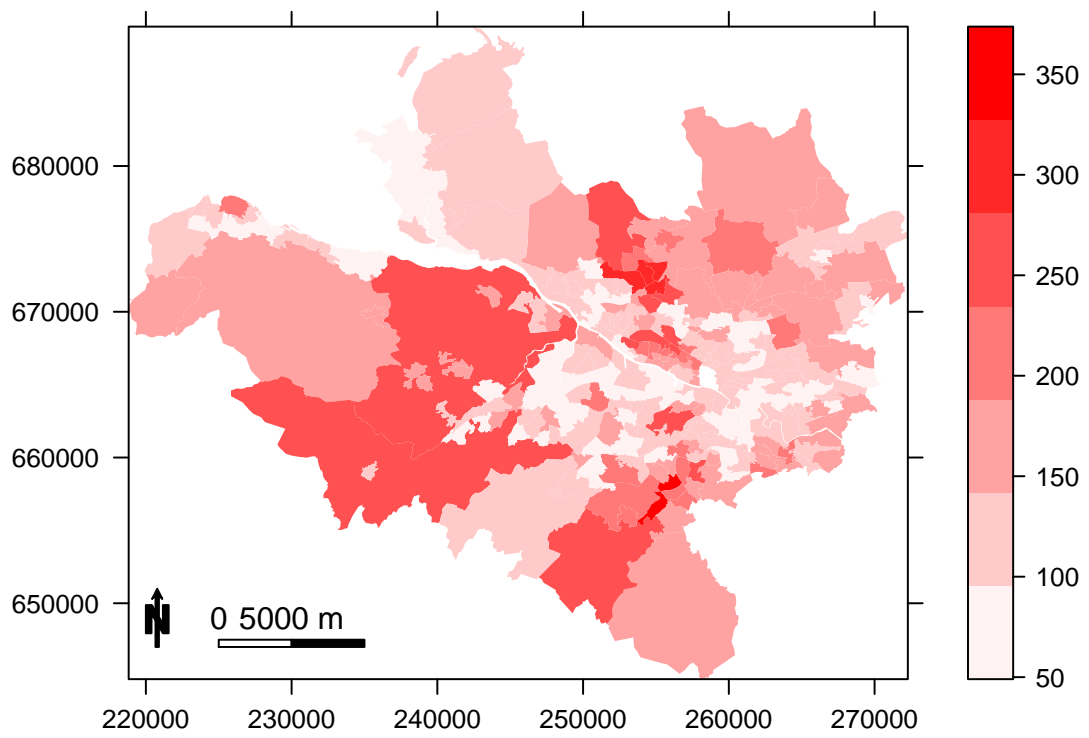
Figure 1: Map displaying median property prices in Greater Glasgow (in thousands).

linear models. Initial plots of the data using the `pairs()` command suggest that the natural log of drive time to a shopping centre is linearly related to the response, and that crime rate has a non-linear relationship to the response. The log transformation of price and drive time to a shop are created using the following commands.

```
> propertydata$logprice <- log(propertydata$price)
> propertydata$logdriveshop <- log(propertydata$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. This is achieved using the following R code:

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

where the first line loads the **splines** library that contains the `ns()` function. Applying the `summary(model)` command to this model shows that 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.

A Moran's I permutation test for spatial autocorrelation was then applied to the residuals from this model based on 10,000 random permutations, using the functionality of the **spdep** package. Code to implement the test is shown below. The first two lines turn the `"SpatialPolygonsDataFrame"` object `spatialhousedata` into an `"nb"` and then a `"listw"` `"nb"` object, which is required by the `moran.mc()` function.

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

        Monte-Carlo simulation of Moran's 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 statistic equals 0.2768 with a corresponding p-value of 0.000099, 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, using the functions outlined in the previous section. We illustrate this by applying model (1) and (4) to the data. The code to implement this model in **CARBayes** is shown below, where the first line creates the binary neighbourhood matrix `W.mat` from the `W.nb` object.

