TADEU VILELA DE SOUZA

ESTIMAÇÃO EM REGRESSÃO INVERSA NO MODELO CAR ESPACIAL

LAVRAS - MG

2017

TADEU VILELA DE SOUZA

ESTIMAÇÃO EM REGRESSÃO INVERSA NO MODELO CAR ESPACIAL

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador
Dr. João Domingos Scalon

LAVRAS - MG
2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a)

Souza, Tadeu Vilela de.
Estimação em regressão inversa no modelo CAR espacial. / Tadeu Vilela de Souza. - 2017
92 p. : il.

Orientador: João Domingos Scalon.
Tese (doutorado) – Universidade Federal de Lavras, 2017.
Bibliografia.

1. Regressão inversa. 2. Dependência espacial. 3. Estimador pontual. 4. Estimador intervalar. 5. Imputação. I. Universidade Federal de Lavras. II. Título.

TADEU VILELA DE SOUZA

ESTIMAÇÃO EM REGRESSÃO INVERSA NO MODELO CAR ESPACIAL

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 17 de fevereiro de 2017.

Prof. Dr. Renato Ribeiro Lima	UFLA
Prof. Dr. Marcelo Silva Oliveira	UFLA
Prof. Dra. Carla Regina Guimarães Brighenti	UFSJ
Prof. Dra. Liliane Lopes Cordeiro	UFV

Dr. João Domingos Scalon
Orientador

LAVRAS - MG
2017

Aos meus pais Maria Anália e Geiel,

pela dedicação, apoio, amor e educação.

À minha irmã Elaine,

pelo incentivo, carinho e por acreditar em mim.

DEDICO.

AGRADECIMENTOS

A Deus, pela oportunidade de estudar e pela força dada para aguentar firme os momentos difíceis.

Aos maiores merecedores da minha gratidão, minha mãe Maria Anália, meu pai Geiel e minha irmã Elaine, pessoas a quem dedico incondicionalmente meus agradecimentos.

A todos os meus familiares e amigos, pelo apoio, carinho, admiração e momentos de alegrias inesquecíveis passados juntos.

Ao meu orientador João Domingos Scalon, pelos conhecimentos e esclarecimentos intelectuais, pela confiança em mim, pela paciência e compreensão das minhas dificuldades.

À minha colega e amiga Liliane, pela indispensável e importante ajuda que recebi dela na realização deste trabalho.

Aos professores membros da minha banca pelas importantes contribuições nesta tese, por serem receptivos e gentis ao receberem o meu convite e por participarem da minha qualificação e defesa.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES), por oferecer estrutura e acolhimento durante esses vários anos de estudo.

Aos professores do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária da UFLA, pelas importantes e úteis contribuições na minha formação durante as suas disciplinas.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pela concessão da bolsa de estudos no começo de meu doutorado.

A todos que contribuíram de forma direta ou indireta para a realização deste trabalho.

RESUMO

Regressão inversa ou calibração estatística é uma técnica estatística utilizada em situações em que, por meio da análise de regressão, deseja-se estimar um valor desconhecido da variável independente dado o valor da variável dependente. Métodos para a estimação pontual e intervalar na regressão inversa para esse valor desconhecido estão disponíveis na literatura. Porém, observa-se que são escassos os métodos que considerem a informação espacial dos dados no processo de estimação na regressão inversa. O objetivo principal desta tese é propor a regressão espacial inversa ou calibração espacial por meio de métodos para a estimação pontual e intervalar do valor desconhecido da variável independente utilizando um modelo que considere a estrutura de dependência espacial em dados de área. Esses estimadores foram construídos a partir de um modelo do erro espacial ou modelo condicional autorregressivo (CAR) e aplicados em dados reais que caracterizam um problema de calibração espacial. Os resultados obtidos mostram que a regressão espacial inversa é apropriada na análise dados de área com dependência espacial, fornecendo um ferramenta útil para casos que configurem a necessidade de se obter o valor de uma variável independente conhecendo-se o valor da variável dependente. Observa-se também que, um grande potencial que esse modelo de regressão espacial inversa tem, está no fato de que ele pode ser um método eficiente de imputação, em casos específicos, de dados faltantes na análise de dados de área.

Palavras-chave: Regressão inversa. Dependência espacial. Estimador pontual. Estimador intervalar. Imputação.

ABSTRACT

Inverse regression or statistical calibration is a statistical technique used in situations where, through regression analysis, it is desired to estimate a unknown value of the independent variable given the value of the dependent variable. Methods for the point and interval estimation in the inverse regression for this unknown value are available in the literature. However, it is observed that there are few methods that consider the spatial information of the data in the estimation process in the inverse regression. The main objective of this thesis is to propose inverse spatial regression or spatial calibration by means of methods for the point and interval estimation of the unknown value of the independent variable using a model that considers the spatial dependence structure in area data. These estimators were constructed using spatial error model or autoregressive conditional model (CAR) and applied to real data that characterize a spatial calibration problem. The results show that the inverse spatial regression is appropriate in the spatial dependence data area analysis, providing a useful tool for cases that configure the need to obtain the value of an independent variable by knowing the value of the dependent variable. It is also observed that a great potential that this inverse spatial regression model has is in the fact that it can be an efficient method of imputation, in specific cases, of missing data in the analysis of area data.

Keywords: Regression inverse. Spatial dependence. Point estimator. Interval estimator. Imputation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Na Estatística, muitas vezes existe o interesse em relacionar duas variáveis de forma que seja possível entender a variação de uma por meio da variação da outra. Particularmente, a análise de regressão simples estuda a relação entre uma variável Y, chamada variável dependente e outra variável X, chamada variável independente. Através de um modelo matemático objetiva-se determinar o valor de Y correpondente a um valor de X. Em contrapartida, em alguns casos específicos, existe interesse em fazer o contrário, ou seja, determinar o valor da variável independente usando o valor conhecido da variável dependente. Esse interesse ocorre, em sua maioria, quando a medição ou a obtenção da variável independente é mais complicada devido a custos financeiros elevados, demanda de tempo, praticidade, etc. Um exemplo típico para ilustrar essa situação é: considere X a concentração de glicose em certa substância; em seguida, um método espectrofotométrico é usado para medir a absorção Y dessa glicose. Essa absorção depende da concentração de glicose X e assim, pode-se ajustar um modelo de regressão da absorção Y como variável dependente em função da concentração de glicose X caracterizando a variável independente. A resposta Y é fácil medir com o método espectrofotométrico, mas a concentração da glicose não é fácil de medir. Como na prática, tem-se interesse em predizer o valor de concentração de glicose X, este é um problema de regressão inversa ou calibração, onde busca-se obter o valor da concentração x_0 da glicose (variável independente) dado um valor y_0 (variável dependente) da absorção medido de forma simples.

Muitos estudos apresentam a calibração como metodologia prática na relação entre as variáveis dependente e independente, sendo o foco principal o processo de estimação de um valor da variável independente x_0 dado um valor da variável dependente y_0.

Grande parte desses estudos assumem que a relação entre as variáveis Y e X é linear e que os erros do modelo são independentes e seguem uma distribuição normal com média zero e variância constante. Utilizando essas suposições, a regressão inversa é utilizada em diversas áreas, como por exemplo, Biologia, Química, Física, Engenharia, Medicina, entre outras. Estimadores inversos pontuais e intervalares já foram propostos na literatura, porém a dependência espacial do fenômeno em estudo não é considerada no processo de obtenção da maioria desses estimadores.

Em algumas áreas de pesquisa que utilizam dados de áreas com contagens como epidemiologia, econometria, experimentos agronômicos e geológicos, etc, quando um modelo de regressão é ajustado por alguma finalidade, muitas vezes observa-se dependência entre os erros depois do processo de ajuste desse modelo. Essa dependência surge devido ao fato de que regiões geograficamente próximas tendem a apresentar valores semelhantes de algum fenômeno observado, ou seja, uma observação realizada em uma posição sofre influência das regiões vizinhas. Portanto, nesse tipo de situação é necessário utilizar métodos da estatística espacial, pois a aplicação dos procedimentos exigidos a uma análise de regressão tradicional não considera a dependência entre os erros.

A suposição de independência dos erros facilita a teoria da estatística, mas modelos de regressão que modelam a dependência espacial de maneira correta podem ser mais realísticos. O uso de modelos de regressão que englobam informação espacial, como o uso de modelos autorregressivos, tendem a proporcionar estimativas paramétricas que melhoram a qualidade do ajuste tendo em vista que, em condições reais, pode existir depedência espacial, o que contribui para o aumento da variação residual e, consequentemente, uma diminuição da precisão da análise.

O foco principal da análise espacial é encontrar um modelo inferencial

que incorpore explicitamente as relações espaciais constituintes de um fenômeno. Nessa perspectiva, Cordeiro (2015) propôs um modelo de calibração espacial que leva em consideração a estrutura de dependência espacial entre áreas vizinhas, considerando o modelo espacial autorregressivo (SAR).

O objetivo dessa tese é propor um modelo de calibração espacial considerando o modelo condicional autorregressivo (CAR). Esse modelo é ajustado buscando modelar a estrutura de dependência espacial que existe em dados de área. A partir dessa modelagem são obtidos estimadores inversos (pontual e intervalar) de um valor x_0 desconhecido da variável independente, dado o valor y_0 conhecido, da variável dependente. Esses estimadores obtidos foram implementados computacionalmente e aplicados em um fenômeno real que caracteriza um problema de calibração espacial.

2 REFERENCIAL TEÓRICO

2.1 Regressão inversa ou calibração estatística

A calibração simples ou regressão inversa simples é caracterizada estatisticamente por duas variáveis X e Y que se relacionam segundo uma função f conhecida. Na primeira etapa do experimento de calibração, amostram-se n observações da variável aleatória Y, a partir de valores prefixados de X, de modo que se possa estimar a função f que relaciona as duas variáveis. Portanto, X e Y podem se relacionar por meio do modelo estatístico $Y_i = f(X_i) + \varepsilon_i$, com $i = 1, 2, ..., n$. Numa segunda etapa, seleciona-se uma amostra aleatória de tamanho k ($k \geq 1$) da variável Y correspondente a um único valor x_0 desconhecido. Procura-se, então, estimar este valor desconhecido x_0 baseado nas informações do ponto de interesse e da função estimada no experimento de calibração.

2.1.1 Estimador pontual

No modelo de regressão linear simples, dado uma variável dependente Y e uma independente X, a relação entre elas pode ser representada por:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.1}$$

em que β_0 e β_1 são parâmetros da relação linear e ε representa o erro de medição. Baseado nesse modelo, pode-se explorar dois métodos mais comuns para o problema de calibração: o método clássico e o método inverso.

Método clássico

O modelo (2.1) com n valores de X pode ser escrito da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{com} \quad i = 1, 2, ..., n, \tag{2.2}$$

em que $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

Para estimar um valor x_0 desconhecido da variável independente em função de y_0, conhecido, tem-se:

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1}, \tag{2.3}$$

em que os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ são obtidos por meio da estimação pelo método de mínimos quadrados ou pelo método de máxima verossimilhança.

Método inverso

O método inverso para calcular o valor x_0, desconhecido, considera a regressão de X em função de Y, assim reescreve-se o modelo 2.1, considerando n valores, da seguinte forma:

$$x_i = \omega_0 + \omega_1 y_i + \varepsilon_i^*, \quad \text{com} \quad i = 1, 2, ..., n, \tag{2.4}$$

em que $\varepsilon_i^* \sim N\left(0, \sigma^2\right)$.

