



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

Duncan Lee
University of Glasgow

Abstract

Conditional autoregressive models are commonly used to represent spatial autocorrelation in data relating to a set of non-overlapping areal units, which arise in a wide variety of applications including agriculture, education, epidemiology and image analysis. Such models are typically specified in a hierarchical Bayesian framework, with inference based on Markov chain Monte Carlo (MCMC) simulation. The most widely used software to fit such models is **WinBUGS** or **OpenBUGS**, but in this paper we introduce the R package **CARBayes**. The main advantage of **CARBayes** compared with the BUGS software is its ease of use, 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. This paper outlines the general class of Bayesian hierarchical models that can be implemented in the **CARBayes** software, describes their implementation via MCMC simulation techniques, and illustrates their use with two worked examples in the fields of house price analysis and disease mapping.

Keywords: Bayesian models, conditional autoregressive priors, **CARBayes**, R.

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 modeling such data, including ecological regression (see Wakefield 2007; Lee, Ferguson, and Mitchell 2009), disease mapping (see Green and Richardson 2002; 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 modeled 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 neighborhood effects, where subjects' behavior is influenced by that of neighboring 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 localized 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 localized 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) (modeling the adjacency structure of the areal units using logistic regression), Reich and Hodges (2008) (variable smoothing via a spatially varying variance in a spatio-temporal setting) and Lee and Mitchell (2012) (modeling the partial correlation between random effects in adjacent areal units as a function of their dissimilarity).

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 provided by 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** (Brezger, Kneib, and Lang 2005), while the R software (R Core Team 2013) has packages **CARramps** (for Gaussian data, Cowles and Bonett 2012), **hSDM** (for binomial data, Vieilledent, Latimer, Gelfand, Merow, Wilson, Mortier, and Jr. 2012) **spatcounts** (for count data including Poisson and zero-inflated Poisson distributions, Schabenberger 2009) and **spdep** (for Gaussian data, Bivand 2013) that can also implement a restricted set of CAR models. These mod-

els can also be implemented in R using integrated nested Laplace approximations (INLA, <http://www.R-INLA.org/>), using the package **INLA** (Rue, Martino, and Chopin 2009).

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 neighborhood matrix; and (2) given the neighborhood 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 modeling.

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 organized as follows. Section 2 outlines the general Bayesian hierarchical model that can be implemented in the **CARBayes** package, while Section 3 gives details about the software. Sections 4 and 5 give two worked examples of using the software, including how to create the neighborhood matrix and produce spatial maps of the results. Finally, Section 6 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 $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_n\}$, which are linked to a corresponding set of responses $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, and a vector of known offsets $\mathbf{O} = (O_1, \dots, O_n)^\top$. The spatial pattern in the response is modeled by a matrix of covariates $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top)^\top$ and a set of random effects $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$, 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 $\mathbf{x}_k^\top = (1, x_{k1}, \dots, x_{kp})$, the first of which corresponds to an intercept term. The general model that **CARBayes** can implement is an extension of a generalized linear model and is given by

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

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 $\mathbf{E}(Y_k) = \mu_k$, 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 $g(\cdot)$, which in this software is either the logit (binomial family), the identity (Gaussian family) or the natural log (Poisson family) function. The vector of regression parameters are denoted by $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$, and non-linear covariate effects can

be incorporated into the above model by including natural cubic spline or polynomial basis functions in \mathbf{X} .

2.2. Level 2: Prior distributions

An independent Gaussian prior is specified for each regression parameter β_j , that is $\beta_j \sim N(m_j, v_j)$ for $j = 0, \dots, p$, 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 uniform prior distribution, that is $\nu^2 \sim U(0, M_\nu)$, where the diffuse specification $M_\nu = 1000$ is the default value. We note that a commonly used alternative prior for variance parameters is the conjugate inverse-gamma distribution, but it is not used here because it is difficult to choose the hyperparameters so that it is non-informative for very small values of ν^2 (for details see [Gelman 2006](#)).

