Gaussian Markov random field spatial models in GAMLSS

Fernanda De Bastiani, Robert A. Rigby, Dimitrios M. Stasinopoulous, Audrey H.M.A. Cysneiros & Miguel A. Uribe-Opazo

Gaussian Markov random field spatial models in GAMLSS

Fernanda De Bastiani[a,b], Robert A. Rigby[c], Dimitrios M. Stasinopoulous[c], Audrey H.M.A. Cysneiros[a] and Miguel A. Uribe-Opazo[d]

[a]Department of Statistics, Universidade Federal de Pernambuco, Recife/PE, Brazil; [b]Department of Statistics, Pontificia Universidad Catolica de Chile, Santiago, Chile; [c]STORM, London Metropolitan University, London, UK; [d]Postgraduate program of agricultural engineering, Universidade Estadual do Oeste do Paraná, Cascavel/PR, Brazil

ABSTRACT

This paper describes the modelling and fitting of Gaussian Markov random field spatial components within a Generalized Additive-Model for Location, Scale and Shape (GAMLSS) model. This allows modelling of any or all the parameters of the distribution for the response variable using explanatory variables and spatial effects. The response variable distribution is allowed to be a non-exponential family distribution. A new package developed in R to achieve this is presented. We use Gaussian Markov random fields to model the spatial effect in Munich rent data and explore some features and characteristics of the data. The potential of using spatial analysis within GAMLSS is discussed. We argue that the flexibility of parametric distributions, ability to model all the parameters of the distribution and diagnostic tools of GAMLSS provide an ideal environment for modelling spatial features of data.

1. Introduction

Since the introduction of the Generalized Additive Model for Location, Scale and Shape (GAMLSS) by Rigby and Stasinopoulos [20], the models have been used in a variety of different fields, such as actuarial science [13], biology, biosciences, energy economics [27], genomics [14], finance, fisheries, food consumption, growth curves estimation [6], and Multicentre Growth Reference Study Group [17,18], marine research, medicine, meteorology, rainfall, vaccines and film studies, [28].

Discrete spatial variation, where the variables are defined on discrete domains, such as regions, regular grids or lattices, can be modelled by Markov random fields (MRF). MRF can be applied in different areas, such as spatial statistics, image analysis, structural time-series analysis, analysis of longitudinal and survival data, spatio-temporal statistics, graphical models and semiparametric models.

Kunsch [16] present many important results for Gaussian Markov random fields (GMRF). Extensive theoretical and practical details of GMRF are provided by Rue and Held [24]. In statistics Besag and Kooperberg [5] considered the Gaussian intrinsic

autoregressive model (IAR), a very important specific case of GMRF models. Wood [30] presents IAR models within a generalized additive model (GAM) framework. There are few papers that use GAMLSS in a spatial framework. Rigby *et al.* [22] commented on the paper 'Beyond mean regression', [15], and presented a simplified analysis of Munich rent data with very few covariates, modelling the μ parameter with a spatial effect using an IAR model term. In this paper we describe in detail the theoretical basis of the GAMLSS implementation of GMRF, develop a package in `R` [10,19] to achieve this and explore the potential of such modelling using the Munich rent data.

Section 2 discusses GAMLSS models and the modelling and fitting of GMRF spatial components within GAMLSS models. In Section 3 we present the full Munich rent data set, the strategy to choose a model and the results for the Munich rent data set. Section 4 investigates the adequacy of the chosen model using residual diagnostic worm plots. The implementation in `R` is described in the appendix and the `R` code used in the analysis is available from the authors at `www.gamlss.org`. Section 5 presents relevant conclusions.

2. Methodology

Section 2.1 defines the GAMLSS framework, while Section 2.2 describes its estimation procedure. Section 2.3 describes how the GMRF models can be incorporated within the GAMLSS framework.

2.1. The GAMLSS framework

GAMLSS provides a very general and flexible system for modelling a response variable. The distribution of the response variable is selected by the user from a very wide range of distributions available in the `gamlss` package in `R`, [20], including highly skewed and kurtotic continuous and discrete distributions. The `gamlss` package includes distributions with up to four parameters, denoted by μ, σ, ν and τ, which usually represent the location (e.g. mean), scale (e.g. standard deviation), and skewness and kurtosis shape parameters, respectively. All the parameters of the response variable distribution can be modelled using parametric and/or non-parametric smooth functions of explanatory variables, thus allowing modelling of the location, scale and shape parameters. Specifically, a GAMLSS model assumes that, for $i = 1, 2, \ldots, n$, independent observations Y_i have probability (density) function $f_Y(y_i|\boldsymbol{\theta}^i)$ conditional on $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})^\top = (\mu_i, \sigma_i, \nu_i, \tau_i)^\top$ a vector of four distribution parameters, each of which can be a function of the explanatory variables. Rigby and Stasinopoulos [20] define an original formulation of a GAMLSS model as follows.

For $k = 1,2,3,4$, let $g_k(.)$ be a known monotonic link function relating the distribution parameter $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kn})^\top$ to predictor $\boldsymbol{\eta}_k = (\eta_{k1}, \ldots, \eta_{kn})^\top$. Then we set

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{1}$$

where $\mathbf{X}_k$ is a known design matrix, $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{J'_k k})^\top$ is a parameter vector of length J'_k, h_{jk} is a smooth nonparametric function of variable X_{jk} and the $\mathbf{x}_{jk}$'s are vectors of length n, for $k = 1,2,3,4$ and $j = 1, \ldots, J_k$.

Model (1) can be written in a random effects form and random effects can also be included in the model for the $n \times 1$ vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\tau}$:

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}, \end{aligned} \tag{2}$$

where here the random effects parameters $\boldsymbol{\gamma}_{jk}$ are assumed to have independent (prior) normal distributions with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \lambda_{jk}^{-1}\mathbf{G}_{jk}^{-1})$ and $\mathbf{G}_{jk}^{-1}$ is the (generalized) inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk}$, where if $\mathbf{G}_{jk}$ is singular then $\boldsymbol{\gamma}_{jk}$ has an improper prior density function proportional to $\exp(-\frac{1}{2}\lambda_{jk}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk})$. Note that the conditional distribution of the response variable (given the random effects parameters) can be any distribution, exponential family or non-exponential family, while the random effects parameters are Gaussian.

Different formulations of the $\mathbf{Z}$'s and the $\mathbf{G}$'s result in different types of additive terms, for example, random effects terms, smoothing terms, time-series terms or spatial terms as presented in Section 2.3. The advantage of modelling spatial data within GAMLSS is that different distributions beside the exponential family can be fitted and also it is possible, if needed, to model spatially any or all the parameters of the distribution.

2.2. Estimation of the model

The log-likelihood function for the GAMLSS model (2) under the assumption that observations of the response variable are independent is given by

$$\ell = \sum_{i=1}^{n} \log f_Y(y_i \mid \mu_i, \sigma_i, \nu_i, \tau_i),$$

where $f_Y(\cdot)$ represents the probability (density) function of the response variable. The penalized log-likelihood function for model (2) is given by

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \lambda_{jk}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk}. \tag{3}$$

We will need estimates for the 'betas', $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^{\top}$, the parameters of the linear part of the model, the 'gammas', $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}, \ldots, \boldsymbol{\gamma}_{J_1 1}, \boldsymbol{\gamma}_{12}, \ldots, \boldsymbol{\gamma}_{J_4 4})^{\top}$, the random effects

parameters, and the 'lambdas' $\boldsymbol{\lambda} = (\lambda_{11}, \ldots, \lambda_{J_1 1}, \lambda_{12}, \ldots, \lambda_{J_4 4})^\top$, the hyper-parameters of the model.

Within the GAMLSS framework the linear parameters $\boldsymbol{\beta}$ and the random effects parameters $\boldsymbol{\gamma}$ are estimated (for fixed values of the smoothing hyper-parameters $\boldsymbol{\lambda}$) by maximizing the penalized likelihood function ℓ_p given by Equation (3). There are two basic algorithms to achieve this, the `RS` and the `CG` algorithms. Both use an iteratively reweighted (penalized) least-squares algorithm. Appendix C of Rigby and Stasinopoulos [20] shows that both algorithms lead, for given $\boldsymbol{\lambda}$ hyper-parameters, to the maximum penalized log-likelihood estimates for the betas and the gammas, i.e. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$. Appendix A.1 of Rigby and Stasinopoulos [20] shows that these estimates are also posterior mode (or MAP) estimates. The hyper-parameters $\boldsymbol{\lambda}$ can be estimated locally, see [21], or globally, see [20]. The local methods are in general a lot faster and easier to implement than the global ones. 'Local' means that the method of estimation of the hyper-parameters applies each time within the `RS` or `CG` GAMLSS algorithms and 'global' means the method is applied outside the `RS` or `CG` GAMLSS algorithms. In addition, for either 'local' or 'global' estimation, there are (at least) three different criteria for estimating the smoothing hyper-parameters:

- minimizing the generalized cross validation (GCV), Wood [30],
- minimizing the generalized Akaike information criteria (GAIC), Akaike [1],
- maximum likelihood (ML).

