CARBayes: An R Package for Bayesian Spatial Modelling with Conditional Autoregressive Priors

Duncan Lee
University of Glasgow

Abstract

This is a vignette for the R package **CARBayes** version 3, and is an updated version of a paper in the Journal of Statistical Software in 2013 Volume 55 Issue 13 with the same title and author. This vignette is required because **CARBayes** has been updated both in terms of syntax and functionality since the paper appeared in the Journal of Statistical Software.

Keywords: Bayesian models, 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, Ferguson, and Mitchell (2009)), disease mapping (see Green and Richardson (2002) and Lee (2011)) and Wombling (see Lu, Reilly, Banerjee, and Carlin (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 including 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 correlated 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 correlated random effects, as part of a Bayesian hierarchical model.

These random effects are typically represented with a Conditional AutoRegressive (CAR, Besag, York, and Mollié (1991)) model, 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, Lei, and Breslow (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 smoothing. Such a uniform level of spatial smoothness for the entire region is unrealistic for real data, which are instead likely to exhibit sub-areas of spatial autocorrelation separated by discontinuities. Such localised spatial smoothing 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. A number of approaches have been proposed for extending the class of CAR priors to deal with localised spatial smoothing, 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, Rushworth, and Sahu (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, Spiegelhalter, Thomas, and Best (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. These 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 CAR models than is possible using the other R packages listed above, as the response data can follow binomial, Gaussian or Poisson distributions. 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.

Therefore the aim of this paper is to present the software **CARBayes**, by outlining the class of models that it can implement and illustrating its use by means of two 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 and five give two worked examples of using the software, including how to create the neighbourhood matrix and produce spatial maps of the results. Finally, Section six contains a concluding discussion, and outlines areas for future development.

2. Bayesian hierarchical models for spatial areal unit data

This section outlines the general Bayesian hierarchical model for spatial areal unit data that can be implemented in the **CARBayes** package.

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 $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and a vector of known offsets $\mathbf{O} = (O_1, \ldots, O_n)^{\mathrm{T}}$. The spatial pattern in the response is modelled by a matrix of covariates $X = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$ 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 that **CARBayes** can implement is an extension of a generalised linear model and 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 $\mathsf{E}(Y_k) = \mu_k$, 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 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 $\nu^2 \sim$ Inverse-Gamma$(0.001, 0.001)$.

CARBayes can implement a number of different random effects models, with the simplest being the independence prior

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

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 required.

Global smoothing CAR priors

Four different conditional autoregressive priors are commonly used for modelling spatial autocorrelation in the statistics literature, the intrinsic and BYM models (both Besag *et al.* (1991)), as well as the alternatives developed by Leroux *et al.* (1999) and Stern and Cressie (1999). Each model is a special case of a Gaussian Markov Random Field (GMRF), and can be written in the general form $\boldsymbol{\phi} \sim \text{N}(\mathbf{0}, \tau^2 Q^{-1})$, 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 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 $w_{kj} = 1$) to be correlated, whereas random effects relating to non-contiguous areal unit are conditionally independent given the values of the remaining random effects. 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

$$\phi_k|\boldsymbol{\phi}_{-k} \sim \text{N}\left(\frac{\sum_{i=1}^n w_{ki}\phi_i}{\sum_{i=1}^n w_{ki}}, \frac{\tau^2}{\sum_{i=1}^n w_{ki}}\right). \tag{3}$$

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 spatially correlated, 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, τ^2 is assigned a conjugate inverse-gamma prior distribution, where the default specification is $\tau^2 \sim \text{Inverse-Gamma}(0.001, 0.001)$. 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, Leroux *et al.* (1999) and Stern and Cressie (1999) proposed alternative CAR priors for modelling varying strengths of spatial autocorrelation, using only a single set of random effects. The model by Leroux *et al.* (1999) is given by

$$\phi_k|\boldsymbol{\phi}_{-k} \sim \mathrm{N}\left(\frac{\rho\sum_{i=1}^n w_{ki}\phi_i}{\rho\sum_{i=1}^n w_{ki}+1-\rho}, \frac{\tau^2}{\rho\sum_{i=1}^n w_{ki}+1-\rho}\right), \tag{4}$$

while the proposal of Stern and Cressie (1999) is

$$\phi_k|\boldsymbol{\phi}_{-k} \sim \mathrm{N}\left(\frac{\rho\sum_{i=1}^n w_{ki}\phi_i}{\sum_{i=1}^n w_{ki}}, \frac{\tau^2}{\sum_{i=1}^n w_{ki}}\right). \tag{5}$$

