Unversidade 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

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Ciências Exatas e da Terra

Orientadora: Fernanda De Bastiani
Coorientador: Dimitrios Stasinopoulos

Recife

2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

O48s Oliveira, Lucas de Miranda
Spatial autoregressive models for areal data within gamlss / Lucas de Miranda Oliveira. – 2019.
66 f.: il., fig., tab.

Orientadora: Fernanda de Bastiani.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2019.
Inclui referências e apêndice.

1. Estatística. 2. Estatística espacial. 3. Autoregressão simultânea. I. Bastiani, Fernanda de (orientadora). II. Título.

310 CDD (23. ed.) UFPE- MEI 2019-120

LUCAS DE MIRANDA OLIVEIRA

SPATIAL AUTOREGRESSIVE MODELS FOR AREAL DATA WITHIN GAMLSS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 25 de julho de 2019.

BANCA EXAMINADORA

Prof.(ª) Fernanda de Bastiani
UFPE

Prof.(º) Getúlio José Amorim do Amaral
UFPE

Prof.(º) Miguel Angel Uribe Opazo
UNIOESTE

ACKNOWLEDGEMENTS

Thank you Lord for the gift of life.
For my family I am grateful for the support this past two years.
Special thanks to my supervisor, Fernanda De Bastiani, for being a reference for me. For trusting me, for making me learn more, and for your dedicated guidance to this work.
I am grateful to my co-supervisor Prof. Dimitrios Mikis Stasinopoulos for his contribution in this work and for his dedication to Science.
Thanks also to Prof. Dr. Robert Rigby for his observations on this work.
I thank the teachers Miguel Uribe Opazo and Getúlio José Amaral for their careful reading of this work and contributions to it.
I thank all the professors of the Statistical Department at UFPE for their contribution to the society and students.
I thank all my friends that I did in the department, especially to César Leonardo, Jairo, Anny, Luís Félix, Adenice, Annabeth, Eduardo, César, Érica, Daniel, Nayara, and João.
My special thanks to Valéria Bittencourt for her empathy and professionalism.
I thank Sthéfano Tavares for his collaboration and for his promptness.
Thanks also to Raphael Botelho for his professionalism.
I thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).
Lastly, I also thank the Centro Nacional de Supercomputação of Universidade Federal do Rio Grande do Sul (CESUP/UFRGS).

ABSTRACT

In spatial data analysis the data are indexed by a set of locations in space, the way this set is defined separates spatial statistics into three areas: Geostatistics, models for Areal data, and Point Process. In this work we will focus on the models for areal data, specifically in the simultaneous autoregressive (SAR) models, which has applications in many fields such as Ecology, Public Health, Texture Analysis and Spatial Econometrics. It is proposed to implement the SAR models within the generalized additive models for location, scale, and shape (GAMLSS), allowing to consider any type of distribution to fit the data, and to model all the parameters of a distributions as function of the explanatory variables. The implementation of this procedure within GAMLSS is made considering the connection between random effects and penalized smoothers, and the relationship of the SAR and conditional autoregressive (CAR) models. An efficient algorithm was implemented to construct the penalty matrix compatible with general scope of penalization methods. Monte Carlo simulation studies were conducted with the purpose of evaluating the properties of the regression coefficients estimators of the SAR-GAMLSS models in the context of finite samples and with different probability distributions for the response variable. The methodology was applied to the analysis of house prices and also to the study of income inequality in the State of Pernambuco, Brazil, considering the spatial structure of the regions in the analysis.

Keywords Simultaneous autoregressive. Spatial statistics. Spatial econometrics. Penalized smoothers.

RESUMO

Na análise de dados espacias os dados são indexados por um conjunto de localizações no espaço, este separa a estatística espacial em três áreas: Geoestatística, modelos para dados de Área e Processos Pontuais. Este trabalho concentra-se nos modelos para dados de área, especificamente nos modelos autoregressivos simultâneos (SAR), que possui diversas aplicações nas áreas de Ecologia, Saúde Pública, Análise de Textura e Econometria Espacial. Propomos a implementação dos modelos SAR nos modelos aditivos generalizados para locação, escala e forma (GAMLSS), permitindo considerar qualquer tipo de função de distribuição para ajuste dos dados, e modelar todos os parâmetros da distribuição como função de variáveis explicativas. O procedimento de implementação nos GAMLSS é feito considerando a conexão existente entre termos de efeitos aleatórios e suavizadores penalizados, e a relação entre os SAR e modelos autoregressivos condicionais (CAR). Um algoritmo eficiente foi implementado para construção da matriz de penalidade compatível com o escopo geral dos métodos de penalização. Estudos de simulação de Monte Carlo foram realizados com o propósito avaliar as propriedades do estimadores dos coeficientes de regressão do modelos SAR-GAMLSS no contexto de amostras finitas, e com distintas funções de probabilidade para a variável resposta. Aplicamos a metodologia à análise dos preços de residências e também ao estudo da desigualdade de renda no Estado de Pernambuco, Brasil, em ambos levando em consideração a estrutura espacial das regiões.

Palavras-chaves Autorregressão Simultânea. Estatística Espacial. Econometria Espacial. Suavizadores Penalizados.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

The first law of geography enunciated by Waldo Tobler states that: "everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). Over time the strengthening of this idea only grew with advancement of technologies in the last decades has allowed some benefits for several areas of knowledge, as is the case of spatial statistics, which studies the phenomena where the locations are relevant for understanding it. In this area, the *observations* are indexed by locations on the map. Areas such as Econometrics, Climatology, Ecology, Health Public and others are incorporating spatial information into their analysis. In Econometrics, for example, the level of economic activity in each city, county, and country, if spatial information is relevant, according to specific criterion, then information on the level of activity of the neighbors leaves the analysis richer (ANSELIN, 2003). In Public Health, on the other hand, location-indexed data help to find factors associated with specific diseases in and it propagation within a region of interest. It also gives a sense of the health location of the individual and its distribution on the map (BHUNIA; SHIT, 2019).

According to Banerjee et al. (2004), let a vector $\boldsymbol{s} = (s_1, s_2)$ of coordinates pairs that taken values in the set $D = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_q\} \subset \mathbb{R}^q$ which is a subset fixed of positive volume. Then, the data in *spatial analysis* are divided into three basic types:

- *areal data*, where $Y(\mathbf{s})$ is a random variable at location $\boldsymbol{s}$, the set D represents a subsets of indexes of pairs of coordinates $\boldsymbol{s}$ that delimit an area or region. For example, a indexes set of cities, districts, counts and others, and $Y(\mathbf{s})$ can represent the number of crimes in a district, the number of workers in a city, rainfall level by region, and other similar variables;
- *point-referecend data* or *geostatistical data*, where $Y(\mathbf{s})$ is a random variable at location $\mathbf{s}$, and $\mathbf{s}$ takes values on a continuous surface $D \subset \mathbb{R}^q$. For example, $Y(\mathbf{s})$ represents the level of soil salinity in a location $\mathbf{s}$;
- *point pattern data*, the set D is random, and s indicates the occurrence of an event on the map, for example indicates regions that record earthquake, terrorist attacks, disease focus and others.

In the context of this master thesis the *areal data* will be treated. An example of this type of spatial data is in Figure 1; On the right side of it, the map of Brazil with its federation units is showed, highlighting the unit corresponding to the state of Pernambuco. On the left side, the map of the state of Pernambuco with 184 municipalities with their respective Gini index values is shown, thus exemplifying the type of spatial

data referring to the *areal data*. One question that can be asked here is: Is the level of inequality (measured by the Gini index) in a city explained by neighboring cities? This question will be will be discussed later in this master thesis.

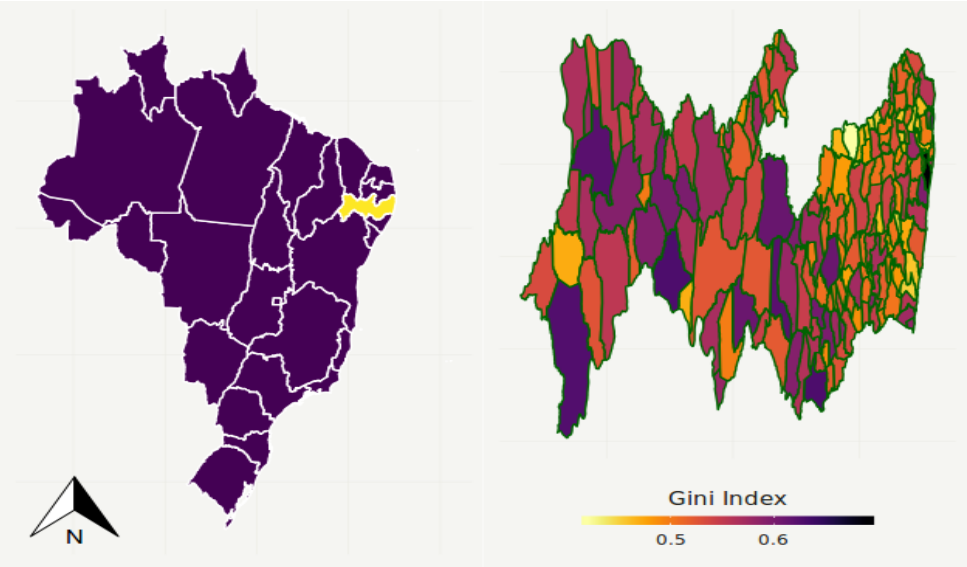

Figure 1 – On the right side showing the map of Brazil with the state of Pernambuco highlighted and map of Pernambuco with the values of Gini index by city on the left side.

Over the decades, statistical models have been developed to work with this type of *spatial data*, as for example the conditional autoregressive (CAR) (BESAG, 1974) models and intrinsic autoregressive (IAR) (BESAG; KOOPERBERG, 1995) models are examples of these models. This work comprised the development of a structure that comprised simultaneous autoregressive (SAR) models (WHITTLE, 1954), which are also models for *areal data* and contains a relationship with the CAR models.

Another point of this master thesis deals with a flexible class of regression models, which are the generalized additive models for location, scale, and shape (GAMLSS) (RIGBY; STASINOPOULOS, 2005). This approach is a flexible alternative for modeling response variables with different probability distribution functions beyond the exponential family. The GAMLSS also allow the modeling of all parameters of the probability distribution of the response variable. And also includes of terms of non-parametric functions in the modeling of these parameters. These models have been applied to various problems. Sá et al. (2018) studied the relationship of wildfires and environmental factors in Portugal, using the functions of the distribution function zero-adjusted Gamma (ZAGA) and zero-inflated Poisson (ZIP) for the response variable. Luo (2013) used the Box-Cox Power Exponential (BCPE) distribution function in the modeling of stock market liquidity. Voudouris et al. (2012) used also the BCPE distribution in the analysis of the problem of film box-office revenues. By comparison of coefficients and a graphical analysis of semiparametric terms, Cajias (2018) checks the accuracy of the GAMLSS models in contrast to the

generalized additive models in Munich's residential market. De Bastiani et al. (2018) show the modelling and fitting of Gaussian Markov random field spatial components within a GAMLSS model.

Structure of the master thesis

In Chapter 2, a review about the models used in the regression analysis and their respective assumptions are presented. It starts with linear models, then generalized linear models (GLM), then present the generalized additive models (GAM) and then the GAMLSS, which are part of our object of study. In Chapter 3, a theoretical review of the models for areal data is presented. 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, the theoretical framework of the proposed implementation. In Chapter 4, it is presented a numerical evaluation of the regression coefficient estimators of the proposed model, comparing it with other models existent in the literature. In Chapter 5 it is presented two applications of the proposed model, one to data set Housing Values in Suburbs of Boston. The an other application related to Spatial Econometrics, where is model the inequality index of Gini for the municipalities in the state of Pernambuco, Brazil is modeled. Finally, in Chapter 6 are presents some final considerations and the appendix with simulation study graphics are shown.

Computational Resources

For the development of this master thesis was used the software `R` (Team, 2019). For the simulation studies, it was used `GAUSS` 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 aspects of regression analysis. Here the contents necessary to understand this master thesis will be exposed in a summarized way. Firsty, a brief explanation about linear models is presented. Secondly the generalized linear models are introduced. Some details about generalized additive models are also presented and finally GAMLSS are presented.

2.1 Linear Models

The regression analysis is the search for the causal relationship between variables, seeking to understand how a response variable can be explained through others variables, called 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 explanatory variables, or covariates, is that the explained variable is a linear and parametric function of the explanatory variables. In contrast, terms of random error are added to the linear function, implying that the response variable is a random variable. For n observations of these variables, this relation can be represented by the following equation:

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

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, which shows 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 residual sum of squares of the model S:

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

It is minimized the sum in relation to the parameters vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, this method is called of *least squares*. Equating $\frac{\partial S}{\partial \boldsymbol{\beta}} = 0$ it is obtained the solution in matrix terms to $\boldsymbol{\beta}$ vector:

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

where $\mathbf{X} = [\mathbf{1} \quad \boldsymbol{x}_1, \ldots \boldsymbol{x}_k]$ denotes an $n \times (k+1)$ design matrix. One of the limitations of this approach is that the distribution function of the response variable is Gaussian, which is often not satisfied in practice, requiring some transformation (such as the logarithmic function) in it. Another point is that, by hypothesis, the variance of the response variable is constant on the regression surface, Fox (2015).

2.2 Generalized Linear Models

The motivation for constructing the generalized linear models class is related to the linear relationship between the response variable and the regression coefficients imposed in the context of linear models with normal errors. Another motivation is due to the context of allowing variance to be modeled. Presented by Nelder e Wedderburn (1972), these models the GLMs allow the flexibility of the distribution function of Y, which now belongs to the *exponential family* ($\mathcal{E}.f.$). Many families of probability distributions are included in the $\mathcal{E}.f.$, such as Bernoulli, Poisson, Gaussian, Exponential, Gamma and others.

The general structure of the GLMs is (MCCULLAGH; NELDER, 1989):

$$
\begin{aligned}
\mathbf{Y} &\sim \mathcal{E}.f.(\boldsymbol{\mu}, \boldsymbol{\phi}) \\
g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta},
\end{aligned}
$$

where $g(.)$ is a monotonic function link (*logit, probit, complementary log-log*), which provides the relationship between the linear predictor and the mean of the response variable distribution. The vector $\boldsymbol{\phi}$ is a vector of constants and $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ is is called a linear predictor, $\boldsymbol{\beta}$ is a vector $(k+1) \times 1$ of unknown parameters and $\mathbf{X}$ is an $n \times (k+1)$ 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 is given by:

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

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:

$$l(\theta,\phi,y_i)=\log f_{\mathrm{Y}}(y_i;\theta,\phi),$$

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

$$\begin{aligned} \frac{\partial l}{\partial\theta} &= \{y_i - b'(\theta)/a(\phi)\}, \\ \frac{\partial^2 l}{\partial\theta^2} &= -b''(\theta)/a(\phi). \end{aligned}$$

From relationships :

$$\begin{aligned} E\left(\frac{\partial l}{\partial\theta}\right) &= 0, \\ E\left(\frac{\partial^2 l}{\partial^2\theta}\right) + E\left(\frac{\partial l}{\partial\theta}\right)^2 &= 0. \end{aligned}$$

And then the mean and variance of the response variable are shown in terms of this parameterization:

$$\begin{aligned} E(Y) &= b'(\theta), \\ Var(Y) &= b''(\theta)a(\phi). \end{aligned}$$

Note that in this approach the limitation is that the response variable belongs to the exponential family and the relationship between the response and explanatory variables is assumed to be linear.

2.3 Generalized Additive Models

The generalized additive models (GAMs) incorporate in the GLM analysis the use of nonparametric techniques in the modeling of the parameters of the distribution of the response variable (HASTIE; TIBSHIRANI, 1990). In the context of GAMs, the linear predictor can be rewritten as a sum of smooth function. The predictor is given by:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + s_1(\mathbf{x_1}) + \ldots + s_J(\mathbf{x_J}), \tag{2.1}$$

where $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of unknown parameters, $\mathbf{X}$ is an $n \times (k+1)$ design matrix, $s(.)$ denotes nonparametric smoothing function for covariate (x). These functions can be P-*splines* smoother (EILERS; MARX, 1996), Ridge Regression (HOERL; KENNARD, 1970), Lasso Regression (TIBSHIRANI, 1996).

