UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Lucas de Miranda Oliveira

Spatial Autoregressive Models for Areal Data within GAMLSS

Recife, 2019

Lucas de Miranda Oliveira

Spatial Autoregressive Models for Areal Data within GAMLSS

Master's thesis submitted to the Graduate Program in Statistics, Department of Statistics, Universidade Federal de Pernambuco as a requirement to obtain a Master's degree in Statistics

Universidade Federal de Pernambuco – UFPE

Departamento de Estatística

Programa de Pós-Graduação em Estatística

Supervisor: Fernanda De Bastiani

Co-supervisor: Dimitrios Mikis Stasinopoulos

Recife

2019

List of Figures

Contents

1 Introduction

The advancement of technologies in the last decades has allowed some benefits for several areas of knowledge, as is the case of georeferenced data. These mean that each die is assigned to a location on the map. Areas such as econometrics, climatology, ecology, health public and others are incorporating spatial information into their analysis. In econometrics, for example, we have the level of economic activity in each city, county, and country, if spatial information is relevant, according to specific criteria, then information on the level of activity of the neighbors leaves the analysis richer. In public health, on the other hand, location-indexed data help to find factors associated with certain diseases in and propagation within a region. It also gives a sense of the health location of the individual and its distribution on the map (BHUNIA; SHIT, 2019).

According to Banerjee, Carlin e Gelfand (2004), georeferenced data are divided into three basic types:

- *point-referecend data* or textitgeostatistical data, where $Y(\boldsymbol{s})$ is a random variable at location $\boldsymbol{s}$, and $\boldsymbol{s}$ takes values on a continuous surface $D \subset \Re^r$, which is a rectangle r-dimensional of positive volume. Note that Y may have discrete or continuous probability distribution function;
- *areal data*, the set $D \subset \Re^r$ is defined as previously, is defined as previously, but is now partitioned into n finite area units, or regions;
- *point pattern data*, the set D is random, and s indicates the occurrence of an event on the map; Determining whether there is a pattern of events or whether they are randomly distributed on the map.

In the context of this dissertation we will worry about the second type shown above, *areal-data*. For decades, statistical models have been developed to work with this type of spatial data. Conditional Autoregressive (CAR)(BESAG, 1974) models and Intrinsic Conditional Autoregressive (ICAR) (BESAG; YORK; MOLLIÉ, 1991) models are examples of these models. Our work comprised the development of a structure that comprised Simultaneous Autoregressive (SAR) models (WHITTLE, 1954), which are also models for *areal data*, in a more flexible regression framework than that found in the literature, allowing the relaxation of some hypotheses of the model. Clearly, we include in the Generalized Additive Models for Location, Scale and Shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005), which is a more flexible alternative for modeling response variables with several probability distribution functions. GAMLSS also allow the modeling of all parameters of the probability distribution of the response variable. And also the

inclusion of terms of non-parametric functions in the modeling of these parameters. For this, the spatial information will be introduced in the GAMLSS through terms of random effects.

Structure of the dissertation

In chapter two, we review the models used in the regression analysis and their respective assumptions. We start with linear models, then by the generalized linear models (GLM) (NELDER; WEDDERBURN, 1972), after we present the Generalized Additive Models (GAM)(HASTIE; TIBSHIRANI, 1990) and then the GAMLSS, which are part of our object of study. In chapter 3, we do a theoretical review of the models for area data that were cited above. To implement the SAR models in the GAMLSS as terms of random effect, it was necessary to study the relationship between this model and the CAR models. In this chapter, we present the theoretical framework of the proposed implementation. In chapter 4, we make a numerical evaluation of the regression coefficient estimators of the proposed model, comparing it with other models existent in the literature, in order to verify their properties in relation to the others. In chapter 5 we make two applications of the proposed model, one to the famous data set Housing Values in Suburbs of Boston (JR; RUBINFELD, 1978). The other application is an application in the area of spatial econometrics, where we model the inequality index of *Gini* for the municipalities in the state of Pernambuco, Brazil. Finally, in chapter six we make some final considerations about the work.

Computational Resources

For the development of this dissertation was used the programming language `R`. And for the writing of this material, we used the LaTeXdocument elaboration system. For the simulation studies, it was used computing cluster belongs to the National Supercomputing Center at Federal University of Rio Grande do Sul (CESUP / UFRGS).

2 Regression Analysis

The purpose of this chapter is to describe important aspects of regression analysis. Here the contents necessary to understand this dissertation will be exposed in a summarized way. We will start with the linear models, then we will present the generalized linear models. Next, we will present the generalized additive models, and finally the generalized additive models for location, scale and shape.

2.1 Linear Models

The regression analysis is the search for the causal relationship between variables, seeking to understand how one variable, called the *response variable*, is affected by another or others, these are denominated *explanatory variables*. The parametric regression approach, which is that of the linear models, has as main characteristic that the relationship between the response variable and the covariates is not a deterministic function of these, $f(x_1, \ldots x_k)$. In contrast, terms of random error are added to the linear function, implying that the response variable is a random variable. The best-known class of linear regression models is:

$$y_i = f(\boldsymbol{x}) + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

with ε_i representing the error term, that are independent and identically distributed (i.i.d), by hypothesis, and has distribution $\varepsilon_i \sim N(0, \sigma^2)$. β_k represents a fixed unknown parameters that is called regression coefficients. Representing the effect of variable x_k on the mean of the response variable, $\mathrm{E}(y|x_1, \ldots, x_k)$. The meaning of linear model here is that the model is linear in the parameters. Usually, the estimation of these unknown parameters of the linear regression model is done using the ordinary least squares method, which minimizes the sum S:

$$S(\beta_0, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})^2.$$

We minimize the sum in relation to the parameters vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$:

$$\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\
\frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_{i1} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\
\vdots &= \quad \vdots \qquad\qquad\qquad \vdots \\
\frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_{ik} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}).
\end{aligned}$$

Equating all the above equations to zero, $\frac{\partial S}{\partial \boldsymbol{\beta}} = 0$ we obtain solution in matrix terms o to vector of unknown parameters:

$$\begin{aligned} & \mathbf{X}^\top \hat{\boldsymbol{\varepsilon}} = \mathbf{0} \\ \Longrightarrow\ & \mathbf{X}^\top (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \\ \Longrightarrow\ & \mathbf{X}^\top \boldsymbol{y} = (\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} \\ \Longrightarrow\ & \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \end{aligned}$$

where $\mathrm{X} = [\mathbf{1} \quad \boldsymbol{x}_1, \ldots \boldsymbol{x}_k]$ denotes the design matrix.

2.2 Generalized Linear Models

The generalized linear models (GLMs) arise to overcome the constraint of linear models consistent with the distribution of the response variable to be Gaussian ,$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$. Presented by Nelder e Wedderburn (1972), these models the GLMs allow the flexibility of the distribution function of $\boldsymbol{Y}$, which now belongs to the *exponential family*. Many families of probability distributions are included in the *exponential family*, such as Gaussian, Exponential, Gamma, Beta, Bernoulli, Poisson and others.

The general structure of the GLMs is:

$$\begin{aligned} \boldsymbol{Y} &\sim \textit{some exponential family } (\boldsymbol{\mu}, \boldsymbol{\phi}) \\ g(\boldsymbol{\mu}) &= \boldsymbol{X}\boldsymbol{\beta}, \end{aligned}$$

where $g(.)$ is a monotonic function link. $\boldsymbol{\phi}$ is a vector of constants and $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ is is called a linear predictor. As previously, $\boldsymbol{\beta}$ is a vector $(k+1) \times 1$ of unknown parameters and $\boldsymbol{X}$ is a $n \times (k+1)$ design matrix. A complete reference in GLMs is found in McCullagh e Nelder (1989). According to these, the likelihood function is constructed from observations y_i of the random variable Y with probability distribution function:

$$f_{\mathrm{Y}}(y_i; \theta, \phi) = \exp\{(y_i\theta - b(\theta))/a(\phi) + c(y_i, \theta)\},$$

where θ is called the *canonical parameter.*, $a(.)$, $b(.)$ and $c(.)$ are specific functions. For Gaussian distribution, we have to:

$$\begin{aligned} f_{\mathrm{Y}}(y_i; \theta, \phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \mu)^2}{2\sigma^2}\} \\ &= \exp\{\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})\} \\ &= \exp\{\frac{y_i\mu - \mu^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})\}, \end{aligned}$$

with $\theta = \mu$, $a(\phi) = \phi = \sigma^2$, $b(\theta) = \theta^2$, $c(\phi, y_i) = -\frac{y_i^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})$. The mean and variance of Y are found when we do:

$$l(\theta, \phi, y_i) = \log f_Y(y_i; \theta, \phi),$$

where $l(\theta, \phi, y_i)$ is the log-likelihood function, which is a function of θ and ϕ. Then we get:

$$\frac{\partial l}{\partial \theta} = \{y_i - b'(\theta)/a(\phi)\}$$
$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi)$$

2.3 GAMLSS

Definition

The generalized additive models for location , scale and shape (GAMLSS) were proposed by Rigby and Stasinopolous in 2005. This is a general class for reponse variables univariates.

In this class of models, the observations of the response variable y_i, by hypothesis, are independent, with $i = 1, 2, ..., n$. And they have probability (or density,in the continuous case) function $f(y_i|\boldsymbol{\theta}^i)$.Conditioned on the vector $\boldsymbol{\theta}^{i\top} = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})$ of p unknown parameters of this function. Each parameter can be modeled through different independent (explanatory) variables and random effects, and generally up to four parameters are modeled. Let $\boldsymbol{y}^\top = (y_1, y_2, \ldots, y_n)$ be a vector of length n of the observations of the response variable and for $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotone link function that associates $\boldsymbol{\theta}_k$ with independent variables and random effects. The original formulation proposed by Rigby e Stasinopoulos (2005) is presented as follows:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.1}$$

where the vector $\boldsymbol{\eta}_k$ is the linear predictor and has length n. Similarly, $\boldsymbol{\theta}_k^\top$=$(\theta_{1k}, \theta_{2k}, ..., \theta_{nk})$ has the same length. In turn, the vector of the parameters $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})$ has dimension $\boldsymbol{J}'_k$, and the matrices of covariates $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ are of orders $n \times \boldsymbol{J}'_k$ and $n \times q_{jk}$. Lastly,the random effects parameter vector γ_{jk} has length $\boldsymbol{J}'_k$ and follows a normal distribution with $\boldsymbol{\gamma}_{jk} \sim N_{qjk}(\mathbf{0}, \mathbf{G}_{\boldsymbol{jk}}^{-\mathbf{1}})$, and $\mathbf{G}_{jk}^{-1}$ is the inverse of a symmetric matrix $q_{jk} \times q_{jk}$, $\mathbf{G}_{jk}$,which may be a function of a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$. And, if $\mathbf{G}_{jk}$ is a singular matrix , then it is understood that $\boldsymbol{\gamma}_{jk}$ has density function that is improper and proportional to $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^\top \boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk})$. Note that in GAMLSS it is possible to model all the parameters of the distribution of the response variable as a linear function of explanatory

variables and/or linear functions of random effects. We emphasize that not all distribution parameters need to be modeled using explanatory variables. For example, set $\mathbf{Z}_{jk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is a identity matrix $n \times n$, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$, for all combinations of j and k in (2.1) we obtain the GAMLSS formulation semiparametric additive. That is:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X_1}\boldsymbol{\beta_1} + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \tag{2.2}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X_2}\boldsymbol{\beta_2} + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \tag{2.3}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X_3}\boldsymbol{\beta_3} + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \tag{2.4}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X_4}\boldsymbol{\beta_4} + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}) \tag{2.5}$$

with vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$ of length n. Here, μ represents the position parameter, the mean, σ of scale, deviation, and ν ,asymmetry, and τ , kurtosis, represent the shape parameters.

Estimation

Primordial in the process of adjustment of the additive components in the GAMLSS structure are: the algorithm *backfitting* and the fact that quadratic penalties in the likelihood function result from the hypothesis of random effects in the linear predictor to follow a normal distribution. In this way, the resulting estimate uses shrinkage matrices-*shrinking*(smoothing) next to the algorithm *backfitting*.

As mentioned in the previous section, in the model (2.1) it is assumed that $\boldsymbol{\gamma}_{jk}$,are independent and have normal distribution with $\boldsymbol{\gamma}_{jk} \sim N_{qjk}(\mathbf{0}, \mathbf{G}_{\boldsymbol{jk}}^{-\mathbf{1}})$. In the GAMLSS framework, the hypothesis of independence between different random effects vectors is essential. However, if, for a k we have two or more random effects vectors that are not independent, we can combine them into a single random effect vector and also their corresponding covariate matrices,$\mathbf{Z}_{jk}$, in a single array of covariates. Rigby e Stasinopoulos (2005) show that, with$\boldsymbol{\lambda}_{jk}$ fixed, $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$ are estimated in the GAMLSS structure by maximizing the penalized likelihood function, l_p,given by:

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^{\top}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk} \tag{2.6}$$

where $l = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}^{\boldsymbol{i}})$ is the logarithm of the likelihood function of the observations given $\boldsymbol{\theta}^i$ for $i = 1, ..., n$. And,It is also shown in Appendix C that the maximization of l_p applied to partial residuals, $\boldsymbol{\epsilon}_{jk}$, to update the estimate of the additive predictor $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ together with an algorithm *backfitting* leads to shrinkage matrix $\mathbf{S}_{jk}$,given below:

$$\mathbf{S}_{jk} = \mathbf{Z}_{jk}(\mathbf{Z}_{jk}^{\top}\mathbf{W}_{kk}\mathbf{Z}_{jk} + \mathbf{G}_{jk})^{-1}\mathbf{Z}_{jk}^{\top}\mathbf{W}_{kk} \tag{2.7}$$

where $k = 1, 2, 3, 4$ and $j = 1, 2, ..., J_k$, with $\mathbf{W}_{kk}$ is an matrix of iterative weights. For different forms of $\mathbf{Z}_{jk}$ and $\mathbf{G}_{jk}$we will have different types of additive terms in the linear predictor $\boldsymbol{\eta}_k$, for $k = 1, 2, 3, 4$.

There are two basic algorithms that are used in GAMLSS tuning. The first is the algorithm `CG`,which is based on the algorithm of Cole and Green (1992, *apud* Stasinopoulos *et al.*, 2008, p. 15). In this, information about the first derivatives and (the expected or approximate value) of the second and the cross-derivatives of the log-likelihood function in relation to the $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ for a distribution with four parameters. However, for many probability functions (density),$f_Y(y|\boldsymbol{\theta})$,the parameters $\boldsymbol{\theta}$ are orthogonal information since the expected values of the log-likelihood function are zero, for example, position and scale models and dispersion family models. This is the case, the `RS` algorithm is more adequate, since it does not use the log-likelihood cross-derivatives (Rigby e Stasinopoulos, 2005).