In both cases 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, that is $\rho \sim \mathrm{U}(0,1)$, while the usual inverse gamma prior is adopted for τ^2. For both (4) and (5) when $\rho = 1$ the intrinsic model proposed by Besag *et al.* (1991) is obtained, while when $\rho = 0$ the only difference is the denominator in the conditional variance. These global CAR models were compared in a recent 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 correlation structure implied by that model, which for (ϕ_k, ϕ_j) is given by

$$\mathrm{COR}(\phi_k, \phi_j|\boldsymbol{\phi}_{-kj}) = \frac{\rho w_{kj}}{\sqrt{(\rho\sum_{i=1}^n w_{ki}+1-\rho)(\rho\sum_{i=1}^n w_{ji}+1-\rho)}}. \tag{6}$$

For non-neighbouring areas (where $w_{kj} = 0$) the random effects are conditionally independent, while for neighbouring areas their partial correlation 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. The paper by Lee and Mitchell (2012) proposes a method for capturing such localised spatial structure, including the identification of boundaries in the random effects surface. The underlying idea is to model the elements of W corresponding to geographically adjacent areas 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 (6), it is straightforward to see that if w_{kj} is estimated as one then (ϕ_k, ϕ_j) are spatially correlated, 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 (although **CARBayes** can also fit binomial and Gaussian specifications). This restriction 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

$$\begin{aligned} w_{kj}(\boldsymbol{\alpha}) &= \begin{cases} 1 & \text{if } \exp(-\sum_{i=1}^{q} z_{kji}\alpha_i) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases}, \\ \alpha_i &\sim \text{Uniform}(0, M_i) \qquad \text{for } i = 1, \ldots, q. \end{aligned} \tag{7}$$

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 (7).

2.3. CAR clustering models

In a forthcoming paper a clustering CAR model is proposed, whose aim is to identify clusters in spatial data. The model was applied to binomial data, where clusters in the proportion surface were identified. Thus the n areal units are partitioned into q clusters each with their own intercept term $(\beta_1, \ldots, \beta_q)$, and if adjacent units (k, i) are in different clusters their probabilities (p_k, p_i) may exhibit very different values. The model is given by

$$\begin{aligned} Y_k|N_k, p_k &\sim \text{Binomial}(N_k, p_k), \\ \ln\left(\frac{p_k}{1-p_k}\right) &= \beta_{X_k} + \phi_k, \\ X_k &\sim \text{Multinomial}(1; 1/q, \ldots, 1/q), \\ \phi_k|\boldsymbol{\phi}_{-k} &\sim \text{N}\left(\frac{\rho\sum_{i=1}^{n} w_{ki}\phi_i}{\rho\sum_{i=1}^{n} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^{n} w_{ki} + 1 - \rho}\right), \\ \beta_i &\sim \text{Uniform}(\beta_{i-1}, \beta_{i+1}) \qquad \text{for } i = 1, \ldots, q, \end{aligned} \tag{8}$$

$$\begin{aligned} \tau^2 &\sim \text{Inverse-Gamma}(0.001, 0.001), \\ \rho &\sim \text{Uniform}(0, 1), \end{aligned}$$

where no covariates are included in the model, only the cluster term represented by β_{X_k} and the spatially smooth CAR prior given by (4). The cluster means $(\beta_1, \ldots, \beta_q)$ are ordered so that $\beta_1 < \beta_2 < \ldots < \beta_q$, which prevents the label switching problem common in mixture models. Thus in the above $\beta_0 = -\infty$ and $\beta_{q+1} = \infty$. Thus the vector $\mathbf{X} = (X_1, \ldots, X_n)$ are cluster indicators, where each indicator can take values $(1, \ldots, q)$. A multinomial prior is specified for each indicator with equal probabilities, which specifies our prior ignorance as $\mathbb{P}(X_k = j) = 1/q$ for all areal units k for $j = 1, \ldots, q$. We note that we do not put any spatial smoothing constraints on the indicator vector $\mathbf{X}$, because similar proportions 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 $\boldsymbol{\phi}$. In addition to this model the **CARBayes** package can fit Gaussian and Poisson variants of this model. The Poisson variant has an option to have an expected count as an offset term as would be common in disease mapping applications, while the Gaussian model does not.