The default method in the GAMLSS software implementation is 'local ML' in which the (smoothing) hyper-parameters (and therefore their corresponding effective degrees of freedom) are estimated automatically using a local maximum likelihood (ML) procedure, see [21]. (This 'local ML' procedure is a penalized quasi-likelihood (PQL) method, Breslow and Clayton [7].)

2.3. Gaussian Markov random fields

An MRF is a set of random variables where a local defined assumption is used to determine their joint (or global) distribution, [2, Section 3]. Their local behaviour is described through Markov properties based on conditional independence assumptions. For example by studying different areas on a map, we would expect neighbourhood areas to be more similar that others far apart. In this case a conditional independence assumption that given the neighbours the occurrence of an event in the area is independent from the event occurring in other areas seems reasonable. Those Markovian assumptions can be presented as an undirected graph $\mathcal{G}$, where each vertex represents an areal unit and each edge connects two areal units and represents a neighbouring relationship, Rue and Held [24]. Areal data are sometimes called lattice data, and often the lattice is a 2-dimensional grid in the plane, either finite or infinite.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph [12,29] that consists of vertices $\mathcal{V} = (1, 2, \ldots, q)$, and a set of edges $\mathcal{E}$, where a typical edge is (m, t), $m, t \in \mathcal{V}$. Undirected is in the sense that (m, t) and (t, m) refer to the same edge. Following Rue and Held [24], a random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ is called a Gaussian MRF (i.e. GMRF) with respect to the graph

$\mathcal{G}$, with mean $\boldsymbol{\mu}$ and (symmetric) precision matrix $\lambda\mathbf{G}$, if and only if its density has the form

$$\pi(\boldsymbol{\gamma}) \propto \exp\left[-\frac{1}{2}\lambda(\boldsymbol{\gamma}-\boldsymbol{\mu})^\top \mathbf{G}(\boldsymbol{\gamma}-\boldsymbol{\mu})\right] \tag{4}$$

and

$$G_{mt} \neq 0 \Longleftrightarrow (m,t) \in \mathcal{E} \quad \text{for } m \neq t,$$

where G_{mt} is the element of matrix $\mathbf{G}$ for row m and column t.

Hence the nonzero pattern of $\mathbf{G}$ determines $\mathcal{G}$. We can read off from $\mathbf{G}$ whether γ_m and γ_t are conditionally independent, because a well-known theorem in this field (Theorem 3.2 of Rue and Held [24]) says that γ_m and γ_t are conditionally independent, given γ_r for all r not equal to m or t, if and only if $G_{mt} = 0$. It also means in practice that the precision matrix is often sparse.

The conditional autoregressive (CAR) model first introduced by Besag [3] and also described in detail in [5] is a GMRF model of the form given in Equation (4) where $\mathbf{G}$ is a non-singular matrix. The CAR model can also be specified by the local definition:

$$\gamma_i|\boldsymbol{\gamma}_{-i} \sim N\left(\sum_j \alpha_{ij}\gamma_j, k_i\right),$$

where $\boldsymbol{\gamma}_{-i} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_q)$ and $\alpha_{ii} = 0$, $\alpha_{ij} = -G_{ij}/G_{ii} (i \neq j)$ and $k_i = 1/(\lambda G_{ii})$ for $i = 1, 2, \ldots, q$. Then using Brook's lemma [8], it can be shown that the joint distribution for $\boldsymbol{\gamma}$ is of the form given in Equation (4) with $\boldsymbol{\mu} = \mathbf{0}$, provided $\alpha_{ij}k_j = \alpha_{ji}k_i$ for all i and j to ensure that matrix $\mathbf{G}$ is symmetric [5]. Specific models in which the precision $\mathbf{G}$ matrix in Equation (4) is singular are called the IAR model. That is, the IAR is a limiting case of CAR in which $\mathbf{G}$ is singular. The IAR model has been used for spatially structured random effects in generalized linear models [2]. Rue and Held [24,30] present the IAR model within a GAM framework. In this context we extend to the GAMLSS framework.

The following shows how a specific IAR model is incorporated in GAMLSS. Let $\mathbf{W}$ be the proximity matrix (which we assume to be symmetric) where the elements w_{ii} are set to 0 and $w_{ij} = 1$ if i and j $(i \neq j)$ share some common boundary, or 0 otherwise. The precision $\mathbf{G}$ matrix can be constructed from $\mathbf{D}_w - \mathbf{W}$, where $\mathbf{D}_w$ is a diagonal matrix with each element in the diagonal being the respective row sum of the proximity matrix. For example, consider the five areas represented in the left size of Figure 1, which is a subsample of the areas in the Munich rent data example analyzed in Section 3. The conditional independence assumptions in this case is described by the undirected graph in the right size of Figure 1. The graph shows that neighbouring areas are connected.

For this example the proximity matrix $\mathbf{W}$ and the $\mathbf{D}_w$ and $\mathbf{G}$ matrices are given by

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{D}_w = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{bmatrix}.$$

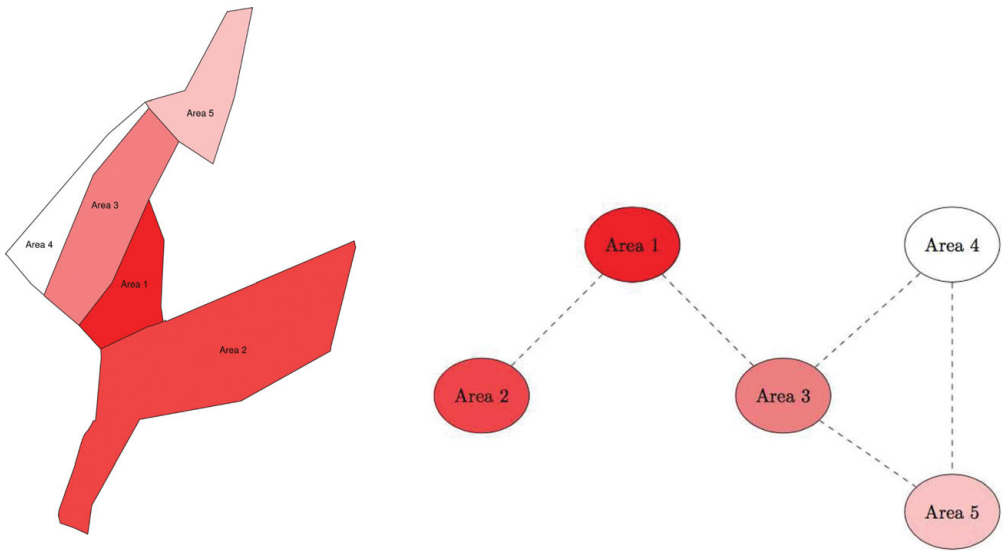

Figure 1. Showing, on the left of the figure, a subset of five regions of the Munich rent data example of Section 3 and, on the right, an undirected graph describing the conditional independence relationship between the five regions.

The effect of the **G** matrix is to bring fitted values from neighbouring regions closer together (rather than to shrink them towards the overall mean as is the case of a simple random effect model term). Note that the matrix **G** is treated within GAMLSS as an extra penalty in the penalized log-likelihood given in Equation (3).

Assume that a response variable and explanatory variables are recorded at observations which belong spatially to one of a set of areas (or regions). Zero, one or more than one observation may be recorded in each region. To incorporate IAR models within the GAMLSS model (2), set $\mathbf{Z}$ to be an index matrix defining which observation belongs to which area, and let $\boldsymbol{\gamma}$ be the vector of q spatial random effects and assume $\boldsymbol{\gamma} \sim N_q(0, \lambda^{-1}\mathbf{G}^{-1})$, where $\mathbf{G}^{-1}$ is the (generalized) inverse of a $q \times q$ matrix, $\mathbf{G}$. In the following IAR model, based on [4], the matrix $\mathbf{G}$ contains the information about the neighbours (adjacent regions), with elements given by $G_{mm} = n_m$ where n_m is the total number of adjacent regions to region m and $G_{mt} = -1$ if region m and t are adjacent, and zero otherwise, for $m = 1, \ldots, q$ and $t = 1, \ldots, q$. This model has the attractive property that conditional on λ and γ_t for all $t \neq m$, then $\gamma_m \sim N(\sum \gamma_t n_m^{-1}, (\lambda n_m)^{-1})$ where the summation is over all regions which are neighbours of region m.