For example, using B-*spline* basis (BOOR et al., 1978), which consists of polynomial pieces that are connected, Eilers e Marx (1996) show that P-*splines* are smoothers with low-rank, with nodes equally spaced and with a difference penalty being applied to the parameters β_j. A fitted curve for n observations of the response variable Y and covariate X can be represented by linear combination $\hat{y}(x) = \sum_{j=1}^{n} \hat{\beta} \mathrm{B}_j(y, x)$, where j denotes the degree of B-splines. The authors propose a penalty in the parameters given by: $\sum_{j=1}^{k-1} (\beta_{j+1} - \beta j)^2$. The advantage of using the P-*splines* is the flexibility given to the fit of the data.

Wood (2017) presents the theory behind the GAMs, as smoothers like cubic splines, Thin plate splines, Duchon splines, Gaussian Markov random fields and as these can be incorporated in Equation 2.1. The computational manipulation of GAMs is also presented with the **mgcv**.

2.4 Generalized Additive Models for Location, Scale, and Shape

The generalized additive models for location, scale, and shape (GAMLSS) were proposed by Rigby and Stasinopoulos in 2005. This is a general class for reponse variables univariates. In this class of models, assumes that the observations of the response variable y_i are independent, with $i = 1, ..., n$. And they have probability (or density, in the continuous case) function $f(y_i|\boldsymbol{\theta}^i)$, conditioned on the vector $\boldsymbol{\theta}^{i\top} = (\theta_{i1}, ..., \theta_{ip})$ of p unknown parameters of this function. All parameters of the distribution function of the response variable can be modeled through different independent (explanatory) variables and random effects. 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 $k = 1, 2, 3, \ldots K$, and let $g_k(.)$ be a known monotonic link function that associates $\boldsymbol{\theta}_k$ with independent variables and random effects through the GAMLSS (RIGBY; STASINOPOULOS, 2005) model given by:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{U}_{jk} \boldsymbol{\gamma}_{jk}, \tag{2.2}$$

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 $\mathbf{X}_k$ and $\mathbf{U}_{jk}$ 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 Gaussian distribution with $\boldsymbol{\gamma}_{jk} \sim \mathrm{N}_{q_{jk}}(\mathbf{0}, \lambda_{jk}^{-1} \mathbf{K}_{jk}^{-1})$, and $\mathbf{K}_{jk}^{-1}$ is the inverse of a symmetric matrix $q_{jk} \times q_{jk}$, $\mathbf{K}_{jk}$, which may be a function of a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$. And, if $\mathbf{K}_{jk}$ is

a singular matrix, then it is understood that $\boldsymbol{\gamma}_{jk}$ has density function that is improper and proportional to $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{K}_{jk}\boldsymbol{\gamma}_{jk})$.

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 random effects, but not all distribution parameters need to be modeled using explanatory variables. For example, set $\mathbf{U}_{jk} = \mathbf{I}_n$, 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 $k = 4$ (for example) in (2.2) is obtained the GAMLSS formulation semiparametric additive given below:

$$\begin{aligned}
g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X_1}\beta_{\mathbf{1}} + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}), \\
g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X_2}\beta_{\mathbf{2}} + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}), \\
g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X_3}\beta_{\mathbf{3}} + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}), \\
g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X_4}\beta_{\mathbf{4}} + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4})
\end{aligned}$$

with vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$ each with length n. Here, μ represents the location parameter, (e.g. the mean), σ of the scale (e.g. the standard deviation), ν and τ are skewness and kurtosis,respectively, and represent the shape parameters.

In this approach assumes the modeling is done for four parameters ($k = 4$) of the distribution function of the response variable. Primordial in the process of fitting 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 estimation uses shrinkage matrices within algorithm backfitting (RIGBY; STASINOPOULOS, 2005).

As mentioned in the previous section, in the model (2.2) it is assumed that $\boldsymbol{\gamma}_{jk}$, are independent and have normal distribution with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \lambda_{jk}^{-1}\mathbf{K}_{jk}^{-1})$. In the GAMLSS framework, the hypothesis of independence between different random effects vectors is essential. However, if for a k two or more random effects vectors are not independent, they can be combined into a single random effect vector and also their corresponding covariate matrices, $\mathbf{U}_{jk}$, 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:

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^{\top}\mathbf{K}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.3}$$

where $l = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}^i)$ is the logarithm of the likelihood function of the observations given $\boldsymbol{\theta}^i$ for $i = 1, ..., n$. It is also shown in Appendix C from Rigby e Stasinopoulos (2005) that the maximization of l_p applied to partial residuals, $\boldsymbol{\epsilon}_{jk}$, to update the estimate of the additive predictor $\mathbf{U}_{jk}\boldsymbol{\gamma}_{jk}$ together with an algorithm backfitting leads to shrinkage matrix $\mathbf{S}_{jk}$, given below:

$$\mathbf{S}_{jk} = \mathbf{U}_{jk}(\mathbf{U}_{jk}^{\top}\mathbf{V}_{kk}\mathbf{U}_{jk} + \mathbf{K}_{jk})^{-1}\mathbf{U}_{jk}^{\top}\mathbf{V}_{kk}, \tag{2.4}$$

where $k = 1, 2, 3, 4$ and $j = 1, 2, ..., J_k$, with $\mathbf{V}_{kk}$ is an matrix of iterative weights. For different forms of $\mathbf{U}_{jk}$ and $\mathbf{K}_{jk}$ different types of additive terms in the linear predictor can be incorporated $\boldsymbol{\eta}_k$, for $k = 1, 2, 3, 4$.

There are two basic algorithms that are used in **gamlss** package in `R` software. The first is the algorithm `CG`, which is based on the algorithm of Cole e Green (1992). 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 $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)^{\top}$ 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, location 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. the fitting process with the `RS` algorithm, presented in Stasinopoulos et al. (2017), is given by the following steps:

- the *outer iteration* which calls;
- the *inner iteration* which calls;
- the *modified backfitting* algorithm.

Given a vector $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \boldsymbol{\nu}_0, \boldsymbol{\tau}_0)$ of initial estimates for the parameters, the *outer iteration* sequentially fits a model for each parameter considering the latest estimates for the other parameters until the convergence of the global deviance. The *inner iteration* is called in each fitting of the parameters and applies a local scoring algorithm. And the *modified backfitting* algorithm uses a good weighted least (WLS) and algorithm a good penalized weighted least squares (PWLS) algorithm (RIGBY; STASINOPOULOS, 2014).

In continuation of the fitting process in the GAMLSS, the vector of hyperparameters $\boldsymbol{\lambda}_{jk}$ can be estimated internally (locally) or globally. In the local estimation, the method of estimation for each $\boldsymbol{\lambda}_{jk}$ is applied each time in the backfitting algorithm of RS or CG algorithms. In globally the method of estimation is is applied outside these algorithms. Three different methods can be applied to estimate hyperparameters:

- Methods based in likelihood as MLE/REML;
- Generalized Akaike information criterion (GAIC);
- Generalized cross validation or validation (GCV) global deviance (VDEV).

The implementation of the **gamlss** package is present in the software `R`, (STASINOPOULOS; RIGBY et al., 2007). for fitting a GAMLSS model. There others packages as **gamlss.add** for fitting extra additive terms in the fitting a parameter of distribution. All distributions within **gamlss** package are found in **gamlss.dist** package (RIGBY et al., 2019 forthcoming).

3 GAUSSIAN MARKOV RANDOM FIELDS

In this section it is presented a definition of Gaussian Markov random fields (GMRF) and its connection with the autoregressive models.

A necessary concept when talking about GMRFs is *conditional independence*. As exemplified in Rue e Held (2005), consider $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ a random vector normally distributed with a $q \times 1$ mean vector $\boldsymbol{\mu}$ and $q \times q$ precision matrix $\mathbf{K}$. Two variables γ_1 and γ_2 are independent if and only if (iff) $\pi_{\gamma_1,\gamma_2}(\gamma_1, \gamma_2) = \pi_{\gamma_1}(\gamma_1)\pi_{\gamma_2}(\gamma_2)$, 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 if and only if (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, according to Edwards (2012). This is due to the fact that γ_1 and γ_2 may be marginally dependent.

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

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

and $k_{ij} \neq 0$ iff $\{i, j\} \in \mathcal{E}$ for all $i \neq j$, $i = 1, \ldots, q$. Hence, the precision matrix $\mathbf{K}$ informs which areas are neighbors, given some neighborhood criterion. For $k_{ij} = 0$, i and j are called conditional independent, by the Markov property. Clearly, for a larger number of neighbors, more dense (or less sparse) will be $\mathbf{K}$.

The first class of GMRF to areal data presented are conditional 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). Also shown is a special case of this class which are *intrinsic* conditional autoregressive models (ICAR). Next, it is presented another approach to the areal data that are simultaneous autoregressive models that are commonly employed in the context of spatial econometrics. A important point to be emphasized is, as Hodges (2016) in Section 5.2, these models were developed to model the variable response, and in the course of time that statisticians began to employ them as distributions of random effects or latent variables.

3.1 Models for areal data

In this section, let assume that the data can be thought of as a realization of a stochastic process $\{\mathbf{Z}(\boldsymbol{s}) : \boldsymbol{s} \in \mathbf{D}\}$, with mean $\mu(\boldsymbol{s})$, $\mathbf{Z}(\boldsymbol{s}) = (\mathrm{Z}(\boldsymbol{s}_1), \ldots, \mathrm{Z}(\boldsymbol{s}_q))^\top$ has a

multivariate normal distribution where the space of variation is discrete (CRESSIE, 1993). Each element of the set $\mathbf{D}$ represents the i-geographic region (unit area), $i = 1, \ldots, q$, where each index represents a set of pairs of coordinates that delimit an area. In the Figure 2, the 49 districts of Columbus city of the state of OHIO, in United States, are displayed, and for each of them a unique code is associated. The set of indices denoting each areal $\mathbf{D}$ can be defined here as $\mathbf{D} = \{1, \ldots, 49\}$ and from this a structure of spatial dependence for the observations can be constructed, considering neighboring regions whose borders touch each other.

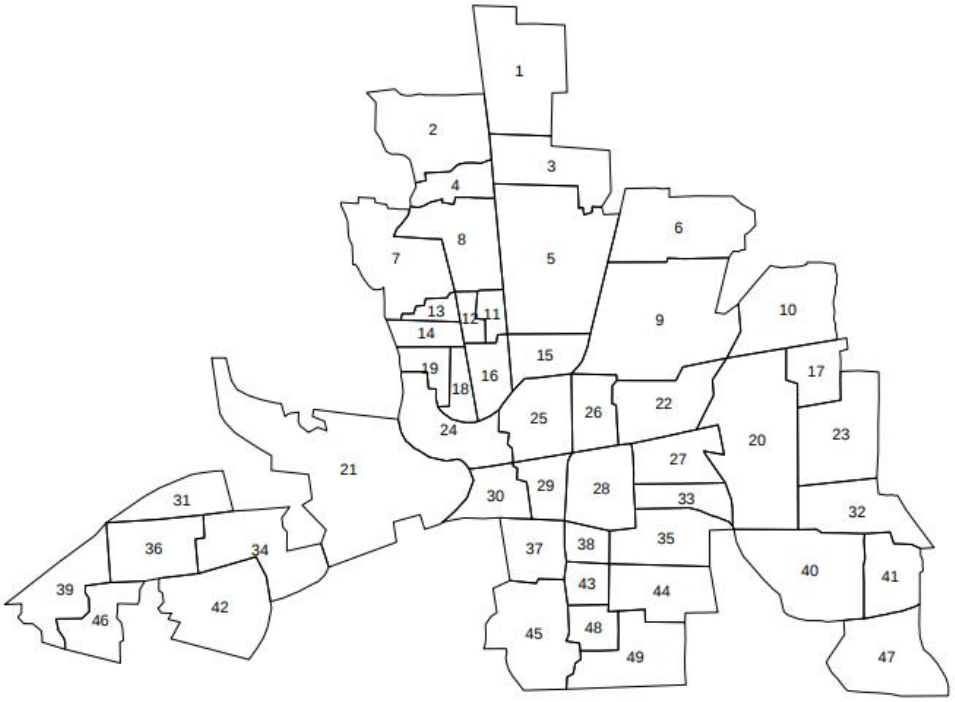

Figure 2 – Indexes for 49 Columbus districts, OHIO, United States.

For example, to construct a neighborhood matrix $\mathbf{W}$ of dimension $q \times q$ given by the number of areas (regions, districts). The elements of this matrix denote the spatial dependence between the regions. Looking at Figure 2, 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, where $w_{i,j}$ represents the line and column of $\mathbf{W}$. By definition the elements of the diagonal of the neighborhood matrix are equal to zero, $w_{ii} = 0$, and symmetry is required, district i is neighbor to j if it is also j is neighbor of i, this is represented by the notation $i \sim j$, two areas are considered neighbors if they share one or more common points at their borders. In this way, the representation of the set of neighbors of district 6 can be given by $\mathrm{N}_6 \equiv \{5, 9\}$. And generally for the above example:

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

Conditional autoregressive models

The Conditional autoregressive models (CAR) are attributed to Besag (1974) and, according to Banerjee et al. (2004), these models have been widely used in recent years in the context of spatial hierarchical models as random effects.

The CAR model is defined as follows (CRESSIE, 1993):

$$\gamma_i|\boldsymbol{\gamma}_{-i} \sim \mathrm{N}\left(\sum_{j=1}^{n} c_{ij}\gamma_j, m_i\right), \tag{3.2}$$

where $\boldsymbol{\gamma}_i$ is a random variable associated with the unit of area i, for $i = 1, \ldots, q$. The vector $\boldsymbol{\gamma}_{-i}$ denotes all realization γ_j except ith. In its turn, $c_{ij}\gamma_j$ is the conditional mean of γ_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:

$$\boldsymbol{\gamma} \sim \mathrm{N}_q(\mathbf{0}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}), \tag{3.3}$$

here, $\mathbf{I}$ is an $q \times q$ identity matrix, and 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 $c_{ii} = 0$. The spatial dependence matrix can be computed as equal to $\mathbf{C} = \rho_c\mathbf{W}$, with ρ_c representing the spatial autocorrelation parameter, and $\mathbf{W}$ is a spatial weight matrix, defined as before. In summary, according to Hoef et al. (2018a), 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;
- $c_{ij}/m_{ii} = c_{ji}/m_{jj}$, for all i and j; and
- $\mathbf{M}$ is a diagonal matrix with positive diagonal elements.

The restriction $1/\lambda_{[1]} < \rho_c < 1/\lambda_{[\mathrm{N}]}$, with $\lambda_{[1]}$ the smallest eigenvalue of $\mathbf{W}$ and $\lambda_{[\mathrm{N}]}$ the highest eigenvalue, is used to obtain a valid covariance matrix. Because the restriction ensures that $(\mathbf{I} - \rho_c\mathbf{W})$ has positive eigenvalues (CRESSIE, 1993).

The intrisic autoregressive model (IAR) Besag e Kooperberg (1995) is a special case of the CAR model when $\rho_c = 1$ implying that the covariance matrix of this model does not exist. According to Besag e Higdon (1999), in IAR models, for $i = 1, \ldots, q$ regions, $\mathbf{Z} \sim \mathrm{N}(0, \mathbf{K}^-)$, where $\mathbf{K}^-$ is the $q \times q$ generalized inverse of matrix $\mathbf{K}$. The elements of the diagonal of matrix $\mathbf{K}$, k_{ii}, are the number of regions adjacent to region i. The non-diagonal elements, k_{ij}, represent the neighborhood relation between two regions, i and j, equal to -1 iff i and j are considered neighbors, and 0 in otherwise.