3. CARBayes

3.1. Obtaining the software

The **CARBayes** software 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 **MASS**, **coda**, **evd**, **lattice**, **MCMCpack**, **Rcpp**, **spam**, **stats4** and **truncdist** packages. Once R and the required packages have been installed, **CARBayes** can be loaded using the following code.

```
R> 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 `library()` command as above: **boot**, **deldir**, **foreign**, **grid**, **maptools**, **Matrix**, **nlme**, **shapefiles**,**sp**, **spdep** and **splines**. These packages have also been loaded for the analyses presented in Sections four and five.

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. In a change from version 1 and 2 of **CARBayes**, version 3 has an additional `family` argument, which means that the binomial, Gaussian and Poisson likelihood variants of the same random

effects models are now called by the same function. This change was made to make the package compatible with other R functions such as `glm()`. The models listed below can be implemented by the software, and in all cases binomial, Gaussian and Poisson variants can be fitted.

1. `independent.re()` - the independent random effects model given by (2), with the likelihood model (1).
2. `iarCAR.re()` - the intrinsic autoregressive model proposed by Besag *et al.* (1991) and given by (3), with the likelihood model (1).
3. `bymCAR.re()` - 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. `lerouxCAR.re()` - the CAR prior proposed by Leroux *et al.* (1999) and given by (4), with the likelihood model (1).
5. `properCAR.re()` - the CAR prior proposed by Stern and Cressie (1999) and given by (5), with the likelihood model (1).
6. `dissimilarityCAR.re()` - the local spatial smoothing model proposed by Lee and Mitchell (2012) and given by (4) and (7), with the likelihood model (1).
7. `clusterCAR.re()` - the CAR model with a cluster component given by (8).

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. The spatial neighbourhood information required to run the CAR models needs to be provided as an $n \times n$ neighbourhood matrix 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. In addition to the functions listed above, the package contains two further functions `combine.data.shapefile()` and `highlight.borders()`. These functions aid in plotting spatial maps of the data, and their use is illustrated in Sections 4 and 5 of this paper. Finally, the package also contains the data files needed to recreate these analyses.

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 after every 1,000 MCMC iterations, 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 - 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.

4.1. 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 **CARBayes** software. The study region is the Greater Glasgow and Clyde health board, which is split into 271 intermediate geographies (IG). These IGs are small areas that have a median area of 124 hectares and a median population of 4,239. The data come in two parts. The first is a comma separated variable (csv) file `housedata.csv`, which contains the response and covariate data as well as a column containing the unique identifier (IG) for each area. The second part of the data is a shapefile, which comprises `shp.shp` containing the polygons, and `dbf.dbf` containing the lookup file linking each area (via IG) to a polygon. These data can be read into R using the following code, provided that the working directory has been set to the location of the data.

```
R> housedata <- read.csv(file="housedata.csv", row.names=1)
R> shp <- read.shp(shp.name="shp.shp")
R> dbf <- read.dbf(dbf.name="dbf.dbf")
```

Note, as these data are all included in the **CARBayes** package they can each be loaded into R using the `data()` function instead (that is using the code `data(housedata)`, `data(shp)`, `data(dbf)`). The structure of `housedata` is shown below using the `head()` function, and from the above `read.csv()` command, the unique identifier (IG) has been turned into the row names of the data frame.

```
R> head(housedata)
```

```
             price crime rooms sales driveshop   type
S02000260 112.250   390     3    68       1.2     flat
S02000261 156.875   116     5    26       2.0     semi
S02000262 178.111   196     5    34       1.7     semi
S02000263 249.725   146     5    80       1.5 detached
S02000264 174.500   288     4    60       0.8     semi
S02000265 163.521   342     4    24       2.5     semi
```

These data are summarised in Table 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

Variable	**Percentiles**				
	0%	25%	50%	75%	100%
House 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.

10,000 people that live there, and the values range between 85 and 1994 with the addition of a single large outlier of 8009. The location of this outlier is the city centre of Glasgow, and the high crime rate is likely to be caused by the large numbers of visitors to this part of the city both during the day and at night. Therefore, as this area has an artificially high crime rate, it is removed from the data set using the following code.

```
R> housedata <- housedata[!rownames(housedata)=="S02000655", ]
```

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%). The next step in the analysis is to combine the data with the shapefile using the **CARBayes** function `combine.data.shapefile()`, which allows spatial maps of the variables in the data.frame `housedata` to be produced. The function requires the row names of `housedata` to appear in the first column of the lookup table in the `dbf` part of the shapefile. We note that `housedata` only relates to a subset of the areas in the shapefile, which contains intermediate geographies for the whole of Scotland. The data and shapefile can be combined with the code

```
R> data.combined <- combine.data.shapefile(data=housedata, shp=shp, dbf=dbf)
```

which produces an object `data.combined` of class `"SpatialPolygonsDataFrame"`, which is an object type from the **sp** package. A spatial map of this response variable can be plotted using the functionality of the **sp** package, using the following R code.