The nonzero pattern of the matrix $\mathbf{G}$ determines the graph $\mathcal{G}$. A nonzero value in matrix $\mathbf{G}$ implies a connection between the two corresponding regions in the graph $\mathcal{G}$ (they are connected neighbours). The zero value in matrix $\mathbf{G}$ implies no connection between the two regions in the graph $\mathcal{G}$ and hence that the corresponding spatial random effects γ_m and γ_t for the two regions are conditionally independent (given the other spatial random effects γ_r for all r not equal to m or t).

The `R` implementation of the above IAR model as a predictor term for any parameter of the distribution of the response variable in a GAMLSS model is achieved by the `R` package `gamlss.spatial` which is described in the appendix. More details about using GAMLSS in `R` are presented in [25].

3. Application to the Munich rent data

Here we use the package `gamlss.spatial` to provide a detailed spatial analysis of a data set on rents for flats in the City of Munich.

3.1. The data

The response variable is the rent, (i.e. the monthly rental price, which remains after having subtracted all running costs and incidentals) of properties in the city of Munich, [15]. We used the Munich rent data in the year 1999, available from data frame `rent99` in the `gamlss.data` package in `R`. The data frame `rent99` has 3082 observations on the following 9 variables:

- `rent`: rent per month (in Euro),
- `rentsqm`: rent per month per square metre (in Euro),
- `area`: living area in square metres,
- `yearc`: year of construction,
- `location`: quality of location: a factor indicating whether the location is average location, (1), good location, (2), or top location, (3),
- `bath`: quality of bathroom: a factor indicating whether the bathroom facilities are standard, (0), or premium, (1),
- `kitchen`: quality of kitchen: a factor indicating whether the kitchen is standard, (0), or premium, (1),
- `cheating`: central heating: a factor indicating a property with central heating, (1), or without central heating, (0),
- `district`: district in Munich (this provides the spatial explanatory variable).

In the data frame `rent99` the variables `location`, `bath`, `kitchen` and `cheating` are declared as factors with reference levels 1, 0, 0 and 0, respectively. The reference level for `cheating` was changed to 1 in the analysis, because most properties have central heating.

The distribution of the monthly rent is asymmetric and skewed towards the right as is shown in Figure 2.

Figure 3 shows plots of the `rent` against each of the above explanatory variables. Although these are bivariate exploratory plots and take no account of the interplay between the explanatory variables, they give an indication of the complexity of this data. The first two explanatory variables, `area` and `yearc`, are continuous. The plot of `rent` against `area` suggests a positive relationship between median rent and area, with an increased variation for larger `area`. The assumption of homogeneity in the variance of the `rent99` variable appears to be violated here. There is also some indication of positive skewness in the distribution of the `rent` variable. The peculiarity of the plot of `rent` against `yearc` is due to the method of data collection. The plot suggests that for houses up to 1960 the median rent price is roughly constant, but for flats constructed after that year there is an increasing trend in the median rent price. The remaining box and whisker plots display how the rent price varies according to the explanatory factors. The median rent price increases as the location changes from average to good and then to top location. The median rent price also increases if the flat has a premium bathroom,

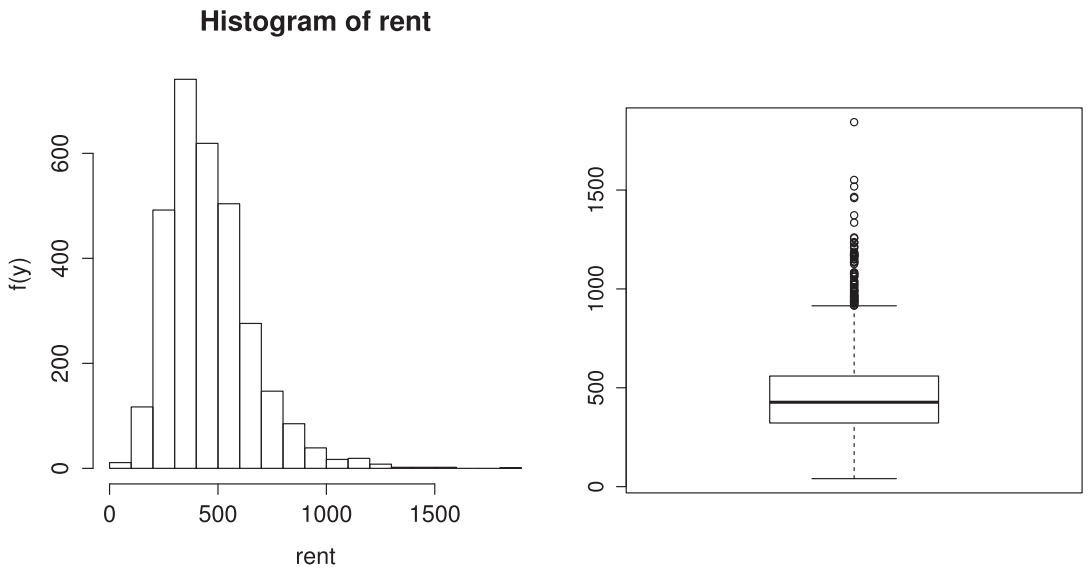

Figure 2. Histogram and box and whisker plots for rent data from the year 1999 in Munich.

a premium kitchen or central heating. There are no surprises in the plots here, but again the problem of skewness is prominent with generally (but not always) longer upper than lower tails.

In summary, any statistical model used for the analysis of the above data should be able to deal with the complexity of the relationship between `rent` and the explanatory variables. The dependence of the median of the response variable `rent` on floor space (`area`) and year of construction (`yearc`) is nonlinear and non-parametric smoothing functions may be needed. Median rent may also depend on interactions between the explanatory variables. There is clear indication of nonhomogeneity of the variance of rent. The variance of the response variable `rent` may depend on its mean and/or explanatory variables. There is clear indication of skewness in the distribution which may also depend on explanatory variables. The median rent (and the variance and skewness of rent) may also depend on the spatial explanatory variable (`district`), which is a key part of the analysis.

3.2. Model selection strategy

This section describes the model selection strategies adopted in this paper. Let $\mathcal{M} = \{\mathcal{D}, \mathcal{L}, \mathcal{T}, \boldsymbol{\lambda}\}$ represent a GAMLSS model as defined in Section 2.1. The components of $\mathcal{M}$ are defined as follows:

$\mathcal{D}$ specifies the distribution of the response variable,
$\mathcal{L}$ specifies the set of link functions for the distribution parameters μ, σ, ν and τ,
$\mathcal{T}$ specifies the terms appearing in the predictors for μ, σ, ν and τ,
$\boldsymbol{\lambda}$ specifies the smoothing hyper-parameters which determine the amount of smoothing of continuous explanatory variables (`area` and `yearc`) and of the spatial effect (`district`).

In the search for an appropriate GAMLSS model for any new data set, all the above four components have to be specified as objectively as possible. The GAMLSS