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

$$\begin{aligned}\theta_k &\sim N(0, \sigma^2), \\ \sigma^2 &\sim U(0, M_\sigma),\end{aligned}\tag{2}$$

where θ_k replaces ϕ_k in the data likelihood (1). The variance parameter is assigned a uniform prior on the interval $(0, M_\sigma)$, where as before the default value is $M_\sigma = 1000$. This specification 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 CAR priors

Four different CAR priors are commonly used for modeling 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 $\phi \sim N(\mathbf{0}, \tau^2 Q^{-1})$, 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 $n \times n$ neighborhood 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 units 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 $f(\phi_k | \phi_{-k})$ for $k = 1, \dots, n$ (where $\phi_{-k} = (\phi_1, \dots, \phi_{k-1}, \phi_{k+1}, \dots, \phi_n)$), 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

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

The conditional expectation is the average of the random effects in neighboring areas, while the conditional variance is inversely proportional to the number of neighbors. The latter is appropriate because if the random effects are spatially correlated, then the more neighbors an area has the more information there is from its neighbors about the value of its random effect. In common with the other variance parameters, τ^2 is assigned a uniform prior on the interval $(0, M_\tau)$, with the default value being $M_\tau = 1000$. 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 CAR 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 modeling varying strengths of spatial autocorrelation, using only a single set of random effects. The model by [Leroux *et al.* \(1999\)](#) is given by

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

while the proposal of [Stern and Cressie \(1999\)](#) is

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

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 $\rho \sim U(0, 1)$, while the usual uniform prior on the interval $(0, M_\tau)$ is adopted for τ^2 . In both cases 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.

Local 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

$$\text{COR}(\phi_k, \phi_j | \phi_{-kj}) = \frac{\rho w_{kj}}{\sqrt{(\rho \sum_{i=1}^n w_{ki} + 1 - \rho)(\rho \sum_{i=1}^n w_{ji} + 1 - \rho)}}. \quad (6)$$

For non-neighboring areas (where $w_{kj} = 0$) the random effects are conditionally independent, while for neighboring 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 localized 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 modeling process. In contrast, if w_{kj} is estimated as zero then no smoothing is imparted between (ϕ_k, ϕ_j) , as they are modeled 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 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ 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 estimate ρ in this model). 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 neighboring populations are very different. This dissimilarity is captured by q non-negative dissimilarity metrics $\mathbf{z}_{kj} = (z_{kj1}, \dots, z_{kjq})$, 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 modeled as

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

$$\alpha_i \sim U(0, M_i) \quad \text{for } i = 1, \dots, q.$$

The q regression parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)$ 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 i th 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 area centroids, would be included in (7).

3. CARBayes

3.1. Obtaining the software

The CARBayes software (Lee 2013) is an add-on package to the statistical software R ($\geq 2.10.0$), and is freely available from the Comprehensive R Archive Network (CRAN, <http://CRAN.R-project.org/package=CARBayes>). In addition to the base implementation of R, it requires the following packages: MASS (Venables and Ripley 2002), coda (Plummer, Best, Cowles, and Vines 2006), spam (Furrer and Sain 2010) and truncdist (Novomestky and Nadarajah 2012). 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 attached or their namespace loaded when package **CARBayes** is loaded, 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 neighborhood 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** (Canty and Ripley 2013; Davison and Hinkley 1997), **deldir** (Turner 2013), **foreign**, **grid**, **maptools** (Bivand and Lewin-Koh 2013), **Matrix** (Bates and Mächler 2013), **nlme** (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2013), **shapefiles** (Stabler 2013), **sp** (Bivand, Pebesma, and Gómez-Rubio 2013), **spdep** and **splines**. These packages have also been loaded for the analyses presented in Sections 4 and 5.