```
R> northarrow <- list("SpatialPolygonsRescale", layout.north.arrow(),
   offset = c(220000,647000), scale = 4000)
R> scalebar <- list("SpatialPolygonsRescale", layout.scale.bar(),
   offset = c(225000,647000), scale = 10000, fill=c("transparent","black"))
R> text1 <- list("sp.text", c(225000,649000), "0")
R> text2 <- list("sp.text", c(230000,649000), "5000 m")
R> spplot(data.combined, c("price"), sp.layout=list(northarrow, scalebar,
   text1, text2), at=seq(min(housedata$price)-1, max(housedata$price)+1,
   length.out=8),  col.regions=c("#FEE5D9", "#FCBBA1", "#FC9272", "#FB6A4A",
   "#EF3B2C", "#CB181D", "#99000D"))
```

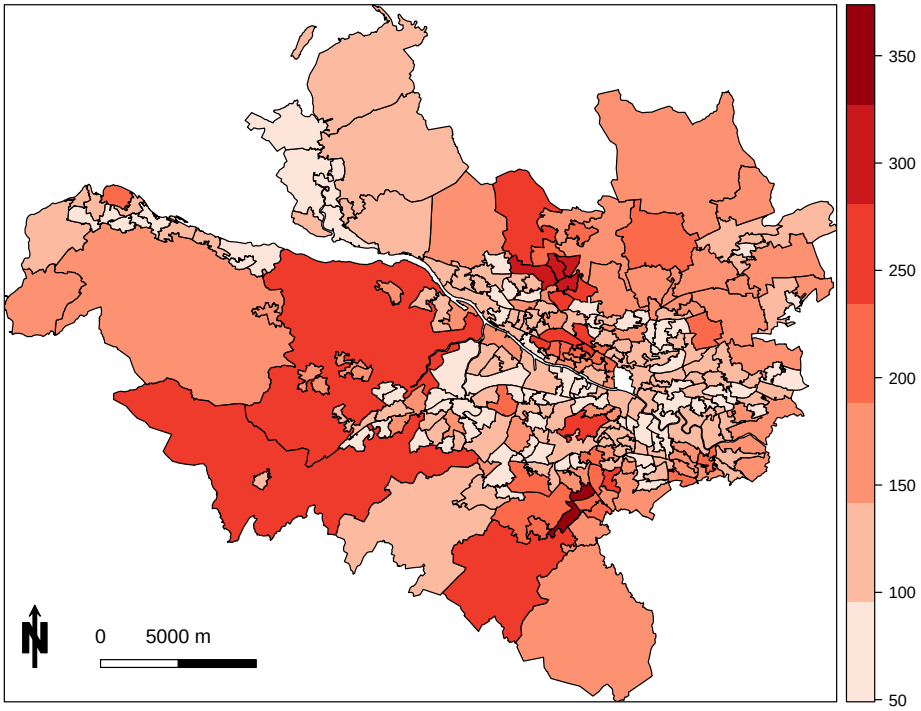

Figure 1: Map displaying median property prices in Greater Glasgow (in thousands).

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.

4.2. 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 linear models. Initial plots of the data using the `pairs()` command suggest that the natural logs of both the crime rate and the drive time to a shopping centre are linearly related to the response, and the transformation of all three variables is achieved using the following commands.

```
R> housedata$logprice <- log(housedata$price)
R> housedata$logcrime <- log(housedata$crime)
R> housedata$logdriveshop <- log(housedata$driveshop)
```

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. This covariate model can be fitted to the data using the following R code:

```
R> form <- housedata$logprice ~ housedata$logcrime + housedata$rooms
   + housedata$sales + factor(housedata$type) + housedata$logdriveshop
R> model <- lm(formula=form)
```

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, 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. Code to implement the test is shown below. The first two lines turn the `"SpatialPolygonsDataFrame"` object `data.combined` into an `"nb"` and then a `"listw" "nb"` object, which is required by the `moran.mc()` function.