In continuation of the adjustment process in the GAMLSS, the vector of hyper-parameters $\boldsymbol{\lambda}_{jk}$ can be estimated internally (locally) or globally. The local estimation method.

3 The Gaussian Markov Random Fields (GMRF)

In this section we present a definition of Gaussian Markov Random Fields (GMRFs) and its connection with the autoregressive models.

A necessary concept when we talk about GMRFs is *conditional independence*. That is, as exemplified in Rue e Held (2005), consider $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ a random vector Normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Two variables γ_1 and γ_2 are independent if and only if $\pi(\gamma_1, \gamma_2) = \pi(\gamma_1)\pi(\gamma_2)$, with $\pi(.)$ representing the density function of the variable. On the other hand, γ_1 and γ_2 are *called conditionally independent* given γ_3 iff $\pi(\gamma_1, \gamma_2|\gamma_3) = \pi(\gamma_1|\gamma_3)\pi(\gamma_2|\gamma_3)$ and the notation is represented by $\gamma_1 \perp\!\!\!\perp \gamma_2|\gamma_3$. Note that independence implies conditional independence, but the reciprocal is not valid. This is due to the fact that γ_1 and γ_2 may be marginally dependent.

Let *G=(V,E)* be an non-directed graph, with $V = \{1, \ldots, q\}$, the set of vertices or nodes representing the q - area units and E is the set of edges that connect these areas. Hence, Rue e Held (2005) define that $\boldsymbol{\gamma} \in R^n$ will be a GMRF with respect to the graph G if its density function is given by:

$$\pi(\boldsymbol{\gamma}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1})(\boldsymbol{\gamma} - \boldsymbol{\mu})), \tag{3.1}$$

and $\Sigma_{ij}^{-1} \neq 0$ if and only if $\{i, j\} \in E$ for all $i \neq j$.

Hence, the symmetric precision matrix $\boldsymbol{\Sigma}^{-1}$ informs which areas are neighbors, given some neighborhood criterion. For $\Sigma_{ij}^{-1} = 0$, we state that i and j are conditional independent, by the property of Markov. Clearly, for a larger number of neighbors, more dense (or less sparse) will be $\boldsymbol{\Sigma}^{-1}$.

The first Markov random fields we present are the class of simultaneously autoregressive (SAR) models that are commonly employed in the context of spatial econometrics. Next, we present another approach to the area data that are conditionally autoregressive models. These models based on the Markov property constitute a special class of spatial models that are suitable for discrete spatial domain Kemp (2007). We thus show the connection between them and then how we link them within GAMLSS. Before proceeding, an important point to be emphasized is, as Hodges (2016) in section 5.2, states in the section, that these models were developed to model the variable response. And it was over time that statisticians began to employ them as distributions of random effects or latent variables.

3.1 Models for Areal Data

In this chapter, we assume that the data can be thought of as a realization of a stochastic process $\{\mathbf{Z}(s) : \mathbf{s} \in \mathbf{D}\}$ where the space of variation is discrete Cressie (1992). Each element of the set $\mathbf{D}$ represents a geographic region (unit area). These models will be similar to time series models. With the Markov property defined not for the time but for space. That is, the value observed in a region depends only on the closest neighbors, while in time series it depends on the most recent past.

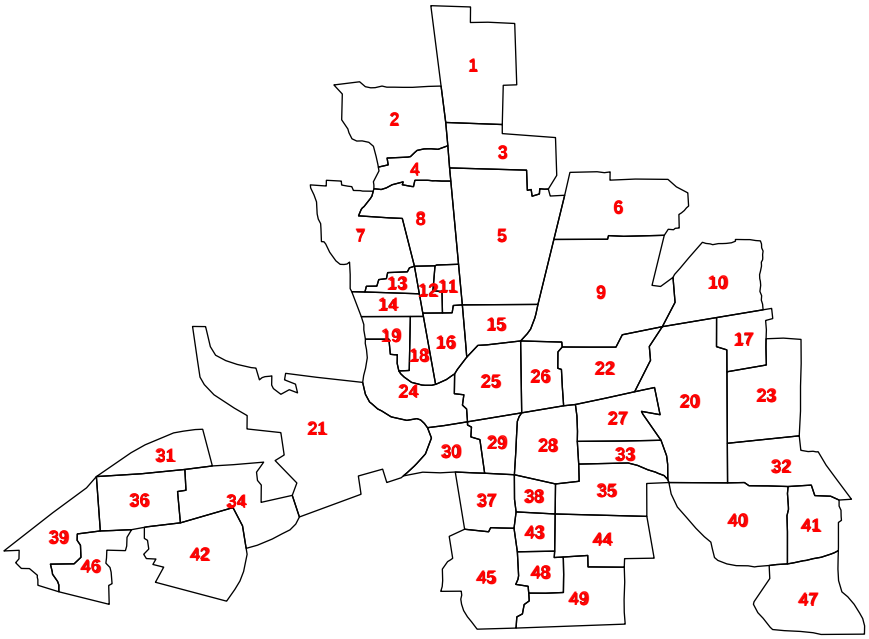

Figure 1 – Columbus districts

In the figure 1, the 49 districts of Columbus city of the state of OHIO are displayed, and for each of them a unique code is associated. The set of indices D can be defined here as $\mathbf{D} = \{1, 2, \ldots, 49\}$, and from this we can construct a structure of spatial dependence for the observations, considering neighboring regions whose borders touch each other. We could, for example, construct a neighborhood matrix $\mathbf{W}$ of dimension $n \times n$ given by the number of areas (regions, districts). The elements of this matrix detonate the spatial dependence between the regions. Looking at 1, we see that the line 6 of the matrix $\mathbf{W}$

which represents the relation of district 6 will have only the elements $w_{6,5} = 1$ and $w_{6,9} = 1$ with non-zero values. By definition the elements of the diagonal of the neighborhood matrix are equal to zero, $w_{ii} = 0$. As symmetry is also required, district i is neighbor to j if it is also j is neighbor of i, this is represented by the notation $i \sim j$. In this way, we can represent the set of neighbors of district 6 by $\mathrm{N}_6 \equiv \{6, 9\}$. And generally for the above example:

$$\mathrm{N}_i \equiv \{k : k \textit{ is a neighbor of } i \}, i = 1, \ldots, 49.$$

Simultaneous autoregressive (SAR) models

The SAR model was introduced by Whittle (1954), which defined a spatial process simultaneously in $\Re^2$ on a countable grid. These models have been studied extensively over the years, and are richly exposed in Cressie (1992), Cressie e Wikle (2011), and most recently in Hoef et al. (2018).These models have application in a diverse amount of scientific areas. As for example in texture, where the authors , Mao e Jain (1992), construct a multiresolution model based on SAR model for texture classification and texture segmentation. SAR models are also commonly used in the ecological data analysis, since they have a certain spatial pattern since they have a certain spatial pattern due to the proximity of the collected observations. Lichstein et al. (2002) analyze breeding habitat relationships for three common Neotropical migrant songbirds with the use of SAR models, with the use of the SAR model, which was adequate for the significance of the autocorrelation parameter.

This model is considered a GMRF with density function given by (3.1). The SAR model with zero mean is given as follows:

$$\gamma_i = \sum_{j=1}^{q} b_{ij}\gamma_j + \varepsilon_i, \quad i = 1, \ldots, q \tag{3.2}$$

which we can rewritten in matrix terms:

$$(\mathbf{I} - \mathbf{B})\boldsymbol{\gamma} = \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\mathbf{I}$ is an identity matrix $q \times q$. The error term is gaussian with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$. In its turn, $\mathbf{B}$ is an spatial dependence matrix with elements b_{ij} which denote the dependence between the area units. Thus, for example, $b_{35} > 0$ this means that the unit of area 3 depends on the unit 5. By convention, area units do not depend on themselves, implying that the elements on the diagonal, b_{ii}, are zero. We have to:

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B}^{\top})^{-1}). \tag{3.4}$$

Thus, for the covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{SAR}}$ to be positive-definite it is sufficient that $(\mathbf{I} - \mathbf{B})^{-1}$ exists (that is, $(\mathbf{I} - \mathbf{B})$ be full rank) and $\boldsymbol{\Lambda}$ be a positive diagonal matrix. In literature, the

definition of $\mathbf{B}$ is to take it as $\mathbf{B} = \rho\mathbf{W}$, where ρ is the parameter that denotes the spatial autocorrelation between the areas. To obtain a precision matrix of a specification of a SAR model we can look directly at the eigenvalues and eigenvectors of the neighborhood matrix $\mathbf{W}$. A sufficient condition for $(\mathbf{I} - \mathbf{B})$ to have inverse, in terms of $\mathbf{W}$, is that the parameter ρ is such that $1/\lambda_{[1]} < \rho < 1/\lambda_{[N]}$, whit $/\lambda_{[1]} < 0$ and $\lambda_{[N]} > 1$ denoting the smallest eigenvalue and higher eigenvalue of $\mathbf{W}$, respectively. Again, this condition is sufficient but not necessary. It is possible to obtain a specification of the SAR model without this condition being met, but in practical terms it is not carried out in this way Hoef et al. (2018). Another choice that can be made for $\mathbf{B}$ is such that $\mathbf{B} = \widetilde{\mathbf{W}}$. Each row of the neighborhood matrix is normalized and such that the sum is equal to 1, That is, the element of the normalized matrix are $\widetilde{w_{ij}} = w_{ij}/w_{i+}$. $\widetilde{\mathbf{W}}$ does not require symmetry, and is called the stochastic row because $\widetilde{\mathbf{W}}\mathbf{1} = \mathbf{1}$ Banerjee, Carlin e Gelfand (2004). If we define, in a similar way, $\mathbf{B} = \alpha\widetilde{\mathbf{W}}$, being α the spatial correlation parameter, (3.2) is modified to:

$$\gamma_i = \alpha \sum_{j \in N_i} \frac{w_{ij}}{\sum_k w_{ik}} \gamma_j + \varepsilon_{ij}, \tag{3.5}$$

with w_{ij} denoting the matrix element of $\mathbf{W}$, and N_i is the set of all indices of regions that are adjacent to region i. One point to note is that unlike the previous version eigenvalues have the restriction of $|\lambda_i| = 1$. And so for $(\mathbf{I} - \alpha\widetilde{\mathbf{W}})$ be full rank it is enough that $\alpha \in (-1, 1)$, and this explains α being denoted as a spatial autocorrelation parameter.

In general, the conditions, according to Hoef, Hanks e Hooten (2018), the conditions that guarantee a valid specification for covariance matrix of a SAR model are listed below:

- The matrix $\mathbf{I} - \mathbf{B}$ be full rank;
- The diagonal elements of $\mathbf{B}$ are zero;
- $\mathbf{\Lambda}$ is a diagonal matrix with positive elements.

These models are used in the area of spatial econometrics, and in this context are known as spatial lag model. As emphasized Cressie e Wikle (2011), the matrix $\mathbf{B}$ is seen in this field as a type of lag operator. Instead of time lag, for time series models, the lag is performed in space. The SAR model with zero mean ($\boldsymbol{\mu} = 0$) is written as $\mathbf{Y} = \mathbf{B}\mathbf{Y} + \boldsymbol{\nu}$, with $\boldsymbol{\nu} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, being this modeling analogous to the autoregressive , AR(1), process of order 1 in the context of time series.

In the next section, another popular auto-Gaussian (normal) model used in the area data modeling will be presented.

Conditionally autoregressive (CAR) models

The Conditionally autoregressive models (CAR) are attributed to Besag (1974). Similar to the SAR models, there is a vast literature covering them. And, according to

Banerjee, Carlin e Gelfand (2004), these models have been widely used in recent years in the context of spatial hierarchical models.

The CAR model is defined as follows Cressie (1992):

$$\mathrm{Z}_i|\ \mathbf{z}_{-i} \sim \mathrm{N}(\sum_{j=1}^{n} c_{ij}\mathrm{z}_j, m_i), \tag{3.6}$$

where Z_i is a random variable associated with the unit of area i. The vector $\mathbf{z}_{-i}$ denotes all realization z_j except ith. In its turn, $c_{ij}\mathrm{z}_j$ is the conditional mean of Z_i and m_i is conditional variance. The element c_{ij} from matrix $\mathbf{C}$ denotes the spatial dependence between the units of area i and j. And $\mathbf{M}$ is diagonal $n \times n$ matrix. Using the brook's lemma and Hammersley-Clifford (1971,*apud* Besag (1974)) theorem, it is shown that joint distribution is given by:

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}), \tag{3.7}$$

here, the elements of the covariance matrix must be symmetric with $c_{ij}m_i^{-1} = c_{ji}{m_j}^{-1}$, for all $i \neq j$, and $c_{ii} = 0$. If, again, we make the spatial dependence matrix equal to $\mathbf{C} = \rho\mathbf{W}$, with ρ representing the spatial autocorrelation parameter as we had in SAR models. To obtain a valid covariance matrix it is sufficient that $1/\lambda_{[1]} < \rho < 1/\lambda_{[\mathrm{N}]}$, whit $\lambda_{[1]}$ the smallest eigenvalue of $\mathbf{W}$ and $\lambda_{[\mathrm{N}]}$ the highest eigenvalue.

In summary, according to Hoef, Hanks e Hooten (2018), four conditions for obtaining a covariance matrix valid for the CAR model:

- The matrix $(\mathbf{I} - \mathbf{C})$ has positive eigenvalues;
- The diagonal elements of $\mathbf{C}$ are zero;
- All elements of $\mathbf{C}$ are symmetrical;
- $\mathbf{M}$ is a diagonal matrix with positive elements.ter. However, the reciprocal is not true.

These autoregressive spatial models, SAR and CAR, are the best known in the literature. And in the next section we will show how the former can be represented uniquely in terms of the latter. However, the reciprocal is not true.

3.2 The relationship between SAR and CAR models

To implement the SAR models in the GAMLSS it is necessary to check the relationship of these with the CAR models. Specifically, we need to know how to write the SAR model as a CAR Model. This relationship between these two models has been the

subject of research by researchers over time. An important result that was found in the literature was the equivalence between these two models when only if and only if their covariance matrices are the same, assuming that the mean was modeled correctly Cressie (1992), this is:

$$(\mathbf{I}-\mathbf{C})^{-1}\mathbf{M} = (\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Lambda}(\mathbf{I}-\mathbf{B}^{\top})^{-1}$$

The results show that any covariance matrix of the SAR model can be expressed as the covariance matrix of a CAR model, but the reverse is not true. In literature, Hoef, Hanks e Hooten (2018) investigate this relationship and makes a generalization for any definite matrix defined, here we follow them.

theoremny positive-definite covariance matrix $\mathbf{\Sigma}$ can be expressed as the covariance matrix of a CAR model $(\mathbf{I}-\mathbf{C})^{-1}\mathbf{M}$ for a unique pair of matrices .