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 names of the functions have the form of `'poisson.lerouxCAR'`, where the first part specifies the likelihood model while the second part after the `'.'` specifies the random effects prior model. The prior models listed below can be implemented by the software, where the `'dist'` in the function name should be replaced by one of `'binomial'`, `'gaussian'` or `'poisson'`.

1. `dist.independent()` – the independence model given by (2).
2. `dist.iarCAR()` – the intrinsic autoregressive model proposed by Besag *et al.* (1991) and given by (3).
3. `dist.bymCAR()` – the BYM model proposed by Besag *et al.* (1991) and given by a linear combination of (2) and (3).
4. `dist.lerouxCAR()` – the CAR prior proposed by Leroux *et al.* (1999) and given by (4).
5. `dist.properCAR()` – the CAR prior proposed by Stern and Cressie (1999) and given by (5).
6. `dist.dissimilarityCAR()` – the local spatial smoothing model proposed by Lee and Mitchell (2012) and given by (4) and (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. The spatial neighborhood information required to run the CAR models needs to be provided as an $n \times n$ neighborhood 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 prints a message to the R console after every 1,000 MCMC iterations, which allows the user to monitor the function's progress. If the fitted model is printed, summary results are shown including 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 modeling 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 Neighborhood 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 value' (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, i.e., using the following code:

```
R> data("housedata", package = "CARBayes")
R> data("shp", package = "CARBayes")
R> data("dbf", package = "CARBayes")
```

The structure of `housedata` is shown below using the `head()` function, and with the above `read.csv()` command, the unique identifier (IG) has been turned into the row names of the data frame.

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


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.

	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 summarized in Table 1, which displays the percentiles of their distributions. 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 British pounds 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 1,994 with the addition of a single large outlier of 8,009. The location of this outlier is the city center of Glasgow, and the high crime rate is likely to be caused by the large number 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 center. 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)
```

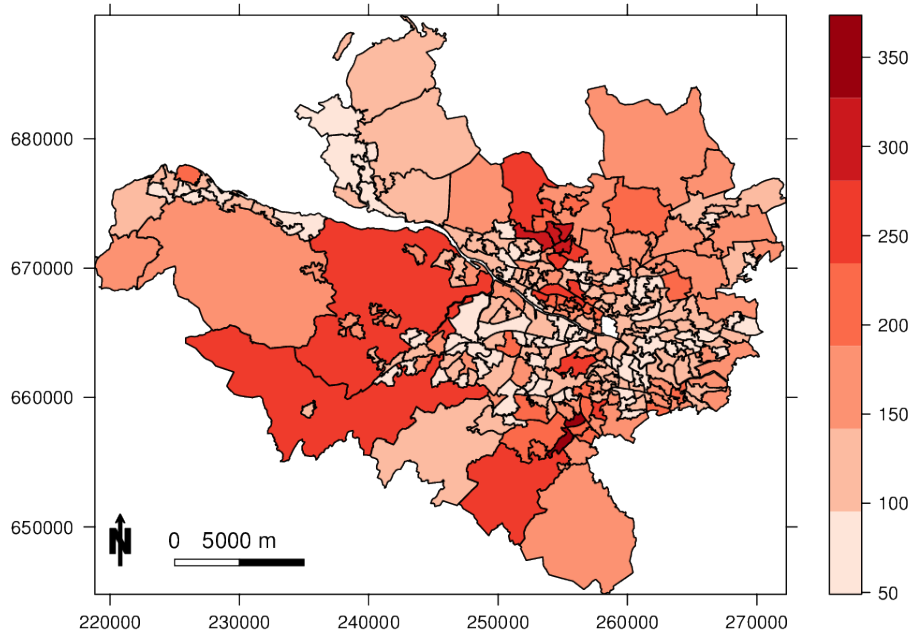


Figure 1: Map displaying median property prices in Greater Glasgow (in thousands).

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, "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"))
```

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 neighboring 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 modeling

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 center are linearly related to the response, and the transformation of the variables is achieved with the following commands.