```
R> W.nb <- poly2nb(data.combined, row.names = rownames(housedata))
R> W.list <- nb2listw(W.nb, style="B")
R> resid.model <- residuals(model)
R> moran.mc(x=resid.model, listw=W.list, nsim=10000)

        Monte-Carlo simulation of Moran's I
data:  resid.model
```

```
weights: W.list
number of simulations + 1: 10001

statistic = 0.2768, observed rank = 10001, p-value = 9.999e-05
alternative hypothesis: greater
```

4.3. 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 (5) to the data, because it allows a direct comparison of the **CARBayes** and BUGS software packages, as the latter has the inbuilt function `car.proper` to implement this model. 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.

```
R> W.mat <- nb2mat(W.nb, style="B")
R> model.spatial <- properCAR.re(formula=form, family="gaussian", W=W.mat,
        burnin=20000, n.sample=100000, thin=10)
```

Inference for this model is based on 8,000 MCMC samples, which were obtained by running the chain for 100,000 samples, with 20,000 being discarded as the burn-in period and the remaining 80,000 being thinned by 10 to reduce the autocorrelation. When the function finished running typing `print(model.spatial)` produces the summary output shown below. 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 fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Proper CAR
Regression equation - housedata$logprice ~ housedata$logcrime + housedata$rooms
+ housedata$sales +   factor(housedata$type) + housedata$logdriveshop

#############
#### Results
#############
Posterior quantiles and DIC

                            Median    2.5%   97.5% n.sample % accept
(Intercept)                 4.7571  4.2723  5.2395     8000    100.0
housedata$logcrime         -0.1121 -0.1718 -0.0525     8000    100.0
housedata$rooms             0.2223  0.1720  0.2727     8000    100.0
housedata$sales             0.0023  0.0017  0.0029     8000    100.0
```

```
factor(housedata$type)flat    -0.2546 -0.3667 -0.1428     8000    100.0
factor(housedata$type)semi    -0.1630 -0.2611 -0.0664     8000    100.0
factor(housedata$type)terrace -0.2912 -0.4151 -0.1704     8000    100.0
housedata$logdriveshop        -0.0030 -0.0590  0.0550     8000    100.0
nu2                            0.0237  0.0139  0.0330     8000    100.0
tau2                           0.0502  0.0239  0.0988     8000    100.0
rho                            0.9800  0.9400  0.9900     8000     25.9

DIC =  -156.0622       p.d =  95.30304
```

Model output

In addition to producing the summary table above, fitting the model returns a list object with the following components, which can be viewed using the `summary(model.spatial)` function as shown below.

```
                Length Class   Mode
formula              3 formula call
samples              5 -none-  list
fitted.values     1350 -none-  numeric
residuals         1350 -none-  numeric
W.summary        72900 -none-  numeric
modelfit             4 -none-  numeric
summary.results     55 -none-  numeric
model                2 -none-  character
accept               5 -none-  numeric
```

The first element of this list is the fixed effects regression model specified by the formula argument. 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 an 8,000×8 matrix containing the MCMC samples for all the regression parameters. The next two elements in the list `fitted.values` and `residuals` comprise matrices of dimension $n \times 5$ (here $n = 270$), which summarise the posterior distribution of the fitted values and residuals respectively. Each row corresponds to a single area, while the columns represent the posterior mean, standard deviation and $50th$, $2.5th$ and $97.5th$ percentiles of the distribution. The `W.summary` element returns the binary neighbourhood matrix W, while `modelfit` gives a selection of model fit criteria. These criteria include the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, and Van der Linde (2002)), DIC_3 (Celeux, Forbes, Robert, and Titterington (2006)), and the Marginal Predictive Likelihood (MPL, Choi, Lawson, Cai, Hossain, Kirby, and Liu (2012)). For further details about Bayesian modelling and model fit criteria see Gelman, Carlin, Stern, and Rubin (2003).

Parameter estimates

The summary output above shows that all covariates exhibit substantial effects on the response except the natural log of the time taken to drive to a shopping centre, as their 95% credible intervals do not include zero. For example, increasing the average number of rooms by one is

Parameter	**CARBayes**	**BUGS**	**% difference**
logcrime	-0.1111	-0.1132	1.8%
rooms	0.2224	0.2221	0.1%
sales	0.0023	0.0023	0%
flat	-0.2542	-0.2501	1.6%
semi	-0.1626	-0.1596	1.8%
terrace	-0.2913	-0.2893	0.7%
driveshop	-0.0013	0.0024	154.2%
nu2	0.0237	0.0247	4.0%
tau2	0.0518	0.0447	13.7%
rho	0.9852	0.9901	0.5%

Table 2: Comparison of the parameter estimates (posterior medians) from the **CARBayes** and the BUGS software packages. The final column displays the absolute percentage difference in the estimates relative to the larger of the two estimates.