Proof. Let $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ and decompose it into $\mathbf{Q} = \mathbf{D}$ - $\mathbf{R}$, where $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} = q_{ii}$, i.e. the elements in diagonal of precision matrix, and $\mathbf{R}$ has elements $r_{ij} = -q_{ij}$ and $r_{ii} = 0$. Let $\mathbf{C} = \mathbf{D}^{-1}\mathbf{R}$ and $\mathbf{M} = \mathbf{D}^{-1}$. Thus, $\mathbf{\Sigma}^{-1} = \mathbf{D}-\mathbf{R} = \mathbf{D}(\mathbf{I}-\mathbf{D}^{-1}\mathbf{R}) = \mathbf{M}^{-1}(\mathbf{I}-\mathbf{C})$, which shows $\mathbf{\Sigma}$ written as a covariance matrix of the CAR model, if the following conditions are attend:

1. $\mathbf{M}$ is strictly diagonal with positive values, so $\mathbf{M}$ and $\mathbf{M}^{-1}$ are positive-definite. By hypothesis,$\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ are positive-definite. Thus, $\mathbf{\Sigma} = (\mathbf{I}-\mathbf{C})\mathbf{M}$ and, by proposition, $(\mathbf{I}-\mathbf{C})^{-1}$ has positive eigenvalues and thus so does $\mathbf{I}-\mathbf{C}$.
2. By construction $m_{ij} = 0$, for $i \neq j$, $m_{ii} = \frac{1}{q_{ii}}$ and the fact that $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ is positive-definite, imply that $q_{ii} > 0$, for $i = 1, 2, 3 \ldots, n$. And consequently, $m_{ii} > 0$.
3. By Proposition 4, $c_{ii} = 0$ by the fact that $\mathbf{C} = \mathbf{D}^{-1}\mathbf{R}$.
4. For all $i \neq j$, $\mathbf{C}$ has elements $c_{ij} = d_{ii}^{-1}r_{ij} = m_{ii}r_{ij}$, and thus $\frac{c_{ij}}{m_{ii}} = r_{ij} = -q_{ij}$.

The symmetry of $\mathbf{Q}$ implies $q_{ij} = q_{ji}$, and we have that $\frac{c_{ij}}{m_{ii}} = \frac{c_{ji}}{m_{jj}}$. □

The proof of the uniqueness of the covariance matrix of the SAR model written as CAR, given by the authors is presented below:

Proof. Assume existence of $\widetilde{\mathbf{C}}$ and $\widetilde{\mathbf{D}}$ other than $\mathbf{C}$ and $\mathbf{D}$, respectively, and satisfying the four conditions in the previous proof. We have that if these matrices also satisfy $\mathbf{\Sigma}_{CAR} = \widetilde{\mathbf{M}}^{-1}(\mathbf{I}-\widetilde{\mathbf{C}})$, then $\text{diag}(\mathbf{M}) = \text{diag}(\widetilde{\mathbf{M}}) = \text{diag}(\mathbf{\Sigma}^{-1})$, by proposition 4, implying that $\mathbf{M} = \widetilde{\mathbf{M}}$, if these are diagonal matrices. From this fact it follows that $\widetilde{\mathbf{C}} = \mathbf{C}$, because $\widetilde{\mathbf{C}} = \mathbf{I} - \widetilde{\mathbf{M}}\mathbf{M}^{-1}(\mathbf{I}-\mathbf{C})$, and so the representation is unique.

□

Besag (1974) provides proof of equivalence between a first order SAR model and the third order CAR model in the context of a rectangular lattice. Be Y_{ij} be a random

variable in the i-th row and j-th column of the grid, and consider the generative process of this variable as a model:

$$\mathrm{Y}_{ij} = \delta_1 \mathrm{Y}_{i-1,j} + \delta_1^{'} \mathrm{Y}_{i+1,j} + \delta_2 \mathrm{Y}_{i,j-1} + \delta_2^{'} \mathrm{Y}_{i,j+1} + \varepsilon_{i,j}, \tag{3.8}$$

where $\varepsilon_{i,j}$ is white noise, admits that the covariance matrix of these are equal to the identity matrix of the same order ($\mathbf{\Lambda} = \mathbf{I}$), thus (3.8) has its representation in the CAR model given by:

$$\begin{aligned}\mathrm{E}(\mathrm{Y}_{i,j}|\{\mathrm{y}_{m,n} : (m,n) \neq (i,j)\}) = (1 + \delta_1^2 + \delta_1^{'2} + \delta_2^2 + \delta_2^{'2})^{-1}\{&(\delta_1 + \delta_1^{'})(\mathrm{y}_{i-1,j}\mathrm{y}_{i+1,j}) \\ &+ (\delta_2 + \delta_2^{'})(y_{i,j-1} + y_{i,j+1}) - (\delta_1\delta_2^{'} + \delta_1^{'}\delta_2)(y_{i-1,j-1} + y_{i-1,j+1}) \\ &- (\delta_1\delta_2 + \delta_1^{'}\delta_2^{'})(y_{i-1,j+1} + y_{i+1,j-1}) - (\delta_1\delta_1^{'})(y_{i-2,j} + y_{i+2,j}) \\ &- (\delta_2\delta_2^{'})(y_{i,j-2} + y_{i,j+2})\}.\end{aligned}$$

This equivalent representation of the SAR model in terms of the CAR model will be very useful as we will see later in the section of the next chapter that deals with the computational implementation of the SAR model for the purpose of this work.

3.3 The implementation of the SAR model in GAMLSS

This section will discuss how to implement the SAR in GAMLSS model. The relationship between the models of discrete space variation (area units,in our case) and nonparametric regression can be found in section 8.2 of Fahrmeir et al. (2013). In turn, the implementation of GMRF in GAMLSS can be seen in Bastiani et al. (2018).

The concept of neighborhood when it comes to units of area varies according to the approach adopted. Here we consider neighbors the units of areas that share the border of these polygons. In addition, if area i is neighbor of j, then j is neighbor of i, exhibiting a symmetric neighborhood relation. According to Fahrmeir et al. (2013), each unit of area ea will have its own regression coefficient γ_i, with $i = 1, \ldots, q$. In order that the coefficients obtained from neighboring regions are more similar, the authors impose a quadratic penalty as follows:

$$\mathrm{PLS}(\lambda) = \sum_{i=1}^{n}(y_i - \gamma_i)^2 + \lambda \sum_{u=2}^{q} \sum_{v \in N(u), v \leq u} (\gamma_u - \gamma_v)^2, \tag{3.9}$$

where $\mathrm{N}(u)$ is the set of all neighbors of area u and λ is smoothing parameter. Rewriting the PLS in matrix form we have to:

$$\mathrm{PLS}(\lambda) = (\boldsymbol{y} - \mathbf{Z}\boldsymbol{\gamma})^\top \mathbf{W}(\boldsymbol{y} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \mathbf{K} \boldsymbol{\gamma}, \tag{3.10}$$

as can be seen in section 9.4 of Stasinopoulos et al. (2017). We have that $\mathbf{Z}$ is a $n \times q$ matrix that associates each observation with its respective unit of area. That is:

$$Z_{i,u} = \begin{cases} 1, \text{if } y_i \text{ belongs region } u, \\ 0, \text{ otherwise .} \end{cases}$$

The $n \times n$ matrix of weights $\mathbf{W}$ is diagonal. The penalty matrix $\mathbf{K}$ has dimension $q \times q$ and has elements:

$$K_{u,v} = \begin{cases} 0, \text{if } u \text{ and } v \text{ if they are not neighbors,} \\ -1, \text{if } i \text{ and } j \text{ are neighbors ,} \\ n_u, \text{ the number of neighbors of } u, \forall u = v. \end{cases}$$

This penalty matrix represents the pseudo-inverse of the covariance matrix of the CAR model. And this represents the reason why the SAR model is not incorporated directly into the GAMLSS, through its covariance matrix. The value of $\hat{\boldsymbol{\gamma}}$ that minimizes (3.10) is $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^\top \mathbf{W} \mathbf{Z} + \lambda \mathbf{K})^{-1} \mathbf{Z}^\top \mathbf{W} \boldsymbol{y}$.

The link between (penalized) smooths, random effects and random fields can be found in section 5.8 of Wood (2017). The author state that the penalty can be a prior distribution as follows:

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \lambda \mathbf{K}^{-1}) \tag{3.11}$$

In this way, the precision model of the SAR model represented by the CAR Model can be incorporated into the GAMLSS models. Where $\boldsymbol{\gamma}$ is a *instrisc* GMRF.

To better exemplify how the SAR Models can be incorporated into the GAMLSS approach, consider figure 1 again from the Columbus districts. From a given configuration of neighbors, and again noting that a first order SAR model is equivalent to the third order CAR model, we obtain the third order neighborhood configuration. Look at figure 2 which is a subset of the districts of Columbus and the indices denoting the areas were remarked.

In order to construct a valid $\mathbf{K}^{-1}$ precision matrix based on the general penalization scope presented above, we can start from the matrix of neighbors $\mathbf{W}$. Each of the seven areas in the figure can be considered as a vertex, as explained in Chapter 2. Neighbors up to third order are obtained when the minimum number of edges between a region is equal

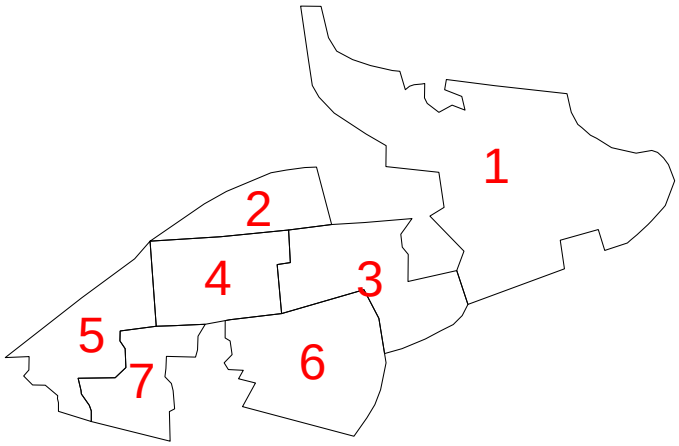

Figure 2 – Columbus Subset of Districts

to $k \leq 3$. Leading to a neighborhood matrix in this case as such:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

where each row in the matrix above represents a represents the neighborhood relation of the regions of figure 2. The element $w_{1,2}$ informs that the area 1 and the area 2 are neighbor of third order.

And so, the precision matrix K, from this neighborhood structure is:

$$\mathbf{K}^{-1} = \begin{bmatrix} 4 & -1 & -1 & -1 & 0 & -1 & 0 \\ -1 & 6 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 6 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 6 & -1 & -1 & -1 \\ 0 & -1 & -1 & -1 & 5 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 6 & -1 \\ 0 & -1 & -1 & -1 & -1 & -1 & 5 \end{bmatrix}$$

Thus each element on the diagonal of $\mathbf{K}^{-1}$ shows how many neighbors the area i has. And the elements outside the diagnoal inform about the neighborhood relation. Since $\mathbf{K}^{-1}$ is a valid matrix for the penalty criteria. set out above.

4 Simulation study

This chapter aims to evaluate the properties of estimators of the coefficients of the GAMLSS-SAR model in the context of finite samples. The simulation study of Monte Carlo differs from the usual in the spatial context. Because in addition to the generation of random variables, we have to introduce spatial dependence in these variables.

To generate the normal response variable in the spatial context, to generate a sample of size $n = 20$, which is equivalent to the number of regions, we simulate a vector of coordinates (x_1, x_2) of size 15 from $\mathcal{U} \sim (0, 1)$ and perform the Delaunay triangulation to obtain the regions (polygons), being realized through the package `gDelaunayTriangulation` in `R`. For a sample of sizes $n = 50$ and $n = 100$, it was necessary to have the coordinate vectors have sizes 31 and 56, respectively. We used 1000 replicas of Monte Carlo for each n sample size. Figure 3 shows each of the regions obtained for the simulation study. To compute spatial dependence we generate n observations with normal probability function and create the covariance matrix of the SAR model from the above polygons and render the values of the areas which are more similar neighbors, we do the cholesky decomposition of this and make the product by the observations generated previously. And so, we created a sample of spatially correlated observations. This medotology follows the work of Haining e Haining (2003). The procedure was performed as follows:

- We obtain the cholesky decomposition for a square matrix $\mathbf{\Sigma}$ of order n such that $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$. Where $\mathbf{L}$ is a lower triangular n by n matrix. The valid matrix taken here is the covariance matrix of the SAR model;

- We generated the n observations of the covariate $\mathbf{x}$ that follows from a uniform $(\mathcal{U})$ probability distribution in the range of 0 to 3;

- For each of the 1000 replicates of monte carlo a vector $\boldsymbol{\varepsilon}$ of length n is generated from uncorrelated normal random variables; and

- We compute our response variable y with spatial dependency for each replicate by doing $\boldsymbol{y} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = \beta_0 + \boldsymbol{x}\beta_1$.

Four different scenarios were computed for estimating the coefficients of regression coefficients based on spatial correlation. We made $\rho = (0.0, 0.2, 0.5, 0.9)$, and in all scenarios we set $\sigma^2 = 1$ and $\boldsymbol{\beta} = (\beta_0, \beta_1) = (2.5, -0.5)$.

The measures used in the comparison of the models in all studies of simulation

were the Relative Bias (%) and the Mean Square Error (MSE):

$$\text{Relative Bias} = \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\beta}_k^{(i)} - \beta_k}{\beta_k},$$

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_k^{(i)} - \beta_k)^2,$$

where m is the number of replicas of Monte Carlo, β_k is the true value of the k-th parameter. $\hat{\beta}_k^{(i)}$ is the i-th estimate of the k-th parameter, and the estimation of parameter β_k is calculated as:

$$\hat{\beta}_k = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_k^{(i)}.$$

4.1 Evaluation in the Normal linear model with spatial dependence

We compared regression coefficients estimators of lag SAR model, Error SAR model, and SAR incorporated into the GAMLSS. the first two models are given, respectively, by:

$$\boldsymbol{y} = \rho\mathbf{W}\boldsymbol{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$
$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u},$$

with $\boldsymbol{u} = \rho\mathbf{W}\boldsymbol{u} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The strict difference of these two models is that in Lag-SAR the spatial dependency is incorporated in the linear trend, whereas the error-SAR model places it in the term of stochastic perturbation.