Agora, o valor x_0 desconhecido da variável independente, em função de y_0, é estimado da seguinte forma:

$$\hat{x}_0 = \hat{\omega}_0 + \hat{\omega}_1 y_0, \tag{2.5}$$

em que as estimativas $\hat{\omega}_0$ e $\hat{\omega}_1$ também são obtidas utilizando o método de mínimos quadrados ou o método de máxima verossimilhança.

Muitos autores discutem a eficiência desses dois estimadores pontuais. Dentre eles pode-se citar Eisenhart (1939), Krutchkoff (1967), Martinelle & Krutchkoff (1970), Shukla (1972), Williams (1969) e Thonnard (2006). O estimador clássico é citado em várias situações como mais eficiente.

Para a escolha de um desses dois métodos em questão, Eisenhart (1939) comparou a análise de variância desses dois estimadores. Na Tabela 1 é apresentada o resultado da análise de variância do estimador clássico e na Tabela 2 é apresentada a análise de variância do estimador inverso. Eisenhart concluiu que o estimador clássico é mais adequado quando o pesquisador seleciona os valores de X com antecedência e depois observa os valores de Y correspondentes. Essa conclusão está baseada na análise da Tabela 2, onde pode-se observar que os valores

da variável x são fixados e estes valores não dependem dos valores observados de Y. A soma de quadrados da regressão, $SQreg$, representa a variabilidade dos valores da variável x observados e esta variabilidade resulta da forma como eles são escolhidos. A $SQreg$ mede a dependência de x em Y, mas esta é uma dependência falsa porque x não depende de Y . Por fim, a $SQreg$ não pode ser interpretada como uma medida do erro dos valores de x, porque os valores dessa variável x são fixados e então não tem um erro. Portanto, como salienta Eisenhart (1939), se os valores da variável x são selecionados e os correspondentes valores de Y são observados, então o estimador clássico é o estimador apropriado para problemas de calibração.

Tabela 1 Análise de variância considerando a regressão de Y em função de x, método clássico.

Fonte de variação	Graus de liberdade	Soma de Quadrados
Modelo	1	$SQreg = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
Erro	$n-2$	$SQerro = \sum_{i=1}^{n} (y_i - \hat{y})^2$
Total	$n-1$	$SQtotal = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Tabela 2 Análise de variância considerando a regressão de x em função de Y, método inverso.

Fonte de variação	Graus de liberdade	Soma de Quadrados
Modelo	1	$SQreg = \hat{\omega}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
Erro	$n-2$	$SQerro = \sum_{i=1}^{n} (x_i - \hat{x})^2$
Total	$n-1$	$SQtotal = \sum_{i=1}^{n} (x_i - \bar{x})^2$

Analisando o número de observações, Shukla (1972) também comparou os dois métodos. A conclusão foi que se o número de observações é pequeno,

então o estimador inverso tem menor erro quadrático médio (EQM). Mas quando um grande número de observações é utilizado, então o estimador clássico é mais vantajoso.

Segundo Thonnard (2006), a maioria dos estatísticos prefere o método clássico, devido à sua consistência e seu EQM de distribuições assintóticas relevantes e que se deve minimizar os erros na direção em que ocorrem, que é na direção de Y. Assim, deve-se utilizar o método clássico.

2.1.2 Estimador intervalar

Além de uma estimativa pontual para o valor de x_0 desconhecido, é importante obter um intervalo que forneça informação da confiança que se pode depositar na estimativa pontual. Nessa perspectiva, existem trabalhos que propõem e discutem a construção de estimadores intervalares para x_0. Podem ser citados Brown (1982), Eisenhart (1939), Fieller (1954), Graybill (1976), Lieberman, Miller e Hamilton (1967), Mathew e Kasala (1994), dentre outros.

Sendo um dos precursores nesse âmbito, Eisenhart (1939), baseado na distribuição de probabilidade t de Student, propôs um estimador intervalar para a estimativa x_0 obtido por meio do método clássico de estimação inversa pontual. Baseado no teorema Student-Fisher que envolve as distribuições t de Student e Qui-quadrado, Eisenhart desenvolveu um intervalo utilizando a distribuição t de Student com $n - 2$ graus de liberdade.

Graybill (1976) apresentou uma técnica para criar um intervalo de confiança para x_0 por meio do método clássico, onde a construção desse intervalo é possível se o parâmetro estimado $\hat{\beta}_1$ para o modelo de regressão é diferente de zero. Para isso, é utilizado um teste de hipóteses que utiliza o teste t de Student

para testar a hipótese nula de que β_1 é igual a zero. Se o teste verificar que a hipótese nula é verdadeira, supõe-se que β_1 é zero e que não existe um intervalo de confiança. Caso contrário, se a hipótese nula for rejeitada, conclui-se que β_1 é diferente de zero e pode-se contruir um intervalo de confiança para x_0 com um coeficiente de confiança ligeiramente inferior a $100(1-\alpha)\%$ (GRAYBILL, 1976).

Em testes com simulação computacional Thonnard (2006) e Cordeiro (2015) geraram intervalos de confiança utilizando o nível de significância $\alpha = 0,05$ e verificaram que a diferença dos coeficientes de confiança do intervalos simulados em relação a $100(1-\alpha)\%$ é muito pequena, sendo que a maior diferença encontrada por ambos foi inferior a $0,5\%$.

2.1.3 Processo de estimação inversa simples (pontual e intervalar)

Neste tópico é apresentado o processo de construção dos estimadores pontual e intervalar proposto por Graybill (1976) para a estimação do valor desconhecido x_0 da variável independente X que é muito difundido entre pesquisadores.

O modelo de calibração linear pode ser formalmente definido por (GRAYBILL, 1976):

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{2.6}$$

$$Y_{0i} = \beta_0 + \beta_1 x_0 + \varepsilon_{0i}, \qquad i = n+1, ..., n+k, \tag{2.7}$$

em que os $\varepsilon_1, ...\varepsilon_n$ e $\varepsilon_{0n+1}, ...\varepsilon_{0n+k}$ são variáveis normais independentes e identicamente distribuídas com média zero e variância constante σ^2. Os valores $x_1, x_2, ...x_n$ são considerados constantes conhecidas, enquanto que β_0, β_1 e σ^2 são parâmetros

desconhecidos, sendo que o interesse principal é estimar o valor x_0 desconhecido. A equação (2.6) diz respeito à primeira etapa do processo de calibração, onde ocorre a estimação dos parâmetros. A equação (2.7) refere-se à segunda etapa, também chamada de calibração propriamente dita, onde estima-se o valor desconhecido x_0.

Segundo Cordeiro (2015), um facilitador computacional da estimação é considerar o modelo de regressão linear simples centrado, onde a variável regressora x é redefinida como o desvio de sua própria média, $x_i - \bar{x}$. Com essa reparametrização o modelo é escrito da seguinte forma:

$$Y_i = \alpha_0 + \alpha_1(x_i - \bar{x}) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \tag{2.8}$$

em que $\alpha_1 = \beta_1$ e $\alpha_0 - \alpha_1\bar{x} = \beta_0$.

Para obter um estimador de x_0 observa-se $k \geq 1$ valores de Y, denotado por $\mathbf{Y_0}$ correspondentes a um único valor x_0 desconhecido (GRAYBILL, 1976). Dessa forma, tem-se uma amostra de tamanho $n + k$, em que x_0 é desconhecido e os valores x_i são distintos. Os k valores de Y, $\mathbf{Y_0} = \{Y_{n+1}, Y_{n+2}, ..., Y_{n+k}\}$ têm distribuição normal com média $\alpha_0 + \alpha_1(x_i - \bar{x})$ e variância σ^2.

A função de verosimilhança do modelo, com a suposição de normalidade é dada por:

$$L(\alpha_0, \alpha_1, \sigma^2, x_0 : (Y_1, x_1), ..., (Y_n, x_n) : Y_{n+1}, ..., Y_{n+k}) = \left(\frac{1}{(2\pi\sigma^2)^{\frac{(n+k)}{2}}} \right) \times$$

$$\exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (Y_i - \alpha_0 - \alpha_1(x_i - \bar{x}))^2 + \sum_{i=n+1}^{n+k} (Y_i - \alpha_0 - \alpha_1(x_0 - \bar{x}))^2 \right] \right\}. \tag{2.9}$$

Obtém-se os estimadores de máxima verossimilhança dos parâmetros α_0, α_1

e σ^2 e da variável independente x_0 igualando a zero as derivadas parciais do *log* da função de verossimilhança dada em (2.9), em relação cada um desses termos e resolvendo o conjunto de equações resultantes. Os estimadores dos parâmetros α_0 e α_1 obtidos a partir dos n primeiros valores $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ são expressos por:

$$\hat{\alpha}_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{2.10}$$

$$\hat{\alpha}_0 = \bar{Y}, \tag{2.11}$$

em que $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ e $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

O estimador da variável independente x_0 obtido usando as $n + k$ observações e baseado nas estimativas de máxima verossimilhança de α_0 e α_1 é expresso por:

$$\hat{x}_0 = \bar{x} + \frac{\bar{Y}_0 - \hat{\alpha}_0}{\hat{\alpha}_1}, \tag{2.12}$$

em que $\bar{Y}_0 = \frac{\sum_{i=n+1}^{n+k} Y_i}{k}$.

O estimador não viesado da variância σ^2 obtido a partir da função de verossimihança é expresso por (GRAYBILL, 1976):

$$\hat{\sigma}^2 = \frac{1}{n+k-3}\left(\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1(x_i - \bar{x}))^2 + \sum_{i=n+1}^{n+k}\left(Y_i - \bar{Y}_0\right)\right). \tag{2.13}$$

Esses estimadores possuem várias propriedades, sendo as principais descritas no seguinte teorema (GRAYBILL, 1976).

Teorema 1: Considerando o modelo expresso pela equação (2.8) e os estimadores de $\alpha_0, \alpha_1, \sigma^2$ e x_0, tem-se que:

1 - $\hat{\alpha}_0$ e $\hat{\alpha}_1$ são independentes;

2 - $\hat{\alpha}_0 \sim N\left(\alpha_0, \frac{\sigma^2}{n}\right)$ e $\hat{\alpha}_1 \sim N\left(\alpha_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$;

3 - $U = (n+k-3)\frac{\hat{\sigma}^2}{\sigma^2}$ tem distribuição qui-quadrado com $(n+k-3)$ graus de liberdade, χ^2_{n+k-3};

4 - $\hat{\sigma}^2$ é independente de α_0, α_1 e $\hat{x}_0$.

Segundo Graybill (1976), $\hat{x}_0$ possui uma distribuição de probabilidade. Porém, ela é muito difícil de ser explicitada e, em contrapartida, ela não é necessária para se obter um intervalo de confiança para x_0.

Buscando construir um intervalo de confiança para x_0 dado Y_0, evidentemente pode-se observar que não há um intervalo útil se α_1 é zero, o que faz com que a linha de regressão linear simples seja horizontal. Sendo α_1 diferente de zero e usando as informações do **Teorema 1**, pode-se construir intervalo de confiança para x_0.