```
W.mat <- nb2mat(W.nb, style="B")
model.spatial <- S.CARleroux(formula=form, data=propertydata, family="gaussian",
W=W.mat, burnin=20000, n.sample=120000, verbose=FALSE, thin=10)
print(model.spatial)
```

Inference for this model is based on 10,000 McMC samples, following a burnin period of 10,000 and the remaining 100,000 samples being thinned by 10 to reduced their autocorrelation. The output from the model can be summarised by the `print(model.spatial)` command. The first part of the output is a description of the model that was fitted, including the likelihood and random effects specifications, as well as the covariates included in the linear predictor. The second part summarises the parameters (except the random effects) by means of posterior medians, 95% credible intervals, and acceptance rates.

Model output

In addition to producing the summary table above, fitting the model returns a list object with the following components:

```
> summary(model.spatial)
```

	Length	Class	Mode
summary.results	65	-none-	numeric
samples	6	-none-	list
fitted.values	270	-none-	numeric
residuals	270	-none-	numeric
modelfit	3	-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 of this list is a summary table of results, which is what is printed by the `print()` function. The next element is a list containing matrices of the thinned and post burnin-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, Spiegelhalter *et al.* (2002)) and the Log Marginal Predictive Likelihood (LMPL, Choi *et al.* (2012)). 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 compatabiltiy with the other functions in the package. Finally, the `formula` and `model` 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 re-created (or similar summarise created for other parameters) using the function

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


$quantiles
               0.5         0.025          0.975
 [1,]   4.239360710   3.974538478   4.533852882
 [2,]  -0.246804128  -0.399486648  -0.099706192
 [3,]  -0.397501364  -0.687444914  -0.116832230
 [4,]  -0.195872093  -0.400359980   0.007575266
 [5,]   0.220290725   0.167046337   0.270358098
 [6,]   0.002274724   0.001621726   0.002877815
 [7,]  -0.252926021  -0.374373889  -0.124341749
 [8,]  -0.168015426  -0.268187198  -0.064834537
 [9,]  -0.294361624  -0.426330175  -0.152823205
[10,]  -0.005722209  -0.063584780   0.049366172


$exceedences
NULL
```

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 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 using the code:

```
> crime.effect <- summarise.lincomb(model=model.spatial, columns=c(2,3,4),
+ quantiles=c(0.5, 0.025, 0.975), distribution=FALSE)
```