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

In the fitted 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 <- paste("logprice ~ logcrime + rooms + sales + factor(type) +",
+ "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`’ object inheriting from class ‘`nb`’, 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 modeling 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 Section 3. 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 neighborhood matrix `W.mat` from the `W.nb` object.

```
R> W.mat <- nb2mat(W.nb, style = "B")
R> model.spatial <- gaussian.properCAR(as.formula(form), data = housedata,
+   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 result is printed, it 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 summarizes the parameters (except for the random effects) by means of posterior medians, 95% credible intervals, and acceptance rates.

```
R> model.spatial

#####
#### Model fitted
#####
Likelihood model - Gaussian (identity link function)
Random effects model - Proper CAR
Regression equation - logprice ~ logcrime + rooms + sales + factor(type) +
logdriveshop

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

              Median    2.5%   97.5% n.sample % accept
(Intercept)      4.7531  4.2710  5.2326   8000    100
logcrime         -0.1114 -0.1721 -0.0508   8000    100
rooms            0.2225  0.1728  0.2731   8000    100
sales            0.0023  0.0017  0.0029   8000    100
factor(type)flat -0.2547 -0.3640 -0.1399   8000    100
factor(type)semi -0.1623 -0.2625 -0.0656   8000    100
factor(type)terrace -0.2900 -0.4144 -0.1661   8000    100
logdriveshop     -0.0019 -0.0577  0.0553   8000    100
nu2              0.0239  0.0144  0.0332   8000    100
tau2             0.0512  0.0239  0.0983   8000    100
```