Simultaneous autoregressive models

The SAR model was introduced by Whittle (1954), which defined a spatial process simultaneously in $\mathbb{R}^2$ on a countable grid. These models have been studied extensively over the years, and are richly exposed in Cressie (1993), Cressie e Wikle (2011), and most recently in Hoef et al. (2018b). These models have application in a diverse amount of scientific areas. In the field of texture analysis, Mao e Jain (1992) construct a multiresolution model based on SAR model for Texture classification and Texture segmentation. The SAR models are also commonly used in quantitative study of the environment and living beings, known as Ecological data analysis, since they have a certain spatial pattern due to the proximity of the collected observations. For example, Lichstein et al. (2002) analyze reproduction 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:

$$\gamma_i = \sum_{j=1}^{q} b_{ij}\gamma_j + \varepsilon_i, \quad i = 1, \ldots, q, \tag{3.4}$$

which can be rewritten in matrix terms:

$$(\mathbf{I} - \mathbf{B})\boldsymbol{\gamma} = \boldsymbol{\varepsilon}, \tag{3.5}$$

where $\mathbf{I}$ is an $q \times q$ identity matrix. The error term is Gaussian with $\boldsymbol{\varepsilon} \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Lambda})$, where $\mathbf{0}$ is a $q \times q$ zero matrix and $\boldsymbol{\Lambda}$ is a $q \times q$ diagonal matrix. 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:

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B}^\top)^{-1}). \tag{3.6}$$

Thus, for the covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{SAR}} = (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I}-\mathbf{B}^\top)^{-1})$ 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 Hoef et al. (2018b), the definition of $\mathbf{B}$ is to take it as $\mathbf{B} = \rho_S\mathbf{W}$, where ρ_s is the parameter that denotes the spatial autocorrelation between the areas. As in CAR models, To obtain a precision matrix of a specification of a SAR model looks directly at the eigenvalues and eigenvectors of the weights 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 $1/\lambda_{[1]} < \rho_S < 1/\lambda_{[\mathrm{N}]}$, with $/\lambda_{[1]} < 0$ and $\lambda_{[\mathrm{N}]} > 1$ 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. (2018b). 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 $\widetilde{w_{ij}} = w_{ij}/w_{i+}$. The matrix $\widetilde{\mathbf{W}}$ does not require symmetry, and is called the stochastic row because $\widetilde{\mathbf{W}}\mathbf{1} = \mathbf{1}$ (BANERJEE et al., 2004). Let, in a similar way, $\mathbf{B} = \alpha\widetilde{\mathbf{W}}$, being α the spatial correlation parameter, (3.4) is modified to:

$$\gamma_i = \alpha \sum_{j \in N_i} \frac{w_{ij}}{\sum_k w_{ik}} \gamma_j + \varepsilon_{ij}, \tag{3.7}$$

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 et al. (2018a), guarantee a valid specification for covariance matrix of a SAR model are listed below:

- The matrix $\mathbf{I} - \mathbf{B}$ is of full rank, i.e. if its determinant is non zero and therefore its rank is q;
- The diagonal elements of $\mathbf{B}$ are zero; and
- $\mathbf{\Lambda}$ is a diagonal matrix with positive elements.

The SAR models are used in the area of Spatial econometrics and under different contexts, which leads to different formulations of the SAR models. As emphasized Cressie e Wikle (2011), the matrix $\mathbf{B}$ is seen in this field as a type of lag operator or backshift operator. This operator applied to an element of a time series produces the element prior to this. Hence, instead of time lag, for time series models, the lag is performed in space. The spatial $\mathrm{SAR}_{\mathrm{lag}}$ model with zero mean ($\boldsymbol{\mu} = 0$) is written as $\mathbf{Y} = \mathbf{B}\mathbf{Y} + \boldsymbol{\nu}$, with $\boldsymbol{\nu} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathrm{SAR}})$, that an equivalent formulation presented in this section. The $\mathrm{SAR}_{\mathrm{error}}$ models assumes that the spatial correlation is present in terms of error, adding to the Ordinary Least Square (OLS) regression model a term to capture this spatial process:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \xi\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon},$$

where ξ is the spatial autoregression coefficient (HAINING; HAINING, 2003).

3.2 The relationship between simultaneous and conditional autoregressive models

To implement the SAR models in the GAMLSS it is necessary to check the relationship of these with the CAR models, because the full conditional distributions for

the SAR random effects have no convenient form (BANERJEE et al., 2004). Specifically, it is necessary know how to write the SAR model as a CAR Model. This relationship between these two models has been the subject of research. 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 equal, assuming that the mean was modeled correctly Cressie (1993), 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 et al. (2018a) investigate this relationship and makes a generalization for any definite matrix defined, here we follow them.

Theorem 1 (Hoef et al. (2018a))**.** *Any 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,* $\mathbf{C}$ *and* $\mathbf{M}$*.*

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:

- $\mathbf{M}$ 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 $\mathbf{I}-\mathbf{C}$;
- 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, $m_{ii} > 0$;
- By condition 2, in CAR models section, $c_{ii} = 0$ by the fact that $\mathbf{C} = \mathbf{D}^{-1}\mathbf{R}$;
- 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 consequently $\frac{c_{ij}}{m_{ii}} = \frac{c_{ji}}{m_{jj}}$. □

The second part to prove is the uniqueness of the covariance matrix of the SAR model written as CAR, given by the authors (HOEF et al., 2018a) 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. Let Y_{ij} be a random variable in the i-th row and j-th column of the grid, and consider the generator 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 and δ_k is the is the k-th regression coefficient. Assume 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:

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

This equivalent representation of the SAR model in terms of the CAR model will be very useful as will showed later in the section of the Chapter 3 that deals with the computational implementation of the SAR model for the purpose of this work.

3.3 The implementation of the SAR model within 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 CAR models in GAMLSS can be seen in De 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 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 have to:

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

as can be seen in Section 9.4 of Stasinopoulos et al. (2017). The matrix $\mathbf{U}$ is a $n \times q$ matrix that associates each observation with its respective unit of area. The idea behind is to place each observation in its respective region. That is:

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

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

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

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 $\hat{\boldsymbol{\gamma}} = (\mathbf{U}^\top \mathbf{V}\mathbf{U} + \lambda \mathbf{K})^{-1} \mathbf{U}^\top \mathbf{V}\boldsymbol{y}$.

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:

$$\boldsymbol{\gamma} \sim \mathrm{N}_q(\mathbf{0}, \lambda \mathbf{K}^-).$$

In this way, the precision model of the SAR model represented by the special case of CAR model, IAR model, can be incorporated into the GAMLSS models, where $\boldsymbol{\gamma}$ is a *intrinsic* GMRF.

To better exemplify how the SAR models can be incorporated into the GAMLSS approach, consider the right side of Figure 1 again, that shows 184 municipalities in Pernambuco with Gini index values. From a given configuration of neighbors, and again noting that a first order SAR model is equivalent to the third order CAR model, it is obtained the third order neighborhood configuration. Look at Figure 3 shows a subset of the 6 cities in state of Pernambuco, on the right side, and the relationship of these cities in terms of graph (*nodes* and *edges*) on left side. In order to construct a valid $\mathbf{K}$ precision matrix based on the general penalization scope presented above, can starts from the matrix of neighbors $\mathbf{W}$. Each of the seven areas in the figure can be considered as a nodes, as explained at the beginning of this chapter. Neighbors up to third order are obtained when

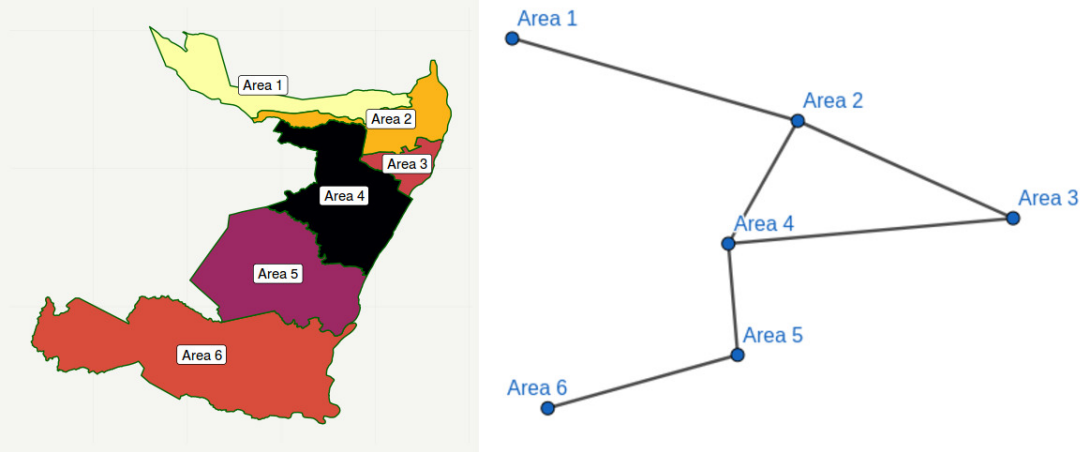

Figure 3 – A subset of six regions of the Pernambuco Gini data example and, on the right, an undirected graph describing the relationship between the six regions

the minimum number of edges between a region is equal to $k \leq 3$. Below is shown the weight matrices for different neighborhood orders, where two areas are neighbors, in terms of graph, if the number of edges between the two areas equals k:

$$\mathbf{W}^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{W}^2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\mathbf{W}^3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix},$$

where each row in the matrices above represents a the neighborhood relation of order k of the regions of Figure 3. For example, the element $w_{1,5}$ in matrix $\mathbf{W}_3$ informs that the area 1 and the area 5 are neighbor of third order. In general, the highest order neighborhood matrices $\mathbf{W}^k$ are constructed with the following elements:

$$w_{i,j}^k = \begin{cases} 0, \text{if } i \text{ and } j \text{ are not neighbors of order } k, \\ -1, \text{if } i \text{ and } j \text{ are neighbors of order } k, \\ n_{ii}, \text{ the number of neighbors of } i \text{ from order } k, \forall i = i. \end{cases}$$

And so, the precision matrix $\mathbf{K}$, from this neighborhood structure is:

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

Thus each element on the diagonal of $\mathbf{K}$ shows how many neighbors has the i, and the elements outside the diagnoal inform about the neighborhood relation. Since $\mathbf{K}$ is a valid matrix for the penalty criteria as shown in Fahrmeir et al. (2013).

The function `sp2precSar()`

Based on algorithm *Brute Force Search* proposed by Anselin e Smirnov (1996), was implemented an algorithm for constructing the compatible penalty matrix. This is a more efficient alternative to the recursive algorithm proposed by Blommestein e Koper (1992). The idea behind the algorithm is to compute neighborhood matrices of higher-orders of neighborhood, for our interest until order 3 for representation of the SAR model as CAR, from the first-order neighborhood matrix. The algorithm follows the following steps given below:

1. Let $\mathbf{A}$ be an accumulation matrix, and set $\mathbf{A} = \mathbf{W}^1$ the first-order contiguity matrix;
2. Compute $\mathbf{W}^2 = \mathbf{W}^1\mathbf{W}^1$ a second-order power matrix;
3. Compute the $\mathbf{P}^2$ matrix, where $p_{ij}^2 = 1$ if the $w_{ij}^2 > 0$, and 0 for otherwise;
4. Update $\mathbf{A} = \mathbf{A} + \mathbf{P}^2$;
5. Compute $\mathbf{W}^3 = \mathbf{W}^2\mathbf{W}^1$ a third-order power matrix;
6. Compute the $\mathbf{P}^3$ matrix, where $p_{ij}^3 = 1$ if the $w_{ij}^3 > 0$, and 0 for otherwise;
7. Return matrix $\mathbf{T} = \mathbf{A} + \mathbf{P}^3$.

From the neighbor relationships established above, the penalty matrix $\mathbf{K}$ is computed from the matrix $\mathbf{T}$, which establishes a neighbor relationship of up to third order, and $\mathbf{K}$ is computed as $\mathbf{K} = \mathbf{D_T} - \mathbf{T}$, where $\mathbf{D_T}$ is a diagonal matrix which denotes the numbers o neighbors up to third order from algorithm above. The implemented function takes two types of spatial object classes in `R` as arguments:

- `SpatialPolygonsDataFrame` objects; and

- `nb` objects which are lists of neighbors.

These two objects are obtained from the **spdep** package. And with the use of these is provided the penalty matrix $\mathbf{K}$. This matrix corresponds to the spatial structure of the SAR model rewritten as a special case of CAR, the IAR model. This matrix is treated as an extra penalty in the penalized log-likelihood, Equation (2.3). Thus, the SAR models, $\boldsymbol{\gamma} \sim \mathrm{N}(0, \lambda^{-1}\mathbf{K}^{-})$, are incorporated into (2.2) as through the terms of random effects, and $\mathbf{U}$ matrix is defined as in Equation (3.3). The purpose of this modeling is to take into account the spatial information in the analysis, if necessary, and to make the observed values of neighboring regions closer.

4 SIMULATIONS STUDY

This chapter aims to evaluate the properties of estimators of the coefficients of the SAR models within GAMLSS approach. Firstly, the methodology was evaluated in the context of finite samples with linear and nonlinear trend under the assumption of normality of the errors. Then, it was evaluated with different probability distribution for the response variable. The simulation consider two different scenarios, (i) simulating the regions, the form of polygons and (ii) the regions based on a *dataset*.

In the study, 1000 replications of Monte Carlo for each $n = (20, 40, 60)$ sample size. The number of polygons generated through the partition of Voronoi is equal to n. According to Okabe et al. (2009), the Voronoi diagram of a set of points subdivide the plan, in this study the plan is $\mathbb{R}^2$. Here three sets of points were used $P_1 = \{1, \ldots, 20\}$, $P_2 = \{1, \ldots, 40\}$ and $P_3 = \{1, \ldots, 60\}$, this sets are called *generators*. The generated regions are $V_1(n_i)$ with $i = 1, \ldots, 20$, $V_2(n_i)$ with $i = 1, \ldots, 40$, and $V_3(n_i)$ with $i = 1, \ldots, 60$. In this study each region contains only one point. The Voronoi diagrams for P_1, P_2, and P_3 are $\mathbb{V}_1\{V_1(n_1), \ldots, V_1(n_{20})\}$, $\mathbb{V}_2\{V_2(n_1), \ldots, V_2(n_{40})\}$, $\mathbb{V}_3\{V_3(n_1), \ldots, V_3(n_{60})\}$. Theses n points are generated randomly from $\mathcal{U} \sim (0, 1)$ in a square area and are then aggregated into cells representing areas as stated above. An example for the generation of Voronoi diagram is given in Figure 4, which shows 50 generated regions. The `voronoi.polygons`() function of the **SDraw** package implemented in the `R` software was used to generate the areas.

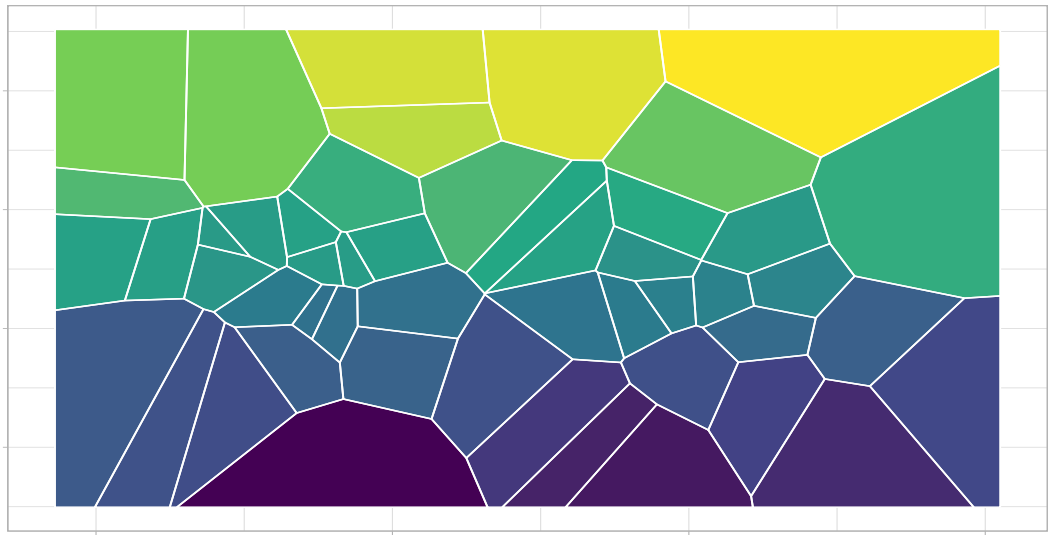

Figure 4 – Plot of 50 areas generated by Voronoi partition.

4.1 Evaluation in the models with spatial dependence and normal errors

Linear Trend

To compute spatial dependence, it was generated n observations following a normal distribution and created the covariance matrix of the SAR models from the above polygons and make spatially correlated response variable values, it is done the cholesky decomposition of this and make the product by the observations generated previously. And so, it was generated a sample of spatially correlated observations. This methodology follows the work of Haining e Haining (2003). The procedure was performed as follows:

1. Obtain the cholesky decomposition for a square matrix $\mathbf{\Sigma}$ of order $n \times n$ 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;

2. Generate the n observations of the vectors of covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ that follows from a continuos uniform ($\mathcal{U}$) probability distribution in the range of 0 to 3;

3. For each of the 1000 replications of Monte Carlo a vector $\boldsymbol{\varepsilon}$ of length n is generated from uncorrelated normal random variables; and

4. Compute the response variable $\boldsymbol{y}$ with spatial dependency for each replicate by doing $\boldsymbol{y} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = \beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2$, a linear trend.