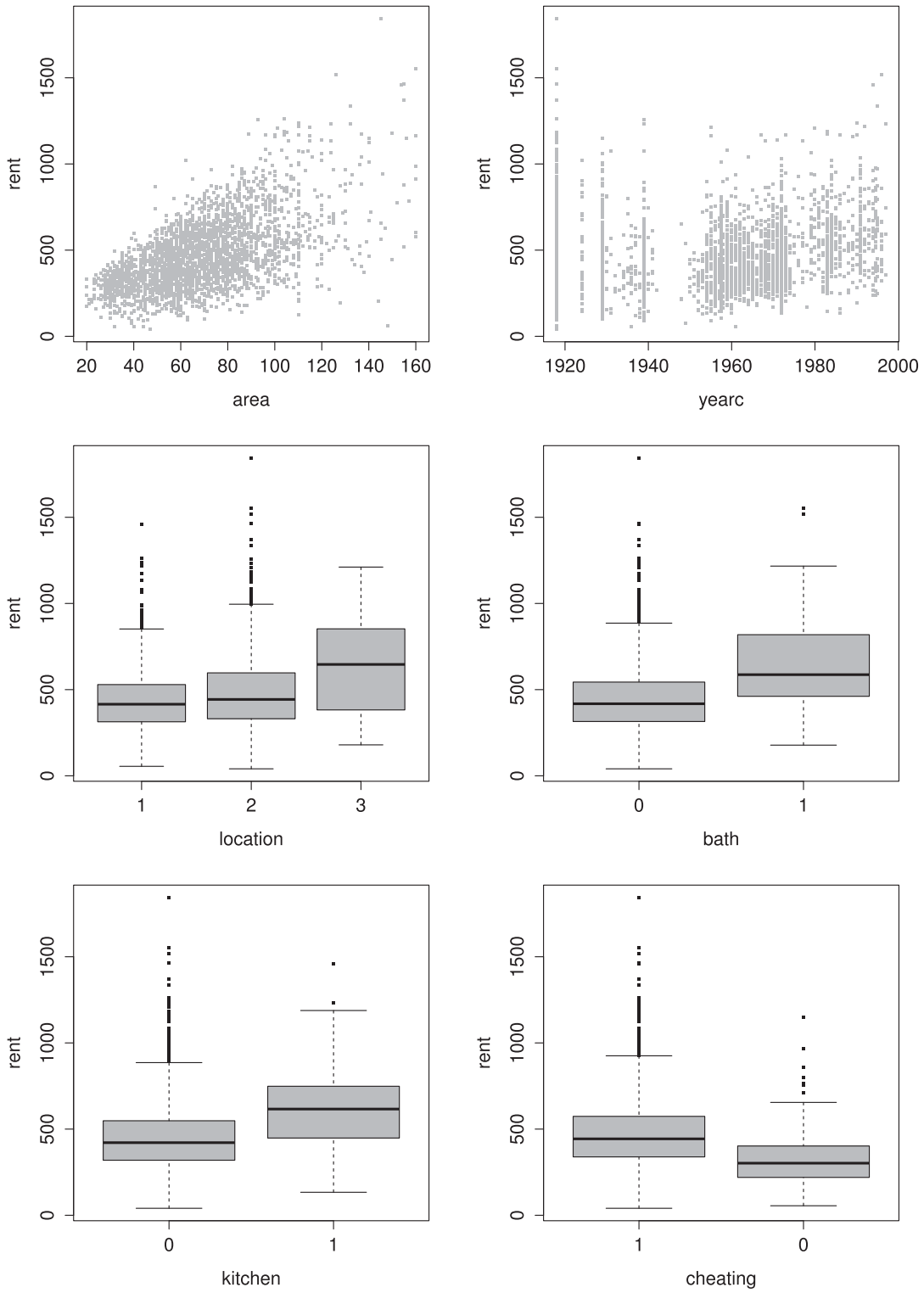

Figure 3. Plot of the `rent99` against explanatory variables `area`, `yearc`, `location`, `bath`, `kitchen` and `cheating`.

framework requires that the empirical researchers have a good understanding of the properties of the distributions from the list of available distributions in the GAMLSS framework.

The selection of the appropriate distribution $\mathcal{D}$ is done in two stages: the fitting stage and the diagnostic stage. The fitting stage involves the comparison of different fitted models

using a generalized Akaike information criterion (GAIC). The diagnostic stage involves the normalized quantile residuals, [11], or 'z-scores', which provide information about the adequacy of the model and can be used in connection with diagnostic plots like worm plots, [26], or other test statistics, e.g. Z-statistics and Q-statistics, [23]. The selection of the link function $\mathcal{L}$ is usually determined by the range of parameters. For a given distribution for the response variable, the selection of the terms $\mathcal{T}$ for the parameters of the distributions is done using a stepwise GAIC procedure.

Preliminary analysis, using distributions defined on the positive real line, indicated that the Box–Cox Cole and Green distribution [9], BCCGo(μ, σ, ν), seems an appropriate distribution for the rent data to use for model selection. The *BCCGo* distribution has a default log link for the median μ. If we use an identity link for μ it implies an additive model for μ and so, for example, changing from an unpopular to a popular district results in a fixed change in median rent, irrespective of how large an area the property has and irrespective of its year. It is more likely that the change in median rent is not a fixed amount but a fixed percentage, implying that a multiplicative model is more appropriate, i.e. a log link for μ. It was found that the log link for μ provided a better fit to the data then the identity link.

Because the fitting time of the spatial GMRF term for `district` in the model is longer that the rest of the terms, first we used a selection procedure for all explanatory variables (apart from `district`) for all distribution parameters (μ, σ and ν) using a generalized Akaike information criterion, GAIC, with penalty equal 4. Then, given the selected model, we tried adding the GMRF term IAR, with penalty equal 2. The reason for the choice of $k = 4$ in GAIC for the selection of terms (excluding the spatial effect) is that several terms have a single parameter and a 5% significance level for a generalized likelihood ratio test for a single parameter being different from zero is based on an (asymptotic) Chi-squared distribution with critical value $\chi^2_{1,0.05} = 3.84 \approx 4$. The spatial term involves many effective parameters being jointly tested and so a lower critical value per effective parameter is appropriate. When choosing whether to select a spatial term, we decided to use the standard AIC with $k = 2$.

The procedure to select the explanatory variables using the *BCCGo* distribution is first to fit an initial starting model and then:

(1) use a forward GAIC selection procedure to select an appropriate model for μ, with σ and ν as constants,
(2) use a forward selection procedure to select an appropriate model for σ, given the model for μ obtained in (1) and for ν fitted as a constant,
(3) use a forward selection procedure to select an appropriate model for ν, given the models for μ and σ obtained in (1) and (2), respectively,
(4) use a backward elimination procedure to select an appropriate model for σ, given the models for μ and ν obtained in (1) and (3), respectively; and
(5) use a backward elimination procedure to select an appropriate model for μ, given the models for σ and ν obtained in (3) and (4), respectively.

The above procedure is executed in `gamlss` using the function `stepGAICAll.A`. The resulting chosen model may contain different explanatory variables for μ, σ and ν. Then from this model we

(i) add the `district` as a spatial effect for μ using the IAR spatial model,
(ii) add the `district` as a spatial effect for μ and σ using the IAR spatial model, and
(iii) add the `district` as a spatial effect for μ, σ and ν using the IAR spatial model.

The (smoothing) hyper-parameters $\boldsymbol{\lambda}$ can be fixed or estimated from the data. The standard way of fixing the (smoothing) hyper-parameters is by fixing their effective degrees of freedom (edf) for smoothing.

The local maximum likelihood estimation method for each λ is the method used in our analysis. Hence, the model terms were selected using the GAIC, while the smoothing parameters (and hence their corresponding edf) were chosen using local maximum likelihood.

3.3. Results

In Section 3.2 we explained the model selection strategy. The final chosen fitted model `m2final` is given by

$$
\begin{aligned}
Y &\sim \text{BCCGo}(\hat{\mu}, \hat{\sigma}, \hat{\nu}),\\
\log(\hat{\mu}) &= 6.06 + h_{11}(\text{yearc}) + h_{21}(\text{area}) + s(\text{district})\\
&\quad + 0.079(\text{if location=2, good}) + 0.211(\text{if location=3, top})\\
&\quad - (0.255 - 0.0038\text{nyearc})(\text{if cheating=0, no central heating})\\
&\quad + (0.146 - 0.0034\text{nyearc} + 0.0023\text{narea})(\text{if kitchen=1, premium})\\
&\quad + 0.067(\text{if bath=1, premium}),\\
\log(\hat{\sigma}) &= 11.811 + h_{12}(\text{yearc}) + 0.0016(\text{area})\\
&\quad + 0.231(\text{if cheating=0, no central heating}),\\
\hat{\nu} &= -12.377 + h_{13}(\text{yearc}) + h_{23}(\text{area}) + 2.381(\text{if kitchen=1, premium}),
\end{aligned}
\tag{5}
$$

where the h functions are smooth non-parametric functions and s is an IAR spatial smoothing function. The distribution BCCGo(μ, σ, ν) has a multiplicative model for the median μ, (resulting from the log link for μ), and *nyearc* and *narea* are, respectively, *yearc* and *area* centred at their means (i.e. subtract their means, 67.37 and 1956.31, respectively). The median μ model includes a spatial term in `district` (using the GMRF model IAR), and provides an improvement (i.e. reduction) in AIC. We also fitted the model with additional spatial effects for σ and ν but the improvement was too small so we opted for the simpler model, in this case, the spatial effect just for μ.

Figures 4–6 display the fitted parametric terms and smooth functions in $\log(\hat{\mu})$ in the final chosen model (5). Their effects are additive for $\log(\hat{\mu})$ and hence multiplicative for the fitted median rent $\hat{\mu}$. The fitted median rent generally increases with area and year of construction (from Figure 4). A good location results in a 8.2% [calculated by $(e^{0.079} - 1) \times 100$] increase in fitted median rent (relative to an average location), while a premium location results in a 23.5% increase, and a premium bathroom results in a 6.9% increase.