estimated to increase the average property price by 24.9%, because the ratio of the average property prices that differ only in having m and $m+1$ rooms is equal to $\exp(0.2223) = 1.249$. Similarly, IGs that predominately comprise flats have lower median property prices by around 22.5% ($1 - \exp(-0.2546) = 0.225$), compared with the baseline category of 'detached'. The above output also shows that the random effects have modelled substantial spatial autocorrelation, as the posterior median for the spatial autocorrelation parameter ρ is 0.9800. The entire posterior distribution (as summarised by the MCMC output) can be viewed using the code

```
R> plot(model.spatial$samples$rho)
```

and the resulting plot is displayed in Figure 2. The matrix of MCMC samples is output as an `"mcmc"` object, which comes from the **coda** package. Plotting this object thus yields a trace plot (left panel) and a density estimate (right panel), and further MCMC diagnostics are available from the **coda** package. The estimated parameters are not highly correlated with each other, for example, the correlations between the regression parameters range between -0.87 and 0.63, with the middle 50% ranging between -0.09 and 0.17. The validity of the parameter estimates from the **CARBayes** software were assessed by fitting the same model in the **BUGS** software. The results of this comparison are displayed in Table 2, which shows the point estimates (posterior medians) from the two software packages as well as the percentage absolute difference relative to the larger of the two estimates. Results are shown for the covariate effects ($\boldsymbol{\beta}$), both variance parameters (τ^2, ν^2), and the correlation parameter (ρ). Overall, the table shows good agreement between the two sets of point estimates, with percentage absolute differences less than two for seven out of the ten parameters. The large disparity between the two software packages over the estimation of the regression coefficient for drive time to a shopping centre is artificial, as both estimates are very close to zero (they differ in the sign). These estimates are accompanied by relatively wide 95% credible intervals, and both software packages suggest that this covariate has no effect on the response. The other biggest difference between the software packages concerns their estimation of the random effects variance τ^2, which is just under 14% larger using **CARBayes**.

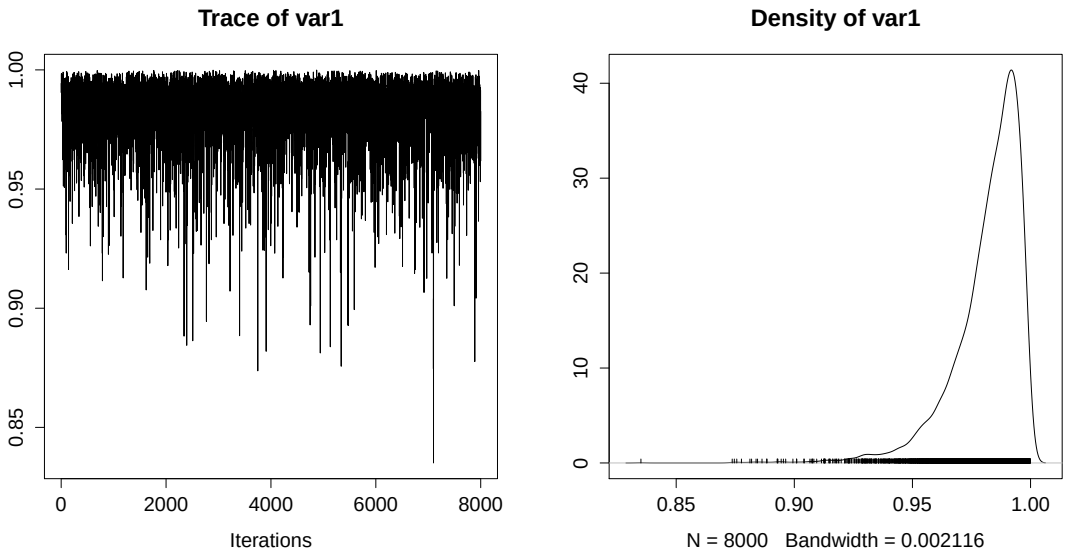

Figure 2: Posterior samples and density plot for ρ.

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.

5. Example 2 - identifying high-risk disease clusters

The second example illustrates the utility of the local CAR 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 in 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).