In Table 1 the scenario is when there is no spatial dependence, $\rho = 0.0$. We note that the estimators of the GAMLSS-SAR model for the coefficients β_0 and β_1 appear to have consistency, note that the biases fall when n increases.

Table 1 – Normal Spatial model with $\rho = 0.0$

$\rho = 0.0, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Estimators	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.694545	7.781784	0.000230	-0.500655	0.131060	0.000012
Error-SAR	2.506123	0.244904	0.000033	-0.502579	0.515909	0.000015
GAMLSS-SAR	2.506767	0.270689	0.000009	-0.503059	0.611771	0.000003
			$n = 50$			
Lag-SAR	2.694636	7.785444	0.000057	-0.4977697	0.4460568	0.000000
Error-SAR	2.503084	0.1233637	0.000001	-0.5010539	0.2107867	0.000000
GAMLSS-SAR	2.503468	0.1387004	0.000000	-0.5013196	0.2639267	0.000000
			$n = 100$			
Lag-SAR	2.590364	3.614572	0.0021991	-0.4984477	-0.310461	0.000000
Error-SAR	2.501593	0.063705	0.000000	-0.5006446	0.128914	0.000000
GAMLSS-SAR	2.502328	0.093130	0.000000	-0.5011519	0.230376	0.000000

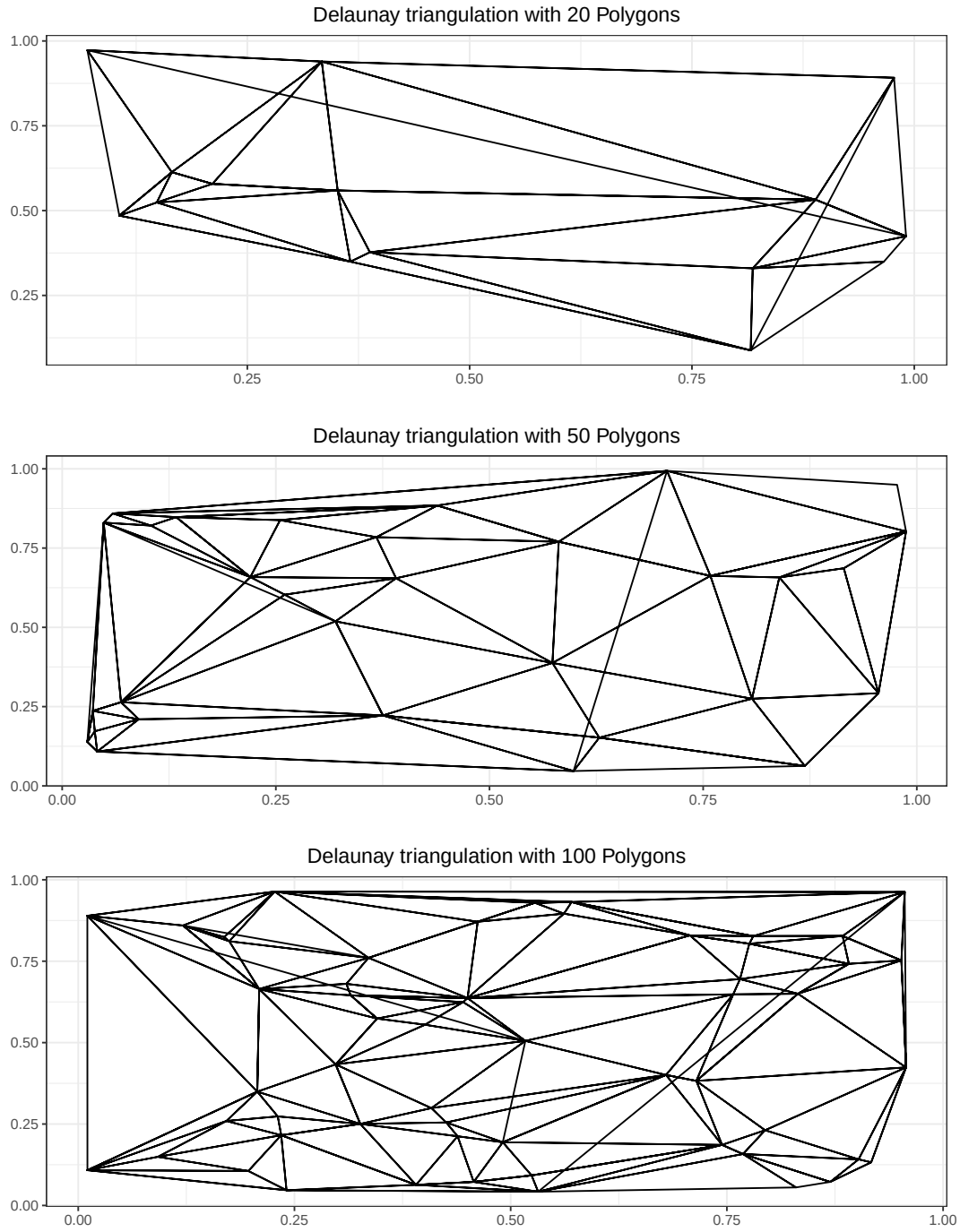

Figure 3 – Delaunay Triangulation for Simulation Study

In general, when we increase the value of the spatial dependence, $\rho = 0.2, 0.5, 0.9$, we verify the continuity of the consistency of the estimators of the coefficients of the GAMLSS-SAR model. This model is an alternative to the Error-SAR and Lag-SAR models.

Table 2 – Normal Spatial model with $\rho = 0.2$

$\rho = 0.2, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.783287	11.33146	0.000997	-0.499925	-0.01489607	0.000008
Error-SAR	2.515448	0.617922	0.000157	-0.504797	0.9594145	0.000062
GAMLSS-SAR	2.518194	0.727754	0.000042	-0.506832	1.366419	0.000021
			$n = 50$			
Lag-SAR	2.54547	1.8188	0.000172	-0.4985324	-0.2935294	0.000000
Error-SAR	2.502714	0.1085535	0.000001	-0.5010358	0.2071629	0.000000
GAMLSS-SAR	2.503066	0.1226597	0.000000	-0.5012648	0.2529573	0.00000
			$n = 100$			
Lag-SAR	2.513741	0.549622	0.001676	-0.499871	-0.02579388	0.000000
Error-SAR	2.50135	0.053999	0.000000	-0.5005187	0.1037453	0.000000
GAMLSS-SAR	2.501972	0.078878	0.000000	-0.5009573	0.191456	0.000000

Table 3 – Normal Spatial model with $\rho = 0.5$

$\rho = 0.5, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.759376	10.37502	0.000525	-0.5002511	0.050225	0.000030
Error-SAR	2.509733	0.3893094	0.000078	-0.5036798	0.735959	0.000036
GAMLSS-SAR	2.511152	0.4460973	0.000022	-0.5047282	0.945630	0.000009
			$n = 50$			
Lag-SAR	2.583673	3.346904	0.000113	-0.4990897	-0.182051	0.000000
Error-SAR	2.501755	0.070213	0.000000	-0.5006813	0.136263	0.000000
GAMLSS-SAR	2.502072	0.082868	0.000000	-0.5009094	0.181871	0.000000
			$n = 100$			
Lag-SAR	2.535689	1.427547	0.000767	-0.4995111	-0.097778	0.000000
Error-SAR	2.500863	0.034524	0.000000	-0.5003425	0.068509	0.000000
GAMLSS-SAR	2.501254	0.050144	0.000000	-0.5006196	0.123916	0.000000

4.2 Evaluation in the Normal linear model with spatial dependence and linear and non-linear trend

In this simulation experiment, we added a nonlinear trend to the predictor of $\boldsymbol{\mu}$. We compute our response variable y with spatial dependency for each replicate by doing $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + f(\boldsymbol{x_2}) + \mathbf{L}\boldsymbol{\varepsilon}$, where $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are generated from $\mathcal{U} \sim (0, 3)$ and $\mathcal{U} \sim (1, 5)$, respectively. The nonlinear trend is computed as follows:

$$f(x_2) = \frac{1}{\sqrt{2\pi 0.5}} \exp - \frac{(x_2 - 5)^2}{1}$$

The objective of introducing the nonlinear trend term, $f(\boldsymbol{x_2})$, was to evaluate the

Table 4 – Normal Spatial model with $\rho = 0.9$

$\rho = 0.9, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.694545	7.78178	0.000230	-0.5006553	0.13106	0.000018
Error-SAR	2.506123	0.244904	0.000032	-0.5025795	0.5159095	0.000014
GAMLSS-SAR	2.506767	0.270689	0.000009	-0.5030589	0.6117707	0.000003
			$n = 50$			
Lag-SAR	2.569618	2.784713	0.000062	-0.4995511	-0.089776	0.000000
Error-SAR	2.501071	0.042826	0.000000	-0.5004079	0.081579	0.000000
GAMLSS-SAR	2.501321	0.052831	0.000000	-0.5005936	0.1187114	0.000000
			$n = 100$			
Lag-SAR	2.524943	0.9977099	0.000292	-0.4997164	-0.05672576	0.000000
Error-SAR	2.500537	0.02147225	0.000000	-0.5002214	0.04428511	0.000000
GAMLSS-SAR	2.500776	0.0310411	0.000000	-0.5003908	0.07815075	0.000000

estimators of β_0 and β_1 of models in this scenario. The modeling done in GAMLSS-SAR for this variable was using *cubic splines*, and in the other models $\boldsymbol{x}_2$ was modeled as a linear term.

Table 5 – Normal Spatial model with $\rho = 0.0$, linear and non-linear trend

$\rho = 0.0, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.693144	7.725762	0.000982	-0.534463	5.015073	0.000109
Error-SAR	2.291171	-8.353153	0.000271	-0.525075	6.892525	0.000171
GAMLSS-SAR	2.260849	-9.566034	0.000288	-0.503671	0.734145	0.000068
			$n = 50$			
Lag-SAR	2.382542	-4.698317	0.000613	-0.504239	0.84781	0.000000
Error-SAR	2.285214	-8.591442	0.000050	-0.509385	1.877011	0.000000
GAMLSS-SAR	2.266537	-9.338515	0.000043	-0.499835	-0.033017	0.000000
			$n = 100$			
Lag-SAR	2.201387	-15.68187	0.000882	-0.490501	-1.899874	0.000000
Error-SAR	2.107953	-11.94453	0.000136	-0.490191	-1.961867	0.000000
GAMLSS-SAR	2.213131	-11.47476	0.000119	-0.499582	-0.083641	0.000000

4.3 Evaluation in the Poisson spatial regression model with spatial dependence

In general, we observed that the GAMLSS-SAR model with dependent variable with Poisson distribution has consistent estimators for the β_0 and β_1 regression coefficients.

Table 6 – Normal Spatial model with $\rho = 0.2$, linear and non-linear trend

$\rho = 0.2, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.517926	0.7170418	0.000468	-0.5234491	4.689823	0.000088
Error-SAR	2.286193	-8.552293	0.000198	-0.531755	6.351000	0.000138
GAMLSS-SAR	2.261336	-9.546546	0.000248	-0.5033237	0.664735	0.000060
			$n = 50$			
Lag-SAR	2.221539	-11.13842	0.000955	-0.5043736	0.874725	0.000000
Error-SAR	2.282965	-8.681399	0.000059	-0.508189	1.637791	0.000000
GAMLSS-SAR	2.264842	-9.406303	0.000045	-0.4997035	-0.059306	0.000000
			$n = 100$			
Lag-SAR	2.005718	-19.77128	0.000521	-0.4922073	-1.558536	0.000000
Error-SAR	2.199386	-12.02455	0.000121	-0.4894004	-2.119924	0.000000
GAMLSS-SAR	2.212441	-11.50237	0.000112	-0.4992283	-0.1543327	0.000000

Table 7 – Normal Spatial model with $\rho = 0.5$, linear and non-linear trend

$\rho = 0.5, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.480585	-0.7765946	0.000152	-0.524263	4.852608	0.000041
Error-SAR	2.292612	-8.295529	0.000062	-0.5342793	6.855854	0.000070
GAMLSS-SAR	2.259231	-9.630753	0.000063	-0.5014855	0.297094	0.000024
			$n = 50$			
Lag-SAR	2.227996	-10.88017	0.000838	-0.5049094	0.9818712	0.000000
Error-SAR	2.283994	-8.640228	0.000055	-0.5091663	1.83326	0.000000
GAMLSS-SAR	2.262535	-9.498594	0.000043	-0.4995397	-0.092059	0.000000
			$n = 100$			
Lag-SAR	1.970019	-21.19924	0.000070	-0.4928455	-1.430909	0.000000
Error-SAR	2.200078	-11.53167	0.000100	-0.4894785	-2.104308	0.000000
GAMLSS-SAR	2.211708	-11.53167	0.000099	-0.4988736	-0.225271	0.000000

For all the spatial dependence scenarios we find that the relative biases and the MSE fall as n increases.

4.4 Evaluation in the Gumbel spatial regression model with spatial dependence

In this section, we analyze the performance of the β_0 and β_1 regression coefficients for the Gumbel regression model with spatial dependence. The Gumbel distribution is given by:

Table 8 – Normal Spatial model with $\rho = 0.09$,linear and non-linear trend

$\rho = 0.9, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Models	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Lag-SAR	2.430456	-2.781767	0.000029	-0.5003117	4.897165	0.000019
Error-SAR	2.304411	-7.82354	0.000010	-0.5388571	7.771426	0.000035
GAMLSS-SAR	2.2588	-9.647982	0.000005	-0.5003117	0.062342	0.000009
			$n = 50$			
Lag-SAR	2.197432	-12.10274	0.000627	-0.5051787	1.035749	0.000000
Error-SAR	2.285445	-8.582197	0.000052	-0.5103991	2.07982	0.000000
GAMLSS-SAR	2.261267	-9.54932	0.000046	-0.4994408	-0.111832	0.000000
			$n = 100$			
Lag-SAR	1.927576	-22.89696	0.000092	-0.4935431	-1.291382	0.000000
Error-SAR	2.199075	-12.03701	0.000007	-0.489009	-2.198203	0.000000
GAMLSS-SAR	2.212065	-11.51739	0.000091	-0.4989038	-0.219245	0.000000

Table 9 – Poisson Spatial with $\rho = 0.0$

$\rho = 0.0, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Estimators	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Poisson-GLM-Lag	2.312969	-7.481239	0.000108	-0.3148691	-37.02618	0.000074
GAMLSS-SAR	2.330345	-6.786186	0.000108	-0.3251679	-34.96642	0.000074
			$n = 50$			
Poisson-GLM-Lag	2.581728	3.269118	0.000007	-0.5399141	7.982827	0.000005
GAMLSS-SAR	2.565316	2.612644	0.000009	-0.5089003	1.780062	0.000011
			$n = 100$			
Poisson-GLM-Lag	2.537855	1.514185	0.000000	-0.4759212	-4.815768	0.000002
GAMLSS-SAR	2.581951	-4.815768	0.000006	-0.4933474	-1.330523	0.000000

$$f(y|\mu,\sigma) = \frac{1}{\sigma}\exp\{(\frac{y-\mu}{\sigma}) - \exp(\frac{y-\mu}{\sigma})\},$$

where $y \in \Re$, $\mu \in \Re$ and $\sigma > 0$.