Primeiramente, calcula-se a variância de $\hat{\varepsilon}_0 = \bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0 - \bar{x})$, dada por:

$$Var\left[\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0-\bar{x})\right] = \sigma^2\left(\frac{1}{k} + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) = \sigma^2 A. \quad (2.14)$$

Com essa variância pode-se construir um variável Z normal padrão, obtende-

se o seguinte resultado:

$$Z = \frac{\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0 - \bar{x})}{\sqrt{Var\left[\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0 - \bar{x})\right]}} = \frac{\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0 - \bar{x})}{\sqrt{\sigma^2 A}} \sim N(0,1). \tag{2.15}$$

Baseado no item 3 do **Teorema 1**, tem-se que:

$$\begin{aligned} U &= (n+k-3)\frac{\hat{\sigma}^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1(x_i - \bar{x}))^2 + \sum_{i=n+1}^{n+k}\left(Y_i - \bar{Y}_0\right)}{\sigma^2} \sim \chi^2_{(n+k-3)}. \end{aligned} \tag{2.16}$$

Antes de prosseguir, torna-se necessário enunciar um importante teorema que relaciona as distribuições de probabilidades normal padrão e qui-quadrado.

Teorema 2: Considere Y e V duas variáveis aleatórias independentes tal que $Y \sim N(0,1)$ e $V \sim \chi^2_k$. Sendo X uma variável aleatória de tal forma que

$$X = \frac{Y}{\sqrt{V/k}},$$

então a variável aleatória X tem distribuição t de Student com k graus de liberdade.

Pelo **Teorema 2**, como $Z \sim N(0,1)$ e $U \sim \chi^2_{(n+k-3)}$ são independentes, segue que:

$$T = \frac{Z}{\sqrt{\frac{U}{n+k-3}}} \sim t_{(n+k-3)} \tag{2.17}$$

e

$$P\left(-t_{(\alpha/2;n+k-3)} \leq T \leq t_{(\alpha/2;n+k-3)}\right) = 1 - \alpha, \tag{2.18}$$

sendo $1-\alpha$ o nível de confiança do intervalo.

De forma equivalente podemos escrever (2.18) da seguinte forma:

$$T^2 \leq t^2_{(\alpha/2;n+k-3)}, \tag{2.19}$$

e, assim, segue que:

$$\frac{\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0-\bar{x})\right)^2}{\hat{\sigma}^2\left(\frac{1}{k}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)} \leq t^2_{(\alpha/2;n+k-3)}, \tag{2.20}$$

ou seja,

$$\left(\bar{Y}_0 - \hat{\alpha}_0 - \hat{\alpha}_1(x_0-\bar{x})\right)^2 - \hat{\sigma}^2\left(\frac{1}{k}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right) t^2_{(\alpha/2;n+k-3)} \leq 0. \tag{2.21}$$

Como o interesse está na variável x_0, expande-se a primeira parte da inequação (2.21), sendo possível identificar x_0 como a única incógnita da inequação resultante, como segue:

$$\left(\hat{\alpha}_1^2 - \frac{\hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)(x_0-\bar{x})^2 - 2\hat{\alpha}_1\left(\bar{y}_0-\hat{\alpha}_0\right)(x_0-\bar{x}) +$$

$$+\left(\left(\bar{y}_0-\hat{\alpha}_0\right)^2 - \hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}\left(\frac{1}{k}+\frac{1}{n}\right)\right) \leq 0. \tag{2.22}$$

A expressão 2.22 trata-se de uma inequação quadrática da seguinte forma $q(x_0-\bar{x}) = a(x_0-\bar{x})^2 + 2b(x_0-\bar{x}) + c \leq 0$, onde a, b e c são identificados

na inequação 2.22. Se os valores de x_0 que satisfazem esta desigualdade formam um intervalo, então tem-se um intervalo de $100\,(1-\alpha)\,\%$ de confiança para x_0 (GRAYBILL, 1976).

O discriminante dessa função quadrática é expresso por $(2b)^2 - 4ac = 4(b^2 - ac)$. De acordo com Graybill (1976), se $b^2 - ac$ é negativo, então o discriminante da função quadrática $q(x_0 - \bar{x})$ é negativo, portanto essa função não pode ser igual a zero (Figura 1**a** e 1**c**). Nesse caso, há um intervalo de confiança infinito e sem utilidade para x_0, $-\infty < x_0 < \infty$ (Figura 1**a**) ou não existe intervalo de confiança para x_0 (Figura 1**c**). Sendo assim, é necessário analisar desigualdade (2.22) nos casos em que o discriminante é positivo.

Se $b^2 - ac$ da função quadrática $q(x_0 - \bar{x})$ é positivo, então o discriminante é positivo, sendo assim existem duas possibilidades para o intervalo de confiança de x_0 (Figura 1**b** e 1**d**). Primeiro, se o coeficiente $a < 0$ (Figura 1**b**), então os valores de x_0 formam dois intervalos infinitos, o que também não é útil. Por fim, no caso em que o discriminante é positivo e o coeficiente $a > 0$, os valores valores de x_0 formam um intervalo de confiança finito e útil para a estimação de x_0 (Figura 1**d**).

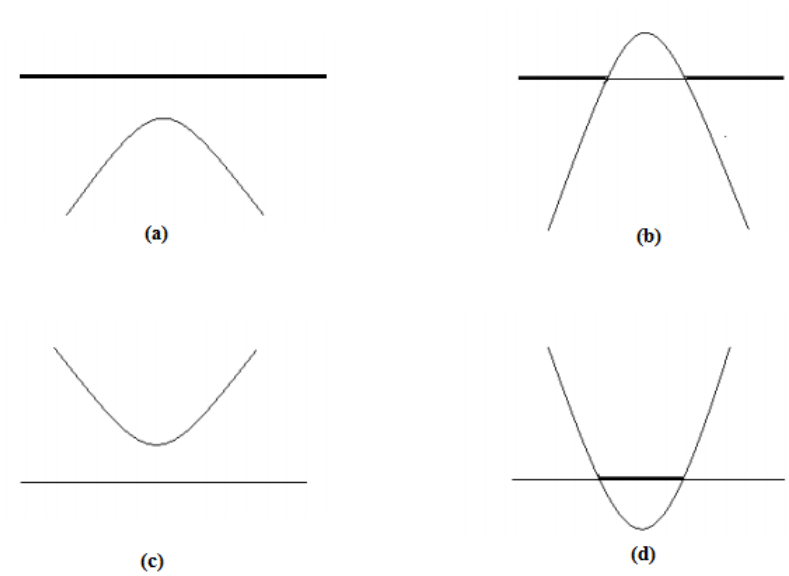

Figura 1 Formas gráficas que uma função quadrática pode assumir dependendo do sinal do discriminante.

Portanto, existe intervalo de confiança para x_0 resultante de $q(x_0 - \bar{x})^2 \leq 0$, se somente se, $a > 0$ e $b^2 - ac > 0$. Ainda, expandindo $b^2 - ac$, tem-se :

$$b^2 - ac = \hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)} \left(\left(\frac{1}{k} + \frac{1}{n} \right) a + \frac{(\bar{y}_0 - \hat{\alpha}_0)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right). \tag{2.23}$$

A partir da expressão (2.23) pode-se notar que se $a \geq 0$, então $b^2 - ac \geq 0$, ou seja, há um intervalo para x_0 se, e somente se, $\hat{\alpha}_1^2 - \frac{\hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \geq 0$. Simplificando a expressão, tem-se que $\frac{\hat{\alpha}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\hat{\sigma}^2} \geq t^2_{(\alpha/2;n+k-3)} = F_{(1;\alpha/2;n+k-3)}$, que caracteriza um teste, de tamanho α, para as hipóteses $H_0 : \alpha_1 = 0 \quad versus \quad H_1 :$

$\alpha_1 \neq 0$.

Portanto, o procedimento para obter o valor de x_0, dado $\bar{Y}_0$ é (GRAYBILL, 1976):

1. Obtém-se a estimativa de x_0 por meio do estimador de máxima verossimilhança, dado por: $\hat{x}_0 = \bar{x} + \frac{\bar{Y}_0 - \hat{\alpha}_0}{\hat{\alpha}_1}$, de onde pode-se obter estimativa pontual de x_0.

Posteriormente, pode-se obter um estimador intervalar para x_0 da seguinte maneira:

1. Realiza-se o teste $H_0 : \alpha_1 = 0 \quad versus \quad H_1 : \alpha_1 \neq 0$, onde rejeita-se H_0 se, e somente se $\frac{\hat{\alpha}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\hat{\sigma}^2} \geq t^2_{(\alpha/2;n+k-3)} = F_{(1;\alpha/2;n+k-3)}$.

2. Se a hipótese H_0 não é rejeitada, assume-se que o modelo é $y_i = \alpha_0 + \in_i$. Portanto, não existe intervalo de confiança para x_0.

3. Se a hipótese H_0 é rejeitada, o limite inferior (LI) e o limite superior (LS) do intervalo de $100\,(1-\alpha)\,\%$ de confiança para x_0 são dados por:

$$LI = \bar{x} + \frac{\hat{\alpha}\,(\bar{y}_0 - \bar{y})}{a} - \frac{\hat{\sigma} t_{(\alpha/2;n+k-3)}}{a} \sqrt{\left(\frac{1}{k} + \frac{1}{n}\right) a + \frac{(\bar{y}_0 - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}}; \quad (2.24)$$

$$LS = \bar{x} + \frac{\hat{\alpha}\,(\bar{y}_0 - \bar{y})}{a} + \frac{\hat{\sigma} t_{(\alpha/2;n+k-3)}}{a} \sqrt{\left(\frac{1}{k} + \frac{1}{n}\right) a + \frac{(\bar{y}_0 - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}}, \quad (2.25)$$

em que $a = \hat{\alpha}_1^2 - \frac{\hat{\sigma}^2 t^2_{(\alpha/2;n+k-3)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Graybill (1976) afirma que este não é um intervalo de $100\,(1-\alpha)\,\%$ de confiança para x_0, mas que tem o coeficiente de confiança ligeiramente menor que $100\,(1-\alpha)\,\%$.

2.1.4 Processo de estimação inversa quadrática (pontual e intervalar)

Na análise de regressão existem situações em que a relação entre a variável dependente Y e a variável independente X não é adequadamente modelada por uma linha reta. Em fenômenos onde isso acontece é necessário ajustar um modelo polinomial, como ilustrado na Figura 2.

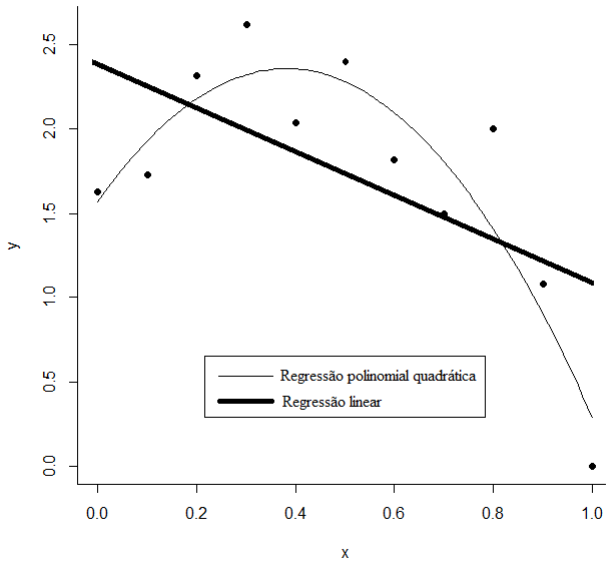

Figura 2 Situação ilustrativa onde a relação entre a variável dependente Y e a variável independente X não é adequadamente modelada por uma linha reta.