The non-linear relationship can be plotted using the code below, and the resulting plot is shown in Figure 2.
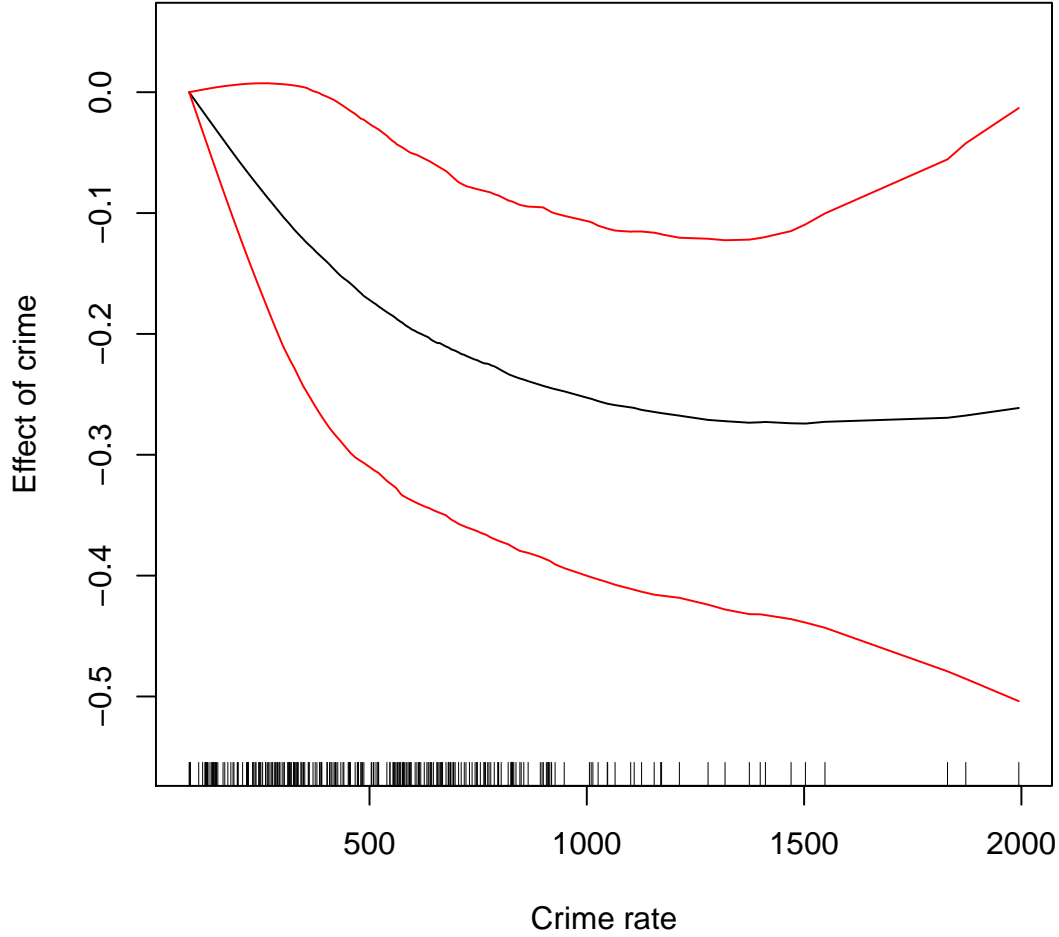
Figure 2: Plot showing the estimated non-linear relationship between crime rate and log-price.

```
ord <- order(propertydata$crime)
plot(propertydata$crime[ord], crime.effect$quantiles[ord ,1], pch=19, ylim=c(-0.55,0.05),
xlab="Number of crimes", ylab="Effect of crime", type="l")
rug(propertydata$crime)
lines(propertydata$crime[ord], crime.effect$quantiles[ord ,2], pch=19, col="red")
lines(propertydata$crime[ord], crime.effect$quantiles[ord ,3], pch=19, col="red")
```

Acceptance rates for the McMC algorithm

The acceptance rate for ρ quantifies the proportion of times the value proposed by the Metropolis updating step was accepted as the new value of the Markov chain. In contrast, due to the conjugacy between the Gaussian likelihood and the prior distributions for $(\boldsymbol{\beta}, \boldsymbol{\phi}, \nu^2, \tau^2)$, Gibbs sampling is employed for updating these parameters, which is the reason for the 100%

acceptance rate. If the likelihood was either binomial or Poisson then Metropolis updating steps would be used for $(\boldsymbol{\beta}, \boldsymbol{\phi})$ instead, and the acceptance rates would then be of interest to the analyst. The obvious acceptance rate of 100% is shown here for consistency of presentation with the summary output across different models.

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 *http://www.sns.gov.uk/* (and are included with the **CARBayesdata** software). 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 the data on respiratory disease risk (included in **CARBayesdata**) and can be loaded with the commands:

```
> data(respiratorydata.spatial)
> respiratorydata <- respiratorydata.spatial@data
> head(respiratorydata)
```

	observed2010	expected2010	incomedep2010	SIR2010
S02000618	105	105.12944	15	0.9987687
S02000613	85	69.41011	22	1.2246054
S02000623	37	87.85767	8	0.4211357
S02000626	90	89.41669	26	1.0065235
S02000636	41	97.55097	8	0.4202931
S02000645	47	84.86336	8	0.5538315

where the second line extracts the dataframe as in example 2, and the `head()` function displays the first 6 rows of the dataframe. The data set contains the numbers of hospital admissions in 2010 in each IG due to respiratory disease (International Classification of Disease tenth revision codes J00-J99), which is stored in the `observed2010` column. However, 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 either internal or external standardisation. For