Here two scenarios were computed for spatial correlation, $\rho = 0.0$ and $\rho = 0.10$. For this values used for ρ, was guaranteed that condition that $\frac{1}{\lambda_{[1]}} < \rho < \frac{1}{\lambda_{[N]}}$, as defined in chapter 3, which implies $\rho \in (-0.39, 0.20)$. Therefore, the two values chosen for ρ were $\rho = (0.0, 0.10)$. Analyzing the behavior of the proposed model when there is no spatial dependence and when there is. The error term parameter from $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2)$ were set as $\sigma^2 = 1$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)\top = (2.5, -0.5, 0.2)\top$. In addition to the SAR evaluation in the GAMLSS context, this was compared to the three regression models in this simulation study, presented in Chapters 2 and 3:

- OLS model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$;
- $\mathrm{SAR}_{\text{lag}}$ model: $\mathbf{Y} = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$;
- $\mathrm{SAR}_{\text{error}}$ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \xi\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

where $\mathbf{X}$ is an $n \times 3$ design matrix with a column of ones and covariates generated $\mathbf{x}_1$ and $\mathbf{x}_2$. The ρ and ξ are spatial autoregression coefficients. The OLS model was fitted using the `gamlss()` function with the default options and without semiparametric terms, fit an OLS model, this function is contained in **gamlss** package. The $\mathrm{SAR}_{\text{lag}}$ and $\mathrm{SAR}_{\text{error}}$ were

estimated using **spdep** package, these models were estimated using the generalized least squares method. For model $\text{SAR}_{\text{gamlss}}$ proposed in this master thesis, the modeling was done as follows:

$$\begin{aligned} \text{Y} &\sim \text{N}(\mu, \sigma^2), \\ \boldsymbol{\mu} &= \beta_0 + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + s(\text{area}), \\ \log(\sigma) &= \beta_0, \end{aligned}$$

where $s(area)$ is an IAR spatial smoother . In the fitting process was used the `ga()` interface in the `gamlss()` that establishes a connection with **mgcv** package (WOOD, 2017). In this is created a base of type IAR with matrix of penalty of given by the function that was we developed `sp2precSAR()`, from the relationship obtained between the autoregressive models SAR and CAR and is in **gamlss.spatial** (De Bastiani et al., 2018). Also in the interface, a low-rank matrix approximation is used to reduce the number of parameters (equivalent to regions) in the IAR smoothing function.

The measures used in the comparison of the models studies with linear trend of simulation were the empirical relative bias (RB-E) (%), the empirical mean square error (MSE-E), and the empirical Akaike's Information Criterion (AIC-E) (AKAIKE, 1974):

$$\begin{aligned} \text{RB-E} &= \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\beta}_k^{(i)} - \beta_k}{\beta_k}, \\ \text{MSE-E} &= \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_k^{(i)} - \beta_k)^2, \\ \text{AIC-E} &= \frac{1}{m}\sum_{i=1}^{m}(-2l(\hat{\theta})^i + 2K), \end{aligned}$$

where m is the number of replications 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 K is the number of parameters. The i-th log-likelihood is given by $l(\hat{\theta})$. The estimation of parameter β_k is calculated as:

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

In the evaluation of the properties of the coefficient estimators in the context of finite sample is expected when the number of elements in the sample increases both RB-E and MSE-E decrease to check the consistency properties of these estimators.

Tables1 show the estimates of the regression coefficients, RB-E, MSE-E from the models for the cases without spatial dependence. Regarding $\text{SAR}_{\text{gamlss}}$, the estimators of the coefficients appear to be asymptotically non-biased with decreasing MSE as the number of observations in the sample increases. Comparing all models, in this context

Table 1 – Estimates, RB-E, MSE-E and AIC-E for β_1 and β_2 of OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ models. The true values of parameters are $\beta_1 = -0.5$, $\beta_2 = 0.2$, $\sigma_2 = 1$, and $\rho = 0.0$.

$\rho = 0.0$		$\beta_1 = -0.5$			$\beta_2 = 0.2$		
Estimators	Estimate	RB-E (%)	MSE-E	Estimate	RB-E (%)	MSE-E	AIC-E
			$n = 20$				
OLS	-0.490675	-1.864954	0.077110	0.204041	2.020496	0.106526	47.02732
SAR_{lag}	-0.465352	-6.929593	0.075745	0.188284	-5.857993	0.102182	48.79764
$\text{SAR}_{\text{error}}$	-0.495872	-0.825654	0.084000	0.206456	3.227956	0.56822	48.1975
$\text{SAR}_{\text{gamlss}}$	-0.487917	-2.416623	0.082355	0.204897	2.448497	0.121333	**46.3445**
			$n = 40$				
OLS	-0.501162	0.232439	0.037316	0.204698	2.349242	0.033008	126.787
SAR_{lag}	-0.490062	-1.987511	0.036305	0.203051	1.525444	0.032233	128.257
$\text{SAR}_{\text{error}}$	-0.501207	0.241482	0.039715	0.205125	2.562331	0.531405	126.1988
$\text{SAR}_{\text{gamlss}}$	-0.499699	-0.060177	0.037839	0.202291	1.145297	0.033934	**123.6795**
			$n = 60$				
OLS	-0.50041	0.082069	0.023055	0.204055	2.027479	0.025307	**165.2223**
SAR_{lag}	-0.493595	-1.281006	0.022545	0.202299	1.149439	0.025146	165.9959
$\text{SAR}_{\text{error}}$	-0.50078	0.155907	0.023512	0.203077	1.538748	0.514603	165.8743
$\text{SAR}_{\text{gamlss}}$	-0.500637	0.127308	0.023325	0.204400	2.200188	0.026084	**165.2223**

of absence spatial dependence all models have RB-E descending when n increases. The SAR_{lag} show more precision in terms of MSE-E for the two estimators of β_1 and β_2, but in general they are more biased. By the criterion of model selection, $\text{SAR}_{\text{gamlss}}$ showed, in general, better performance among the compared models. The boxplots for β_0 are in the Appendix A of this paper.

Figure 5 shows the boxplots for the parameter estimates of β_1 and β_2, with no spatial correlation, and for each sample size n. The green line marks the true value of the parameter, showing that the estimators are close to the true value in both parameters, generally, for all models.

In the context of spatial dependence, Tables 2 show the estimates of the regression coefficients, RB-E, MSE-E from the analyzed models. The $\text{SAR}_{\text{gamlss}}$ shows consistent estimators for the regression coefficients. As expected, empirical AIC shows that spatial models are preferable when in the context of finite sample and low degree of spatial dependence.

Figure 6 shows the boxplots, in context of spatial dependence, for the estimates from models for β_1 and β_2. Showing bias for the SAR_{lag} model when n increases. The OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ show a low dispersion of the estimates to the true parameter values.

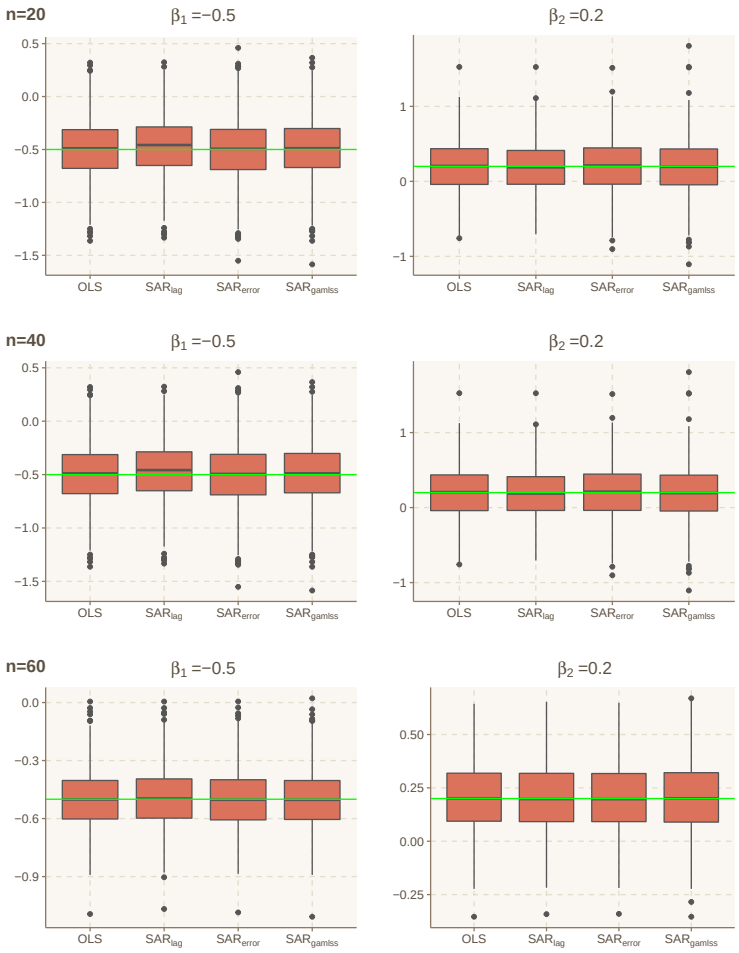

Figure 5 – Boxplots of parameters estimates for β_1 and β_2 from models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.0$.

Nonlinear Trend

In order to evaluate the flexibility of the $\text{SAR}_{\text{gamlss}}$, non-linear trend response variables were simulated and computed as follows:

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

2. Generate the n observations of the vectors of covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ that follows from a uniform ($\mathcal{U}$) probability distribution in the range of 0 to 3;

3. For each of the 1000 replications of Monte Carlo a vector $\boldsymbol{\varepsilon}$ of length n is generated from uncorrelated normal random variables; and

4. Compute response variable $\boldsymbol{y}$ with spatial dependency for each replicate by doing $\boldsymbol{y} = f(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{L}\boldsymbol{\varepsilon}$, where $f(\mathbf{x}_1, \mathbf{x}_2)$ is a nonlinear trend.

Table 2 – Estimates, RB-E, MSE-E and AIC-E for β_1 and β_2 of OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ models. The true values of parameters are $\beta_1 = -0.5$, $\beta_2 = 0.2$, $\sigma_2 = 1$, and $\rho = 0.1$.

$\rho = 0.10$		$\beta_1 = -0.5$			$\beta_2 = 0.2$		
Estimators	Estimate	RB-E (%)	MSE-E	Estimate	RB-E (%)	MSE-E	AIC-E
			$n = 20$				
OLS	-0.492488	-1.502369	0.067175	0.206019	3.009448	0.129656	50.64832
SAR_{lag}	-0.525209	5.041705	0.07565	0.217459	8.729509	0.109321	51.22088
$\text{SAR}_{\text{error}}$	-0.493619	-1.276258	0.066422	0.21036	5.180058	0.547488	49.93418
$\text{SAR}_{\text{gamlss}}$	-0.489522	-2.095593	0.072134	0.207878	3.938787	0.124604	**48.73013**
			$n = 40$				
OLS	-0.508173	1.634647	0.043178	0.207593	3.796664	0.04145	147.8509
SAR_{lag}	-0.510037	2.007308	0.037387	0.192001	-3.999287	0.033117	136.6093
$\text{SAR}_{\text{error}}$	-0.498122	-0.375633	0.034929	0.201562	0.780873	0.522299	**133.1951**
$\text{SAR}_{\text{gamlss}}$	-0.500475	0.09499	0.038303	0.20072	0.360108	0.036884	135.0565
			$n = 60$				
OLS	-0.500009	0.001867	0.022554	0.201955	0.977709	0.029762	183.1435
SAR_{lag}	-0.534155	6.831008	0.024201	0.208804	4.402074	0.027921	172.1048
$\text{SAR}_{\text{error}}$	-0.5008	0.159991	0.019707	0.202116	1.057873	0.510827	**171.4649**
$\text{SAR}_{\text{gamlss}}$	-0.500024	0.004879	0.021631	0.202108	1.054092	0.029622	178.0091

The nonlinear trend is computed by:

$$f(\mathbf{x}_1, \mathbf{x}_2) = 10\pi\sigma_{x_1}\sigma_{x_2}\{1.2\exp(-(\mathbf{x}_1 - 0.2)^2/\sigma_{x_1}^2 - (\mathbf{x}_2 - 0.3)^2/\sigma_{x_1}^2) + 0.8\exp(-(\mathbf{x_1} - 0.7)^2/\sigma_{x_1}^2 - (\mathbf{x_2} - 0.8)^2/\sigma_{x_2}^2)\},$$

where $\sigma_{x_1} = 0.3$ and $\sigma_{x_2} = 0.4$. This methodology is based on Durbán et al. (2012). Where the authors generated responses with a non-linear spatial trend in the coordinates. As before, two scenarios were evaluated, with spatial dependence and without, with $\rho = (0.0, 0.10)$. The models used were OLS, SAR_{lag}, and $\text{SAR}_{\text{error}}$, these were defined as in the previous Section in 3. Note that these models consider the linear trend for μ. The $\text{SAR}_{\text{gamlss}}$ used here is given by:

$$\begin{aligned} \text{Y} &\sim \text{N}(\mu, \sigma^2), \\ \mu &= \beta_0 + h_{11}(\mathbf{x_1}) + h_{21}(\mathbf{x_2}) + s(\text{region}), \\ \log(\sigma) &= \beta_0, \end{aligned}$$

where s is an IAR spatial smoothing function and $h(.)$ are smooth terms of type P-Splines (EILERS; MARX, 1996) to capture the non-linearity of the covariates $\mathbf{x_1}$ and $\mathbf{x_2}$. Other forms for nonlinear trend modeling can also be considered, such as an interaction between covariates (e.g. $s(\mathbf{x}_1, \mathbf{x}_2)$). To evaluate and compare the peformances of the models, the AIC-E was computed as before.

Figure 7 shows the boxplots for empirical AIC from fitted models without spatial dependence in this scenario. It is noticed that $\text{SAR}_{\text{gamlss}}$ has a better estimate of AIC-E in

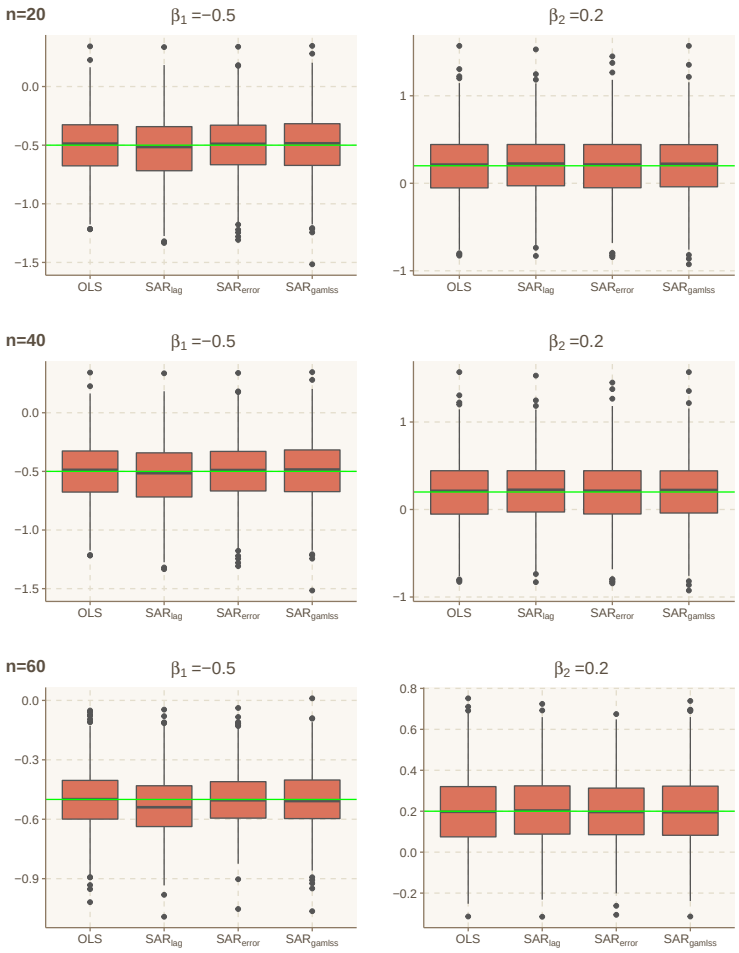

Figure 6 – Boxplots of parameters estimates for β_1 and β_2 from models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.10$.

relation to the other models. In relation to the models that consider linear trend, OLS is that it has better performance among these in this type of scenario.

For scenario with dependence spatial, Figure 8 shows the boxplots for empirical AIC. For all sample sizes, $\text{SAR}_{\text{gamlss}}$ has best estimate of AIC-E in relation to the other models due the flexibility of P-*splines* smooth terms. The SAR_{lag} and $\text{SAR}_{\text{error}}$ models performed slightly better than the OLS model. It can be attributed to the fact that spatial dependence is now present. And as shown in Table 4.1, the $\text{SAR}_{\text{error}}$ has better performance, generally, than SAR_{lag}.