```
rho          0.9853  0.9420  0.9979    8000    60
DIC = -153.3135      p.d = 96.831
```

Model output

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

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

	Length	Class	Mode
formula	3	formula	call
samples	5	-none-	list
fitted.values	1350	-none-	numeric
random.effects	1350	-none-	numeric
residuals	1350	-none-	numeric
W.summary	72900	-none-	numeric
DIC	1	-none-	numeric
p.d	1	-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 five elements are matrices containing the thinned and post burn-in MCMC samples for each set of parameters. For example, `samples.beta` is an $8,000 \times 8$ matrix containing the MCMC samples for all the regression parameters. The next three elements in the list `fitted.values`, `random.effects` and `residuals` comprise matrices of dimension $n \times 5$ (here $n = 270$), which summarize the posterior distribution of the fitted values, random effects 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 DIC element displays the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, and Van der Linde 2002), which is a Bayesian measure of overall model fit used for model comparison. This quantity trades off the overall fit to the data against the effective number of parameters in the model, in a similar way to the AIC and BIC criteria. The list also contains `p.d`, which is the estimated effective number of parameters in the model. The DIC criterion is used for comparing the overall fit of multiple models applied to the same data, and lower values indicate a better fitting model. For further details about Bayesian modeling see Gelman, Carlin, Stern, and Rubin (2003).

Parameter estimates

The printed 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 center, as their 95% credible intervals do not include zero. For example, increasing the average number of rooms by one is estimated to increase the average property price by 24.9%, because the ratio of the average

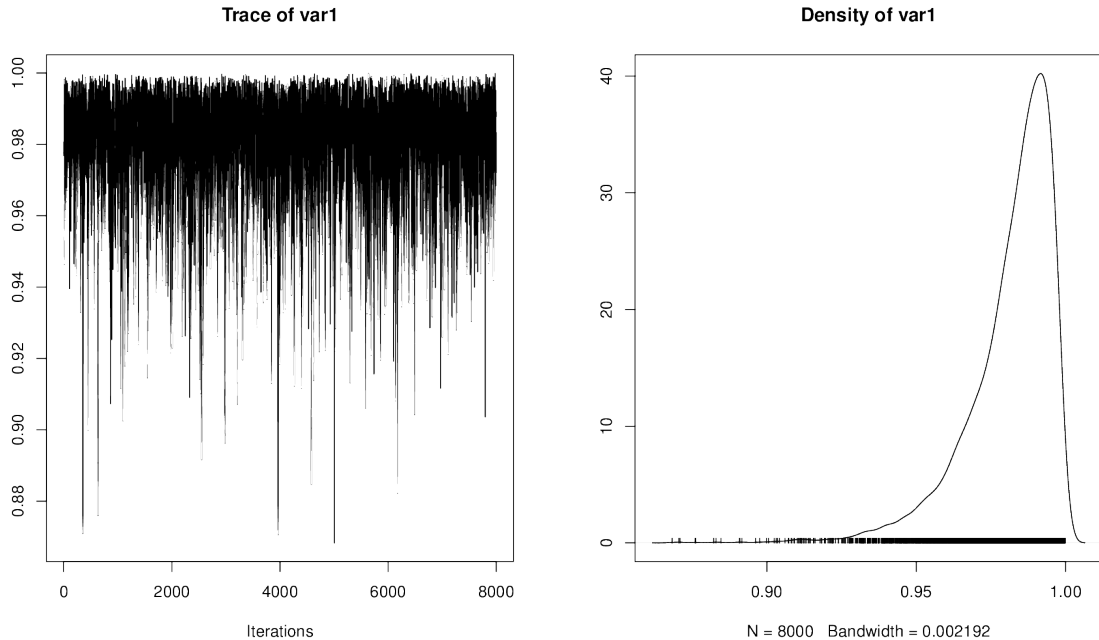


Figure 2: Posterior samples and density plot for ρ .

property prices that differ only in having m and $m + 1$ rooms is equal to $\exp(0.2224) = 1.249$. Similarly, IGs that predominately comprise flats have lower median property prices by around 22.4% ($1 - \exp(-0.2542) = 0.224$), compared with the baseline category of ‘detached’. The above output also shows that the random effects have modeled substantial spatial autocorrelation, as the posterior median for the spatial autocorrelation parameter ρ is 0.9852. The entire posterior distribution (as summarized 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 returned as an ‘mcmc’ object, which is defined in 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 (β), 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 center is artificial, as both estimates are very close to zero