Muitas vezes, o modelo polinomial usado é o de grau dois, denominado modelo polinomial quadrático. Na regressão polinomial quadrática assume-se que a relação entre a variável independente X e a variável dependente Y é expressa por:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \varepsilon_i, \quad \text{com} \quad \varepsilon_i \sim N\left(0, \sigma^2\right), \tag{2.26}$$

em que $i = 1, 2, ..., n$.

Na forma matricial o modelo (2.26) apresenta-se da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim \boldsymbol{NM}\left(\mathbf{0}, \mathbf{I}\boldsymbol{\sigma}^{\mathbf{2}}\right), \tag{2.27}$$

em que:

$\mathbf{Y}$ é um vetor $n \times 1$ cujas componentes correspondem às n observações;

$\mathbf{X}$ é uma matriz de dimensão $n \times 3$ denominada matriz de incidência;

$\boldsymbol{\beta}$ é um vetor 3×1 cujos elementos são os parâmetros da regressão;

$\boldsymbol{\varepsilon}$ é um vetor de dimensão $n \times 1$ cujas componentes são os erros.

Expandindo em termos matriciais tem-se:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{2.28}$$

Os parâmetros β_0, β_1 e β_2 da equação (2.26) podem ser estimados usando o método de mínimos quadrados ou o método de máxima verossimilhança. Dessa

forma, tem-se que o estimador do vetor de parâmetros β_0, β_1 e β_2 é dado por:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}. \tag{2.29}$$

A variância desse estimador é:

$$Var\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2\left(\mathbf{X^T X}\right)^{-1}. \tag{2.30}$$

Baseado nisso, alguns autores como Kirkup e Mulholland (2004) e Oliveira e Aguiar (2009) discutiram o problema de calibração (ou regressão inversa) para dados univariados na regressão polinomial quadrática. Segundo esses autores, uma vez que tenha sido realizado o ajuste do modelo, com o objetivo de obter o estimador para x_0 , observa-se $k \geq 1$ valores de Y para um valor x_0 desconhecido. Dado a média $\bar{y}_0$ dos valores observados de Y para um valor x_0, o estimador pontual clássico de mínimos quadrados de x_0 é obtido pela equação (2.31):

$$\hat{x}_0 = \frac{\hat{\beta}_1 \pm \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}{2\hat{\beta}_2}, \tag{2.31}$$

com a raiz sendo positiva quando a função y_i é crescente e negativa quando a função y_i é decrescente.

A variância de $\hat{x}_0$ pode ser calculada a partir da expansão em série de Taylor da equação (2.31) em torno do ponto $P(\beta_0, \beta_1, \beta_2, E(\bar{y}_0))$. Considerando a função $\hat{x}_0 = f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \bar{y}_0)$, calcula-se a variância, desprezam-se os termos de ordem superior e a correlação entre $\hat{y}$ e os coeficientes estimados, obtendo-se

assim a equação (2.32) (OLIVEIRA; AGUIAR, 2009):

$$\begin{aligned} Var(\hat{x}_0) &= \left(\frac{\partial \hat{x}_0}{\partial \bar{y}_0}\right)^2 Var(\bar{y}_0) + \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0}\right)^2 Var(\hat{\beta}_0) + \\ &+ \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\right)^2 Var(\hat{\beta}_1) + \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Var(\hat{\beta}_2) + 2\hat{\beta}_0 \left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\right)^2 Cov(\hat{\beta}_0, \hat{\beta}_1) + \\ &+ 2\left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Cov(\hat{\beta}_0, \hat{\beta}_2) + 2\left(\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1}\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2}\right)^2 Cov(\hat{\beta}_1, \hat{\beta}_2), \end{aligned} \tag{2.32}$$

em que as variâncias e as covariâncias dos parâmetros são obtidas da equação (2.30) e a variância de uma resposta y_0 é estimada pela variância da regressão. Se $\bar{y}_0$ é a média de k observações dependentes, tem-se:

$$\hat{\sigma}^2_{\bar{y}_0} = \frac{\hat{\sigma}^2}{k}, \tag{2.33}$$

em que

$$\hat{\sigma}^2 = \frac{\mathbf{Y^T Y} - \hat{\boldsymbol{\beta}}^{\mathbf{T}} \mathbf{X^T Y}}{n-3}. \tag{2.34}$$

Os termos correspondentes às derivadas parciais de $\hat{x}_0$ em relação a cada um dos parâmetros da equação (2.32) são expressos por:

$$\frac{\partial \hat{x}_0}{\partial \hat{\beta}_0} = \frac{-1}{\sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}; \tag{2.35}$$

$$\frac{\partial \hat{x}_0}{\partial \hat{\beta}_1} = \frac{-1 + \hat{\beta}_1 \Big/ \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}{2\hat{\beta}_2}; \tag{2.36}$$

$$\frac{\partial \hat{x}_0}{\partial \hat{\beta}_2} = \frac{\hat{\beta}_1 - \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}{2\hat{\beta}_2^2} - \frac{\left(\hat{\beta}_0 - \bar{y}_0\right)}{\hat{\beta}_2\sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}; \tag{2.37}$$

$$\frac{\partial \hat{x}_0}{\partial \bar{y}_0} = \frac{1}{\sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_2\left(\hat{\beta}_0 - \bar{y}_0\right)}}. \tag{2.38}$$

Para obter um intervalo de $100(1-\alpha)\%$ de confiança para uma nova observação x_0, o erro padrão é multiplicado pelo quantil $(1-\alpha/2)$ da distribuição t de Student bicaudal para os graus de liberdade da calibração $(n-3)$ (KIRKUP; MULHOLLAND, 2004). Dessa forma, tem-se os seguintes limites para o intervalo:

$$LI = \hat{x}_0 - t\left(1-\alpha/2; n-3\right)\sqrt{Var\left(\hat{x}_0\right)} \tag{2.39}$$

e

$$LS = \hat{x}_0 + t\left(1-\alpha/2; n-3\right)\sqrt{Var\left(\hat{x}_0\right)}. \tag{2.40}$$

2.2 Estatística Espacial

Segundo Assunção (2001), a estatística espacial é uma área da estatística que estuda metodologias para coleta, descrição, visualização e análise de dados, associados a posições geográficas, que possam ser modelados como processos estocásticos.

Pode-se então, definir a estatística espacial como um conjunto de métodos e modelos que usam explicitamente a referência espacial (coordenadas), associada a cada observação, para descrever os padrões existentes nessas observações e

estabelecer, preferencialmente, de forma quantitativa, os relacionamentos entre as diversas variáveis geográficas envolvidas na análise.

Conforme Bailey et al. (1995), Cressie (1991) e Câmara et al. (2004), a análise estatística espacial considera, entre outros, três tipos básicos de observações geográficas. São eles: *configurações pontuais*, *superfícies contínuas* e *áreas com contagens e taxas agregadas (lattice)*.

- *Configurações pontuais*: são fenômenos expressos por meio de ocorrências identificadas como pontos localizados no espaço. Exemplo desses tipos de fenômenos são a localização de casos de uma doença em uma região geográfica e a localização de indivíduos de uma determinada espécie.

- *Superfícies contínuas (Geoestatística)*: são fenômenos que se distribuem continuamente em uma região. Usualmente, esse tipo de observação é resultante de levantamento de fenômenos naturais. Exemplos desse tipo de fenômeno são as medidas de concentração de um elemento químico no solo, medidas de precipitação e de temperatura.

- *Áreas com contagens e taxas agregadas (lattice)*: são fenômenos associados aos dados de levantamentos populacionais, como censos, e que originariamente referem-se a indivíduos localizados em áreas específicas no espaço. Normalmente, esses dados são agregados em unidades de análises, usualmente delimitadas por polígonos fechados, tais como setores censitários, municípios e microrregiões. Um exemplo desse tipo de fenômeno pode ser a contagem do número de pessoas portadoras do vírus HIV por município.

Para Guimarães (2004), a incorporação da informação espacial nos dados

pode complementar a análise clássica destes, uma vez que, na análise clássica em geral, as realizações das variáveis aleatórias são independentes. Logo, observações vizinhas não sofrem influência umas das outras, o que não acontece na análise espacial, pois esta considera, em suas estimações, as correlações existentes entre as observações.

Em cada um do tipos de observações geográficas existe o interesse em se obter alguma estatística que quantifique a medida de dependência espacial do conjunto de dados. Para cada um dos três tipos básicos de observações geográficas citados existem diferentes métodos estatísticos para analisá-los considerando a dependência espacial. Neste trabalho, o enfoque está na análise de dados de área.

2.3 Análise de dados de área

Segundo Assunção (2001), análise de dados de área associa o mapa geográfico de uma região R a uma base de dados. Esse mapa é dividido em sub-regiões A_i, com $i = 1, 2, ..., n$, de forma que $\bigcup_{i=1}^{n} A_i = R$ e $A_i \cap A_j = \emptyset$, se $i \neq j$, qualquer que seja a forma de A_i e de R. Essas divisões geográficas que resultam nas áreas são, em geral, de caráter político e geofísico, geralmente caracterizadas por bairros, municípios, estados ou setores censitários. Em dados de área, não se conhece a localização exata do evento. Dessa forma a análise é feita no valor agregado a cada área do estudo procurando-se identificar padrões espaciais de distribuição nos valores observados.

A forma inicial de apresentação de dados de áreas é o uso de mapas coloridos do fenômeno de interesse na região. Se houver padrão espacial, espera-se encontrar padrões ou cores parecidas geograficamente próximas, como mostrado na Figura 3.

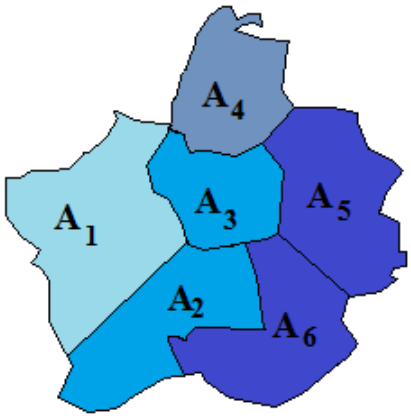

Figura 3 Mapa para ilustrar a apresentação de um mapa da área R.

As técnicas de análise de dados de área foram desenvolvidas para tentar identificar regiões onde a distribuição dos valores possa apresentar um padrão específico associado a sua localização espacial. A informação que se busca é o quanto o valor de uma variável em uma determinada área assemelha-se com os valores da mesma variável localizada em sua vizinhança próxima e o quanto é similar ou dissimilar aos valores de sua vizinhança distante (SILVA et al., 2011).

2.4 Análise de autocorrelação espacial

Para Cliff & Ord (1981), a autocorrelação espacial pode ser entendida como a tendência a que o valor de uma variável associada a uma determinada localização assemelha-se mais aos valores de suas observações vizinhas do que ao restante das localizações do conjunto amostral.