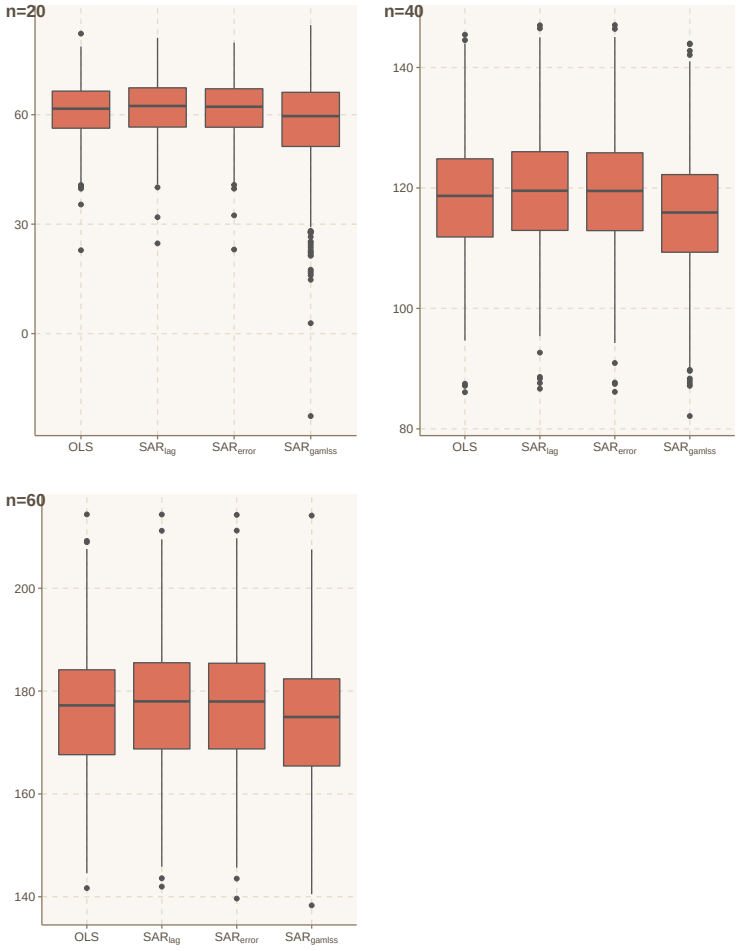

Figure 7 – Boxplots of AIC estimates for models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.0$ and nonlinear trend.

n	OLS	SAR_{lag}	$\text{SAR}_{\text{error}}$	$\text{SAR}_{\text{gamlss}}$
		$\rho = 0.0$		
20	61.16245	61.86282	61.65316	**57.79662**
40	118.4104	119.2874	119.1686	**115.6146**
60	176.3101	177.275	177.2223	**173.8598**
		$\rho = 0.10$		
20	62.94609	63.29105	63.23864	**58.64176**
40	124.4764	122.0848	121.9803	**118.4058**
60	187.3604	181.7326	181.765	**179.5834**

Table 3 – AIC estimates for models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.10$ and nonlinear trend

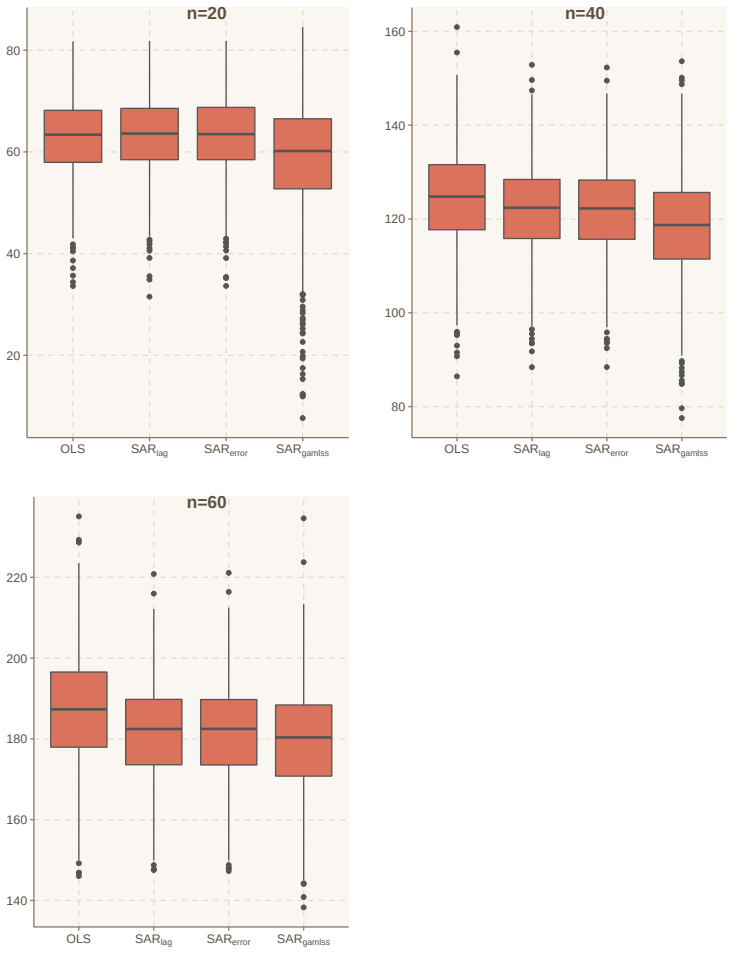

Figure 8 – Boxplots of AIC estimates for models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.10$ and nonlinear trend

4.2 Simulation study based on real data

Simulations studies based on real data are used to test the properties of coefficient estimators in spatial models as seen in Lee e Lee (2012), and Alam, Rönnegård e Shen (2015). Here the data set used was from Boston Housing Data (HARRISON; RUBINFELD, 1978), this dataset will be analyzed in more detail in the Chapter 5. The response variable is the median value of owner-occupied homes in $1000s$ in 506 census track of Boston, Massachusetts, United States. Here, 506 observations of response variable were simulated using the covariates `RM` and `RAD`, that are average number of rooms and index of accessibility to radial highways, respectively. Figure 9 shows the spatial structure in the districts of Boston with the median values of of owner-occupied homes.

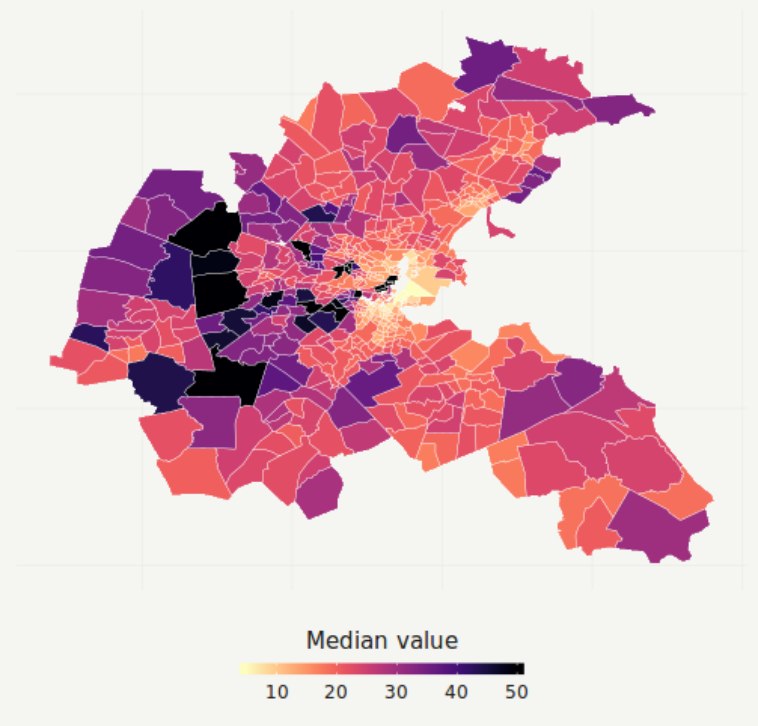

Figure 9 – Median values of owner-occupied homes in suburbs of Boston.

The probability distribution used for *variable response* was Sinh-Arcsinh (SHASH) distribution,Y$\sim$ SHASH(μ, σ, ν, τ), (JONES; PEWSEY, 2009) which is given by:

$$f(y|\mu, \sigma\, \nu, \tau) = \frac{c}{\sqrt{2\pi}\sigma(1+z^2)^{1/2}} e^{-r^2/2},$$

where:

$$r = \frac{1}{2}\left\{\exp\left[\tau \sinh^{-1}(z)\right] - \exp\left[-\nu \sinh^{-1}(z)\right]\right\}$$

and:

$$c = \frac{1}{2}\left\{\tau \exp\left[\tau \sinh^{-1}(z)\right] + \nu \exp\left[-\nu \sinh^{-1}(z)\right]\right\},$$

where $z = (y - \mu)/\sigma$, for $-\infty < y < \infty$, $\mu = (-\infty, +\infty)$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$ (STASINOPOULOS et al., 2017). Figure 10 shows the SHASH distribution density for different parameter values. This distribution function is implemented in **gamlss**, (RIGBY et al., 2019 forthcoming). The μ is the median, σ is a scaling parameter, ν is the left tail heaviness and τ is the right tail heaviness parameter.

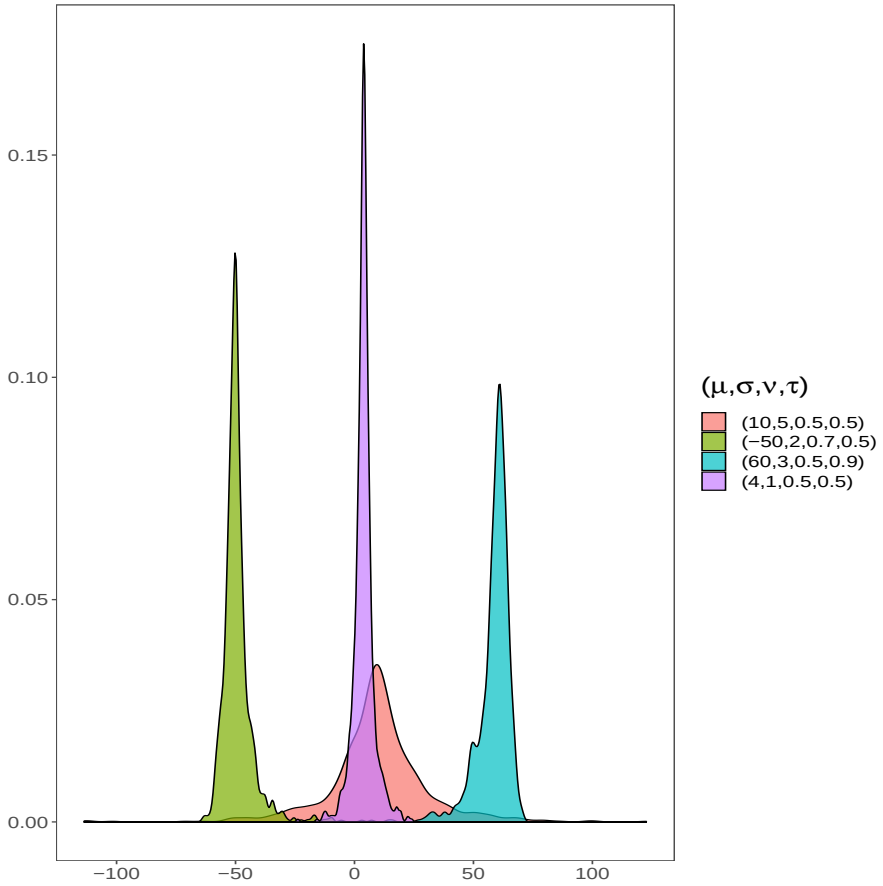

Figure 10 – Plot of Sinh-Arcsinh distribution

In the simulation scheme, 1000 replications of Monte Carlo were made. The regression coefficients were $(\beta_0, \beta_1, \beta_2) = (1, 0.5, 0.5)$. The Models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$, and $\text{SAR}_{\text{gamlss}}$, in Equation 4.1, were used and compared in the simulation.

1. Obtain the cholesky decomposition for a square matrix $\mathbf{\Sigma}$ of order $n \times n$ 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 from structure spatial in Boston;

Table 4 – Estimates, RB-E, MSE-E and AIC-E for β_1 and β_2 of OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ models. The true values of parameters are $\rho = 0.10$, $\beta_1 = -0.5$, $\beta_2 = 0.5$, and $\sigma = 1, \nu = 0.5, \tau = 0.5$, from SHASH distribution with linear trend.

$\rho = 0.10$	OLS	SAR_{lag}	$\text{SAR}_{\text{error}}$	$\text{SAR}_{\text{gamlss}}$
		$\beta_1 = 0.5$		
Estimate	0.495889	0.495932	0.494812	0.496347
RB-E (%)	-0.822197	-0.813572	-1.037651	-0.730605
MSE-E	0.072518	0.073034	0.073171	0.039514
		$\beta_2 = 0.5$		
Estimate	0.499481	0.501614	0.499404	0.500277
RB-E (%)	-0.103826	0.322838	-0.119161	0.055324
MSE-E	0.000718	0.001455	0.073171	0.00055
AIC-E	2881.795	2883.787	2882.81	**2778.225**

2. For each of the 1000 replications of monte carlo a vector $\boldsymbol{\varepsilon}$ of length n is generated from uncorrelated SHASH random variables; and

3. Compute response variable $\boldsymbol{y}$ with spatial dependency for each replicate by doing $\boldsymbol{y} = \beta_0 + \beta_1 \texttt{RM} + \beta_2 \texttt{RAD} + \mathbf{L}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \text{SHASH}(\mu, \sigma, \nu, \tau)$

The $\text{SAR}_{\text{gamlss}}$ model used in this simulation scenario was:

$$\begin{aligned} \text{Y} &\sim \text{SHASH}(\mu, \sigma, \nu, \tau), \\ \mu &= \beta_0 + \beta_1 \texttt{RM} + \beta_2 \texttt{RAD} + s(\text{region}), \\ \log(\sigma) &= \beta_0, \end{aligned} \tag{4.1}$$

as before, s is an IAR spatial smoothing function.

Table 4 shows the simulation results based on Boston Housing data. The $\text{SAR}_{\text{gamlss}}$ model displays a smaller RB-E for β_1 and β_2, compared to the other models. The $\text{SAR}_{\text{gamlss}}$ also shows a better performance in relation to MSE-E and AIC-E. This result is attributed to the fact that $\text{SAR}_{\text{gamlss}}$ allows any distribution function to response variable.

The main conclusions obtained from the simulation studies carried out are given below:

- In finite sample context with normal errors and linear trend and without and low spatial dependence, the OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ models exhibit behavior similar to low bias and MSE-E.

- In context of nonlinear trend, $\text{SAR}_{\text{gamlss}}$ models is preferable to allow flexible terms that model non-linear relations and spatial dependence.

- $\text{SAR}_{\text{gamlss}}$ models should also be considered when the response variable is suspected to be non-Gaussian.

5 APPLICATIONS

5.1 Boston Housing data

To illustrate the use of the GMRF within the GAMLSS models with the spatial structure being represented by a SAR model, a known set of data was used. It is the hedonic pricing data of Harrison e Rubinfeld (1978). 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, but with Gaussian distribution for response variable. These data are used in the literature to verify robust estimation as in Subramanian e Carson (1988), it was also used in nonparametric estimation (PACE, 1993) and new spatial regression models (SIMLAI, 2014). A comparison is made between the model estimated by these authors (PACE; GILLEY, 1997) and the model proposed in this paper, this comparison is relevant by the fact that some premises of this model may not be true. Another relevant point is that fittings of models can be compared in which the spatial term is parameterized and another that the term is smoothing. This section is divided as follows. Firstly, it is presented the variables. Then, was fitted a model with spatial configuration as in Pace e Gilley (1997), and check the goodness of fit is checked. Finally, the $\text{SAR}_{\text{gamlss}}$ model is fitted considering different continuous distribution on the real line.