Parameter	CARBayes	BUGS	% difference
logcrime	-0.1111	-0.1132	1.8
rooms	0.2224	0.2221	0.1
sales	0.0023	0.0023	0.0
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
ν^2	0.0237	0.0247	4.0
τ^2	0.0518	0.0447	13.7
ρ	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.

(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**. Currently, the **CARBayes** package is much slower than the **BUGS** software (the model ran in this section runs 17 times faster in **BUGS**), but a re-engineering of **CARBayes** using C++ is planned for the near future, which will make the speeds more comparable.

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 $(\beta, \phi, \nu^2, \tau^2)$, 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 (β, ϕ) 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 the file `respiratorydata.csv`. Note, the shapefiles are those used for the property price analysis. These data sets can be read in using code similar 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 standardization. For these data we use external standardization, based on age and sex standardized rates for the whole of Scotland. These expected numbers are stored in the `expected2010` column, and the simplest measure of disease risk is the standardized 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 localized spatial structure in these disease data, with numerous different locations where high and low risk areas border each other. This

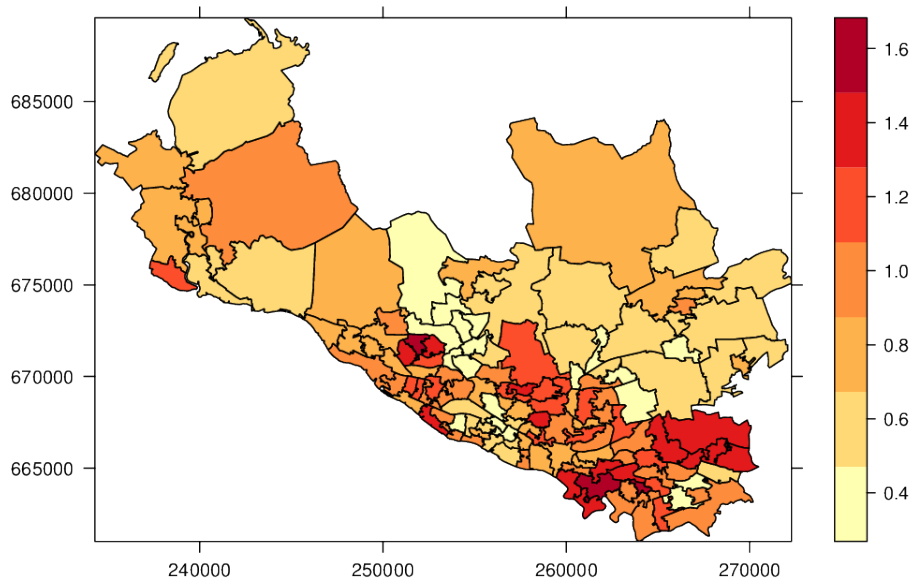


Figure 3: Map displaying the SIR for respiratory disease risk in the Northern half of Greater Glasgow in 2010.

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 (i.e., 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`.

5.2. Spatial modeling with CARBayes

Let the observed and expected numbers of hospital admissions be denoted by $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{E} = (E_1, \dots, E_n)$ 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 localized 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 <- "observed2010 ~ offset(log(expected2010))"
R> model.dissimilarity <- poisson.dissimilarityCAR(as.formula(form),
+   data = respdata, W = W.mat, Z = list(Z.income = Z.income), rho = 0.99,
+   fix.rho = TRUE, 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 $\mathbf{R} = (R_1, \dots, R_n)$). 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 \(2012\)](#). Printing the result produces the following summary output.