5.1. 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 **CARBayes** 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 are contained in `respiratorydata.csv`, Note, the shapefiles are those used for the property price analysis. These data sets can be read in using similar code to that presented in Section 4, and the respiratory disease data are read into a data frame called `respdata`. They can be viewed using `head(respdata)`, which gives the following output.

```
           observed2010 expected2010 incomedep2010
S02000618           105    105.12944            15
S02000613            85     69.41011            22
S02000623            37     87.85767             8
S02000626            90     89.41669            26
S02000636            41     97.55097             8
S02000645            47     84.86336             8
```

In common with the previous example these data are contained in the **CARBayes** package, and can be loaded into R using the `data()` function. They contain 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 `respdata` using the code below, which also creates the spatial objects that are required for the analysis (see Section 4 for details).

```
R> respdata$SIR2010 <- respdata$observed2010 / respdata$expected2010
R> data.combined <- combine.data.shapefile(data=respdata, shp=shp, dbf=dbf)
R> W.nb <- poly2nb(data.combined, row.names = rownames(respdata))
R> 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 4 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

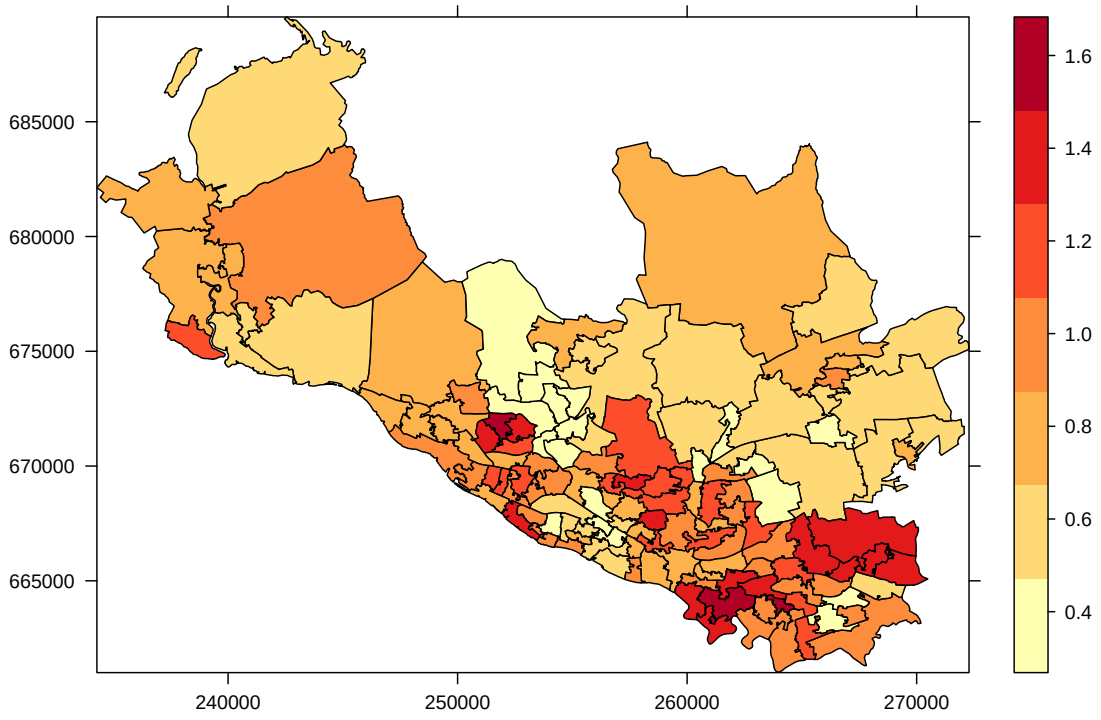

Figure 3: Map displaying the SIR for respiratory disease risk in the northern half of Greater Glasgow in 2010.

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`.

5.2. 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.

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

The function to implement the localised CAR model is called `poisson.dissimilarityCAR()`, 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.

```
R> form <- respdata$observed2010 ~ offset(log(respdata$expected2010))
R> model.dissimilarity <- dissimilarityCAR.re(formula=formula, family="poisson",
W=W.mat, Z=list(Z.incomedep=Z.incomedep), rho=0.99, burnin=20000,
n.sample=100000, thin=10)
```

Inference for this model is based on 8,000 MCMC samples, which were obtained by running the chain for 100,000 samples, with 20,000 being discarded as the burn-in period and the remaining 80,000 being thinned by 10 to reduce the autocorrelation. 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 $\mathbf{R} = (R_1, \ldots, R_n)$). The arguments `rho=0.99` and `fix.rho=TRUE` fix ρ to enforce strong global spatial autocorrelation, which is altered locally by estimating the elements of W as zero, for further details see Lee and Mitchell (2012). When the function finishes typing in `print(model.dissimilarity)` produces the following summary output.

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Localised CAR
```

```
Dissimilarity metrics -  Z.incomedep
Regression equation - respdata$observed2010 ~ offset(log(respdata$expected2010))

#############
#### Results
#############
Posterior quantiles and DIC

             Median    2.5%   97.5% n.sample % accept alpha.min
(Intercept) -0.2198 -0.2410 -0.1999     8000     60.5        NA
tau2         0.1351  0.0827  0.1934     8000    100.0        NA
Z.incomedep  0.0516  0.0470  0.0616     8000     61.0    0.0158

DIC =  1057.79       p.d =  99.39471
```