these data we use external standardisation, based on age and sex standardised rates for the whole of Scotland. These expected numbers are stored in the `expected2010` column, and the simplest measure of disease risk is the Standardised Incidence Ratio (SIR), which is the ratio of the observed to the expected numbers of hospital admissions. The SIR is added to the `respiratorydata` and `respiratorydata.spatial` objects using the code below, which also creates the spatial objects that are required for the analysis.

```
> respiratorydata$SIR2010 <- respiratorydata$observed2010 / respiratorydata$expected2010
> respiratorydata.spatial@data$SIR2010 <- respiratorydata$SIR2010
> W.nb <- poly2nb(respiratorydata.spatial, row.names = rownames(respiratorydata))
> W.mat <- nb2mat(W.nb, style="B")
```

A map of the SIR for these data is displayed in Figure 3, which was created using similar code to that provided in Section 5 for mapping the median property price data. Values of the SIR above one relate to areas exhibiting above average risks, while values below one correspond to below average risks. The figure shows evidence of localised spatial structure in these disease data, with numerous different locations where high and low risk areas border each other. This in turn suggests that boundaries are likely to be present in these data, 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 (are in receipt of a combination of means tested benefits), because it is well known that socio-economic deprivation plays a large role in determining people's health. The income data for each IG are contained in the `incomedep2010` column in `respdata`.

6.2. Spatial modelling with CARBayes

Let the observed and expected numbers of hospital admissions be denoted by $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{E} = (E_1, \ldots, E_n)$ 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, which can be created from the vector of area level income deprivation scores using the following code.

```
> Z.incomedep <- as.matrix(dist(cbind(respiratorydata$incomedep2010,
+ respiratorydata$incomedep2010), method="manhattan", diag=TRUE, upper=TRUE)) *
+ W.mat / 2
```

The function to implement the localised CAR model is called `S.CARdissimilarity()`, and it takes the same arguments as the global CAR models except that it additionally requires the dissimilarity metrics. These are required in the form of a list of $n \times n$ matrices, and the model is run using the following code.
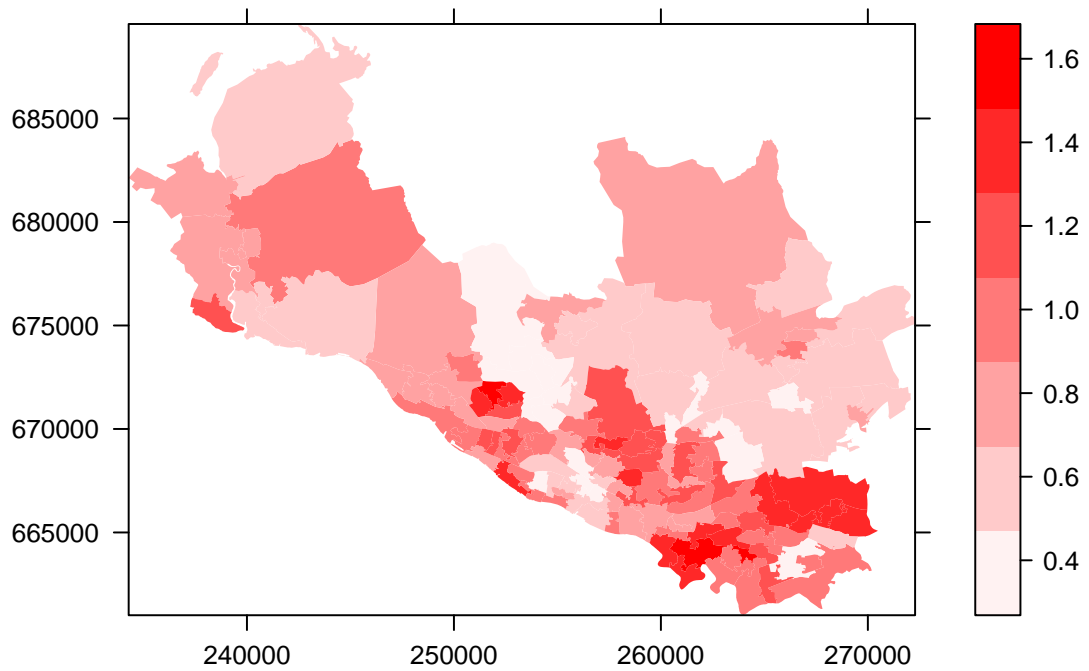
Figure 3: Map displaying Standardised Incidence Ratio for the northern part of Greater Glasgow.

```
formula <- observed2010 ~ offset(log(expected2010))
model.dissimilarity <- S.CARdissimilarity(formula=formula, data=respiratorydata,
    family="poisson", W=W.mat, Z=list(Z.incomedep=Z.incomedep), burnin=20000,
n.sample=40000, verbose=FALSE)
print(model.dissimilarity)
```