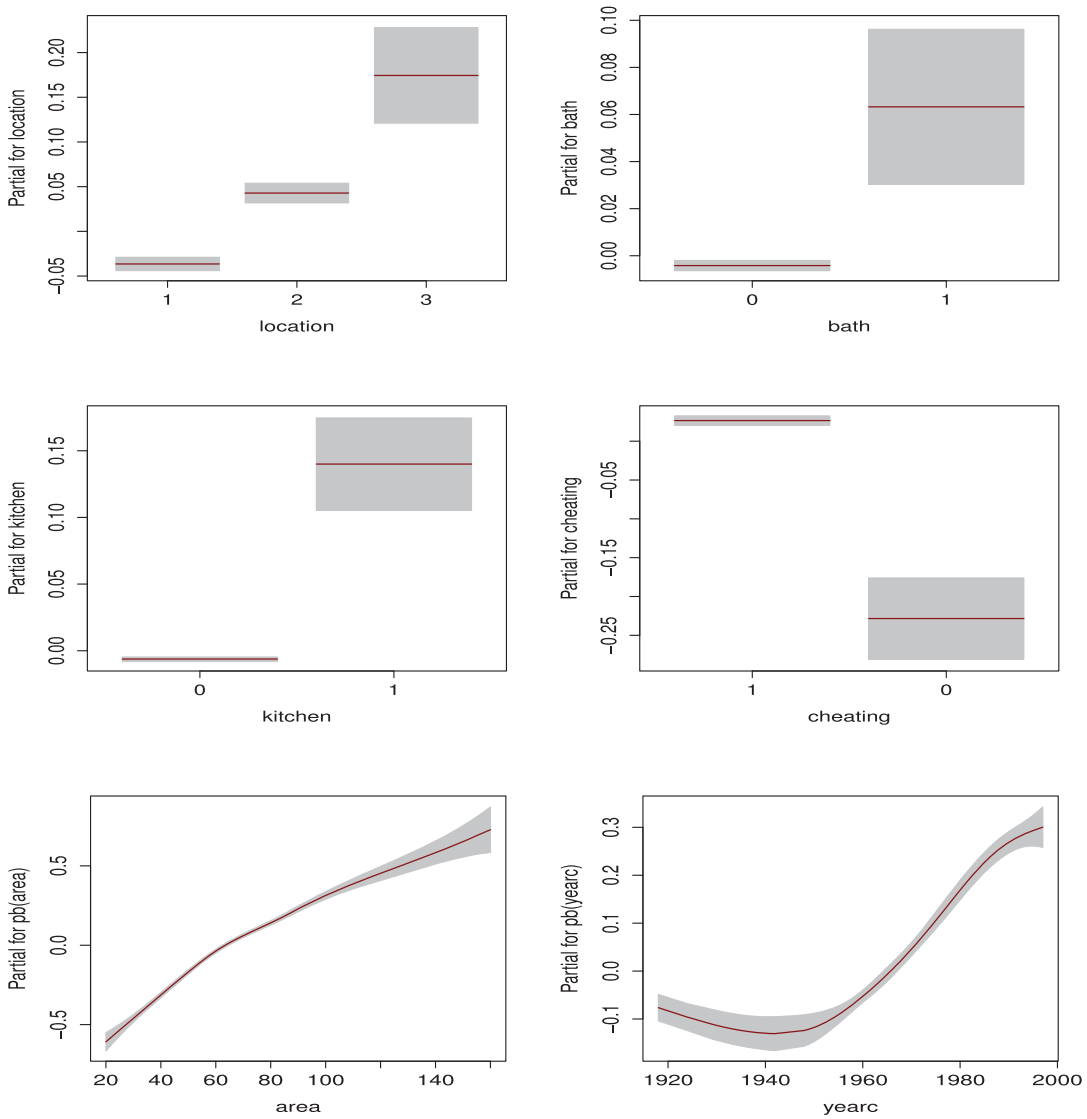

Figure 4. Term plots for μ.

The effect on median rent of no central heating depends on the year of construction. No central heating results in a 22.5% decrease in median rent for the average year of construction, and a higher % decrease for older properties. The effect of a premium kitchen on median rent depends on both year of construction and area of the property, resulting in a 15.6% increase in median rent for a property with average year of construction and average area, and a higher % increase for older or larger properties.

Figure 6 shows the `district` effect on $\log(\hat{\mu})$ where we can see that the rent prices are higher in the centre and southeast regions than in the north and west regions of the Munich city. Relative to the baseline district a region with the best district has a 10.5% [i.e $(e^{0.10} - 1) \times 100$] higher fitted median rent, while a region with worst district has a 9.5% [i.e $(1 - e^{-0.10}) \times 100\%$] lower fitted median rent (assuming all other explanatory variables including location type are fixed).

Figure 7 shows the fitted parametric terms and smooth function in $\log(\hat{\sigma})$, in the final chosen model (5). Figure 7 shows that the fitted $\hat{\sigma}$ (the approximate coefficient of variation of rent) increases with area but decreases with year of construction. No central heating results in a 26.1% increase in $\hat{\sigma}$. [It should be noted that if the total effective degrees of

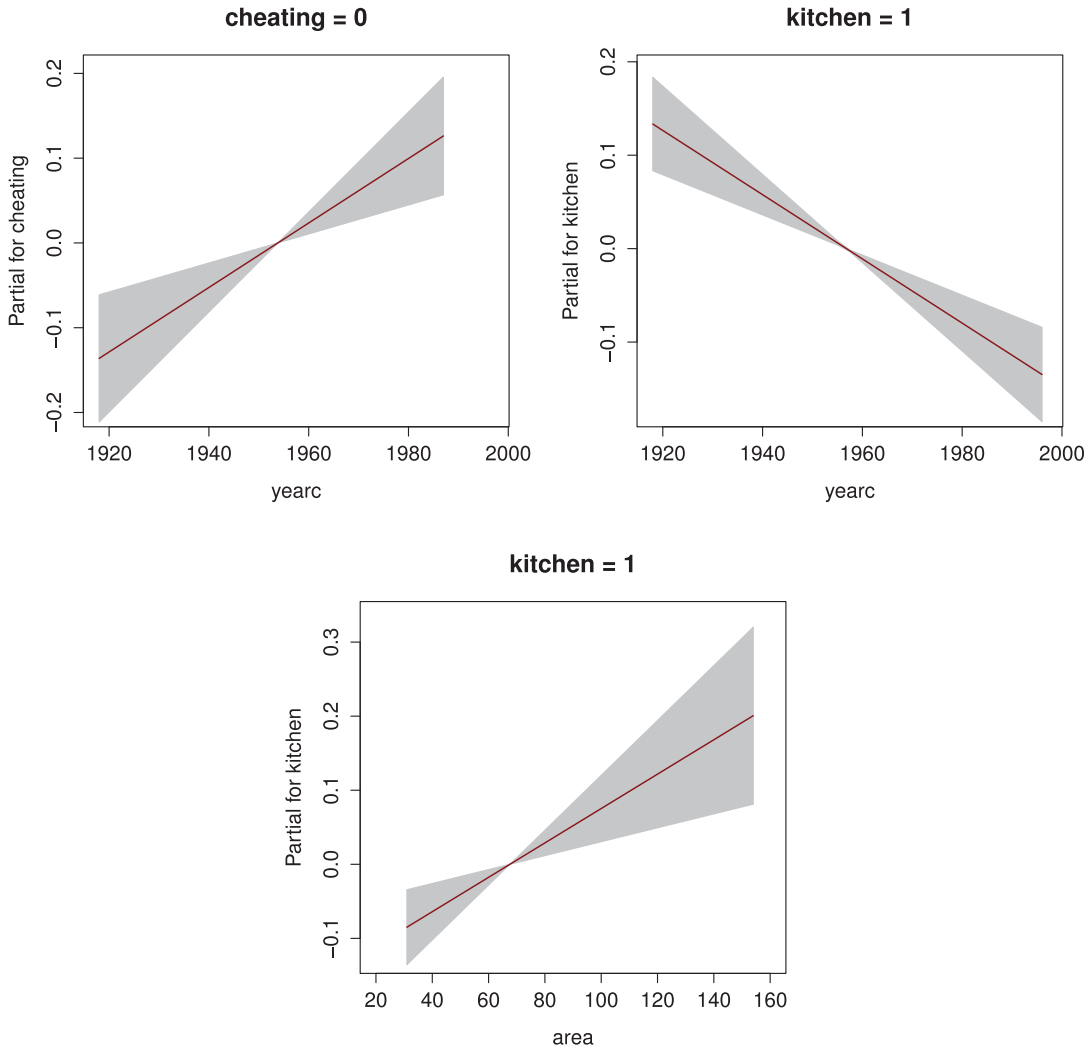

Figure 5. Term plots of the interactions for μ.

freedom used in the model for μ is high relative to the sample size, then this can result in negative bias in $\hat{\sigma}$. This was not the case in the fitted model (5)].