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$W.summary$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.

```
R> border.locations <- model.dissimilarity$W.summary$W.posterior
R> risk.estimates <- model.dissimilarity$fitted.values[ ,3] /
   respdata$expected2010
R> data.combined@data <- data.frame(data.combined@data, risk.estimates)
R> boundary.final <- highlight.borders(border.locations=border.locations,
   ID=rownames(respdata), shp=shp, dbf=dbf)
R> boundaries = list("sp.points", boundary.final, col="white", pch=19,
   cex=0.2)
R> northarrow <- list("SpatialPolygonsRescale", layout.north.arrow(),
   offset = c(220000,647000), scale = 4000)
R> scalebar <- list("SpatialPolygonsRescale", layout.scale.bar(),
   offset = c(225000,647000), scale = 10000, fill=c("transparent","black"))
R> text1 <- list("sp.text", c(225000,649000), "0")
R> text2 <- list("sp.text", c(230000,649000), "5000 m")
R> spplot(data.combined, c("risk.estimates"), sp.layout=list(northarrow,
   scalebar, text1, text2, boundaries), scales=list(draw = TRUE),
```

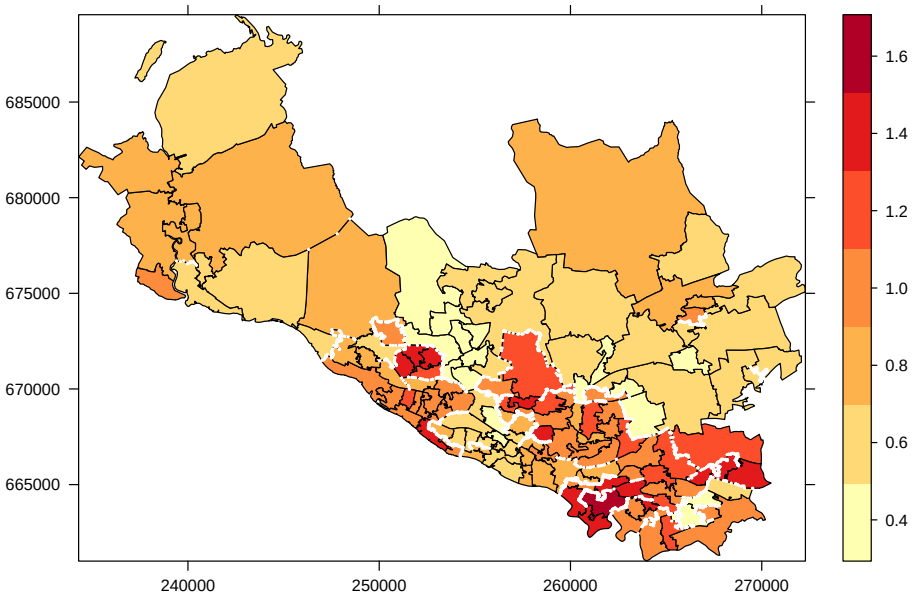

Figure 4: Map displaying the estimated spatial pattern in disease risk and the location of the boundaries

```
at=seq(min(risk.estimates)-0.1, max(risk.estimates)+0.1, length.out=8),
col.regions=c("#FFFFB2", "#FED976", "#FEB24C", "#FD8D3C", "#FC4E2A",
"#E31A1C", "#B10026"))
```

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. The result of these commands are displayed in Figure 4, which shows the fitted risk surface and the locations of the boundaries (denoted by white 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.

6. Discussion

This paper has presented 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 in 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 examples being Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi, and Songini (1995) and Knorr-Held (2000).

Acknowledgements

The data and shapefiles used in this paper were provided by the Scottish Government.

Funding

The research and the software package described in this paper were supported by the Economic and Social Research Council (ESRC), grant RES-000-22-4256.

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.

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

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*, **to appear**, DOI:10.1111/biom.12156.

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/