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 $\mathbf{R} = (R_1, \ldots, R_n)$). When the function finishes, typing in `print(model.dissimilarity)` as above produces summary output 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 had no effect in identifying boundaries in the response (random effects) surface. A brief description is given in Section 2.2, 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 an $n \times n$ 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 $\mathbf{R}$) using the following code. The first line saves the matrix of border locations, while the second and third add the estimated risk values to the `data.combined` object. The next two lines identify the boundary points (using the **CARBayes** function `highlight.borders()`), and format them to enable plotting. The remaining commands relate to the plotting, and are similar to those used to produce the earlier spatial maps.

```
border.locations <- model.dissimilarity$localised.structure$W.posterior
risk.estimates <- model.dissimilarity$fitted.values / respiratorydata$expected2010
respiratorydata.spatial@data$risk.estimates <- risk.estimates
boundary.final <- highlight.borders(border.locations=border.locations,
spdata=respiratorydata.spatial)
boundaries = list("sp.points", boundary.final, col="black", pch=19, cex=0.2)
northarrow <- list("SpatialPolygonsRescale", layout.north.arrow(),
offset = c(220000,647000), scale = 4000)
scalebar <- list("SpatialPolygonsRescale", layout.scale.bar(),
offset = c(225000,647000),  scale = 10000, fill=c("transparent","black"))
text1 <- list("sp.text", c(225000,649000), "0")
text2 <- list("sp.text", c(230000,649000), "5000 m")
spplot(respiratorydata.spatial, c("risk.estimates"), sp.layout=list(northarrow,
scalebar, text1, text2, boundaries), scales=list(draw = TRUE),
```

```
at=seq(min(risk.estimates)-0.1, max(risk.estimates)+0.1, length.out=8),
col.regions=hsv(0,seq(0.05,1,length.out=7),1), col="transparent")
```

The result of these commands are displayed in Figure 4, which shows the fitted risk surface and the locations of the boundaries (denoted by black dots). The model has identified 103 boundaries in the risk surface, which is 28.6% of the total number of borders in the study region. 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 demarkation between the low risk 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 model proposed by Lee and Mitchell (2012). The response data can be binomial, Gaussian or Poisson, with the canonical link functions logit, the identity and natural log respectively. The availability of areal unit data has grown dramatically in recent times, due to the launch of freely available online databases such as Neighbourhood Statistics in the UK (see *http://www.neighbourhood.statistics.gov.uk* and *http://www.sns.gov.uk/*), and Surveillance Epidemiology and End Results (SEER, *http://seer.cancer.gov/*) in the USA. 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.

A number of other software packages can also fit conditional autoregressive models to spatial data, including BUGS, BayesX and R packages **CARramps**, **hSDM**, **INLA**, **spatcounts** and **spdep**. However, these software packages either can only fit a limited selection of CAR models, or require a degree of programming which may be beyond some users of spatial data. Thus a gap in the market exists for user friendly software that can fit a wide class of CAR models, which was the motivation behind the **CARBayes** software. The user friendly features of **CARBayes** have been illustrated by the two worked examples presented in Sections 4 and 5, which include: (i) models can be implemented using a single function call; (ii) the spatial information required by the models is straightforward to create from a shapefile; (iii) only a small number of arguments are required to run a *default* analysis; and (iv) the software reports on the progress of model fitting, and produces a summary table of the results when it has finished.