Table 10 – Poisson Spatial with $\rho = 0.2$

$\rho = 0.2, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Estimators	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Poisson-GLM-Lag	2.449203	-2.031865	0.000119	-0.449054	-10.1892	0.000049
GAMLSS-SAR	2.419278	-3.228862	0.000119	-0.4474102	-10.51796	0.000049
			$n = 50$			
Poisson-GLM-Lag	2.537198	1.4879	0.000001	-0.4903463	-1.93073	0.000000
GAMLSS-SAR	2.545802	1.832098	0.000000	-0.4882551	-2.348976	0.000000
			$n = 100$			
Poisson-GLM-Lag	2.494323	-0.227061	0.000001	-0.4889286	-2.214278	0.000000
GAMLSS-SAR	2.496011	-0.159556	0.000001	-0.4884424	-2.311515	0.000000

Table 11 – Poisson Spatial with $\rho = 0.5$

$\rho = 0.5, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Estimators	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Poisson-GLM-Lag	2.551156	2.046232	0.000042	-0.4974498	-0.5100495	0.000003
GAMLSS-SAR	2.55452	2.180812	0.000050	-0.494387	-1.122606	0.000003
			$n = 50$			
Poisson-GLM-Lag	2.533883	1.355303	0.000003	-0.497273	-0.545446	0.000002
GAMLSS-SAR	2.538595	1.543814	0.000001	-0.493733	-1.25345	0.000002
			$n = 100$			
Poisson-GLM-Lag	2.533883	1.355303	0.000003	-0.497273	-0.5454465	0.000002
GAMLSS-SAR	2.538595	1.543814	0.000001	-0.493733	-1.253454	0.000002

Table 12 – Poisson Spatial with $\rho = 0.9$

$\rho = 0.9, \sigma = 1$	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
Estimators	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$n = 20$			
Poisson-GLM-Lag	2.610163	4.4065	0.000020	-0.5725064	14.50127	0.000000
GAMLSS-SAR	2.590584	3.623374	0.000043	-0.5608624	12.17247	0.000001
			$n = 50$			
Poisson-GLM-Lag	2.506273	0.2509324	0.000001	-0.4831396	-3.372076	0.000000
GAMLSS-SAR	2.506159	0.2463548	0.000001	-0.4852572	-2.948556	0.000000
			$n = 100$			
Poisson-GLM-Lag	2.518183	0.727305	0.000000	-0.509590	1.91802	0.000002
GAMLSS-SAR	2.520564	0.822546	0.000000	-0.508557	1.711508	0.000002

Table 13 – Estimates of SAR-GAMLSS model coefficients for response variable with Gumbel distribution with spatial dependence

	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
n	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$\rho = 0.0$			
20	2.416102	-3.3559	0.003240	-0.530625	6.12507	0.000812
50	2.501353	0.05412987	0.000015	-0.5054212	1.084232	0.000002
100	2.517094	0.6837562	0.000009	-0.5090158	1.803166	0.000004
			$\rho = 0.2$			
20	2.396089	-4.156442	0.000317	-0.5219365	4.387303	0.000057
50	2.495277	-0.1889155	0.000080	-0.5054232	1.084637	0.000010
100	2.509855	0.3941852	0.000002	-0.5111031	2.220623	0.000003
			$\rho = 0.5$			
20	2.388944	-4.442252	0.000145	-0.5190735	3.814698	0.000001
50	2.493605	-0.255812	0.000026	-0.5124728	2.494566	0.000007
100	2.498865	-0.045383	0.000000	-0.5057683	1.153652	0.000007
			$\rho = 0.9$			
20	2.357355	-5.705787	0.0000001	-0.5055049	1.10098	0.000003
50	2.495929	-0.1628443	0.000125	-0.5180981	3.619616	0.000106
100	2.490843	-0.3662853	0.000010	-0.5004848	0.09695771	0.000016

4.4.1 Simulation Based on `Columbus` Dataset

In this section we will do the simulation study considering the spatial structure of the columbus data set. We have three parameters to estimate: $\beta_0, \beta_1, \sigma^2$. Three scenarios will be evaluated: the regression coefficients will be varied, for fixed σ and ρ. Then we will vary σ, to ρ and fixed betas. And finally, we'll vary ρ to fixed β_0, β_1 and σ^2.

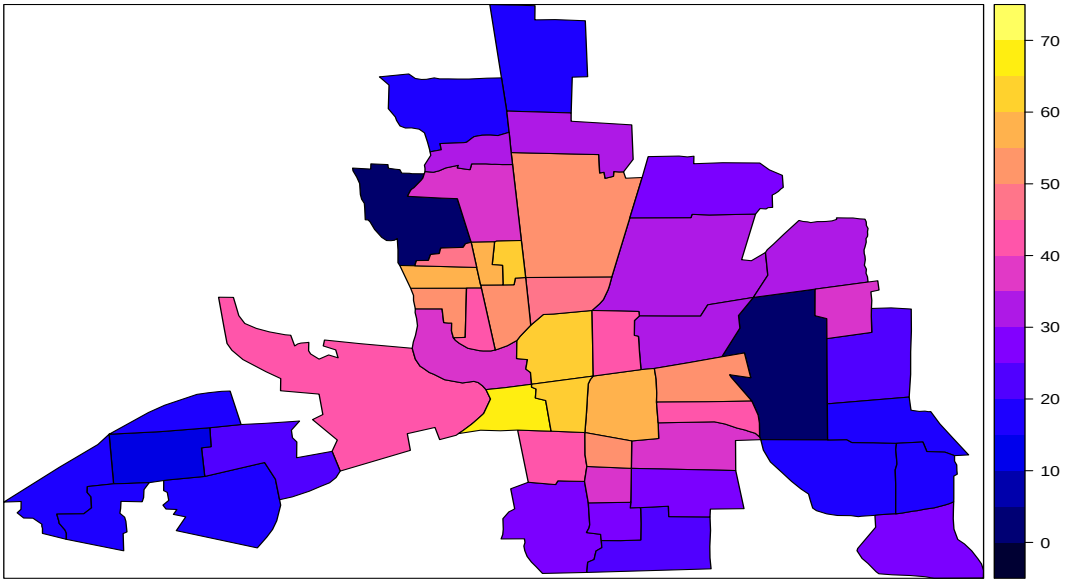

Figure 4 – Plot of `Columbus` polys.

	Estimate	Relative Bias (%)	MSE
	$(\beta_0, \beta_1, \sigma, \rho) = (1.0, 0.5, 1, 0.4)$		
$\hat{\beta}_0$	0.996631	-0.336921	0.000002
$\hat{\beta}_1$	0.501135	0.227048	0.000000
	$(\beta_0, \beta_1, \sigma, \rho) = (2.5, 1.0, 1, 0.4)$		
$\hat{\beta}_0$	2.500037	0.001494	0.000000
$\hat{\beta}_1$	1.000897	0.089712	0.000001
	$(\beta_0, \beta_1, \sigma, \rho) = (1.0, 0.5, 2.0, 0.4)$		
$\hat{\beta}_0$	1.007683	0.7682966	0.000000
$\hat{\beta}_1$	0.498203	-0.3594192	0.000004
	$(\beta_0, \beta_1, \sigma, \rho) = (1.0, 0.5, 2.0, 0.9)$		
$\hat{\beta}_0$	0.999666	-0.0334229	0.000000
$\hat{\beta}_1$	0.500627	0.125458	0.000002

5 Applications

5.1 Boston Housing Data

To illustrate the use of the GMRF in the GAMLSS models with the spatial structure being represented by the spatial configuration of the SAR model, we used a classic set of data in the literature. It's hedonic pricing data of Jr e Rubinfeld (1978), in its corrected version. In this article, the authors analyzed the demand for clean air through a hedonic price model for residences in Boston. Considering the spatial structure, Pace e Gilley (1997) estimates a parametric SAR model and obtains a more accurate prediction of the parameters. However, the comparison we make here is made relevant by the fact that some premises of this model may not be true, As, for example, the distribution of the response variable follows the normal distribution. Another relevant point is that we can compare fittings of models in which the spatial term is parameterized and another that the term is smoothing.

This section is divided as follows. First we present the database with descriptive measures. Then, we fit a model as without spatial configuration as in Pace e Gilley (1997) and check the diagnostic part of the model. The same is done considering the spatial model of the authors. And lastly, Finally, we fitting the gamlss models with spatial configuration SAR as smoothing.

Description of the variables

The independent variables used in the database are: crime per capita in the town (`CRIM`); the units occupied by the owners, proportionally, which were built before the year 1940 (`AGE`); nitric oxides (`NOX`); borders Charles River(`CHAS`),which is a factor indicating if the property is near the Charles River; number of rooms (`RM`); proportion of residential land zoned for lots over 25,000 (`ZN`),proportion of non-retail business acres per town,(`INDUS`); pupil-teacher ratio by town (`PTRATIO`); index of accessibility to radial highways (`RAD`); full-value property-tax rate per 10,000 (`TAX`); proportion of blacks by town (`B`), lower status of the population (`LSTAT`); weighted mean of distances to five Boston employment centres (`DIS`); latitude of census tract (`LAT`); and longitude (`LONG`) of census tract.The response variable is the logarithm of the median corrected value of household values in USD 1000's (`PRICE`).

As we can see in figure 5, in the left chart, the data appear to be symmetrical. In the chart to the right the box plot of this variable is displayed, and shows the presence of many points considered as outliers. The symmetry of the data appears again, we can see

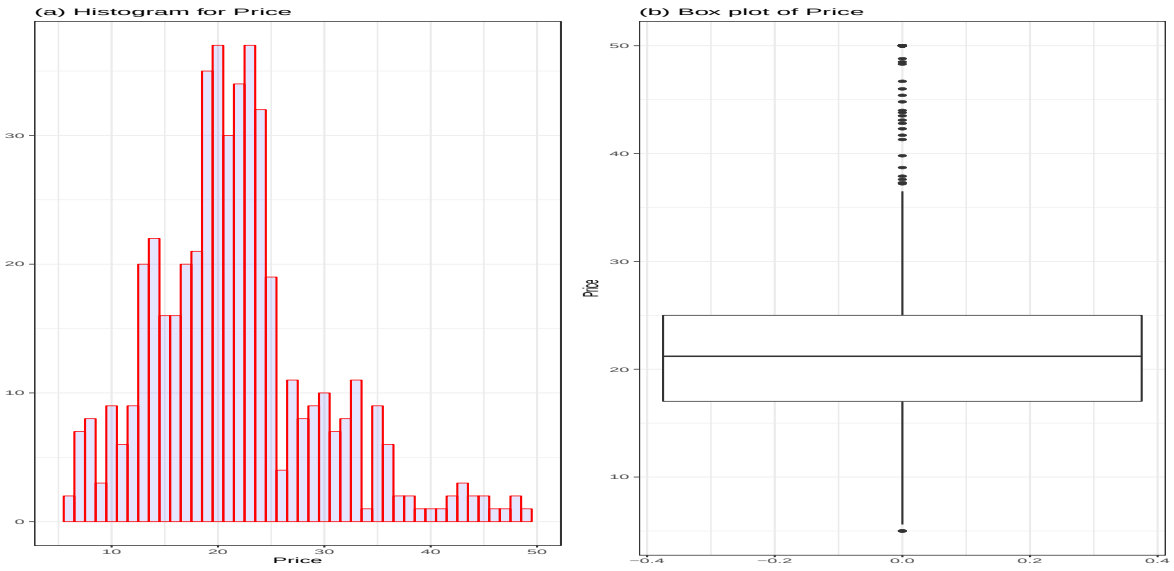

Figure 5 – Histogram (left) and Boxplot (right) of `price`.

that they are little dispersed.On the other hand, when we look at the summary measures we see homogeneous data with coefficient of variation equal to 0.135. The skewness is -0.334 and kurtosis is 3.808. Therefore a probability distribution that differs from the Gaussian can be required to model the respose variable, `PRICE`.

The plot of the response variable against explanatory variables is given by figure 6 Except `CHAS`, all variables are continuous. For this, the median price increases when the residence is located near the Charles River.Indicating, once again, that the data the distribution of the response variable needs to deal with skewness. When we look at the median to `RM`, we note a linear positive relation. When we look at the median to MR, we note a linear positive relation. Differently, the median relationship of the response variable with `LSTAT` is linearly negative. For all other variables, the relation a complex relation is drawn, thus requiring some non-parametric function in the modeling. Another point to note here is the homoscedasticity hypothesis, present in linear models, that appears to be violated. Thus, it is necessary to model the dispersion parameter as a function of explanatory variables.