Figure 8 shows that the fitted $\hat{\nu}$ (the skewness parameter) in Equation (5) decreases with area but increases with year of construction. Note that decreasing $\hat{\nu}$ increases the positive skewness of the fitted distribution for rent. A premium kitchen results in an increase of 2.4 in $\hat{\nu}$. Hence larger older properties with a standard kitchen have a more positively skew fitted distribution for rent.

4. Residual diagnostics

We check the adequacy of the fitted model using (normalized quantile) residuals, [11]. If the model is correct then the true residuals have a standard normal distribution. Figure 9 displays a worm plot, [26], for the residuals of the chosen fitted model. The worm plot is a detrended normal QQ plot of the residuals which indicates a reasonable fit to the data, since over 95% of the points lie within the elliptical (dashed) 95% pointwise interval bands.

In order to investigate the adequacy of the chosen model for different combinations of the two continuous explanatory variables `yearc` and `area`, we cut each explanatory

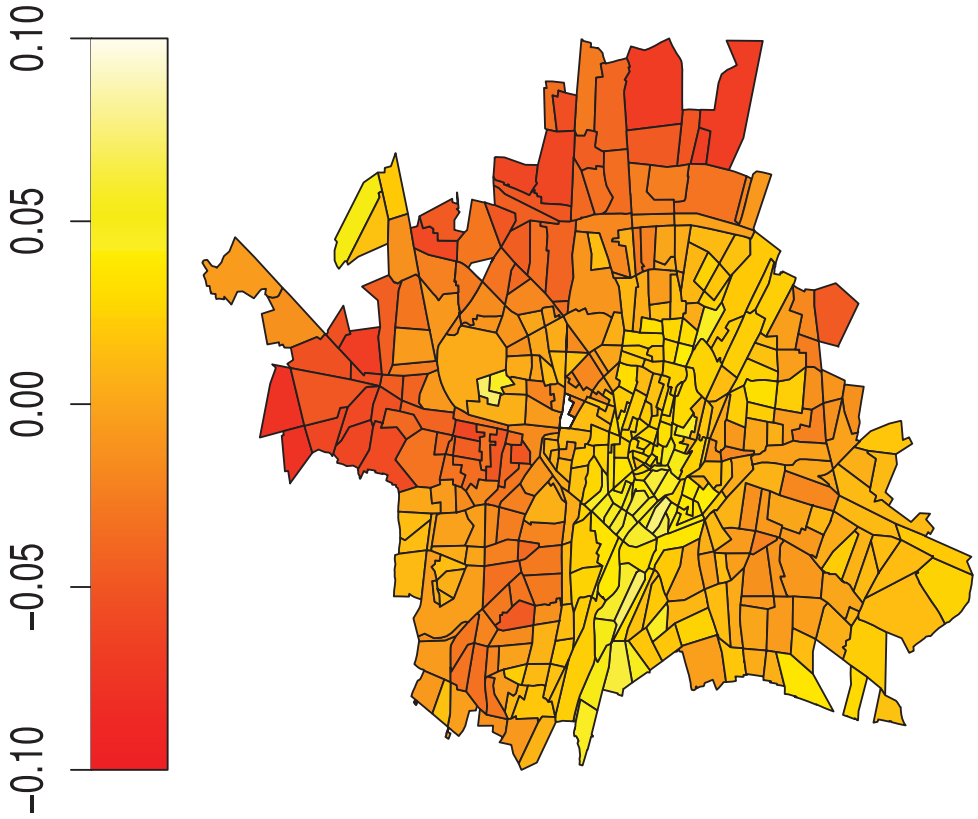

Figure 6. The fitted spatial effect for μ for the chosen model with spatial effect.

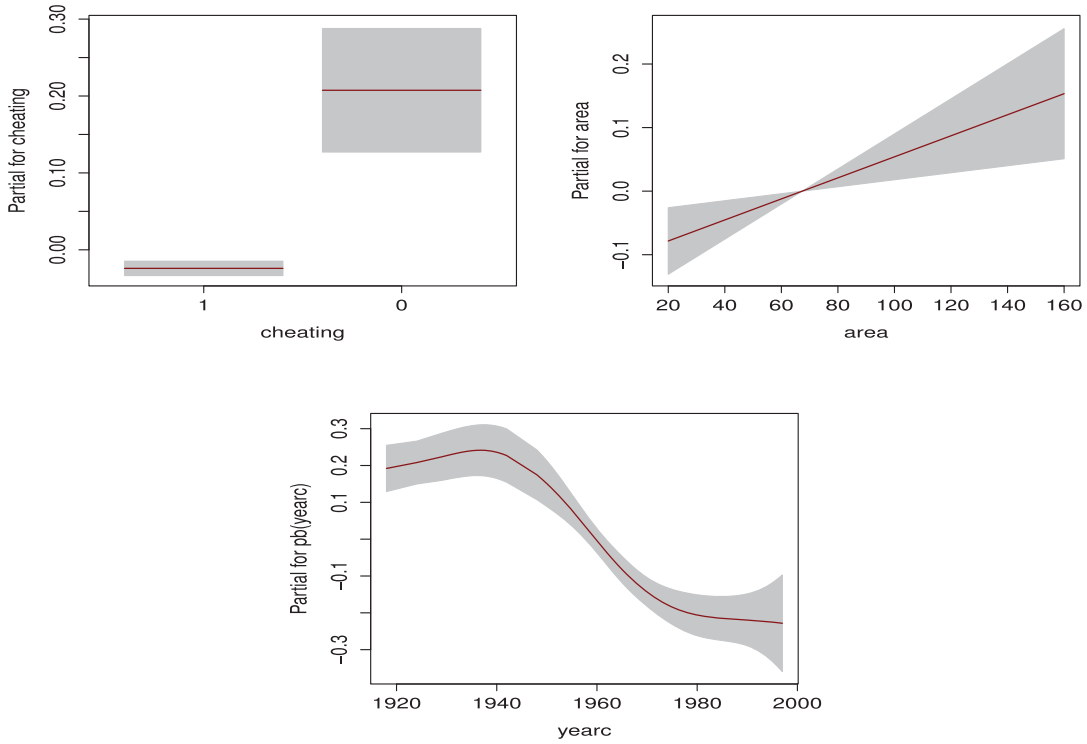

Figure 7. Term plots for σ.

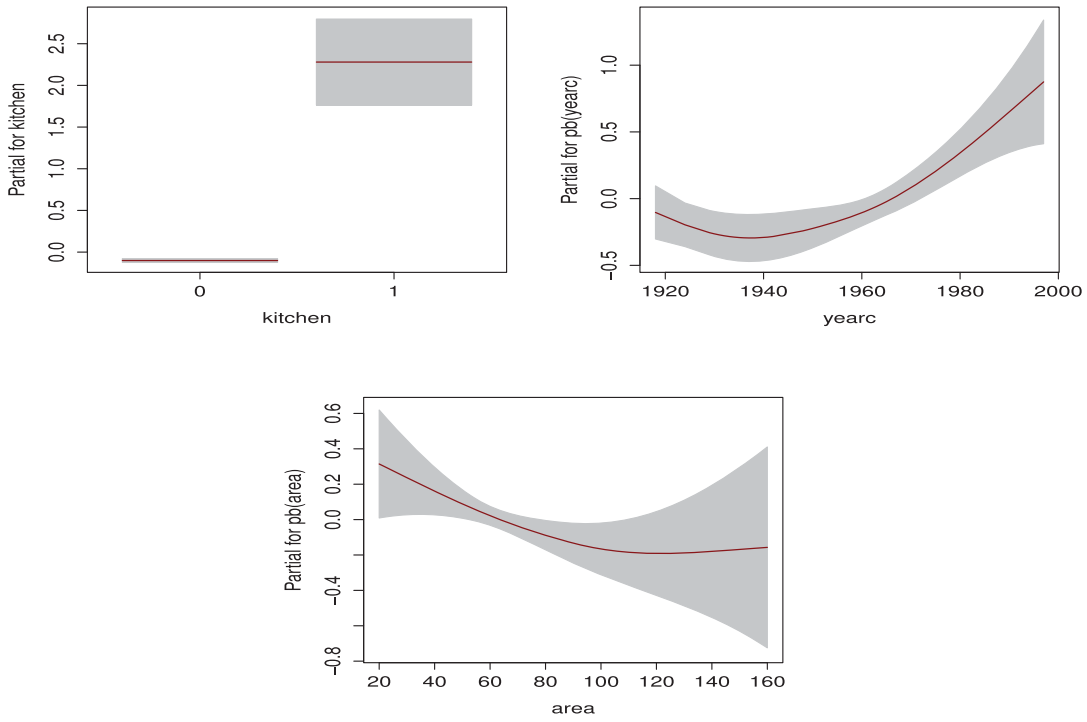

Figure 8. Term plots for ν.