Description of the variables

The data consist of 506 observations per census tract, with 14 variables related to the structure of the households, the location and socioeconomic characteristics. The variables used in the analysis are given below:

- `PRICE`: The response variable is the logarithm of the median corrected value of household values in USD 1000's;
- `CRIM`: Crime per capita in the town;
- `AGE`: Proportion of owner-occupied units built prior to 1940;
- `NOX`: Nitric oxides concentration (parts per 10 million);
- `CHAS`: Borders Charles River, which is a factor indicating if the property is near the Charles River or not;
- `RM`: Average number of rooms per dwelling;

- `ZN`: proportion of residential land zoned for lots over 25,000 sq.;
- `INDUS`: proportion of non-retail business acres per town;
- `PTRATIO`: pupil-teacher ratio by town;
- `RAD`: index of accessibility to radial highways;
- `TAX`: full-value property-tax rate per $10,000;
- `B`: $1000(\mathtt{B}_k - 0.63)^2$ where $\mathtt{B}_k$ is the proportion of black people by town;
- `LSTAT`: % lower status of the population in the town;
- `DIS`: weighted mean of distances to five Boston employment centres ;
- `LAT`: latitude of census tract; and
- `LONG`: longitude of census tract.

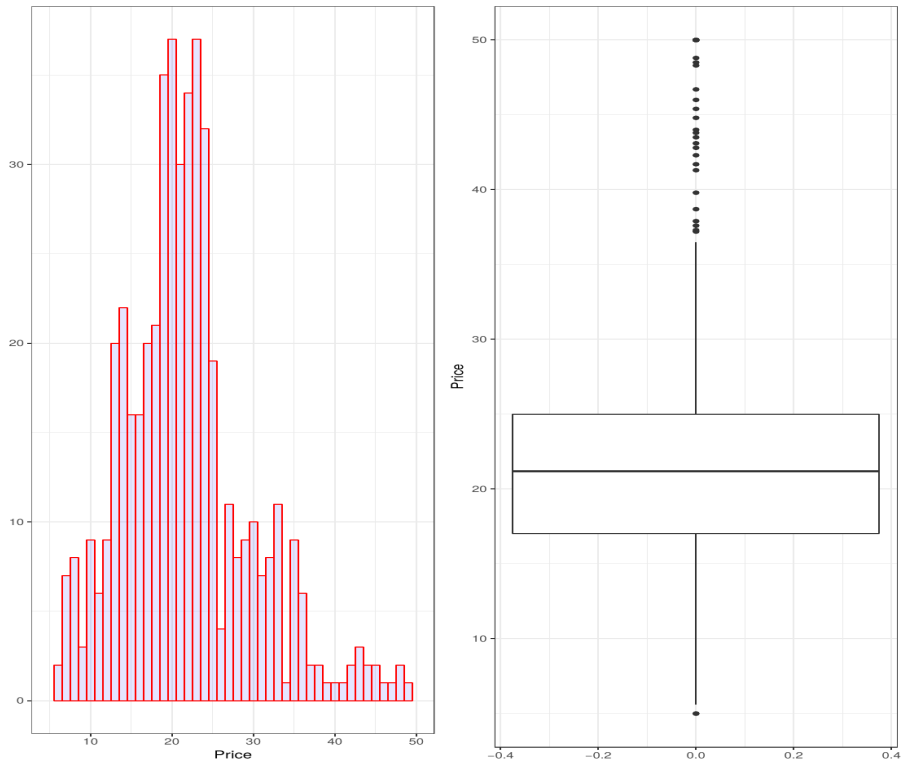

Figure 11 – Histogram (left) and Boxplot (right) of `price` from Boston housing data.

As can be see in Figure 11 the data appears to be symmetrical. In the box plot of this variable is displayed, Figure 11b, shows the presence of many points considered as outliers. The symmetry of the data appears again, showing that they are little dispersed. On the other hand, looking at the summary measures is verified a homogeneous data with coefficient of variation equal to 0.135. The skewness is -0.334 and kurtosis is 3.808.

Therefore a nonormal probability distribution that differs from the Gaussian can be required to model the respose variable, PRICE indicating the need to use a distribution function capable of dealing with these characteristics.

The plot of the response variable against explanatory variables is given by figure 12. For this, the median price increases when the residence is sited near the Charles River. Looking at Figure 12f for the median to RM, note a linear positive relation. The median relationship of the response variable with LSTAT, Figure 12m, appears be linearly negative. For all other variables, the relation a complex relation is drawn, as for example the plot of median against the RAD, shown in Figure 12j. 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.

Comparison

The final fitted model by Pace e Gilley (1997), under the assumption that the data follow a normal distribution for response variable, is shown in (5.1). They collected the location of tracts in terms of latidude and longitude. And so they added spatial information to the hedonic price model of Harrison e Rubinfeld (1978) and the modeling of median household prices was performed as follows:

$$\begin{aligned}\log(\texttt{PRICE}) = {} & \beta_0 + \beta_1\texttt{CRIM} + \beta_2\texttt{AGE} + \beta_3\texttt{NOX} + \beta_4\texttt{CHAS} + \beta_5\texttt{RM} \\ & + \beta_6\texttt{ZN} + \beta_7\texttt{INDUS} + \beta_8\texttt{PTRATIO} + \beta_9\texttt{RAD} + \beta_{10}\texttt{TAX} \\ & + \beta_{11}\texttt{B} + \beta_{12}\texttt{LSTAT} + \beta_{13}\texttt{DIS} + \beta_{14}\texttt{LAT} + \beta_{15}\texttt{LONG} \\ & + +\beta_{16}\texttt{LAT}^2 + \beta_{17}\texttt{LONG}^2 + +\beta_{18}\texttt{LAT*LONG}, \end{aligned} \tag{5.1}$$

in the above equation the price of residences is has a linear relation in all the explanatory variables, including the interaction between the coordinates (LAT*LONG). To better analyze this model, it was performed the analysis of residuals, in Figure 13 to verify the suitability of the model with the Worm plot Buuren e Fredriks (2001) was used, and this one way of ascertaining the adequacy of the regression residuals. The Worm plot is a detrended QQ plot for verify the fitting of data, because showing the differences between two distributions, conditioned to covariate values. This plot 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 dotted horizontal line).

The distribution Box-Cox t, BCT(μ, σ, ν, τ), distribution was chosen in a preliminary analysis, based on Generalized Akaike Information Criterion (GAIC) (PAN, 2001) with penalty $k = 2$, which is equivalent to the standard Akaike Information Criterion (AIC).

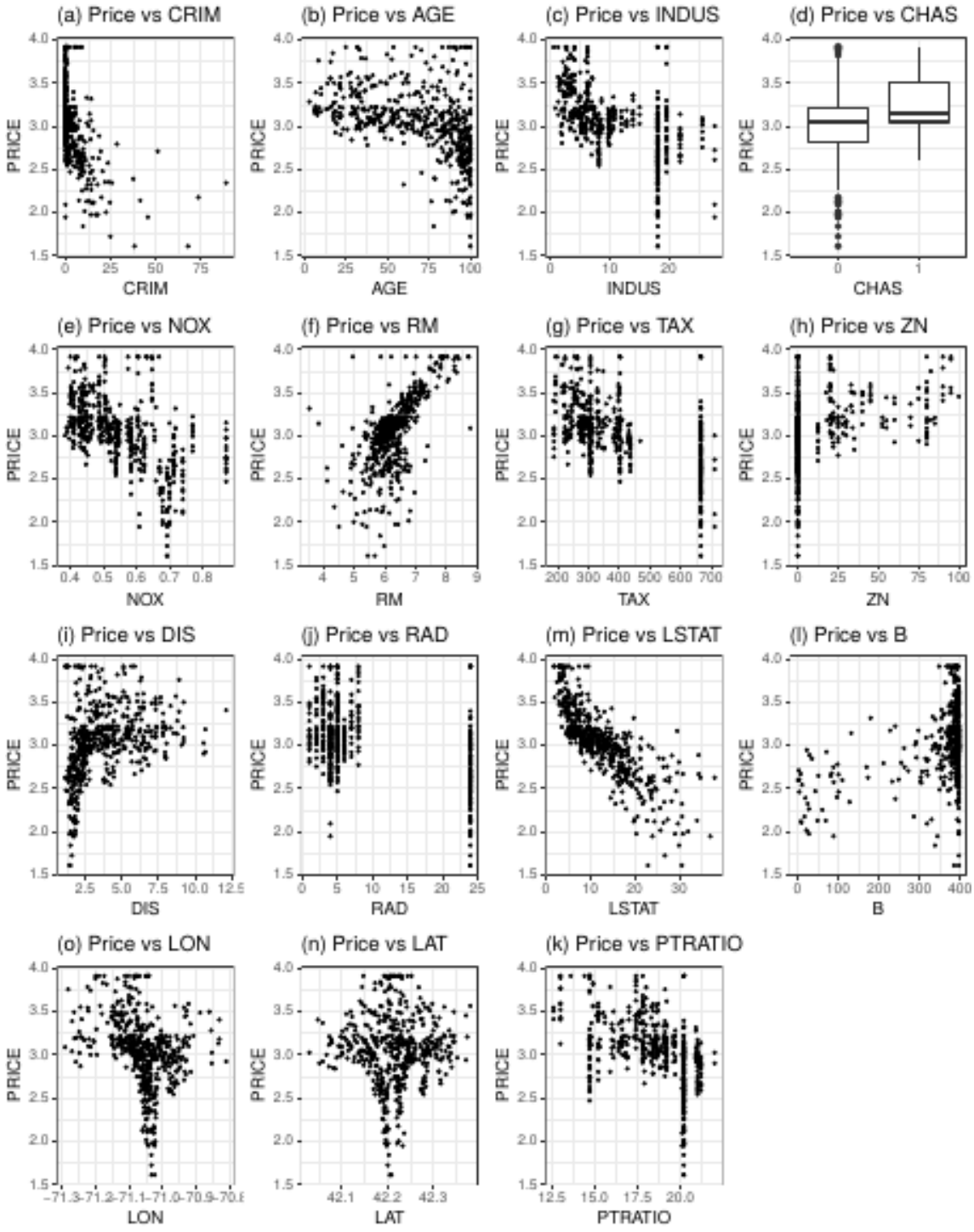

Figure 12 – Plot of `Price` against exploratory variables, of Boston Housing data

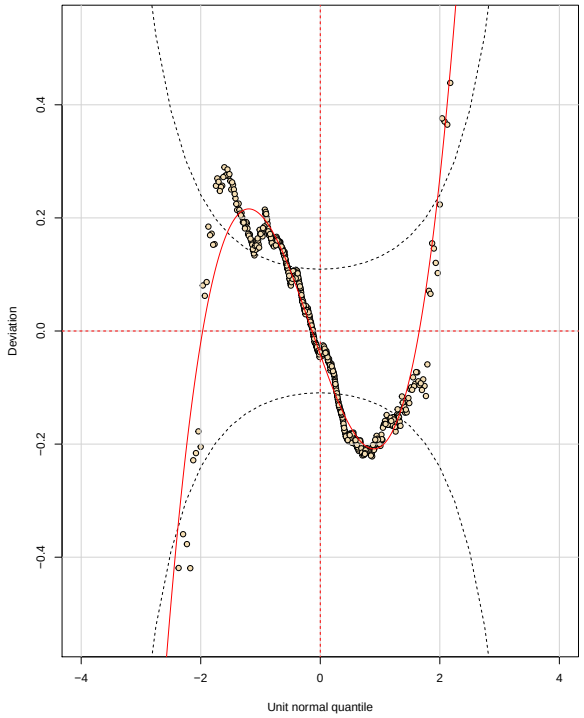

Figure 13 – Worm plot of the model Pace e Gilley (1997) for the Boston Housing data

This was done through the function `chooseDist()`, from **gamlss** package. The explanatory variables were selected based on GAIC, using the `stepGAICAll.A()` function from same package. The parameter $\mu > 0$ corresponds to the median in BCT, and $\sigma(\frac{\tau}{\tau-2})^{0.5}$ is approximate the coefficient of variation (when $\sigma > 0$ is small, $\nu > 0$ and τ is moderate or larger)(RIGBY et al., 2019 forthcoming), ν and τ control skewness and kurtosis, respectively. Figure 14 shows a plot of the BCT distribution for different values of the parameters.

The SAR model in the approach GAMLSS

Thus, the final fitted model was:

$$
\begin{aligned}
Y &\sim \text{BCT}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}), \\
\hat{\mu} &= -43.36 - h_{11}(\texttt{LSTAT}) - h_{21}(\texttt{NOX}) + h_{31}(\texttt{RM}) \\
&\quad - h_{41}(\texttt{CRIM}) - h_{51}(\texttt{LON}) + s(\text{census tract}), \\
\log(\hat{\sigma}) &= -2.4325078 + h_{12}\texttt{TAX} - h_{12}\texttt{DIS} + -0.0014336\texttt{B}, \\
\log(\hat{\nu}) &= -0.5126, \\
\log(\hat{\tau}) &= 1.67 - h_{14}(\texttt{CRIM}),
\end{aligned}
$$

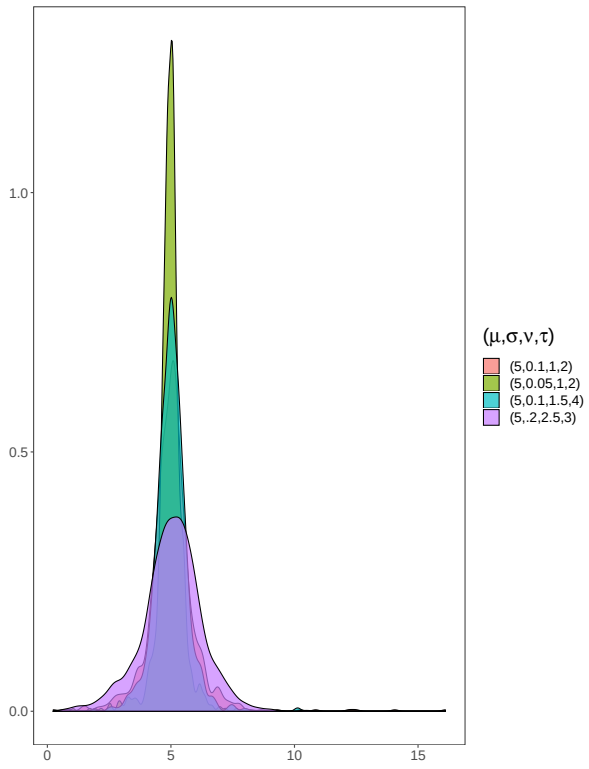

Figure 14 – Plot of Box-Cox T distribution

where the functions h are P-*splines* functions, and s is IAR spatial smoothing function with penalty matrix provided by the `sp2precSar()` function, which corresponds to the covariance of the SAR model rewritten as an IAR model. The spatial IAR smoother was employed only in median (μ) modeling. The term plots shows the partial effect of the variables used on the model parameter, in this case μ, σ and τ. In Figure 15, `LSTAT` has a decreasing effect on μ, indicating the idea that the larger the number of people with lower status in the census tract the more devalued are the real estate in this area. For the higher levels nitric oxides concentration (`NOX`) in the census tract, there is a depreciates the value of real estate. The average number of rooms has a positive effect on μ for already large residences, with number of rooms greater than 6.

In relation to fitted scale parameter, σ, Figure 16 shows the effects of explanatory variables in modeling. The weighted distances to five Boston employment centers (`DIS`) has negative effect on σ, which can be thought of as the prices of the residences become more equal. The same is true for variable `B`, indicating that the greater number of black people in the census tract reduces the variability of house prices. And the increase in `TAX` implies the price of residences with greater variability. Figure 17 shows the partial for the τ, indicating that the number of crimes has a negative effect on house price kurtosis.

In Figure 18 the map of predicted values by census tract is shown. Note that the places where the residences have higher values are in the center-west region of the map,

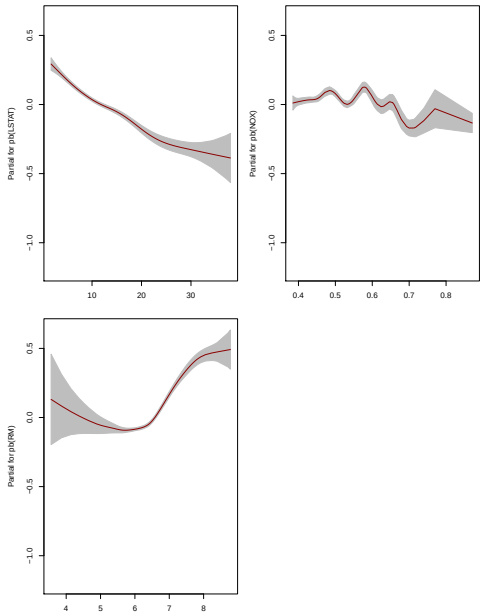

Figure 15 – Term plot of model `mfinal.spatial` for $\hat{\mu}$.