```
R> model.dissimilarity

#####
#### Model fitted
#####
Likelihood model - Poisson (log link function)
Random effects model - Localised CAR
Dissimilarity metrics - Z.incomedep
Regression equation - observed2010 ~ offset(log(expected2010))

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

              Median    2.5%   97.5% n.sample % accept alpha.min
(Intercept) -0.2202 -0.2410 -0.1996   8000   61.4      NA
tau2         0.1383  0.0836  0.1982   8000  100.0      NA
Z.incomedep  0.0516  0.0467  0.0621   8000   61.3   0.0158

DIC = 1057.11      p.d = 99.54436
```

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 summarized in the element of the output list called `W.posterior` (obtained with the code `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, "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 = 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 demarcation between the low risk city center/west end of Glasgow in the middle of the region and the deprived neighboring 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

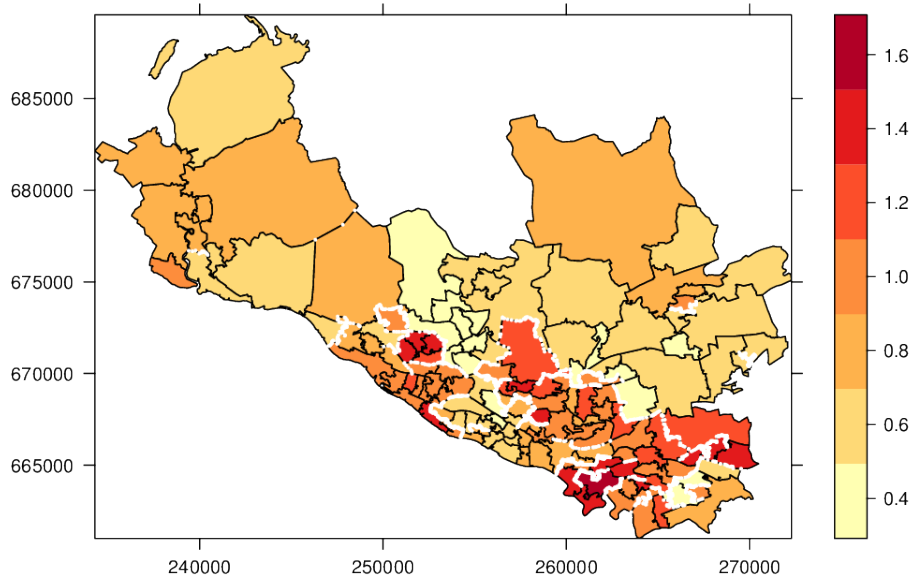


Figure 4: Map displaying the estimated spatial pattern in disease risk and the location of the boundaries.

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 CAR models to spatial areal unit data, as well as the localized 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 Neighborhood 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 fueled a growth in modeling 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 CAR 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.

As previously mentioned, future development for the software will re-engineer it in C++ (currently it is written exclusively in R), which should result in a dramatic reduction in the computing time required to fit the models. In addition, 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**, Bakar and Sahu 2013). The development of statistical modeling 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).

Acknowledgments

The author gratefully acknowledges the valuable comments and suggestions made by both the editor and the referees, which have significantly improved this paper. The data and shapefiles used in this paper were provided by the Scottish Government.

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

References

- Bakar KS, Sahu SK (2013). *spTimer: Spatio-Temporal Bayesian Modelling Using R*. R package version 0.8, URL <http://CRAN.R-project.org/package=spTimer>.
- Bates D, Mächler M (2013). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.0-14, URL <http://CRAN.R-project.org/package=Matrix>.
- 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**(21–22), 2433–2443.
- Besag J, Higdon D (1999). “Bayesian Analysis of Agricultural Field Experiments.” *Journal of the Royal Statistical Society B*, **61**(4), 691–746.
- Besag J, York J, Mollié A (1991). “Bayesian Image Restoration with Two Applications in Spatial Statistics.” *The Annals of the Institute of Statistics and Mathematics*, **43**(1), 1–59.
- Bivand R (2013). *spdep: Spatial Dependence: Weighting Schemes, Statistics and Models*. R package version 0.5-65, URL <http://CRAN.R-project.org/package=spdep>.
- Bivand R, Lewin-Koh N (2013). *mapttools: Tools for Reading and Handling Spatial Objects*. R package version 0.8-27, URL <http://CRAN.R-project.org/package=maptools>.
- Bivand RS, Pebesma E, Gómez-Rubio V (2013). *Applied Spatial Data Analysis with R*. 2nd edition. Springer-Verlag.
- Brewer M, Nolan A (2007). “Variable Smoothing in Bayesian Intrinsic Autoregressions.” *Environmetrics*, **18**(8), 841–857.

- Brezger A, Kneib T, Lang S (2005). *BayesX: Analyzing Bayesian Structured Additive Regression Models*. URL <http://www.jstatsoft.org/v14/i11/>.
- Canty A, Ripley BD (2013). *boot: Bootstrap R (S-PLUS) Functions*. R package version 1.3-9, URL <http://CRAN.R-project.org/package=boot>.
- Cowles K, Bonett S (2012). *CARramps: Reparameterized and Marginalized Posterior Sampling for Conditional Autoregressive Models*. R package version 0.1.2, URL <http://CRAN.R-project.org/package=CARramps>.
- Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.