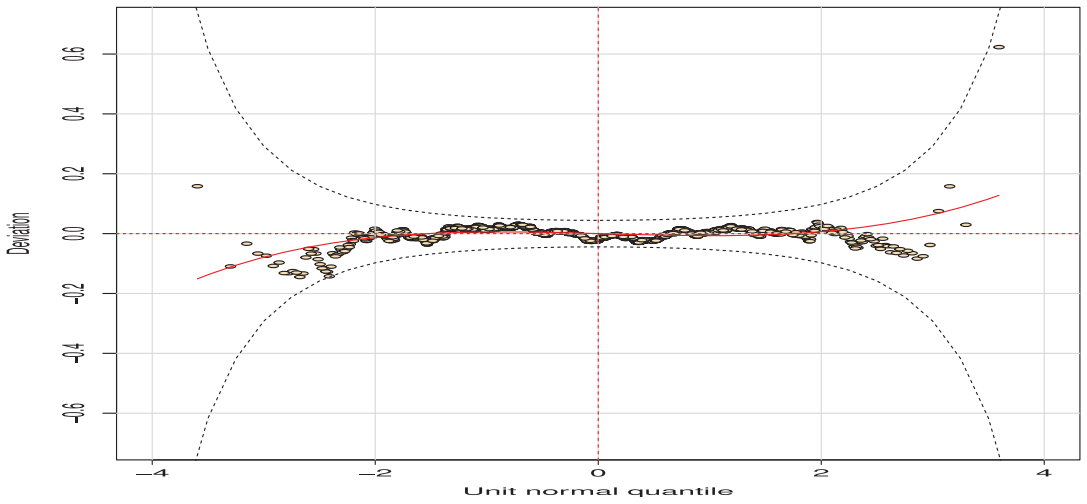

Figure 9. Worm plot of the residuals for the chosen final model `m2final`.

variable into four non-overlapping intervals with equal numbers of observations giving 16 joint intervals and obtain a worm plot (i.e detrented QQ plot) for cases in each of the 16 joint intervals. This is a way of highlighting failures of the model within different joint ranges of the two explanatory variables. Figure 10 shows the result, (obtained by a single `worm` command in the `gamlss` package), where above the plot the four intervals for `yearc` are displayed and to the right of the plot the four intervals of `area` are displayed. The worm plots generally indicate a reasonable fit to the data in the 16 joint intervals. Similarly, Figure 11 displays the worm plots for combinations of the two explanatory variables `yearc` and `kitchen`, which also indicate a reasonable fit to the data. For the sake of

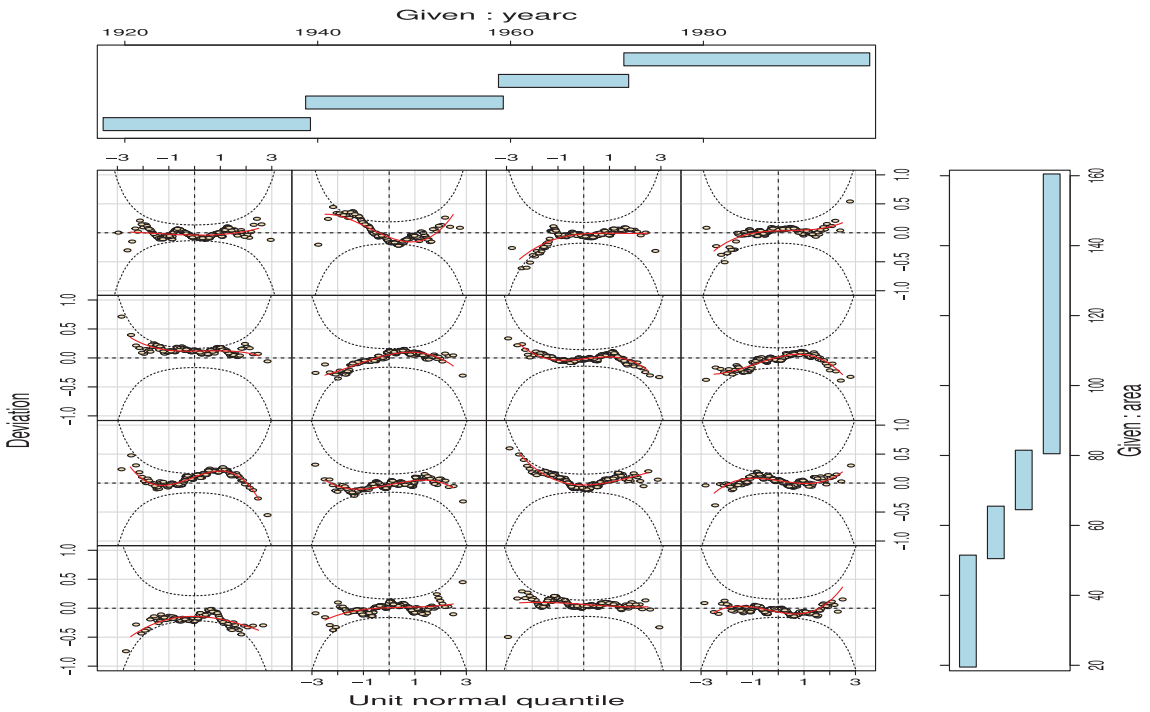

Figure 10. Worm plot of the residuals split by the `yearc` and `area` variables for the final model.

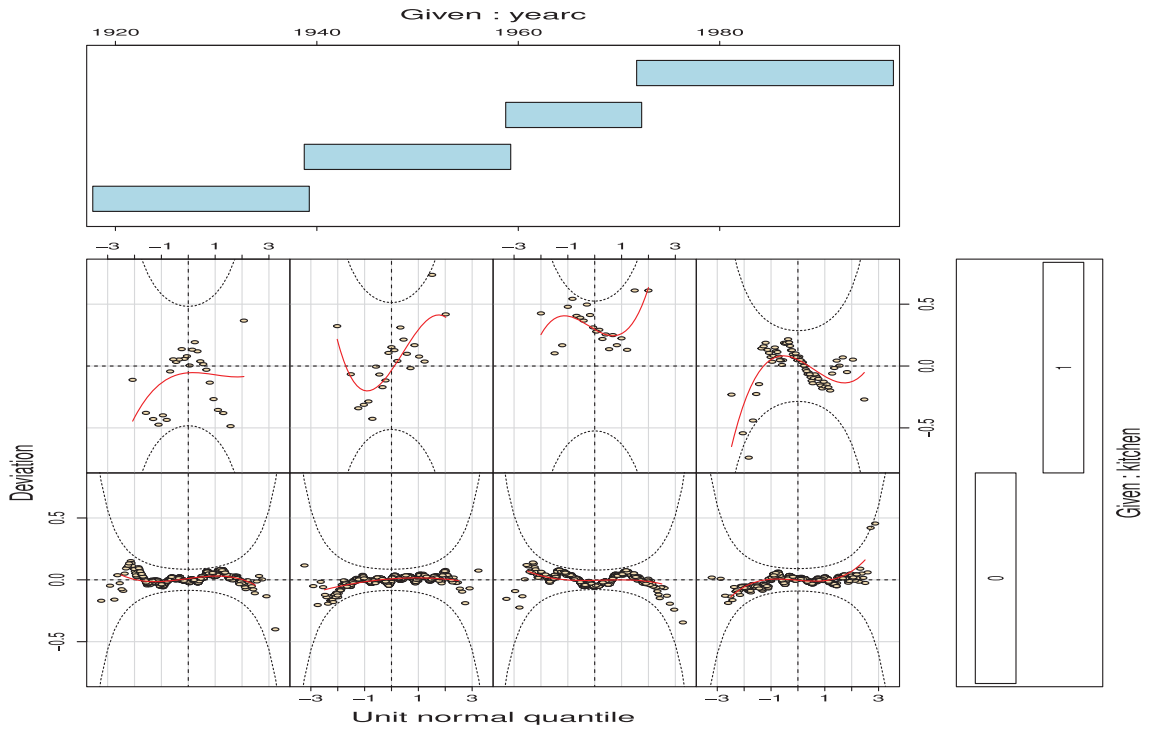

Figure 11. Worm plot of the residuals split by the `yearc` and `kitchen` variables for the final model.

brevity, the worm plots for individual explanatory variables and for other combinations of two explanatory variables were omitted here, but they also indicated a reasonable fit to the data.