A presença de autocorrelação espacial evidencia a existência de dependência espacial, que é um conceito de extrema importância para a compreensão dos fenômenos espaciais. Nesse contexto, para contabilizar a presença de dependência

espacial, segundo Waller & Gotway (2004), na análise de dados de área, o grau de dependência espacial ou similaridade é avaliado por meio da autocorrelação espacial. Essa avaliação pode ser feita utilizando várias estatísicas como: índice de Moran (Bailey & Gatrell, 1995), estatísticas Gi e Gi* (Getis & Ord, 1992), índice de Geary (Rosenberg et al., 1999), dentre outras. Essas técnicas possibilitam estimar o quanto o valor observado de uma variável aleatória em uma determinada localização é dependente dos valores desta mesma variável nas localizações vizinhas. Destaca-se como metodologia muito difundida em estudos envolvendo análise de dados de área o índice de Moran. A aplicação desse índice depende da definição de uma matriz de vizinhança ou matriz de proximidade espacial. A matriz de vizinhança espacial e o índice de Moran estão apresentados, respectivamente, nas duas próximas seções.

2.5 Matriz de vizinhança espacial

A matriz de vizinhança espacial é um dos componentes presentes na estimação da variabilidade espacial de dados de área. Também conhecida como matriz de proximidade espacial, ela é definida de várias formas na literatura. Segundo Assunção (2001), Câmara et al. (2004) e Collins, Babyak e Moloney (2006), dado um conjunto de n áreas $\{A_1, A_2, ..., A_n\}$ constrói-se a matriz de vizinhança espacial W $(n \times n)$, onde cada elemento w_{ij} representa uma medida de proximidade entre A_i e A_j. Waller & Gotway (2004) apresentam três diferentes critérios que podem ser usados na construção da matriz W. São eles:

a) $w_{ij} = 1$, se o centróide de A_i está a uma determinada distância do centróide de A_j, caso contrário $w_{ij} = 0$;

b) $w_{ij} = 1$, se A_i compartilha um lado comum com A_j, caso contrário $w_{ij} = 0$;

c) $w_{ij} = l_{ij}/l_i$, em que l_{ij} é o comprimento da fronteira entre A_i e A_j e l_i

é o perímetro de A_i.

Uma das formas comumente empregadas para a construção da matriz W consiste em adotar o citério (b). Nesse critério, a diagonal principal da matriz W possui todos os elementos iguais a zero, por definição. O elemento w_{ij} da matriz assume o valor $w_{ij} = 1$, caso os polígonos i e j sejam vizinhos, ou seja, façam fronteira e $w_{ij} = 0$, caso i e j não sejam vizinhos. A seguir será dado um exemplo de matriz de vizinhança espacial. Seja a região apresentada na Figura 4.

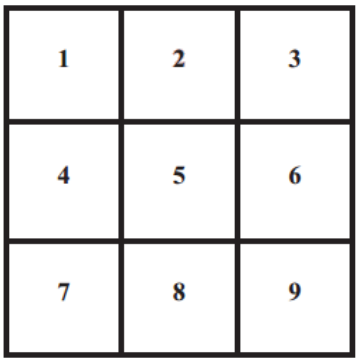

Figura 4 Região hipotética utilizada como exemplo para a construção de uma matriz de vizinhança espacial.

Usando a parcela 4 como referência, tem-se que as parcelas 1, 5 e 7 seriam consideradas vizinhas e na matriz de vizinhança W os elementos w_{41}, w_{45} e w_{47} receberiam o valor 1. Já os demais elementos w_{4i} $(i = 2, 3, 4, 6, 8, 9)$ receberiam o valor 0, pois não são vizinhas da parcela 4. Agora, tomando como referência a parcela 5, seriam vizinhas de 5 as parcelas 2, 4, 6 e 8. Portanto, reberiam o valor 1 os elementos w_{52}, w_{54}, w_{56} e w_{58} e 0 os elementos w_{5i} $(i = 1, 3, 5, 7, 9)$. Assim, a matriz de vizinhança espacial W dessa região hipotética, utilizando o critério (b), é dada por:

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

A matriz W, com elementos 0 e 1, é conhecida como matriz de vizinhança não normalizada. Alguns autores tais como Cressie (1991), Druck et al. (2004) e Waller & Gotway (2004), para facilitar o processamento computacional de indicadores como o índice de Moran, recomendam a normalização das linhas da matriz W dividindo cada elemento da matriz pela soma total da linha e assim, obtém-se a soma dos pesos de cada linha igual a 1. Essa nova matriz W*, designada de matriz normalizada, possui todas as linhas com a soma igual a 1. Por sua vez, a matriz W original é simétrica, o que não acontece para a matriz W* (YWATA; ALBUQUERQUE, 2011).

2.6 Índice de Moran

Moran (1948) propôs uma medida para avaliar o grau de autocorrelação de variáveis espacialmente referenciadas. Essa medida, conhecida como coeficiente ou índice de Moran, pode ser calculada comparando-se os pares adjacentes das observações com o seu desvio em relação a média das observações, utilizando, segundo Waller e Gotway (2004), a seguinte fórmula:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}, \tag{2.41}$$

em que:

n é o número de áreas ou de observações;

Y_i é a variável aleatória na área i;

Y_j é a variável aleatória na área j;

$\bar{Y}$ a média amostral da variável aleatória em toda região, dada por $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

w_{ij} são os elementos da matriz de proximidade espacial.

Os valores do coeficiente de Moran podem ser positivos, assim como, negativos, podendo assumir qualquer valor no conjunto dos números reais (WALLER; GOTWAY, 2004). Porém, o mais comum são valores dentro do intervalo $[-1, 1]$. Segundo Plant (2012), quando existe homogeneidade entre as parcelas próximas, o índice I tende a ser positivo, enquanto que se as parcelas próximas forem dissimilares, o coeficiente tende a ser negativo. Um valor próximo de zero indica ausência de autocorrelação espacial.

Calculado o índice de Moran, é necessário estabelecer a sua validade submetendo tal valor a um teste de significância, ou seja, é preciso verificar se os valores encontrados representam correlação espacial significativa ou não. Para avaliar a significância do índice é preciso associá-lo a uma distribuição amostral. Segundo Cliff e Ord (1981), o mais comum é associar esse índice à distribuição normal. Sob a suposição de normalidade, o valor esperado da estatística de Moran na ausência de autocorrelação espacial é dado por:

$$E\left[I\right] = -\frac{1}{n-1}, \tag{2.42}$$

e a variância é expressa por:

$$Var\left[I\right] = \frac{n^2 S_1 - nS_2 + 3{S_0}^2}{\left(n-1\right)\left(n+1\right){S_0}^2} - \left(\frac{1}{n-1}\right)^2, \tag{2.43}$$

em que: $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2$, $S_2 = \sum_{i=1}^{n}\left(w_{i+} + w_{+j}\right)$, sendo $w_{i+} = \sum_{j=1}^{n} w_{ij}$ e $w_{+j} = \sum_{i=1}^{n} w_{ij}$.

A significância da estatística de Moran pode ser avaliada por meio da estatística teste expressa por:

$$z = \frac{I - E\left[I\right]}{\sqrt{Var\left[I\right]}}. \tag{2.44}$$

O valor de z obtido na equação 2.44, corresponde a um quantil da distribuição normal padronizada, que está associado a um valor-p. O índice de Moran será considerado significativo se o valor-p for inferior ao valor nominal de significância pré-definido.

Segundo Anselin (2005), Waller e Gotway (2004) e Ywata & Alburquerque (2011), a presença da dependência espacial nos resíduos pode ser analisada calculando a estatística de Moran dos resíduos no modelo Gauss-Markov ordinário conforme a equação seguinte:

$$I_{res} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\left[\frac{\mathbf{u^T W u}}{\mathbf{u^T u}}\right], \tag{2.45}$$

em que $\mathbf{u}$ representa o vetor de resíduos obtidos no ajuste do modelo de regressão linear simples, $\mathbf{W}$ é a matriz de vizinhança, w_{ij} são os elementos da matriz de vizinhança e n é o número de áreas da região em estudo.

O índice de Moran dos resíduos segue uma distribuição normal assintótica com média e variância dadas pelas equações (2.46) e (2.47), respectivamente

(CLIFF & ORD, 1981):

$$E\left[I_{res}\right] = \frac{n}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}}\left[\frac{tr\left(\boldsymbol{MW}\right)}{n-p}\right] \tag{2.46}$$

e

$$Var\left[I_{res}\right] = \left(\frac{n}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}}\right)^{2}\left[\frac{tr\left(\boldsymbol{MW}(\boldsymbol{MW})^{T}\right) + tr\left((\boldsymbol{MW})^{2}\right) + (tr(\boldsymbol{MW}))^{2}}{(n-p)(n-p+2)} - \left[E\left(I_{res}\right)\right]^{2}\right], \tag{2.47}$$

em que $\mathbf{M} = \mathbf{I_n} - \mathbf{X}\left(\mathbf{X^T X}\right)^{-\mathbf{1}}\mathbf{X^T}$ é matriz de projeção, com $\mathbf{I_n}$ sendo a matriz identidade.

A partir da estatística I_{res}, pode-se construir um teste para a hipótese nula de presença de independência espacial. Por sua vez, a rejeição da hipótese nula implica evidências da existência de dependência espacial no modelo. A estatística de Moran é assintoticamente distribuída e é representada pela seguinte equação:

$$z = \frac{I_{res} - E\left[I_{res}\right]}{\sqrt{Var\left[I_{res}\right]}}. \tag{2.48}$$

O valor de z obtido na equação (2.48), corresponde a um quantil da distribuição normal padronizada, que está associado a um valor-p. O índice de Moran será considerado significativo se o valor-p for inferior ao valor nominal de significância pré-definido.

2.7 Modelos autorregressivos

Dado um conjunto de regiões geográficas, as observações coletadas em regiões mais próximas geralmente tendem a ter características similares, em com-

paração com regiões distantes. Do ponto de vista estatístico, este fenômeno é atribuído ao fato da autocorrelação entre as observações recolhidas em regiões mais próximas ser mais elevada do que as regiões que estão distantes. Assim, este processo espacial observado ao longo de uma rede ou de um conjunto de regiões, geralmente é modelado usando modelos autorregressivos (KYUNG; GHOSH, 2010).

Segundo Cressie (1991), os modelos autorregressivos assumem que a resposta Y de cada lugar é uma função não só da variável explicativa nesse local, mas também dos valores das respostas dos vizinhos, isto é, a estrutura autorregressiva dos modelos requer uma definição de dados de vizinhança.

Os dois modelos mais comumente utilizados na análise de regressão de dados com dependência espacial, são: *Simultaneous Autoregressive Model* (SAR) e *Spatial Error Model* (SEM), esse segundo também conhecido como *Conditional Autoregressive Model* (CAR). Ambas abordagens relacionam os dados de um determinado local com uma combinação linear de valores vizinhos, que representam a estrutura autorregressiva (COLLINS; BABYAK; MOLONEY, 2006).

Câmara et al. (2002) salientam que, na prática, a distinção entre os dois tipos de modelos de regressão espacial com parâmetros globais é difícil, pois, apesar da diferença nas suas motivações, eles são muito próximos em termos formais. No entanto, o modelo CAR é utilizado quando o resíduo resultante de um modelo de regressão convencional possui dependência espacial, constatada pelo estatística de Moran. Outros fatores que corroboram a utilização desse modelo são a ausência de variáveis explicativas ou variáveis não-observáveis, erros de medida e heterocedasticidade.