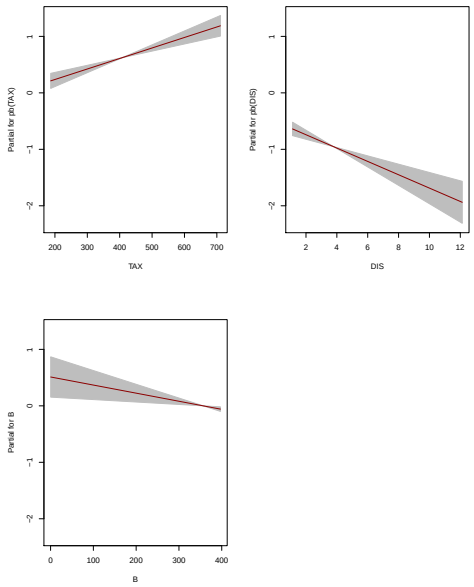

Figure 16 – Term plot of model `mfinal.spatial` for $\log(\hat{\sigma})$.

and these places have as local neighbors with high values as well.

In the diagnostic analysis, looking at Figure 19 which indicates a reasibable fit to the data, since over 95% of points lie within the elliptical (dashed) 95% interval bands. The worm plot gives us a sense of how appropriate for data our fitted model is. In figure 19 all points are within the 95% confidence band between the two elliptic curves, showing that this model specification is adequate.

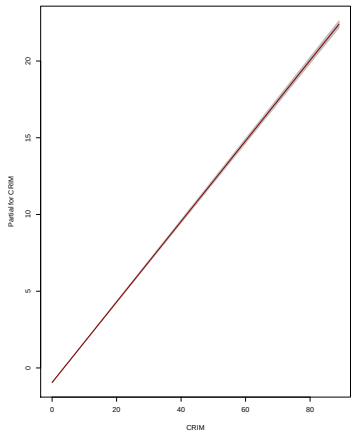

Figure 17 – Term plot of model `mfinal.spatial` for $\log(\hat{\tau})$.

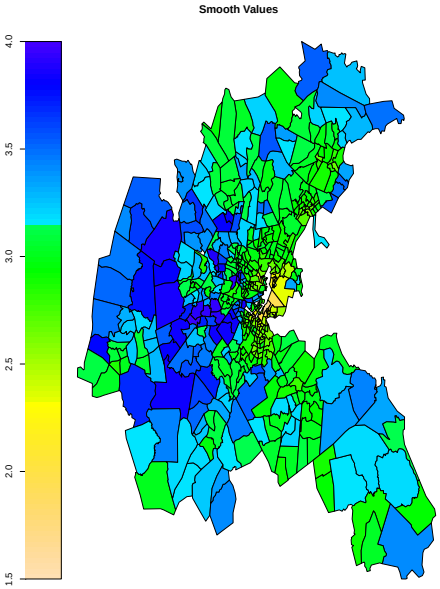

Figure 18 – Fitted values of $\hat{\mu}$ from `final model`

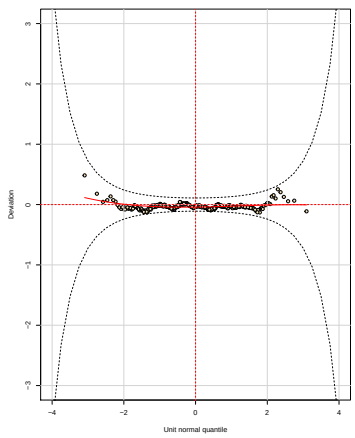

Figure 19 – Worm plot of the `mfinal.spatial`

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:

$$\mathcal{I}_{gini} = \frac{1}{2d^2\psi}\sum_{j=1}^{m}\sum_{k=1}^{m} d_j d_k |z_j - z_k|,$$

where ψ is average income in society, d is the total of persons in society. And here the income data are ordered and subdivided into j classes, and thus the absolute difference of the pairs of income, $|z_j - z_k|$, is computed.

The set of variables that affect the coefficient of gini is described below, according to 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);
- `GDP` is the gross domestic product for the current year of 2010;
- `POP_TOT` is the number of inhabitants of that city in 2010;
- `PEA` is the number of economically active people in the population for the year of 2010;
- `POP_elderly` is the number of old-aged people in the population for the year of 2010;
- `POP_young` is the number of young people in the population for the year of 2010;
- `TX_illiterate` is the proportion of illiterate people in the population for the year of 2010;
- `TX_unem` is the proportion of unemployed people in the population for the year of 2010;
- `PBF` which is a financial aid to poor families with pregnant women and children and adolescents between 0 and 17 years old and extremely poor, benefits range from R$20.00 to R$ 182.00, for the year of 2010; and
- `BPC` consists of a minimum wage income (R$ 510,00) for the elderly and deficient who cannot and cannot be supported by their families, were collected in Ministério do Desenvolvimento Social (MDS) for year of 2010.

In the descriptive analysis were found problems related to the correlation of some of the explanatory variables. The variance inflation factor (VIF) was used to select variables that could incur modeling problems. In order to avoid potential problems of multicollinearity, was 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 `ELDeYOUNG` was produced by the ratio between `POP_elderly` and `POP_young`.

The selection of a suitable model $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L}\}$ for the data of the Gini was through the following components of $\mathcal{M}$:

1. $\mathcal{D}$: A distribution function is specified for the response variable;
2. $\mathcal{G}$: Specifies the set of link functions for the modeling of parameter;
3. $\mathcal{T}$: Denotes the terms used in modeling for each parameter; and
4. $\mathcal{H}$: Specifies the smoothing hyperparameters which determine amount of smoothing in the $h_{jk}()$ and of spatial effect.

The selection of an appropriate distribution for variable was performed by comparing different models using the AIC, this stage is called a fitting stage. The other stage of selection of model is the diagnostic stage, this involves the use of worm plots of normalized quantile residuals (i.e. z-scores) to verify the distribution function used. The distribution function chosen with these two selection stages was the Beta distribution with parameterization proposed by Ferrari e Cribari-Neto (2004).

The probability density function (pdf) of a Beta Y variable is given by:

$$f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1,$$

where the parameters satisfy $0 < \mu < 1$, $\phi > 1$ with $E(y) = \mu$, $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$ (RIGBY et al., 2019 forthcoming).

In the selection of the link functions (the component $\mathcal{G}$), the criterion used in the choice was the interval of the parameters. The criterion used in the choice was the interval of the parameters. Therefore, the logit function was chosen for μ and ν modeling and the log was chosen to σ.

For the $\mathcal{T}$ component, which is selection of the terms in the model, a GAIC procedure was used (with $k = 4$) and selection strategy using Beta (μ,σ) distribution is given below:

1. use a backward GAIC selection procedure to select an appropriate model for μ, with σ fitted as constant;

2. use a forward selection procedure to select an appropriate model for σ , given the model for μ obtained in (1)

3. use a backward selection procedure to select an appropriate model for σ , given the model for μ obtained in (1)

4. use a backward selection procedure to select an appropriate model for μ, given the models for σ obtained in (1)

The smoothing parameters, component $\mathcal{H}$, were fitted using local maximum-likelihood method, and were used for both parameters μ and σ. After selecting the explanatory variables and the distribution for response variable, two models were fitted: one with spatial effect for μ, and other for μ and σ. In both, the spatial effect is a IAR spatial model with penalty matrix from SAR. The fitted model is given by:

$$\begin{aligned} Y &\sim \text{BE}(\hat{\mu}, \hat{\sigma}), \\ \log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\} &= 0.07462 + h_{11}(\texttt{TX\_unem}) + h_{21}(\texttt{GDP}) + h_{31}(\texttt{PBF}) + s(\text{city}), \\ \log\left\{\frac{\hat{\sigma}}{1-\hat{\sigma}}\right\} &= -1.99515 - 0.12557\texttt{GDP} - 0.0003409\texttt{BPC} + s(\text{city}), \end{aligned}$$

where the functions h are P-*splines* functions, and s is IAR spatial smoothing function with penalty matrix which corresponds to the covariance of the SAR model rewritten as an IAR model, as before, provided by function `sp2precSar()`. Here the IAR spatial smoother was statistically significant in both the modeling of the mean and the scale parameter.

The partial effects of the explanatory variables on the μ and σ parameters can be seen in term plots. Figure 20 shows that `TX_unem` has a positive effect on the μ to some extent and then this effect is decreasing, showing that if the unemployment rate increases enough everyone will be poor to the same extent reducing income inequality. The `GDP` has a decreasing effect in μ for small values, then there is a positive effect to indicate an increase in the mean of the inequality to a certain level, finally if the GDP grows too much, the societies become richer and the average of the Gini index decreases .

Figure 21 shows the partial effect of explanatory variables `GDP` and `BPC`. In both the partial effect is decreasing linearly, indicates that the increase of `GDP` and `BPC` imply in higher concentration of the Gini index.

Figure 22 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 plot shows the vast majority of points on the red line, indicating a reasonable fitting. Figures 23 shows the worm plot in four intervals of `GDP`, the worm plot graph on top right side shows that the model fitted for this range of `GDP` variable did not

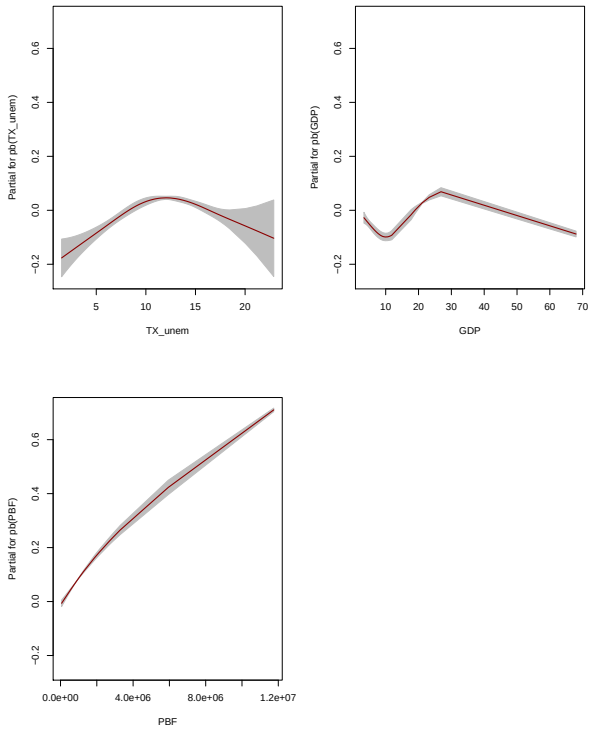

Figure 20 – Term plots for $\log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\}$

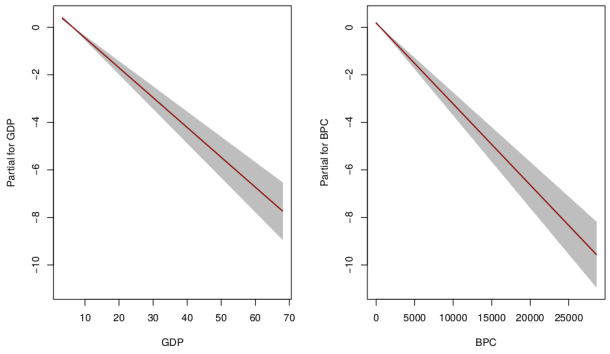

Figure 21 – Term plots for $\log\left\{\frac{\hat{\sigma}}{1-\hat{\sigma}}\right\}$

have a good fit. shows that the model adjusted for this range of the GDP variable did not have a good fit. evidencing the a reasonable fit. The Figure 24 indicates the worm plot for three intervals of variables `Tx_unem` and `PBF` indicates a mean too high of residuals, but in general a acceptable fit for the data.

The predicted values from the fitted model for the μ of Gini coefficient are shown in Figure 25. High coefficient values are found in the cities located to the east and west in the state of Pernambuco. These cities have greater economic activities, and therefore the concentration of income are greater in these cities.

In this chapter, the $\text{SAR}_{\text{gamlss}}$ models were applied to real data for different problems.

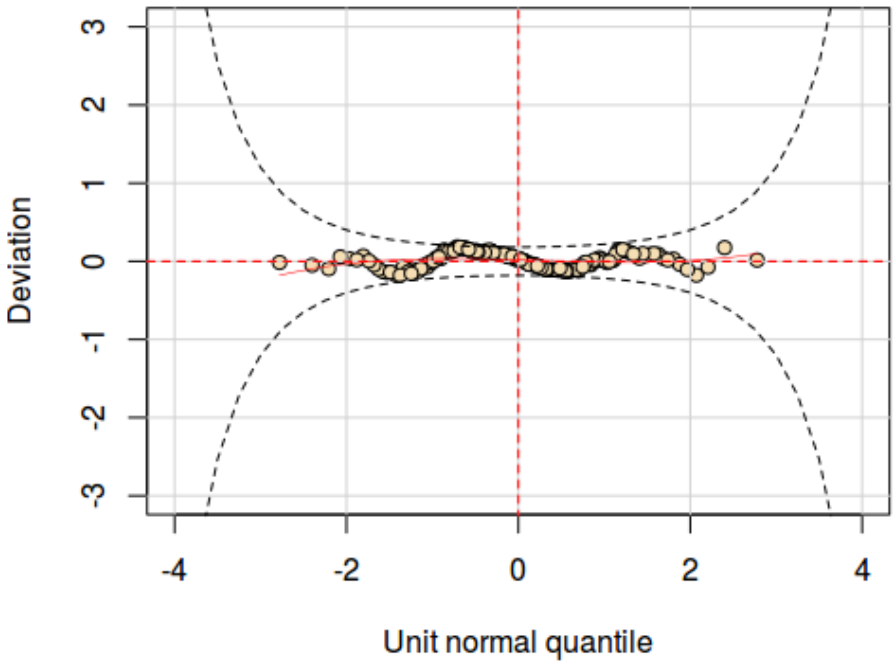

Figure 22 – Worm plot of $\text{SAR}_{\text{gamlss}}$ model for `Gini`

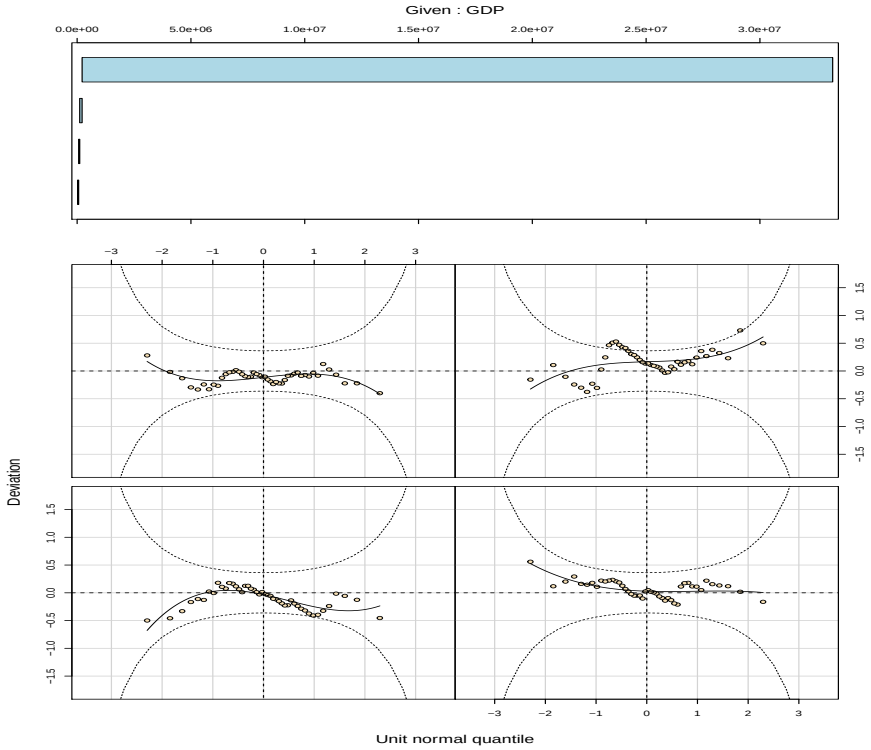

Figure 23 – Worm plot of residuals by levels of `GDP` for the final model fitted

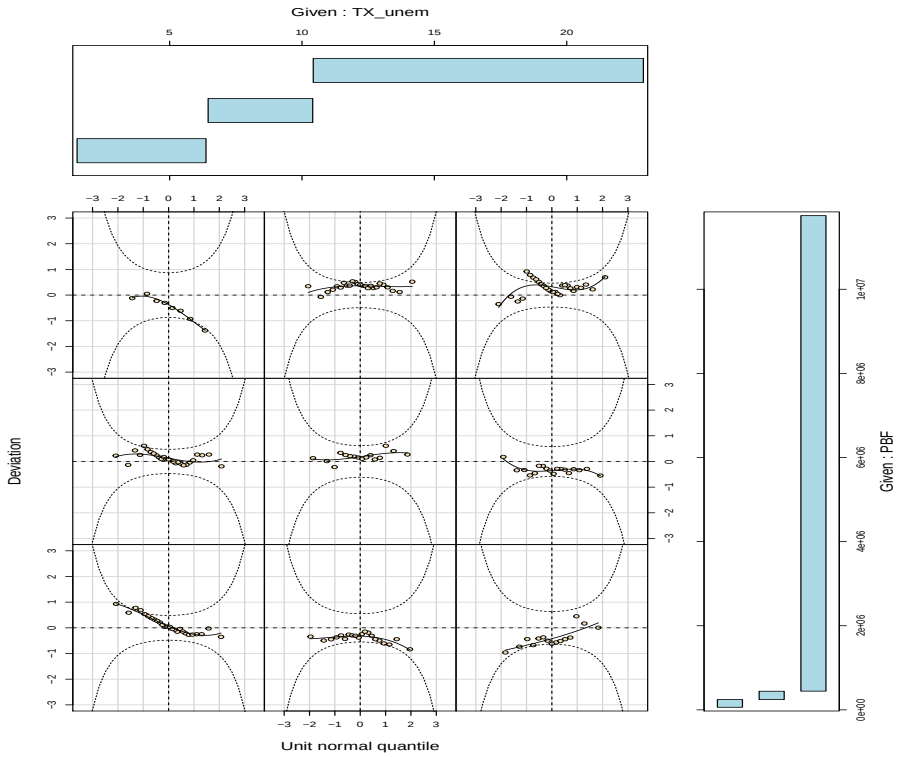

Figure 24 – Worm plot of residuals by levels of `Tx_unem` and `PBF` from the final model fitted