5. Conclusions

We have shown that the GAMLSS framework provides a platform to fit, compare and check spatial models for the parameters of the distribution of a response variable which may be non-exponential family. This includes continuous response variable distributions which are highly positively or negatively skewed and/or have high or low kurtosis (i.e. leptokurtic or platykurtic), discrete count distributions that are overdispersed (e.g. negative binomial) or have excess zeros (eg zero inflated negative binomial), or mixed continuous–discrete distributions (e.g. zero-inflated gamma and inflated beta). The spatial analysis shown in this paper can be applied to other data sets that have geographical information specifying the neighbours of each region.

We would like to finish by emphasizing that looking at a single statistical model in isolation is not good practice. Any chosen model should be able to stand up to scrutiny and that involves being able to compare it with alternative models and checking its assumptions. The data for this article can be found in the package `gamlss.data`. The commands to fit the model and plot the results in the paper are in a vignette distributed with the package `gamlss.spatial`.

Acknowledgments

We acknowledge the partial financial support from Fundação Araucária of Paraná State, Capes, CNPq and FACEPE from Brazil.

References

[1] H. Akaike, *Information measures and model selection*, Bull. Int. Stat. Instit. 50 (1983), pp. 277–290.
[2] S. Banerjee, B.P. Carlin, and A.E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2014.
[3] J. Besag, *Spatial interaction and the statistical analysis of lattice systems (with discussion)*, J. R. Statist. Soc, Ser. B 36 (1974), pp. 192–236.
[4] J. Besag and D. Higdon, *Bayesian analysis of agricultural field experiments (with discussion)*, J. R. Statist. Soc: Ser. B (Statist. Methodol.) 61 (1999), pp. 691–746.
[5] J. Besag and C. Kooperberg, *On conditional and intrinsic autoregressions*, Biometrika 82 (1995), pp. 733–746.
[6] E. Borghi, M. de Onis, C. Garza, J. Van den Broeck, E.A. Frongillo, L. Grummer-Strawn, S. Van Buuren, H. Pan, L. Molinari, R. Martorell, A.W. Onyango, and J.C. Martines, *Construction of the World Health Organization child growth standards: Selection of methods for attained growth curves*, Stat. Med. 25 (2006), pp. 247–265.
[7] N.E. Breslow and D.G. Clayton, *Approximate inference in generalized linear mixed models*, J. Am. Statist. Ass. 88 (1993), pp. 9–25.
[8] D. Brook, *On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems*, Biometrika 51 (1964), pp. 481–483.
[9] T.J. Cole and P.J. Green, *Smoothing reference centile curves: The LMS method and penalized likelihood*, Stat. Med. 11 (1992), pp. 1305–1319.
[10] F. De Bastiani and M. Stasinopoulos, gamlss.spatial: Package to fit spatial data in gamlss, *R package version 1.4*, 2015. Available at `http://www.gamlss.org/`.

[11] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Statist. 5 (1996), pp. 236–244.
[12] D. Edwards, *Introduction to Graphical Modelling*, Springer, New York, 2000.
[13] G. Heller, D. Stasinopoulos, R. Rigby, and P. De Jong, *Mean and dispersion modelling for policy claims costs*, Scand. Actuar. J. 4 (2007), pp. 281–292.
[14] M. Khondoker, C. Glasbey, and B.A. Worton, *Comparison of parametric and nonparametric methods for normalising cDNA microarray data*, Biometr. J. 49 (2007), pp. 815–823.
[15] T. Kneib, *Beyond mean regression*, Statist. Model. 13 (2013), pp. 275–303.
[16] H.R. Kunsch, *Gaussian Markov random fields*, J. Fac. Sci. Univ. Tokyo, Sect. IA. Math. 26 (1979), pp. 53–73.
[17] Multicentre Growth Reference Study Group WHO. Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development, Geneva: World Health Organization, 2006.
[18] Multicentre Growth Reference Study Group WHO. *Child Growth Standards: Head circumference-for-age, arm circumference-for-age, triceps circumference-for-age and subscapular skinford-for-age: Methods and development*, Geneva: World Health Organization, 2007.
[19] R Core Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, 2015. Available at `http://www.R-project.org/`.
[20] R.A. Rigby and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape (with discussion)*, Appl. Stat. 54 (2005), pp. 507–554.
[21] R.A. Rigby and D.M. Stasinopoulos, *Automatic smoothing parameter selection in GAMLSS with an application to centile estimation*, Statist Methods Med. Res, SAGE Publ. 23 (2013), pp. 318–332.
[22] R.A. Rigby, D.M. Stasinopoulos, and V. Voudouris, *Discussion: A comparison of GAMLSS with quantile regression*, Statist. Model. 13 (2013), pp. 335–348.
[23] P. Royston and E.M. Wright, *Goodness-of-fit statistics for age-specific reference intervals*, Stat. Med. 19 (2000), pp. 2943–2962.
[24] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall, New York, 2005.
[25] D.M. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris, and F. De Bastiani, *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC The R Series, Boca Raton, FL, 2017.
[26] S. van Buuren and M. Fredriks, *Worm plot: A simple diagnostic device for modelling growth reference curves Statistics in Medicine*, Statist. Med. 20 (2001), pp. 1259–1277.
[27] V. Voudouris, D.M. Stasinopoulos, R.A. Rigby, C. Di Maio, *The Aceges laboratory for energy policy: Exploring the production of crude oil energy policy*, Energy Policy 39 (2011), pp. 5480–5489.
[28] V. Voudouris, R. Gilchrist, R.A. Rigby, J. Sedgwick, and D.M Stasinopoulos, *Modelling skewness and kurtosis with the BCPE density in GAMLSS*, J. Appl. Stat. 39 (2012), pp. 1279–1293.
[29] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, New York, 2009.
[30] S. Wood, *Generalized Additive Models. An Introduction with R*, Chapman and Hall/CRC Press, Boca Raton, FL, 2006.

Appendix. The R implementation of GMRF spatial model within GAMLSS

Here we explain the implementation of the important GMRF submodel, the IAR model described in Section 2.3, within GAMLSS. The IAR model is implemented in the package `gamlss.spatial` thought the function `gmrf()`. A new package was needed because of several dependencies of the function `gmrf()` on existing but not standard `R` packages. Also `gmrf()` is the first function of a series of additive term spatial functions that are in progress. The function `gmrf()` fits an IAR term within the predictor of any distribution parameter in a GAMLSS model. There are two methods implemented for estimating the (smoothing) hyper-parameter λ. The two different methods

should produce identical results and can be seen as PQL methods [7]. The method is selected by the argument `method` of the function `gmrf()`. There are two possible values for the method:

(i) `method = "Q"` which estimates the spatial IAR (smoothing) hyper-parameter λ by minimizing the Q-function, see [21], which is a way to minimize the local marginal likelihood function,
(ii) `method = "A"` which estimates the spatial IAR (smoothing) hyper-parameter λ using the 'alternating' method to minimize the local marginal likelihood, see [21].

To perform the analysis, we need the matrix **G**, which has the information about the relationships between the areas, showing if they are neighbouring areas or not. If two polygons areas have at least a single point in common, then they are treated as neighbours. The function `gmrf()` accepts three different ways to pass the geographical information:

(i) `polys`, is a `R` list comprising the region label followed by coordinates of points in two columns in matrix form defining the boundary for each area,
(ii) `neighbour`, is a `R` list comprising each region label followed by its neighbouring region labels,
(iii) `precision`, is a `R` matrix containing the **G** matrix.

For instance, for a simple model (with no explanatory variables and modelling only the location parameter) the information can be given in three different ways:

```
fit <- gamlss(rent~gmrf(district, polys=polys), data=rent99),
```

or

```
fit <- gamlss(rent~gmrf(district, neighbour=neighbour),
       data=rent99),
```

or

```
fit <- gamlss(rent~gmrf(district, precision=precision),
       data=rent99).
```

If the `polys` information is given, then the function `gmrf()` will automatically compute the matrix **G** to do the analysis. The same happens if the `neighbour` information is given. The fastest way to estimate the spatial IAR model in (2), as described in Sections 2.2 and 2.3, is to give the `precision` (i.e. matrix **G**) since no extra calculations are needed.

Extra utility functions available within the package to obtain the matrix **G** before performing the analysis (to speed up the fitting) are:

`polys2nb()` which creates `neighbour` list from the polygon `polys` information and
`nb2prec()` which creates the matrix **G** from the `neighbour` information.

When fitting several models for model selection this saves time. To plot the fitted values of a fitted `gmrf` object the function `draw.polys()` is available. For more details about the `gamlss.spatial` package, see the help file in `http://cran.r-project.org/`.