A seguir, é apresentada uma descrição desses dois modelos e do processo de estimação desses parâmetros. Nesta tese, a calibração espacial será abordada utilizando-se o modelo autorregressivo condicional CAR.

2.8 Modelo espacial autorregressivo - SAR

O modelo SAR assume que o processo autorregressivo é dado pela variável resposta (autocorrelação espacial inerente). Este inclui o termo ρW que é um vetor de *lags* espaciais para modelar a autocorrelação espacial na variável resposta Y. O modelo pode ser representado da seguinte forma (ANSELIN, 1999):

$$Y = X\beta + \rho WY + \varepsilon, \tag{2.49}$$

em que:

Y é um vetor $n \times 1$ dos valores observados;

X é uma matriz $n \times p$ de incidência das variáveis explicativas;

β é um vetor $p \times 1$ dos parâmetros;

ρ é o coeficiente espacial autorregressivo;

W é a matriz de proximidade espacial;

ε é um vetor $n \times 1$ de erros aleatórios inerentes a cada observação que seguem um distribuição normal com média zero e variância constante, $\varepsilon \sim N\left(0, I\sigma^2\right)$.

A ideia básica no modelo (2.49) é incorporar a autocorrelação espacial como componente do modelo e é utilizado quando se deseja explicar a variável dependente Y a partir dela mesma e de outras variáveis explicativas. Porém, o fato de Y depender dos seus próprios *lags* Y espaciais pode implicar que também dependa dos *lags* espaciais do vetor de covariáveis, incorrendo no problema de reflexão (*reflexion problem*), apontado por Manski (1993). A consequência prática é que a inclusão de *lags* espaciais do vetor de covariáveis pode ocasionar uma matriz de delineamento X com alto grau de multicolinearidade (YWATA; ALBUQUERQUE, 2011).

2.8.1 Estimação dos parâmetros do modelo SAR

A estimação dos parâmetros no modelo SAR via mínimos quadrados ordinários produz estimativas inconsistentes. Sendo assim, pode-se utilizar máxima verossimilhança para essa estimação, a partir da hipótese de que o vetor de resíduos ε possui distribuição normal multivariada com média nula e matriz de covariâncias $\sigma^2 I$ (YWATA; ALBUQUERQUE, 2011). Pode-se então escrever o modelo (2.49) da seguinte forma (ANSELIN, 1999):

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon, \tag{2.50}$$

em que o vetor de variáveis observadas Y possui distribuição (condicional a X) normal multivariada, com média condicional:

$$E[Y] = (I - \rho W)^{-1} X\beta \tag{2.51}$$

e matriz de variância condicional dada por:

$$VAR(Y) = \sigma^2 (I - \rho W)^{-1} \left[(I - \rho W)^{-1} \right]^T. \tag{2.52}$$

Assim, segundo Anselin (1999), Militino, Ugarte e Reinaldos (2004) e Plant (2012), a função de log-verossimilhança $\ln L\left(\rho, \beta, \sigma^2\right)$ do modelo SAR é definida como:

$$\begin{aligned} \ln L\left(\rho, \beta, \sigma^2 | X, Y\right) &= -\frac{n}{2} \ln\left(2\pi\sigma^2\right) + \ln\left(I - \rho W\right) - \\ &- \left(\frac{1}{2\sigma^2}\right) \left(Y - X\beta - \rho WY\right)^T \left(Y - X\beta - \rho WY\right). \end{aligned}$$

Por fim, para obter os estimadores dos parâmetros do modelos SAR usando

o método de máxima verossimilhança, deriva-se a (2.53) em relação aos parâmetros e iguala-se a zero, resolvendo o sistema de equações resultantes.

Segundo Ord (1975), a estimação por máxima verossimilhança de um modelo espacial autorregressivo consiste em explorar a decomposição do Jacobiano $J = \left|\frac{\partial \varepsilon}{\partial Y}\right| = |I - \rho W|$ em termos de autovalores γ_i da matriz W. Dessa forma:

$$J = \left|\frac{\partial \varepsilon}{\partial Y}\right| = \ln |I - \rho W| = \ln \left[\prod_{i=1}^{n} (1 - \rho \gamma_i)\right], \tag{2.53}$$

em que γ_i são os autovalores da matriz W. Da equação (2.53) obtém-se um polinômio que não possui solução única. Portanto, o parâmetro ρ na equação é estimado por métodos iterativos, usando métodos de aproximação numérica como o algoritmo de Newton-Raphson ou de Gauss-Newton.

2.8.2 Estimação espacial inversa via modelo SAR

Em seu trabalho, Cordeiro (2015), baseado no método de Graybill e utilizando o modelo autorregressivo SAR, propôs um estimador pontual e intervalar para o valor x_0 desconhecido considerando a dependência espacial dos dados. Para incorporar a dependência espacial considerou-se o modelo autorregressivo SAR com uma única variável independente, dado por:

$$Y = X\beta + \rho WY + \varepsilon, \tag{2.54}$$

em que Y é um vetor de variáveis dependentes, X é uma matriz de variáveis independentes, β é um vetor de parâmetros, ρ é um coeficiente espacial autorregressivo, W é uma matriz de proximidade espacial e ε é um vetor de erros aleatórios não correlacionados que seguem uma distribuição normal com média zero e vari-

ância comum. Para a construção do estimadores espaciais inversos, os parâmetros do modelo (2.54) foram estimados por meio da máxima verossimilhança, a partir da hipótese de que o vetor de resíduos ε possui distribuição normal multivariada com média zero e matriz de covariâncias $\sigma^2 I$.

O estimador pontual da variável independente desconhecida x_0, em função de y_0, é dado pela seguinte expressão:

$$x_0 = \frac{y_0 - \rho\left(\mathbf{W_{00}Y_0} + \mathbf{W_{0c}Y_c}\right) - \beta_0}{\beta_1}, \tag{2.55}$$

em que há uma partição no vetor de variáveis dependentes Y e na matriz de proximidade espacial W, sendo que $\boldsymbol{Y_c}$ é o vetor correspondente às unidades da amostra onde os valores de Y e X são conhecidos, $\boldsymbol{Y_0}$ é o vetor correspondente às unidades onde os valores de X não são conhecidos, $\boldsymbol{W_{0c}}$ a matriz de proximidade espacial correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra e as unidades pertencentes à amostra, e $\boldsymbol{W_{00}}$ correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra. Maiores detalhes sobre o estimador (2.55) e sobre o estimador intervalar podem ser encontrados em Cordeiro (2015).

2.9 Modelo autorregressivo condicional - CAR

O modelo CAR assume que o processo autorregressivo é encontrado no termo referente ao erro, a dependência espacial ocorre a partir dos erros. Segundo Diniz Filho et al. (2003), isso é mais realista se a autocorrelação espacial não é totalmente explicada pela inclusão de variáveis explicativas (dependência espacial induzida), como exemplo, se uma variável explicativa espacialmente estruturada não foi considerada. Nesse caso, os efeitos da autocorrelação espacial são associ-

ados ao termo do erro espacial e o modelo CAR tem a seguinte especificação:

$$Y = X\beta + u, \tag{2.56}$$

em que:

Y é um vetor $n \times 1$ dos valores observados da variável dependente;

X é uma matriz $n \times p$ de incidência das variáveis explicativas;

β é um vetor $p \times 1$ dos parâmetros;

u é um vetor $n \times 1$ de erros espacialmente dependentes;

Os erros da equação (2.56) apresenta a seguinte estrutura autorregressiva:

$$u = \lambda W u + \varepsilon, \tag{2.57}$$

em que:

λ é um parâmetro espacial autorregressivo;

W é a matriz de vizinhança espacial;

u é o vetor $n \times 1$ do erro espacialmente dependente;

ε é um vetor $n \times 1$ dos erros inerentes a cada observação.

O vetor de resíduos ε possui distribuição normal multivariada, com média nula e matriz de covariâncias $\sigma^2 I$. O coeficiente escalar λ indica a intensidade da autocorrelação espacial entre os resíduos obtidos no ajuste do modelo de regressão linear simples, ou seja, esse parâmetro mensura o efeito médio dos erros dos vizinhos em relação ao resíduo da região em estudo.

Em termos de componentes individuais, o modelo CAR pode ser expresso

por:

$$\mathbf{y_i} = \beta_0 + \sum_{i=1}^{p} \mathbf{x_i}\beta_\mathbf{i} + \lambda \left(\sum_{j=1}^{n} \mathbf{w_{ij}}\mathbf{u_j} \right) + \varepsilon_\mathbf{i}. \tag{2.58}$$

Em contraposição ao modelo SAR, nos modelos CAR a autocorrelação espacial aparece nos termos do erro, de forma que a variável resposta não se apresenta como função direta dos seus vizinhos. Outra diferença e uma vantagem do modelo CAR em relação aos modelos SAR é que os coeficientes no vetor $\boldsymbol{\beta}$ podem ser estimados consistentemente via mínimos quadrados ordinários.

2.9.1 Estimação dos parâmetros do modelo CAR

Os parâmetros do modelo espacial CAR podem ser estimados pelos métodos de Mínimos Quadrados Ordinários ou Generalizados e pelo método da Máxima Verossimilhança. O processo de estimação utilizando esses métodos são descritos e discutidos a seguir.

2.9.1.1 Estimação pelo método dos mínimos quadrados ordinários

Segundo Ywata e Albuquerque (2011), os coeficientes no vetor β podem ser estimados de uma forma consistente utilizando mínimos quadrados ordinários (*ordinary least squares*-OLS), obtendo-se

$$\hat{\beta}_{ols} = \left(X^T X\right)^{-1} X^T Y. \tag{2.59}$$

Porém, a matriz de covariância das estimativas de $\hat{\beta}_{ols}$ não será $\sigma^2\left(X^T X\right)^{-1}$, devido aos erros correlacionados. Sendo assim, matriz de covariância de $\hat{\beta}_{ols}$ é dada por:

$$Var\left[\hat{\beta}_{ols}\right] = \left[X^T X\right]\left[X^T \Omega^{-1} X\right]^{-1}\left[X^T X\right], \tag{2.60}$$

em que $\Omega = Var\left[u\right] = \sigma^2(I - \lambda W)^{-1}\left[(I - \lambda W)^{-1}\right]^T = \sigma^2\Theta$. Essa matriz Ω depende do coeficiente autorregressivo λ e da variância σ^2 e, segundo Ywata e Albunquerque (2011), as estimativas desses dois parâmetros podem ser obtidas consistentemente a partir da estimação de um modelo SAR sem variáveis exógenas, via máxima verossimilhança, para os resíduos $\hat{u} = Y - X\hat{\beta}_{ols}$, dado por $\hat{u} = \lambda W\hat{u}$. Com as estimativas $\hat{\lambda}$ e $\hat{\sigma}^2$ obtém-se o estimador para a matriz de covariância de $\hat{\beta}_{ols}$, dada por:

$$\hat{Var}\left[\hat{\beta}_{ols}\right] = \left[X^T X\right]\left[X^T \hat{\Omega}^{-1} X\right]^{-1}\left[X^T X\right], \tag{2.61}$$

onde $\hat{\Omega} = \hat{\sigma}^2(I - \hat{\lambda}W)^{-1}\left[(I - \hat{\lambda}W)^{-1}\right]^T$.