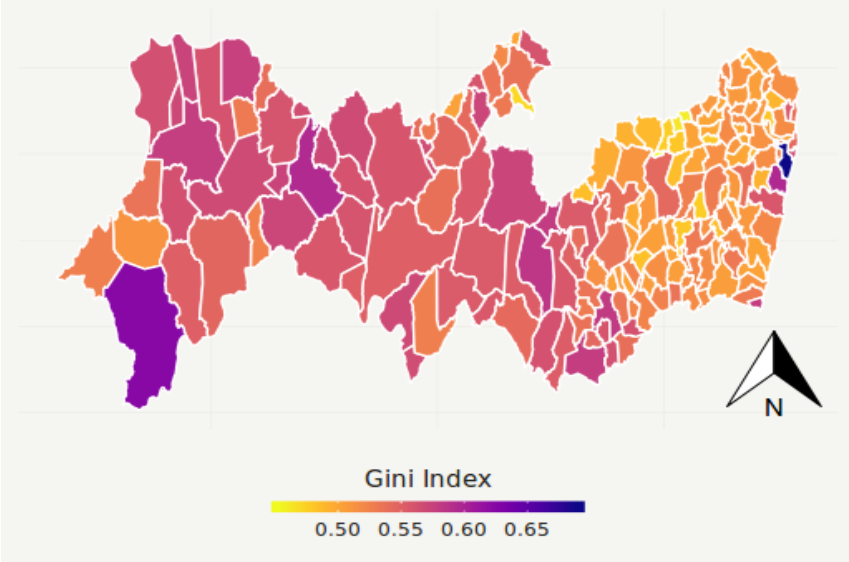

Figure 25 – Fitted values of $\hat{\mu}$ by Cities in Pernambuco

The first application to the Boston Housing data, showed a comparison made to the Pace e Gilley (1997) model that considers normal errors, and indicated that the authors' model may not be adequate for this data set. The flexibilization of the distribution function of the response variable, its modeling and the incorporation of spatial effects and semiparametric terms bring a reasonable fit for these data. The second problem was a modeling for the Gini data. The $\text{SAR}_{\text{gamlss}}$ was applied to take into account the spatial effect contained in these data and also a more flexible distribution function for the data. Through the worm plots the acceptable fit for these data was verified.

6 CONCLUSION

The objective of this work was to introduce the class of simultaneous autoregressive spatial models within the GAMLSS approach. For this, 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. The conclusions of simulations results, in finite sample context with normal errors and linear trend and without and low spatial dependence, the the $\text{SAR}_{\text{gamlss}}$ have similiar behavior to SAR_{lag}, and $\text{SAR}_{\text{error}}$. The results also showed that the $\text{SAR}_{\text{gamlss}}$ is preferable in the context of nonlinear trend and when the response variable is non-Gaussian. Two applications were performed showing the importance of the employability of this tool in the field of Spatial Econometrics. The flexibilization of the distribution function of response variable, and the incorporation of spatial effects and semiparametric terms bring a reasonable fit for Boston House data. The second application was a modeling for the Gini index, the flexibility of the GAMLSS allowed the modeling of location and scale parameters of distribution for Gini index with spatial effects in both. Further, diagnostic analysis confirmed a reasonable fit for these data. The spatial analysis performed in this master thesis can be applied to other data sets that have georeferenced information and neighbourhood information.

REFERENCES

AKAIKE, H. A new look at the statistical model identification. In: *Selected Papers of Hirotugu Akaike*. [S.l.]: Springer, 1974. p. 215–222.

ALAM, M.; RÖNNEGÅRD, L.; SHEN, X. Fitting conditional and simultaneous autoregressive spatial models in hglm. *The R Journal*, v. 7, n. 2, p. 5–18, 2015.

ANSELIN, L. Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, Sage Publications, v. 26, n. 2, p. 153–166, 2003.

ANSELIN, L.; SMIRNOV, O. Efficient algorithms for constructing proper higher order spatial lag operators. *Journal of Regional Science*, Wiley Online Library, v. 36, n. 1, p. 67–89, 1996.

BANERJEE, S. et al. *Hierarchical modeling and analysis for spatial data*. [S.l.]: Chapman and Hall/CRC, 2004.

BARROS, R. P. d.; CARVALHO, M. d.; FRANCO, S.; MENDONÇA, R. A queda recente da desigualdade de renda no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2007.

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.

BESAG, J.; HIGDON, D. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 61, n. 4, p. 691–746, 1999.

BESAG, J.; KOOPERBERG, C. On conditional and intrinsic autoregressions. *Biometrika*, Oxford University Press, v. 82, n. 4, p. 733–746, 1995.

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

BLOMMESTEIN, H. J.; KOPER, N. A. Recursive algorithms for the elimination of redundant paths in spatial lag operators. *Journal of Regional Science*, Wiley Online Library, v. 32, n. 1, p. 91–111, 1992.

BOOR, C. D.; BOOR, C. D.; MATHÉMATICIEN, E.-U.; BOOR, C. D.; BOOR, C. D. *A practical guide to splines*. [S.l.]: springer-verlag New York, 1978. v. 27.

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.

CAJIAS, M. Is there room for another hedonic model? the advantages of the gamlss approach in real estate research. *Journal of European Real Estate Research*, Emerald Publishing Limited, v. 11, n. 2, p. 204–245, 2018.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

CRESSIE, N. *Statistics for spatial data*. [S.l.]: J. Wiley, 1993. (Wiley series in probability and mathematical statistics: Applied probability and statistics). ISBN 9780471002550.

CRESSIE, N.; WIKLE, C. *Statistics for spatio-temporal data, vol. 465*. [S.l.]: Wiley, 2011.

De Bastiani, F.; RIGBY, R. A.; STASINOPOULOUS, D. M.; CYSNEIROS, A. H.; URIBE-OPAZO, M. A. Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 1, p. 168–186, 2018.

DURBÁN, M. et al. Sar models with nonparametric spatial trends. a p-spline approach. *Estadística Española*, v. 54, n. 177, p. 89–111, 2012.

EDWARDS, D. *Introduction to graphical modelling*. [S.l.]: Springer Science & Business Media, 2012.

EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, JSTOR, p. 89–102, 1996.

FAHRMEIR, L.; KNEIB, T.; LANG, S.; MARX, B. *Regression: models, methods and applications*. [S.l.]: Springer Science & Business Media, 2013.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of applied statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

FOX, J. *Applied regression analysis and generalized linear models*. [S.l.]: Sage Publications, 2015.

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

HARRISON, D.; 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.

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

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

HOEF, J. et al. On the relationship between conditional (car) and simultaneous (sar) autoregressive models. *Spatial Statistics*, Elsevier, v. 25, p. 68–85, 2018a.

HOEF, J. et al. Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, Wiley Online Library, v. 88, n. 1, p. 36–59, 2018b.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis Group, v. 12, n. 1, p. 55–67, 1970.

JONES, M.; PEWSEY, A. Sinh-arcsinh distributions. *Biometrika*, Oxford University Press, v. 96, n. 4, p. 761–780, 2009.

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

LEE, W.; LEE, Y. Modifications of reml algorithm for hglms. *Statistics and Computing*, Springer, v. 22, n. 4, p. 959–966, 2012.

LICHSTEIN, J. W.; SIMONS, T. R.; SHRINER, S. A.; FRANZREB, K. E. Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, Wiley Online Library, v. 72, n. 3, p. 445–463, 2002.

LUO, J. o. Distribution characteristics of stock market liquidity. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 392, n. 23, p. 6004–6014, 2013.

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.

MCCULLAGH, P.; NELDER, J. *Generalized Linear Models, Second Edition*. [S.l.]: Chapman & Hall, 1989. (Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series). ISBN 9780412317606.

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.

OKABE, A.; BOOTS, B.; SUGIHARA, K.; CHIU, S. N. *Spatial tessellations: concepts and applications of Voronoi diagrams*. [S.l.]: John Wiley & Sons, 2009. v. 501.

PACE, R. K. Nonparametric methods with applications to hedonic models. *The Journal of Real Estate Finance and Economics*, Springer, v. 7, n. 3, p. 185–204, 1993.

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.

PAN, W. Akaike's information criterion in generalized estimating equations. *Biometrics*, Wiley Online Library, v. 57, n. 1, p. 120–125, 2001.

RAY, D. *Development economics*. [S.l.]: Springer, 2008.

RIGBY, R. et al. *Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2019 forthcoming.

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.

RIGBY, R. A.; STASINOPOULOS, D. M. Automatic smoothing parameter selection in gamlss with an application to centile estimation. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 23, n. 4, p. 318–332, 2014.

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

SÁ, A. et al. Exploring fire incidence in portugal using generalized additive models for location, scale and shape (gamlss). *Modeling Earth Systems and Environment*, Springer, v. 4, n. 1, p. 199–220, 2018.

SIMLAI, P. Estimation of variance of housing prices using spatial conditional heteroskedasticity (sarch) model with an application to boston housing price data. *The Quarterly Review of Economics and Finance*, Elsevier, v. 54, n. 1, p. 17–30, 2014.

STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2017.

SUBRAMANIAN, S.; CARSON, R. T. Robust regression in the presence of heteroskedasticity. *Advances in Econometrics*, JAI Press, v. 7, p. 85–138, 1988.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. *Economic geography*, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.

VOUDOURIS, V.; GILCHRIST, R.; RIGBY, R.; SEDGWICK, J.; STASINOPOULOS, D. Modelling skewness and kurtosis with the bcpe density in gamlss. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012.

WHITTLE, P. On stationary processes in the plane. *Biometrika*, JSTOR, p. 434–449, 1954.

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

APPENDIX A – SIMULATIONS STUDY

Complementary results for chapter 4 on simulation studies are presented in this appendix. The boxplots for the β_0 parameter are shown for the SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ models, with different values of n, and also in the presence of spatial dependence and not. Figures 26 and 27 show the inconsistency of the β_0 estimator from SAR_{lag}.

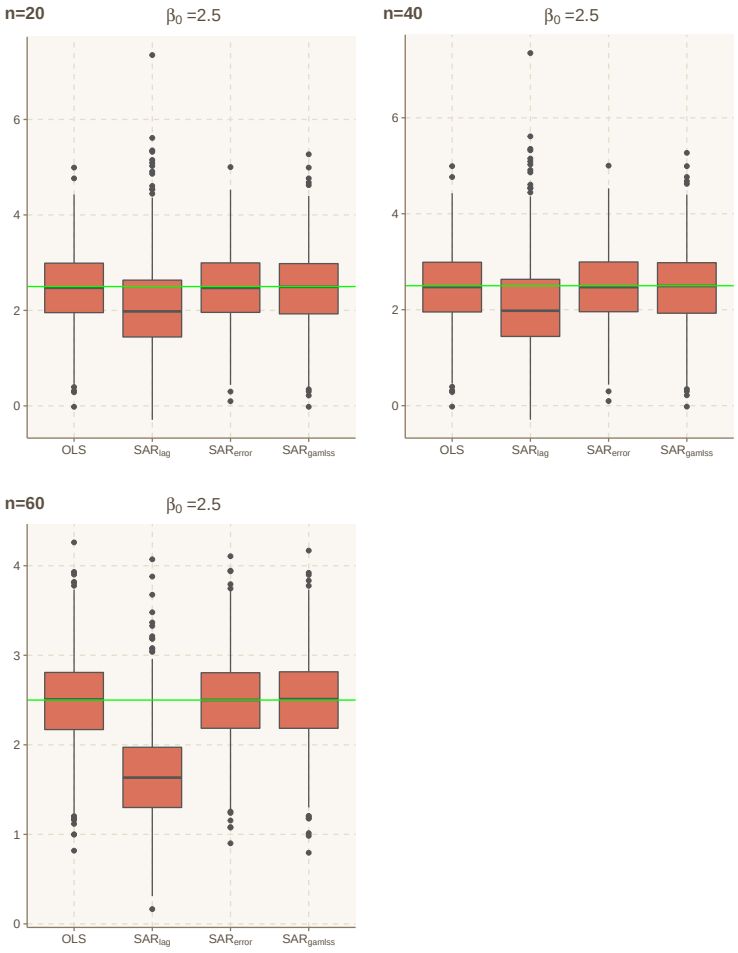

Figure 26 – Boxplots of parameters estimates for β_0 from models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.0$.

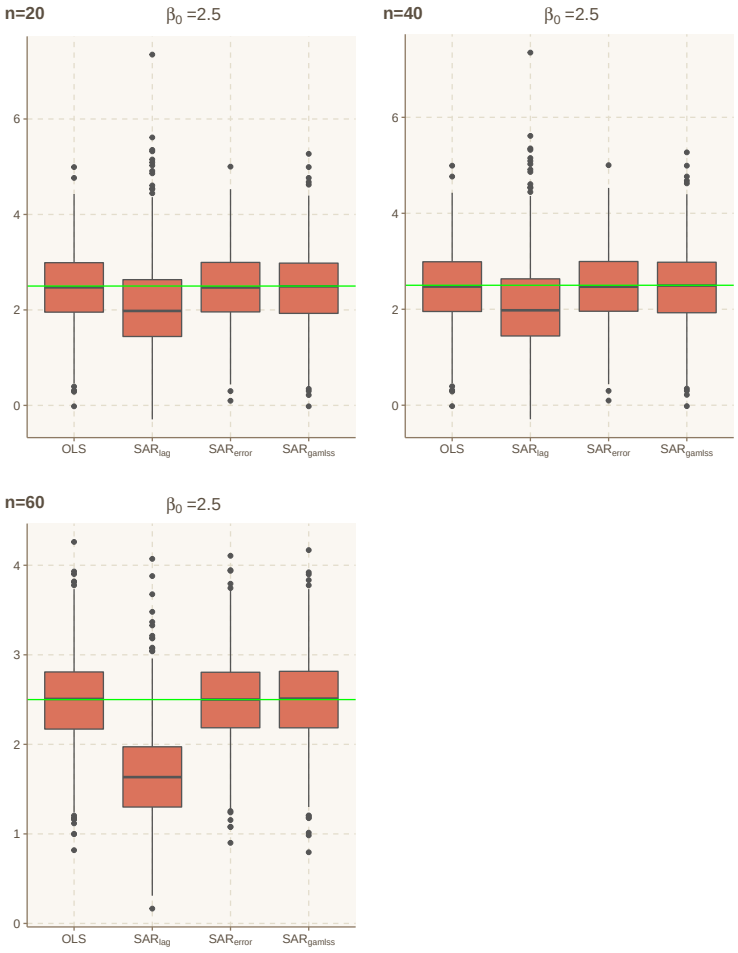

Figure 27 – Boxplots of parameters estimates for β_0 from models OLS, SAR_{lag}, $\text{SAR}_{\text{error}}$ and $\text{SAR}_{\text{gamlss}}$ with $\rho = 0.10$.