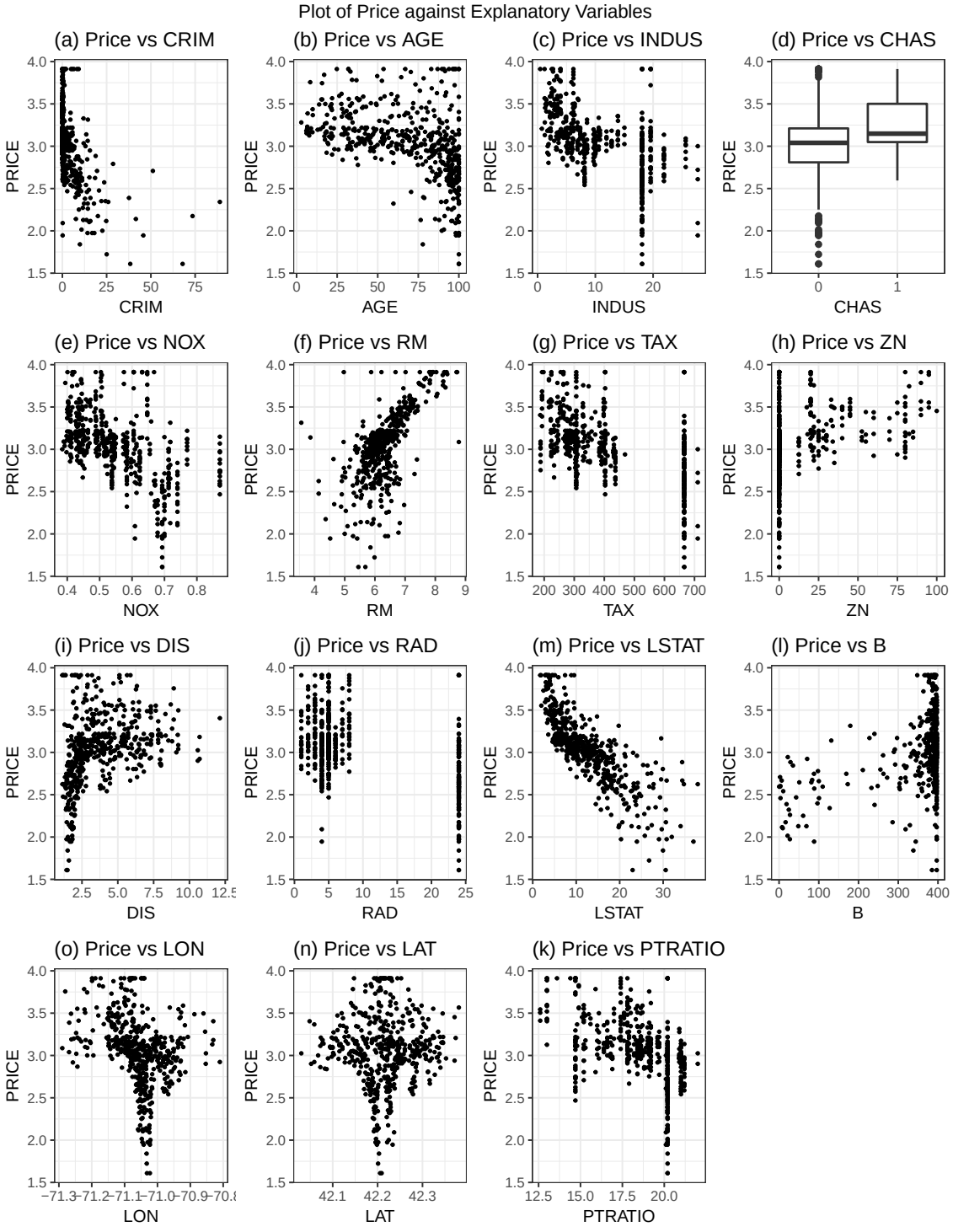

Figure 6 – Plot of `Price` against exploratory variables.

Comparative

In the first estimated model we ignored the spatial configuration of the data, and as in Pace e Gilley (1997),the explanatory variables were:

$$\begin{aligned}\log(\texttt{PRICE}) = \beta_0 &+ \beta_1\texttt{CRIM} + \beta_2\texttt{AGE} + \beta_3\texttt{NOX} + \beta_4\texttt{CHAS} + \beta_5\texttt{RM} \\ &+ \beta_6\texttt{ZN} + \beta_7\texttt{INDUS} + \beta_8\texttt{PTRATIO} + \beta_9\texttt{RAD} + \beta_{10}\texttt{TAX} \\ &+ \beta_{11}\texttt{B} + \beta_{12}\texttt{LSTAT} + \beta_{13}\texttt{DIS} + \beta_{14}\texttt{LAT} + \beta_{15}\texttt{LONG} \\ &+ +\beta_{16}\texttt{LAT}^2 + \beta_{17}\texttt{LONG}^2 + +\beta_{18}\texttt{LAT*LONG}\end{aligned}$$

To better analyze this model, we perform a residuals analysis in figure 7 to verify the suitability of the model.

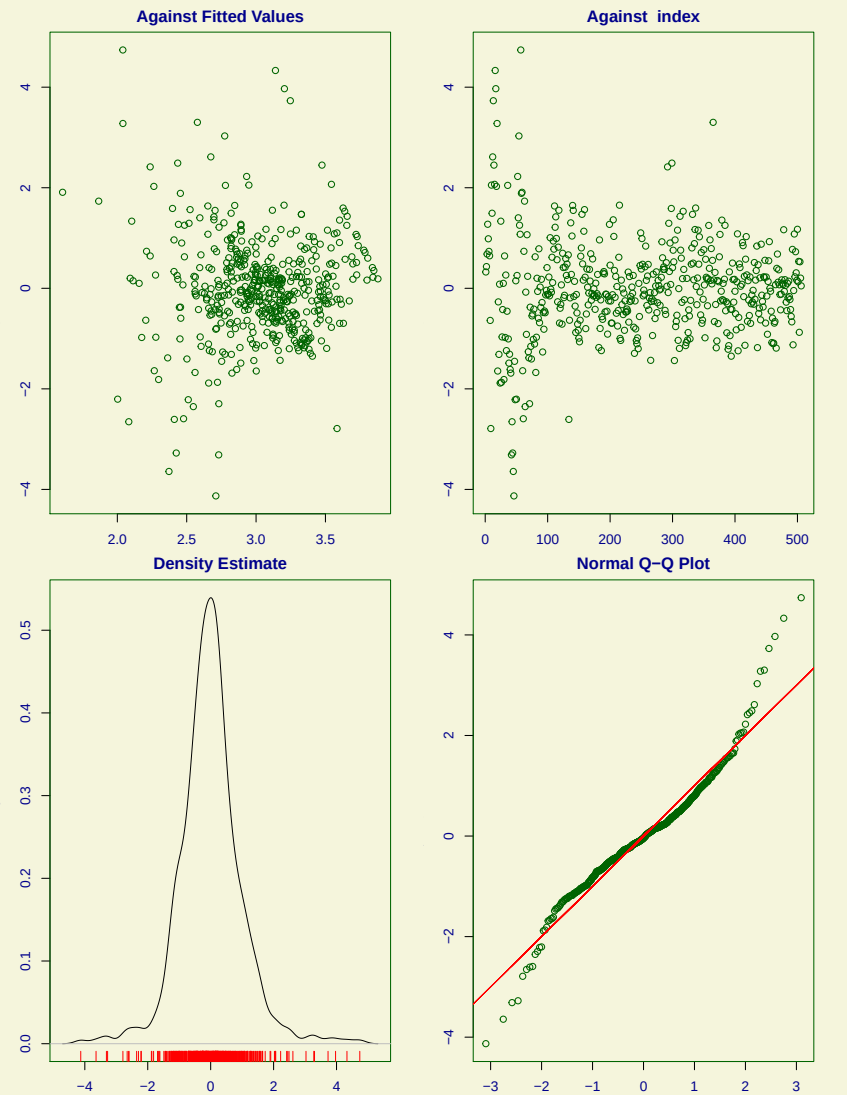

Figure 7 – Residual plots of the non-spatial model in pace1997using

In the lower left chart, in the Normal QQ plot, we noticed some problem with the tails of the distribution. That is checked by the shape parameter estimates. The asymmetry coefficient is 0.383 and the kurtosis coefficient is 6.41. That is, we can rule out the normality hypothesis for the residuals.

In addition, the simple Worm plot Buuren e Fredriks (2001) which is one way of ascertaining the adequacy of the regression residuals, is given by figure 8.

This graph shows the non-adequacy of the distribution of the response variable, showing that the ordered residuals are far from their approximate expected values (indicated by the red horizontal line).

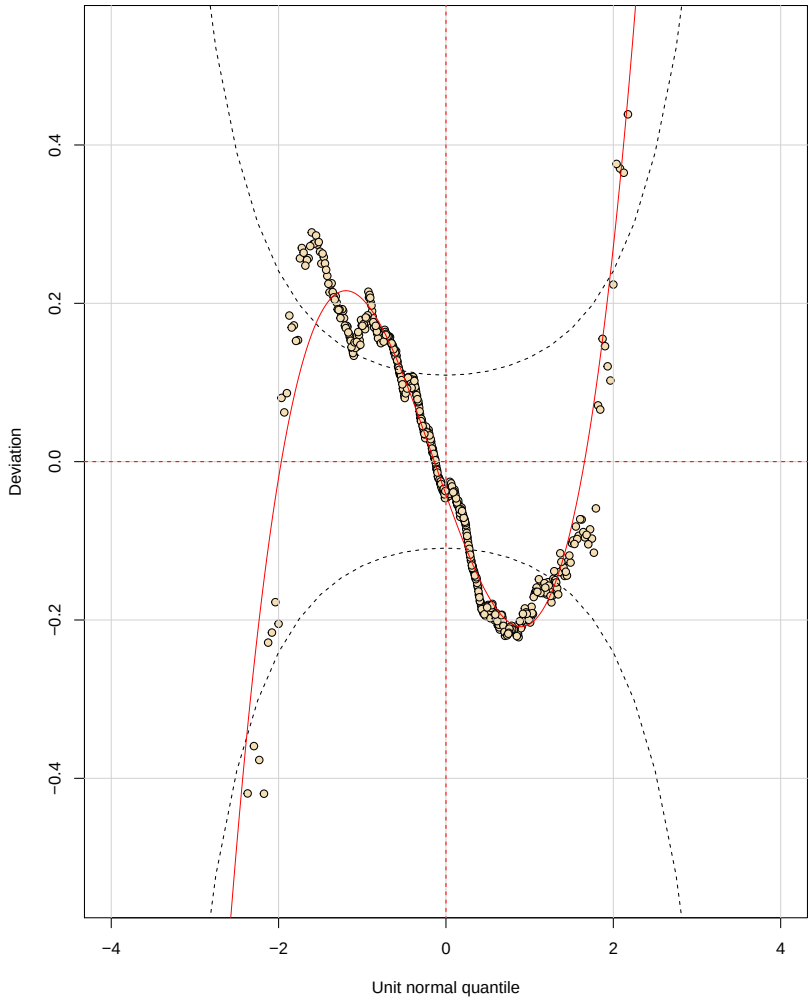

Figure 8 – Worm Plot of pace1997using Model

Parametric SAR

The approach used for the SAR model in the field of spatial econometrics as spatial lag model uses the coefficient of spatial autocorrelation as another parameter to be estimated by the model as it is exposed in chapter 2.

The following interpretation, once again, is of the non-suitability of the model by the residue analysis present in the figure 9. On the left side, Fitted Values against Residuals, the dots do not seem randomly distributed totally, especially for low quantities of fitted values. On the right side, in turn, QQ plot reveals that the distribution of residues does not resemble one that comes from normal. Revealing that the inferences made by the authors may be wrong.

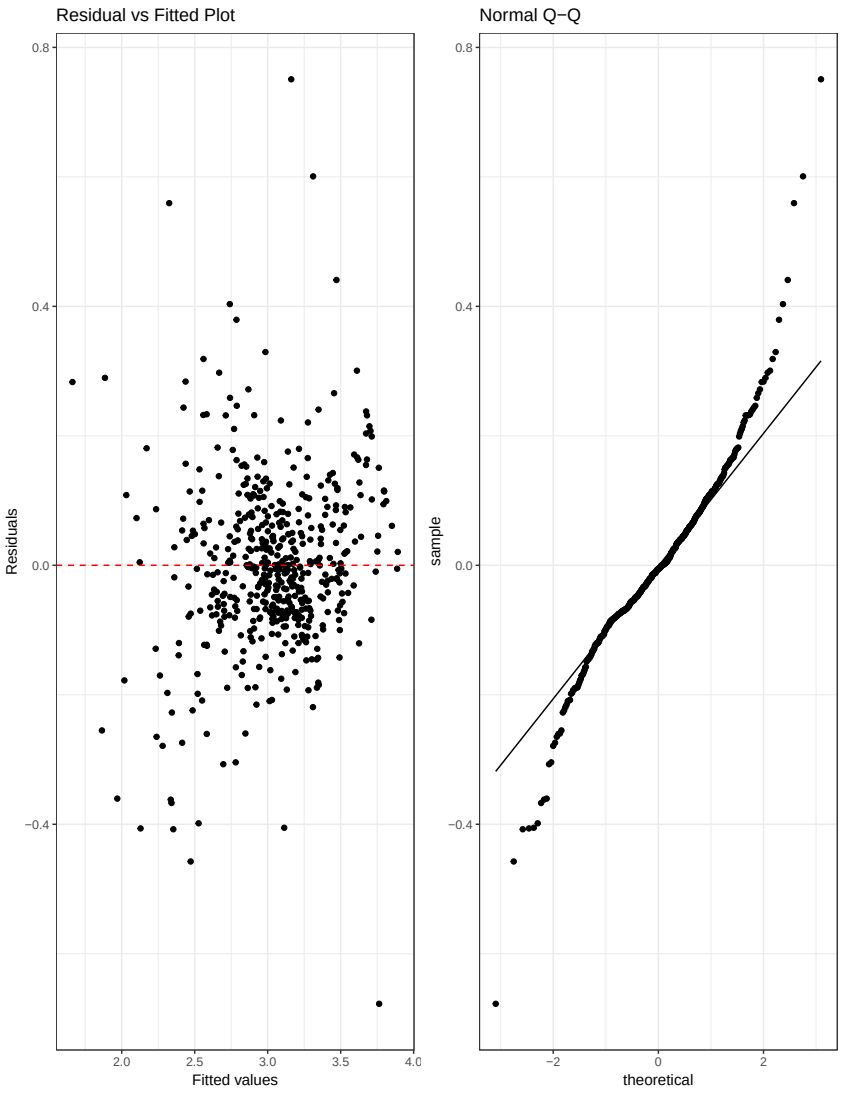

Figure 9 – Fitted Values Vs Residuals and Normal Q-Q plot

The SAR model in the approach GAMLSS

The model chosen was Box-Cox t distribution in a preliminary analysis, based on Generalized Akaike Information Criterion (GAIC) with penalty $k = 2$, Which is equivalent to the standard Akaike Information Criterion (AIC), We also performed a selection variables based on GAIC. Thus, the adjusted model was:

$$
\begin{aligned}
Y &\sim \text{BCT}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}),\\
\hat{\mu} &= -43.36 - 0.0215\text{s(LSTAT)} - 0.543564\text{s(NOX)} + 0.157325\text{s(RM)}\\
&\quad - 0.012759\text{s(CRIM)} + \text{s(region)} - 0.647570\ \text{s(LON)}\ ,\\
\log(\hat{\sigma}) &= -2.4325078 + 0.0018871\ \text{s(TAX)} - 0.0014336\text{B},\\
\log(\hat{\nu}) &= -0.5126,\\
\log(\hat{\tau}) &= 1.67 - 0.01865\text{s(CRIM)}.
\end{aligned}
$$

In figure 10 the map of predicted values by district is displayed. In it, we can verify that the places where the residences have higher values are in the center-west region of the map, and these places have as local neighbors with high values as well.