2.9.1.2 Estimação pelo método dos mínimos quadrados generalizados

Modelos lineares com variáveis exógenas e resíduos correlacionados, como o modelo CAR, podem ter os parâmetros estimados utilizando-se estimadores de mínimos quadrados ordinários de forma consistente, mas não eficiente. Portanto, existem outros estimadores lineares que produzem variâncias menores (YWATA; ALBUQUERQUE, 2011).

Para o modelo CAR, o estimador linear com variância mínima é o estimador de mínimos quadrados generalizados (*generalized least squares* - GLS), dado por:

$$\hat{\beta}_{gls} = \left[X^T \Theta^{-1} X\right]^{-1}\left[X^T \Theta^{-1} Y\right]. \tag{2.62}$$

A matriz Θ não é conhecida, uma vez que ela depende do parâmetro desconhecido λ. Utiliza-se então o estimador de mínimos quadrados generalizados exequíveis (*feasible generalized least squares* - FGLS), para o qual os erros das estimativas de mínimos quadrados ordinários são utilizados para obter uma estimativa consistente da matriz de covariância Ω. Assim, o estimador pode ser expresso como:

$$\hat{\beta}_{fgls} = \left[X^T \hat{\Theta}^{-1} X\right]^{-1} \left[X^T \hat{\Theta}^{-1} Y\right], \tag{2.63}$$

em que $\hat{\Theta} = (I - \hat{\lambda}W)^{-1}\left[(I - \hat{\lambda}W)^{-1}\right]^T$, sendo $\hat{\lambda}$ estimador de máxima verossimilhança do modelo SAR sem variáveis exógenas, a partir dos resíduos $\hat{u} = Y - X\hat{\beta}_{ols}$.

Portanto, uma alternativa para a estimação dos parâmetros do modelo CAR por meio de mínimos quadrados é dada pelos seguintes passos (YWATA; ALBUQUERQUE, 2011):

i) Obter a estimativa de mínimos quadrados ordinários $\hat{\beta}_{ols} = \left(X^T X\right)^{-1} X^T Y$;

ii) Calcular os resíduos $\hat{u} = Y - X\hat{\beta}_{ols}$;

iii) Estimar os parâmetros λ e σ^2, via máxima verossimilhança, para o modelo SAR em $\hat{u}$, $\hat{u} = \lambda W\hat{u} + e$, como descrito posteriormente;

iv) Calcular a estimativa $\hat{\Theta} = (I - \hat{\lambda}W)^{-1}\left[(I - \hat{\lambda}W)^{-1}\right]^T$;

v) Obter a estimativa $\hat{\beta}_{fgls} = \left[X^T \hat{\Theta}^{-1} X\right]^{-1} \left[X^T \hat{\Theta}^{-1} Y\right]$;

vi) Obter a estimativa para a matriz de covariância de $\hat{\beta}_{fgls}$, $V\hat{a}r\left[\hat{\beta}_{fgls}\right] =$

$\left[X^T\hat{\Omega}^{-1}X\right]^{-1}$.

O processo para se obter a estimativa final para o vetor β não deve ser interrompido no passo (v), pois uma vez obtida uma estimativa $\hat{\beta}_{fgls}$, pode-se obter um novo vetor $\hat{u} = Y - X\hat{\beta}_{fgls}$. Para este novo vetor $\hat{u}$, estimam-se novamente os parâmetros λ e σ^2, repetindo-se em seguida (iv) e (v). Este processo deve ser efetuado repetidamente até que os valores no vetor $\hat{\beta}_{fgls}$ atinjam a convergência, finalizando, então, as estimações com o passo (vi).

Estimação dos parâmetros λ e σ^2 via máxima verossimilhança do modelo SAR simples

Considerando o modelo SAR $u = \lambda Wu + e$ no item iii do algoritmo descrito anteriormente, tem-se que $e = Bu$, onde:

$$B = I - \lambda W. \tag{2.64}$$

Dado que $e \sim N\left(0, \sigma^2 I\right)$, a função de log-verossimilhança para λ e σ^2 é representada pela equação

$$\ln L(\lambda, \sigma^2) = -\left(\frac{n}{2}\right)\ln\left(2\pi\sigma^2\right) + \ln|B| - \frac{u^T B^T B u}{2\sigma^2}, \tag{2.65}$$

em que $|B|$ é o Jacobiano da transformação de e para y dado por $J = \left|\frac{\partial e}{\partial y}\right| = |I - \lambda W|$.

Segundo Ord (1975), de (2.65) obtém-se os estimadores de máxima veros-

similhança

$$\hat{\sigma}^2 = n^{-1} u^T B^T B u, \tag{2.66}$$

e $\hat{\lambda}$ sendo o valor de λ que maximiza

$$\ell\left(\lambda, \hat{\sigma}^2\right) = const - \frac{n}{2} \ln\left(\hat{\sigma}^2 |B|^{-2/n}\right). \tag{2.67}$$

Nessa forma, B é um polinômio de ordem n, onde λ é estimado por métodos iterativos. A forma explícita desses polinômios é conhecido para algumas configurações regulares de pontos de amostragem, mas a avaliação de λ torna-se muito demorada (mesmo computacionalmente) quando n não é pequeno.

A dificuldade da análise do Jacobiano $|B| = |I - \lambda W|$ pode ser minimizada utilizando-se os autovalores $\gamma_1, ..., \gamma_n$ da matriz W. Decompondo $|B|$ em termos desses autovalores tem-se que:

$$|B| = \prod_{i=1}^{n} (1 - \lambda \gamma_i). \tag{2.68}$$

A vantagem de (2.68) é que os autovalores são determinados apenas uma vez, então a estimativa de λ é o valor de λ que minimiza

$$\left\{ \prod_{i=1}^{n} (1 - \lambda \gamma_i) \right\}^{-2/n} \left(u^T u - 2\lambda u^T u_L + \lambda^2 u_L^T u_L\right), \tag{2.69}$$

onde $u_L = Wu$.

A obtenção de $\hat{\lambda}$ é feita utilizando-se métodos computacionais, onde o tempo envolvido no processo depende do tamanho de n. Um procedimento computacional para tal fim é descrito no Apêndice A.

2.9.1.3 Estimação pelo método da máxima verossimilhança

Combinando-se as equações (2.56) e (2.57) obtém-se:

$$Y = X\beta + (I - \lambda W)^{-1}\varepsilon, \tag{2.70}$$

em que ε possui distribuição normal com média zero e covariância $I\sigma^2$. Dessa forma, o vetor da variável resposta Y possui distribuição normal multivariada com média condicional igual a

$$E\left[Y|X\right] = X\beta, \tag{2.71}$$

e matriz de variância condicional

$$VAR[Y|X] = \sigma^2(I - \lambda W)^{-1}\Big[(I - \lambda W)^{-1}\Big]^T = \sigma^2\Phi. \tag{2.72}$$

Partindo-se da distribuição de Y obtém-se a função de log-verossimilhança condicional da seguinte forma (LESAGE; PACE, 2009):

$$\ln L(\lambda, \sigma^2, \beta|X, Y) = -\left(\frac{n}{2}\right)\ln\left(2\pi\sigma^2\right) + \ln|I - \lambda W| - \frac{\varepsilon^T\varepsilon}{2\sigma^2}, \tag{2.73}$$

sendo $\varepsilon = (I - \lambda W)\,(Y - X\beta)$.

Maximizando-se a função (2.73) em relação aos parâmetros do modelo, encontram-se as estimativas para os coeficientes e para a variância dos resíduos.

Os estimadores dos parâmetros e variância do erro são obtidos pelo processo de derivação da equação (2.73) em relação a cada um desses parâmetros e resolução dos sistemas de equações geradas por esse procedimento.

Dessa forma, como mostra Ord (1975) e Anselin (1988), o estimador de máxima verossimilhança para β é

$$\hat{\beta} = \left(X^{*^T} X^*\right)^{-1} X^{*^T} Y^*, \tag{2.74}$$

em que X e Y podem ser escritos como $X^* = (I - \lambda W)\, X$ e $Y^* = (I - \lambda W)\, Y$.

A variância desse estimador é

$$Var\left[\hat{\beta}\right] = \sigma^2 . \left[X^T (I - \lambda W)^T (I - \lambda W)\, X\right]^{-1}. \tag{2.75}$$

Derivando-se (2.73) em relação ao parâmetro σ^2 obtém-se o estimador da variância residual dado por:

$$\hat{\sigma}^2 = \frac{[(I - \lambda W)(Y - X\beta)]^T [(I - \lambda W)(Y - X\beta)]}{n}. \tag{2.76}$$

Calculando a esperança desse estimador, tem-se que

$$\begin{aligned} E\left[\hat{\sigma}^2\right] &= E\left[\frac{[(I - \lambda W)(Y - X\beta)]^T [(I - \lambda W)(Y - X\beta)]}{n}\right] \\ &= \frac{1}{n} E\left[[(I - \lambda W)(Y - X\beta)]^T [(I - \lambda W)(Y - X\beta)]\right] \\ &= \frac{1}{n} E\left[\varepsilon^T \varepsilon\right] \\ &= \frac{1}{n} E\left[SQ\mathrm{Res}\right] \end{aligned} \tag{2.77}$$

Este resultado mostra que a estimativa (2.76) para $\hat{\sigma}^2$ é viesada. Na construção de um estimador intervalar para x_0, na regressão espacial inversa, é necessário um valor não viesado para σ^2. Dessa forma, segundo Anselin (1988) é

sugerido a utilização do seguinte estimador não viesado

$$\hat{\sigma}^2 = \frac{SQ\mathrm{Res}}{n - 2tr(S) + tr(S^T S)}, \tag{2.78}$$

sendo $S = \lambda W + H\left(I - \lambda W\right)$, com $H = X\left(X^T X\right)^{-1} X^T$.

3 MATERIAL E MÉTODOS

Neste trabalho, o problema da regressão espacial inversa ou calibração espacial é abordado propondo-se um estimador pontual e um estimador intervalar para um valor x_0 desconhecido da variável independente X. Para a construção desses estimadores considerou-se o modelo de regressão espacial com parâmetros globais, conhecido como *"modelo do erro espacial"*, também chamado de *"conditional autoregressive model* (CAR)"ou *"spatial error model* (SEM)"ou ainda.

A implementação computacional desses estimadores, bem como os cálculos e análises estatísticas, foram realizados utilizando funções desenvolvidas e funções já existentes em bibliotecas do *software* R (R CORE TEAM, 2016).

Nas próximas seções deste capítulo estão apresentadas as etapas para a construção dos estimadores propostos.

3.1 Ajuste e estimação dos parâmetros do modelo CAR

O processo de análise espacial para dados de área inclui duas etapas: análise exploratória e modelagem.

A análise exploratória permite descrever as distribuições das variáveis e a existência de dependência espacial das mesmas para que seja possível a tomada de decisão sobre qual modelo será ajustado. Portanto, foram calculadas estatísticas descritivas clássicas das variáveis buscando entender a distribuição das mesmas, e

o índice de Moran nas variáveis e nos resíduos para detectar a presença de dependência espacial nesses componentes.

Feita a análise exploratória, para modelar a dependência espacial ajustou-se o modelo CAR considerando uma variável dependente e uma única variável independente, por meio dos métodos de mínimos quadrados generalizados exequíveis ou máxima verossimihança.