- Furrer R, Sain SR (2010). “**spam**: A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields.” *Journal of Statistical Software*, **36**(10), 1–25. URL <http://www.jstatsoft.org/v36/i10/>.
- Gavin J, Jennison C (1997). “A Subpixel Image Restoration Algorithm.” *Journal of Computational and Graphical Statistics*, **6**(2), 182–201.
- Gelman A (2006). “Prior Distributions for Variance Parameters in Hierarchical Models.” *Bayesian Analysis*, **1**(3), 515–533.
- 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**(420), 1055–1070.
- Knorr-Held L (2000). “Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk.” *Statistics in Medicine*, **19**(17–18), 2555–2567.
- Lawson A, Clark A (2002). “Spatial Mixture Relative Risk Models Applied to Disease Mapping.” *Statistics in Medicine*, **21**(3), 359–370.
- Lee D (2011). “A Comparison of Conditional Autoregressive Models Used in Bayesian Disease Mapping.” *Spatial and Spatio-Temporal Epidemiology*, **2**(2), 79–89.
- Lee D (2013). *CARBayes: Spatial Areal Unit Modelling*. R package version 1.6, URL <http://CRAN.R-project.org/package=CARBayes>.
- Lee D, Ferguson C, Mitchell R (2009). “Air Pollution and Health in Scotland: A Multicity Study.” *Biostatistics*, **10**(3), 409–423.
- Lee D, Mitchell R (2012). “Boundary Detection in Disease Mapping Studies.” *Biostatistics*, **13**(3), 415–426.
- Leroux B, Lei X, Breslow N (1999). “Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence.” In ME Halloran, D Berry (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 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**(4), 433–452.
- Lunn D, Spiegelhalter D, Thomas A, Best N (2009). “The BUGS Project: Evolution, Critique and Future Directions.” *Statistics in Medicine*, **28**(25), 3049–3082.
- Ma H, Carlin B (2007). “Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis.” *Bayesian Analysis*, **2**(2), 281–302.
- Novomestky F, Nadarajah S (2012). *truncdist: Truncated Random Variables*. R package version 1.0-1, URL <http://CRAN.R-project.org/package=truncdist>.
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-111, URL <http://CRAN.R-project.org/package=nlme>.
- Plummer M, Best N, Cowles K, Vines K (2006). “**coda**: Convergence Diagnosis and Output Analysis for MCMC.” *R News*, **6**(1), 7–11. URL <http://CRAN.R-project.org/doc/Rnews/>.
- R Core Team (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**(3), 790–799.
- Rue H, Martino S, Chopin N (2009). “Approximate Bayesian Inference for Latent Gaussian Models using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society B*, **71**(2), 319–392.
- Schabenberger H (2009). *spatcounts: Spatial Count Regression*. R package version 1.1, URL <http://CRAN.R-project.org/package=spatcounts>.
- Spiegelhalter D, Best N, Carlin B, Van der Linde A (2002). “Bayesian Measures of Model Complexity and Fit.” *Journal of the Royal Statistical Society B*, **64**(4), 583–639.
- Stabler B (2013). *shapefiles: Read and Write ESRI Shapefiles*. R package version 0.7, URL <http://CRAN.R-project.org/package=shapefiles>.
- Stern H, Cressie N (1999). “Inference for Extremes in Disease Mapping.” In AB Lawson, A Biggeri, D Böhning, E Lesaffre, JF Viel, R Bertollini (eds.), *Disease Mapping and Risk Assessment for Public Health*, pp. 63–84. John Wiley & Sons.
- Turner R (2013). *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation*. R package version 0.1-1, URL <http://CRAN.R-project.org/package=deldir>.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.
- Vieilledent G, Latimer AM, Gelfand AE, Merow C, Wilson AM, Mortier F, Jr JAS (2012). *hSDM: Hierarchical Bayesian Species Distribution Models*. R package version 1.0, URL <http://CRAN.R-project.org/package=hSDM>.

- Wakefield J (2007). “Disease Mapping and Spatial Regression with Count Data.” *Biostatistics*, **8**(2), 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**(2), 311–324.
- Womble W (1951). “Differential Systematics.” *Science*, **114**(2961), 315–322.

Affiliation:

Duncan Lee
School of Mathematics and Statistics
15 University Gardens
University of Glasgow
Glasgow
G12 8QQ, United Kingdom
E-mail: Duncan.Lee@glasgow.ac.uk
URL: <http://www.gla.ac.uk/schools/mathematicsstatistics/staff/duncanlee/>