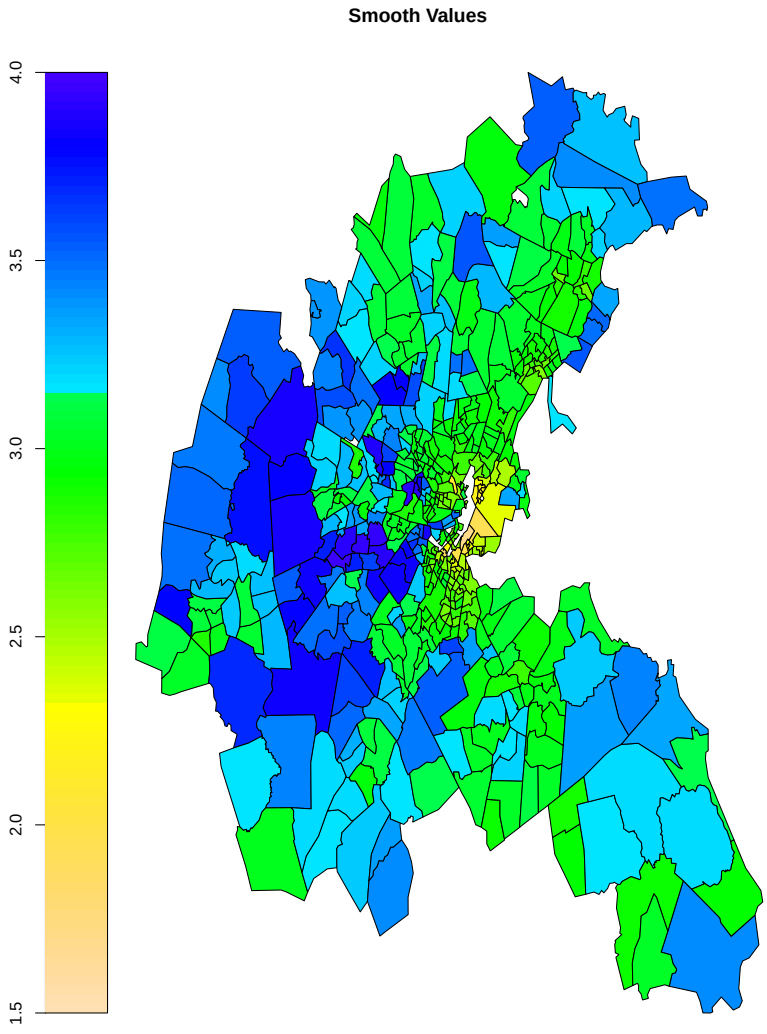

Figure 10 – Fitted values of `PRICE` from Boston Data

Look at figure 11 as residuals is now more well-behaved. The flexibilization of the distribution of the response variable allows a greater gain in the predictive capacity of the

model. The residuals in fact seem to come from the normal distribution, looking at for example the Normal QQ plot at the bottom right of the graph.

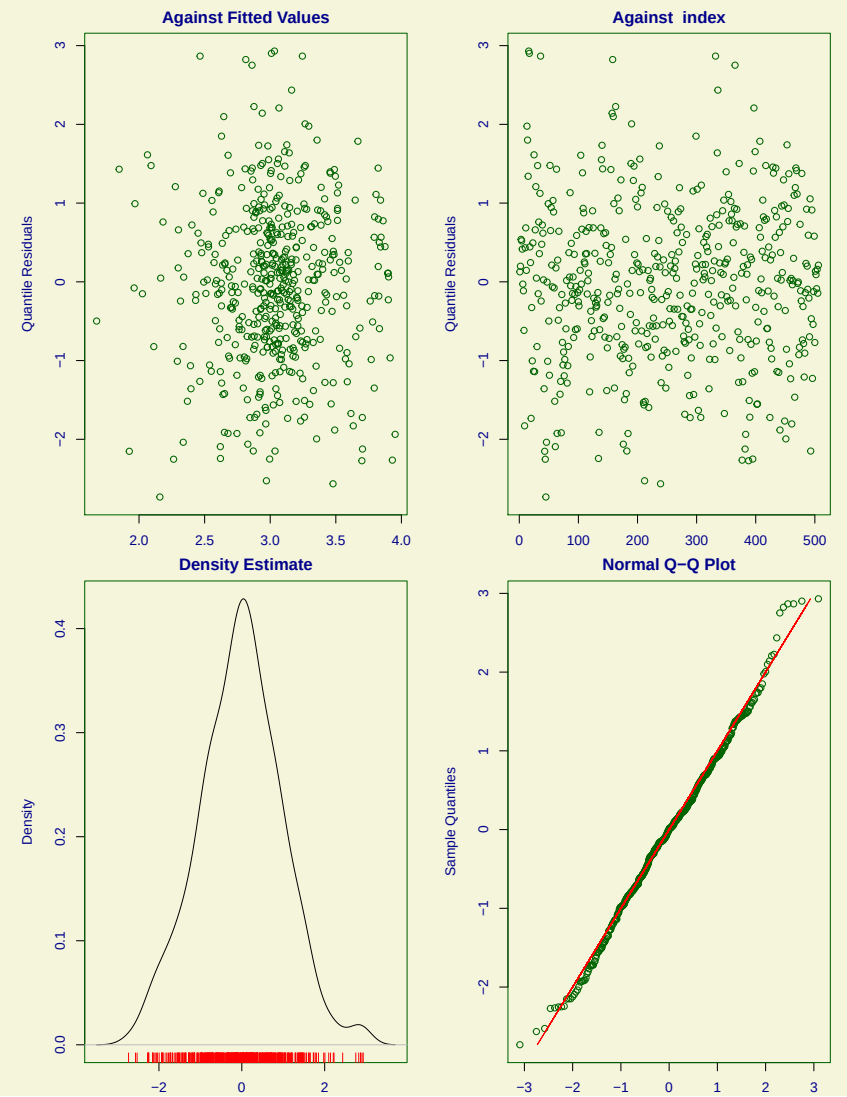

Figure 11 – Residual plots of the SAR approach in GAMLSS

The worm plot gives us a sense of how well our model is adjusted. In figure 12 All points are within the 95% confidence band between the two elliptic curves, showing that this model specification is adequate.

The Term plot shows the partial effect of the variables used on the model parameter, in this case μ. In figure 13, `LSTAT` has a decreasing effect on mu, unlike `RM`. In turn, `NOX` has a complicated effect on μ, sometimes decreasing, or increasing, and this relationship is not very clear.

We compared in figure 14 the spatial distribution of the fitted values between the approaches SAR within GAMLSS and Spatial SAR. Note that the maps are quite similar, with very few sites with different fitted values.

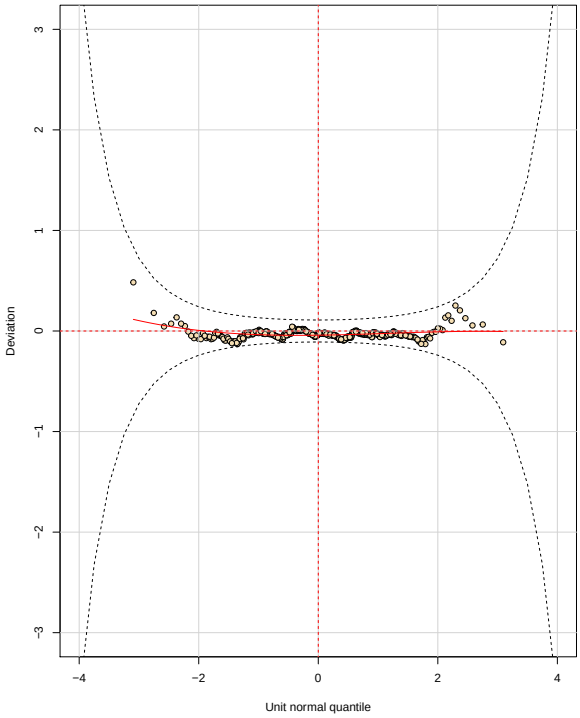

Figure 12 – Worm Plot of the SAR approach in GAMLSS

5.2 Gini Data

This application consists of an analysis of the determinants of the Income inequality index in the State of Pernambuco in Brazil, measured by the GINI index, incorporating the space in the analysis.

The Gini coefficient is a measure of the degree of income distribution in a society. This measure ranges between 0, perfect equality, and 1 that extreme inequality. According to Ray (2008), this coefficient is given by:

$$\mathrm{G} = \frac{1}{2n^2\mu}\sum_{j=1}^{m}\sum_{k=1}^{m} n_j n_k |y_j - y_k|,$$

here the income data are ordered and subdivided into j classes, and thus the absolute difference of the pairs of income,$|y_j - y_k|$, is computed. The set of variables that affect the coefficient of gini is described below, here we follow the work of Barros et al. (2007).

`Gini` is the index of gini collected in 2010 for all the cities of Pernambuco, collected from the portal Ipeadata of Instituto de Pesquisa Econômica Aplicada (IPEA). `PIB` is the gross domestic product for the current year. `POP_TOT` is the number of inhabitants of that city in 2010. `PEA` is the number of economically active people in the population. `POP_IDOSA` is the number of old-aged people in the population. `POP_JOV` is the number of young people in the population. `TX_ANALF` is the proportion of illiterate people in the

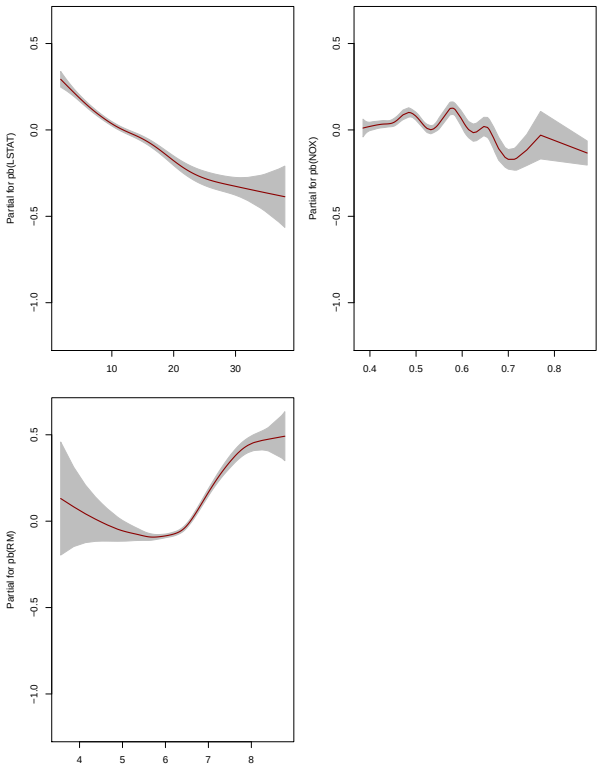

Figure 13 – Term plot of model `mfinal.spatial`

population. `TX_DESEMP` is the proportion of unemployed people in the population. All these variables were collected from Ipeadata. Already `PBF` which is a benefit received by poor families and `BPC` that a retirement benefit for poor people, were collected in Ministério do Desenvolvimento Social (MDS). In the descriptive analysis we find problems related to the correlation of some of the explanatory variables. In order to avoid potential problems of multicollinearity, we have decided to do as follows. The variables `PIB` and `POP_TOT` were joined by the ratio forming the variable `Pibcap` which is the gross domestic product per municipality. On the other hand, the variable `IDoJOV` was produced by the ratio between `POP_IDOSA` and `POP_JOV`.

As in the previous section, we did select variables based on GAIC and the distribution of the response variable was based on the AIC. The fitted model is given by:

$$
\begin{aligned}
Y &\sim \text{BEZI}(\hat{\mu}, \hat{\sigma}, \hat{\nu}),\\
\log(\frac{\hat{\mu}}{1-\hat{\mu}}) &= 0.089 - 0.001605\ \text{s(Pibcap)} + 9.092e^{-08}\ \text{s(PBF)}\\
&\quad + \text{s(city)},\\
\log(\hat{\sigma}) &= 2.060 + 0.14\text{Pibcap} + 6.65\text{IDoJOV} + 0.000421\text{BPC}\\
\log(\frac{\hat{\nu}}{1-\hat{\nu}}) &= -27.02.
\end{aligned}
$$

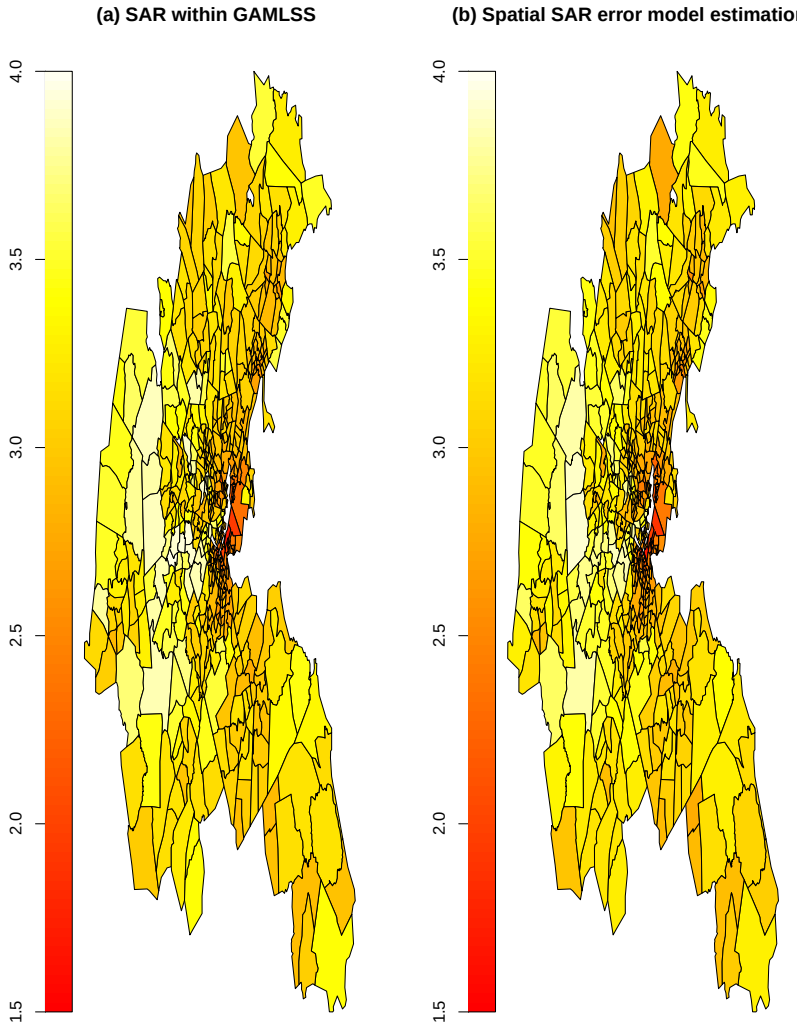

Figure 14 – Comparative Plot for fitted values of `PRICE` between SAR within GAMLSS and Spatial SAR

Figure 15 shows the residuals of the regression, for the diagnostic analysis of the adequacy of the fitted model. The estimated density (lower left) looks like the Gaussian distribution. And the normal q-q plot shows the vast majority of points on the red line, which is a good indicator of normality.