Future development for the software will focus on moving into the spatio-temporal domain, because there is relatively little existing software (especially in R) that can fit spatio-temporal models for areal unit data (an example for geostatistical data is **spTimer**). The development of statistical modelling techniques for such data is also in its infancy, with prominent early
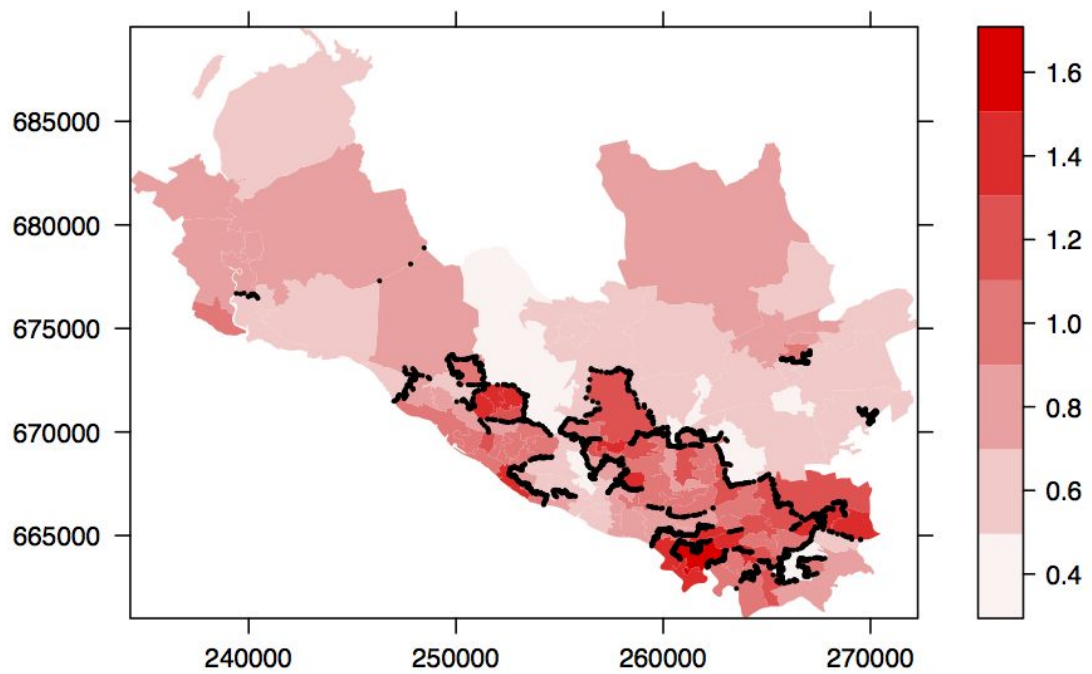
Figure 4: Map displaying estimated risk and locations of the boundaries for the northern part of Greater Glasgow.

examples being Bernardinelli *et al.* (1995) and Knorr-Held (2000).

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.

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

Choi J, Lawson A, Cai B, Hossain M, Kirby R, Liu J (2012). "A Bayesian latent model with spatio-temporally varying coefficients in low birth weight incidence data." *Statistical Methods in Medical Reseaerch*, **21**, 445–456.

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

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

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

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, 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, 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 (2014). "Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies." *Technical report*, arXiv:1412.4479.

Leroux B, Lei X, Breslow N (1999). *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*, chapter Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds), pp. 135–178. 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.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reich B, Hodges J (2008). "Modeling Longitudinal Spatial Periodontal Data: A Spatially-Adaptive Model with Tools for Specifying Priors and Checking Fit." *Biometrics*, **64**, 790–799.

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.

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.

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

Affiliation:

Duncan Lee
School of Mathematics and Statistics
15 University Gardens

University of Glasgow
Glasgow
G12 8QQ, Scotland
E-mail: Duncan.Lee@glasgow.ac.uk
URL: http://www.gla.ac.uk/schools/mathematicsstatistics/staff/duncanlee/