Another relevant indicative of good residuals behavior as shown by the worm plot in figure 16.

In this we see that all points are randomly distributed over the red line between the elliptic curves. With higher adjusted values of the *gini* spatially distributed in the economically richest regions in the state.

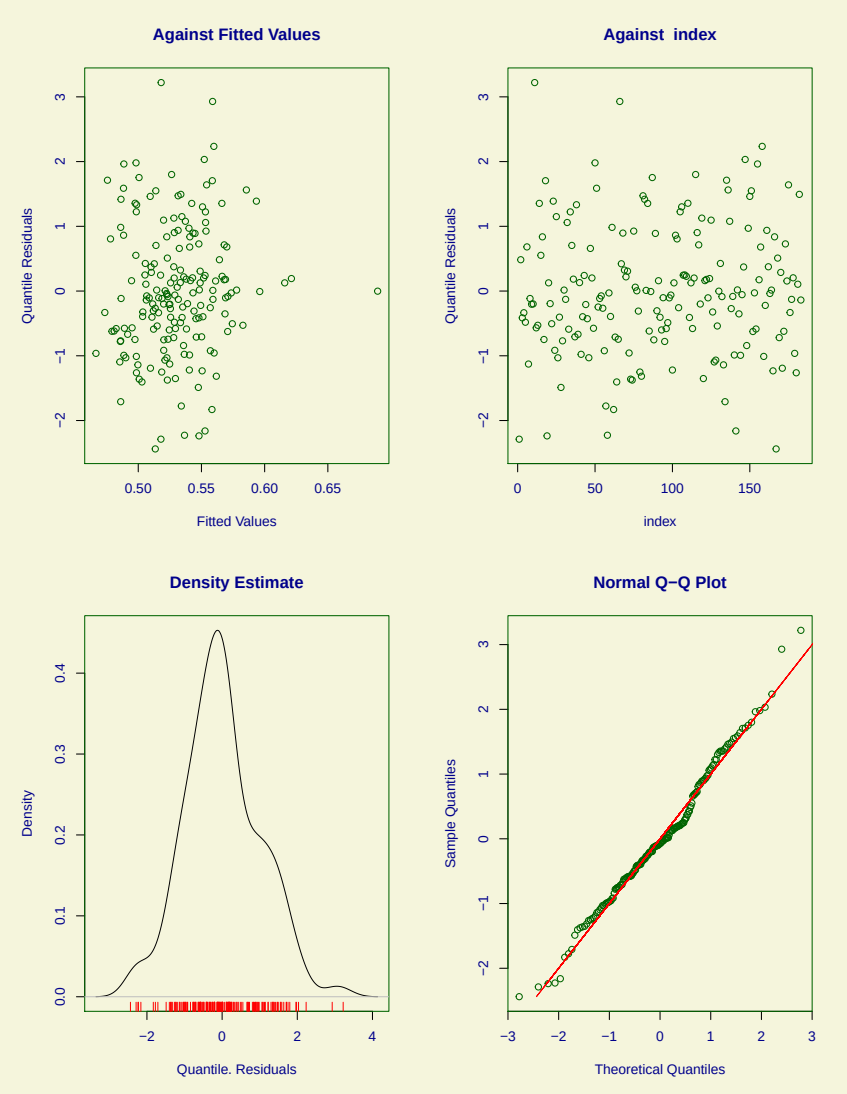

Figure 15 – Residual plots of model for `Gini`

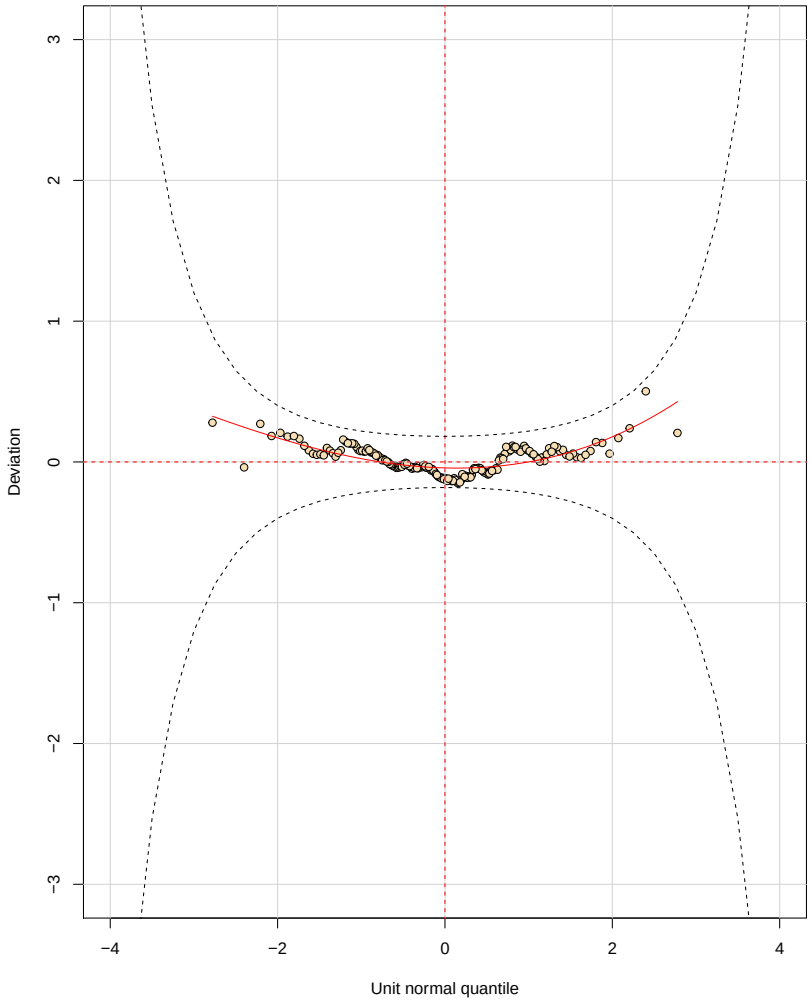

Figure 16 – Worm Plot of model for `Gini`

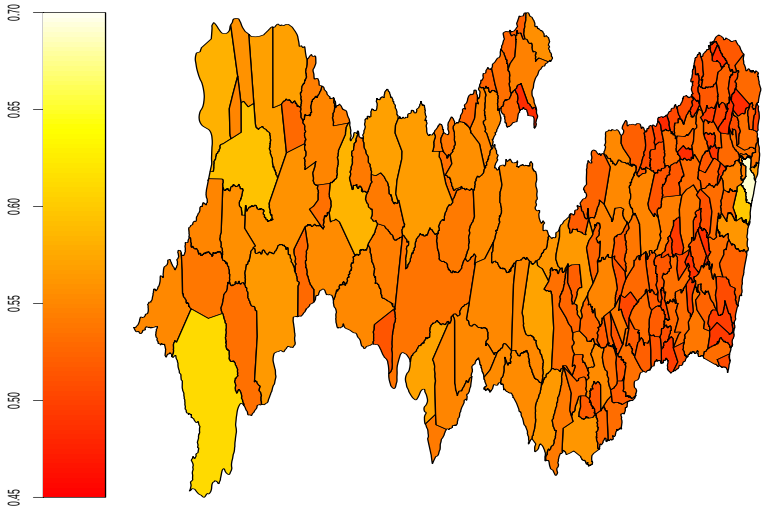

Figure 17 – Fitted values of `Gini` for Cities in Pernambuco

6 Conclusion

The objective of this work was to introduce the class of simultaneous autoregressive spatial models within the GAMLSS environment. For this, we show the relationship of those with the class of CAR models and starting from this relation in terms of covariance matrices is that the objective is reached. For these last ones meet the requirements of the general scope necessary to have penalization, through the precision matrix for the modeling in the GAMLSS. We make two applications showing the importance of the employability of this tool in the field of spatial econometrics.

Bibliography

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis for spatial data*. [S.l.]: Chapman and Hall/CRC, 2004. Citado 3 vezes nas páginas 5, 16, and 17.

BARROS, R. P. d. et al. A queda recente da desigualdade de renda no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2007. Citado na página 41.

BASTIANI, F. D. et al. Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 1, p. 168–186, 2018. Citado na página 19.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 192–236, 1974. Citado 4 vezes nas páginas 5, 16, 17, and 18.

BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, Springer, v. 43, n. 1, p. 1–20, 1991. Citado na página 5.

BHUNIA, G. S.; SHIT, P. K. *Geospatial Analysis of Public Health*. [S.l.]: Springer, 2019. Citado na página 5.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001. Citado na página 36.

CRESSIE, N. Statistics for spatial data. *Terra Nova*, Wiley Online Library, v. 4, n. 5, p. 613–617, 1992. Citado 4 vezes nas páginas 14, 15, 17, and 18.

CRESSIE, N.; WIKLE, C. *Statistics for spatio-temporal data, vol. 465*. [S.l.]: Wiley, 2011. Citado 2 vezes nas páginas 15 and 16.

FAHRMEIR, L. et al. *Regression: models, methods and applications*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 19.

HAINING, R. P.; HAINING, R. *Spatial data analysis: theory and practice*. [S.l.]: Cambridge University Press, 2003. Citado na página 23.

HASTIE, T.; TIBSHIRANI, R. *Generalized additive models*. [S.l.]: Wiley Online Library, 1990. Citado na página 6.

HODGES, J. S. *Richly parameterized linear models: additive, time series, and spatial models using random effects*. [S.l.]: Chapman and Hall/CRC, 2016. Citado na página 13.

HOEF, J. M. V.; HANKS, E. M.; HOOTEN, M. B. On the relationship between conditional (car) and simultaneous (sar) autoregressive models. *Spatial Statistics*, Elsevier, v. 25, p. 68–85, 2018. Citado 3 vezes nas páginas 16, 17, and 18.

HOEF, J. M. V. et al. Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, Wiley Online Library, v. 88, n. 1, p. 36–59, 2018. Citado 2 vezes nas páginas 15 and 16.

JR, D. H.; RUBINFELD, D. L. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, Elsevier, v. 5, n. 1, p. 81–102, 1978. Citado 2 vezes nas páginas 6 and 33.

KEMP, K. *Encyclopedia of Geographic Information Science*. [S.l.]: SAGE publications, 2007. Citado na página 13.

LICHSTEIN, J. W. et al. Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, Wiley Online Library, v. 72, n. 3, p. 445–463, 2002. Citado na página 15.

MAO, J.; JAIN, A. K. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern recognition*, Elsevier, v. 25, n. 2, p. 173–188, 1992. Citado na página 15.

MCCULLAGH, P.; NELDER, J. *Generalized Linear Models, Second Edition*. Chapman & Hall, 1989. (Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series). ISBN 9780412317606. Disponível em: <http://books.google.com/books?id=h9kFH2_FfBkC>. Citado na página 8.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado 2 vezes nas páginas 6 and 8.

PACE, R. K.; GILLEY, O. W. Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics*, Springer, v. 14, n. 3, p. 333–340, 1997. Citado 2 vezes nas páginas 33 and 35.

RAY, D. *Development economics*. [S.l.]: Springer, 2008. Citado na página 41.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado 3 vezes nas páginas 5, 9, and 10.

RUE, H.; HELD, L. *Gaussian Markov random fields: theory and applications*. [S.l.]: CRC press, 2005. Citado na página 13.

STASINOPOULOS, M. D. et al. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2017. Citado na página 19.

WHITTLE, P. On stationary processes in the plane. *Biometrika*, JSTOR, p. 434–449, 1954. Citado 2 vezes nas páginas 5 and 15.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: Chapman and Hall/CRC, 2017. Citado na página 20.

Appendix

APPENDIX A – Simulation study

Table 14 – Estimates of Error-SAR model coefficients for response variable with Gumbel distribution with spatial dependence

	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
n	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$\rho = 0.0$			
20	1.95105	-21.957960	0.000799	-0.506226	1.245196	0.000083
50	1.929136	-22.834570	0.000591	-0.4987259	-0.2548137	0.000143
100	1.925965	-22.96138	0.001056	-0.497137	-0.5726039	0.000024
			$\rho = 0.2$			
20	1.935941	-22.562340	0.000004	-0.5195563	3.911262	0.000162
50	1.92839	-22.864380	0.000636	-0.498269	-0.346206	0.000005
100	1.925621	-22.975170	0.001050	-0.4969426	-0.611478	0.000023
			$\rho = 0.5$			
20	1.9262	-22.95202	0.000795	-0.5007165	0.1433092	0.000190
50	1.92747	-22.90119	0.000570	-0.4981766	-0.3646709	0.000002
100	1.924723	-23.01106	0.001029	-0.4964779	-0.7044252	0.000021
			$\rho = 0.9$			
20	1.92177	-23.12919	0.000401	-0.510348	2.069748	0.000158
50	1.926856	-22.92575	0.000530	-0.4980767	-0.3846626	0.000001
100	1.924231	-23.03077	0.001019	-0.4961915	-0.7616996	0.000020

Table 15 – Estimates of Lag-SAR model coefficients for response variable with Gumbel distribution with spatial dependence

	$\beta_0 = 2.5$			$\beta_1 = -0.5$		
n	Estimate	Relative Bias (%)	MSE	Estimate	Relative Bias (%)	MSE
			$\rho = 0.0$			
20	2.459709	-1.611631	0.000466	-0.4934892	-1.302159	0.000069
50	2.181977	-12.72092	0.000069	-0.4931703	-1.365941	0.000137
100	2.040016	-18.39937	0.001387	-0.4925906	-1.481872	0.000023
			$\rho = 0.2$			
20	2.374765	-5.009411	0.000078	-0.5091985	1.839696	0.000164
50	2.156633	-13.73467	0.000844	-0.4938229	-1.235424	0.000004
100	2.16183	-13.52679	0.000838	-0.4936548	-1.269042	0.000002
			$\rho = 0.5$			
20	2.373859	-5.045658	0.000515	-0.4906988	-1.860246	0.000203
50	2.16183	-13.52679	0.000838	-0.4936548	-1.269042	0.000002
100	2.038342	-18.46633	0.001380	-0.4919674	-1.606514	0.000020
			$\rho = 0.9$			
20	2.398297	-4.068131	0.003481	-0.4960725	-0.7854928	0.000105
50	2.161705	-13.53178	0.000824	-0.493536	-1.292801	0.000001
100	2.037964	-18.48145	0.001379	-0.4916867	-1.662655	0.000002