A matriz de vizinhança espacial W necessária nas duas etapas de análise exploratória e modelagem foi construída adotando-se o seguinte o critério: se A_i compartilha um lado comum com A_j então $w_{ij} = 1$, caso contrário $w_{ij} = 0$. Em seguida, a matriz W foi normalizada conforme é citado na seção (2.5). Por fim, para uma notação mais leve nos cálculos feitos e expressos na obtenção dos estimadores, denota-se a matriz normalizada apenas por W.

3.2 Construção dos estimadores pontual e intervalar

As equações de estimação pontual foram desenvolvidas com base na abordagem clássica apresentada na seção 2.1.1. Para essa abordagem considerou-se a regressão de Y em função de X , considerando o modelo CAR com uma única variável independente, caracterizando uma regressão simples.

O estimador intervalar foi obtido por meio da metodologia proposta por Graybill (1976), conforme a seção 2.1.13. Para isso, assim como em Kato (2008) e Thomas-Agnan (2013) fizeram em seus trabalhos sobre predição utilizando modelos espacias, considerou-se dois tipos áreas: o primeiro formado pelas unidades observadas pertencentes à amostra utilizada na estimação do modelo e o segundo por aquelas que não pertencem à essa amostra.

3.3 Aplicação dos estimadores

A metodologia desenvolvida foi aplicada aos dados descritos por Anselin (1988) e que estão disponíveis na biblioteca "spdep"do software R (R CORE TEAM, 2016).

4 RESULTADOS METODOLÓGICOS

Neste capítulo é apresentado o processo metodológico proposto para a obtenção dos estimadores propostos.

4.1 Estimador pontual do valor desconhecido x_0 da variável independente X

Na estimação de x_0 relaciona-se dois tipos de unidades: as unidades observadas pertencentes à amostra utilizada na estimação do modelo com as que não pertencem a essa amostra. As unidades observadas pertencentes à amostra são as n áreas que são usadas na estimação do modelo de regressão espacial, onde se conhecem os valores da variável dependente Y, denotada por Y_c, e da variável independente X, denotada por X_c. As unidades que não pertencem à amostra são as k áreas onde não se conhece os valores da variável independente, denotada por X_0, mas os valores da variável dependente, denotada por Y_0, são conhecidos. Dessa forma os vetores das variáveis dependente e independente são particionados, respectivamente, em $Y = (Y_c, Y_0)$ e $X = (X_c, X_0)$. Da mesma forma, particiona-se a matriz de vizinhança espacial da seguinte forma:

$$W = \left(\begin{array}{c|c} W_{cc} & W_{c0} \\ \hline W_{0c} & W_{00} \end{array} \right), \tag{4.1}$$

em que:

- W_{cc} é uma matriz $n \times n$ correspondente à estrutura de vizinhança das unidades pertencentes à amostra;
- W_{c0} é uma matriz $n \times k$ correspondente à estrutura de vizinhança entre as unidades pertencentes à amostra e as unidades não pertencentes à amostra;
- W_{0c} é uma matriz $k \times n$ correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra e as unidades pertencentes à amostra;
- W_{00} é uma matriz $k \times k$ correspondente à estrutura de vizinhança entre as unidades não pertencentes à amostra.

As áreas sombreadas na Figura 5 são unidades que não são utilizadas na fase de estimação e ajuste do modelo, ou seja, não são usadas na estimação dos parâmetros. No entanto são utilizadas na fase de estimação do valor x_0 da variável independente X.

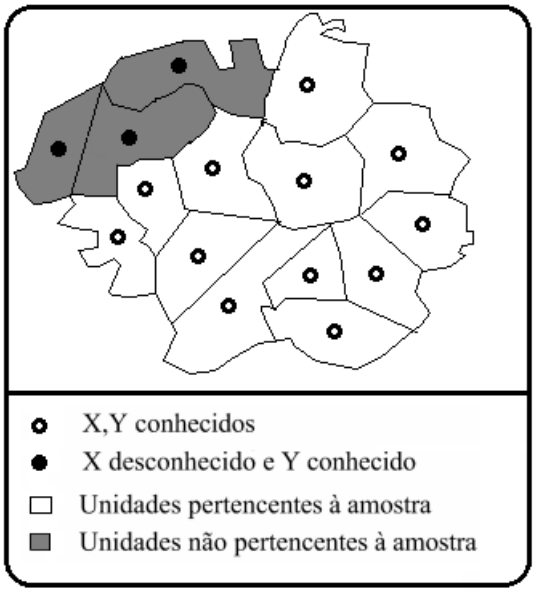

Figura 5 Caracterização do problema de estimação de um valor desconhecido da variável independente.

O modelo CAR na forma matricial é escrito da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \lambda \boldsymbol{Wu} + \boldsymbol{\varepsilon}, \tag{4.2}$$

e considerando a forma particionada, tem-se que:

$$\begin{bmatrix} \boldsymbol{Y_c} \\ \boldsymbol{Y_0} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{X_c} \\ 1 & \boldsymbol{X_0} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_c \end{bmatrix} + \lambda \left(\begin{array}{c|c} \boldsymbol{W_{cc}} & \boldsymbol{W_{c0}} \\ \hline \boldsymbol{W_{0c}} & \boldsymbol{W_{00}} \end{array} \right) \begin{bmatrix} \boldsymbol{u_c} \\ \boldsymbol{u_0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_c} \\ \boldsymbol{\varepsilon_0} \end{bmatrix}. \tag{4.3}$$

Depois de se ajustar o modelo CAR utilizando os n valores da variável dependente Y e da variável independente X que são conhecidos e usando as estimativas dos parâmetros λ, β_0 e β_1, o modelo 4.3 fica da seguinte forma:

$$\begin{bmatrix} \boldsymbol{Y_c} \\ \boldsymbol{Y_0} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{X_c} \\ 1 & \boldsymbol{X_0} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_c \end{bmatrix} + \hat{\lambda} \left(\begin{array}{c|c} \boldsymbol{W_{cc}} & \boldsymbol{W_{c0}} \\ \hline \boldsymbol{W_{0c}} & \boldsymbol{W_{00}} \end{array} \right) \begin{bmatrix} \hat{\boldsymbol{u}}_c \\ \hat{\boldsymbol{u}}_{\boldsymbol{0}} \end{bmatrix}. \tag{4.4}$$

Dessa expressão, pode-se obter:

$$\boldsymbol{Y_0} = \boldsymbol{X_0}\hat{\beta} + \hat{\lambda}\left(\boldsymbol{W_{0c}}\hat{\boldsymbol{u}}_{\boldsymbol{c}} + \boldsymbol{W_{00}}\hat{\boldsymbol{u}}_{\boldsymbol{0}}\right). \tag{4.5}$$

Expandindo (4.5) em termos matriciais tem-se que:

$$\begin{bmatrix} y_{01} \\ \vdots \\ y_{0k} \end{bmatrix} = \begin{bmatrix} 1 & x_{01} \\ \vdots & \vdots \\ 1 & x_{0k} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} + \hat{\lambda} \begin{bmatrix} w_{0c_{11}} & \cdots & w_{0c_{1n}} \\ \vdots & \vdots & \vdots \\ w_{0c_{k1}} & \cdots & w_{0c_{kn}} \end{bmatrix} \begin{bmatrix} \hat{u}_{c1} \\ \vdots \\ \hat{u}_{cn} \end{bmatrix} +$$

$$+\hat{\lambda} \begin{bmatrix} w_{00_{11}} & \cdots & w_{00_{1k}} \\ \vdots & \vdots & \vdots \\ w_{00_{k1}} & \cdots & w_{00_{kk}} \end{bmatrix} \begin{bmatrix} \hat{u}_{01} \\ \vdots \\ \hat{u}_{0k} \end{bmatrix}. \tag{4.6}$$

Assim, em termos individuais:

$$y_{0i} = \hat{\beta}_0 + x_{0i}\hat{\beta}_1 + \hat{\lambda}\left(\boldsymbol{w_{0c_i}}\hat{\boldsymbol{u}}_{\boldsymbol{c}} + \boldsymbol{w_{00_i}}\hat{\boldsymbol{u}}_{\boldsymbol{0}}\right). \tag{4.7}$$

Portanto, o estimador para um determinado valor desconhecido x_{0i} da variável X é expresso por:

$$\hat{x}_{0i} = \frac{y_{0i} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}}\hat{\boldsymbol{u}}_{\boldsymbol{c}} + \boldsymbol{w_{00_i}}\hat{\boldsymbol{u}}_{\boldsymbol{0}}\right) - \hat{\beta}_0}{\hat{\beta}_1}. \tag{4.8}$$

É necessário observar que os componentes u_{0i} do vetor $\boldsymbol{u_0}$ são desconhe-

cidos, pois $\boldsymbol{u_0} = \boldsymbol{Y_0} - \boldsymbol{X_0}\beta$ e os valores da variável independente X_0 são desconhecidos. Dessa forma, para se obter a estimativa de x_{0i} por meio da expressão (4.8) é necessário estimar o valor de $\boldsymbol{u_0}$. Alguns métodos podem ser utilizados, como:

1. Média do erros conhecidos (observados), $\hat{u}_{0i} = \bar{u}_c = \frac{\sum_{i=1}^{n} u_{ci}}{n}$, em que u_{ci} são os valores do vetor $\boldsymbol{u_c} = \boldsymbol{Y_c} - \boldsymbol{X_c}\beta$.

2. Média dos erros conhecidos das áreas que fazem fronteira com a área onde se deseja estimar o erro, $\hat{u}_{0i} = \bar{u}_v = \frac{\sum_{i=1}^{n} u_{vi}}{n}$, em que u_{vi} são os valores dos erros vindos das áreas que fazem fronteira com a área onde se deseja estimar o erro.

3. Estimar x_{0i} desconsiderando a dependência espacial por meio do estimador inverso clássico porposto por Graybill (1976), e depois estimar o erro $\hat{u}_{0i} = y_{0i} - \hat{x}_{0i}\hat{\beta}_1 - \hat{\beta}_0$.

4. Estimar x_{0i} usando o estimador espacial inverso proposto por Cordeiro (2015), e obter o erro $\hat{u}_{0i} = y_{0i} - \hat{x}_{0i}\hat{\beta}_1 - \hat{\beta}_0$.

O primeiro método citado possui facilidade de se calcular apenas uma média, ou seja, a obtenção dos valores de $\boldsymbol{u_c}$ é fácil, porém essa média utiliza valores que pertencem a locais espacialmente distantes do local onde se deseja fazer a estimação do valor desconhecido da variável independente. Isso faz com que lugares próximos e lugares distantes influenciam de mesma intensidade, o que quase sempre não é verdade em dados espaciais.

O segundo método também é de fácil utilização visto que calcula-se apenas uma média. Além disso, ele corrige o problema de influência espacial homogênea independente da distância apresentado no primeiro método. Essa correção

acontece pelo fato de que apenas valores de regiões que fazem fronteira com a área em estudo são considerados no cálculo dessa média. Porém, pode acontecer que a área em estudo possua apenas uma região vizinha, de forma que a média seja obtida usando apenas a informação de um local. Dessa forma, não é confiável fazer inferência com média calculada utilizando apenas um valor.

A estimação usando o quarto método pode ocasionar a remoção da dependência espacial entre os valores de $\boldsymbol{u_c}$ devido ao fato de se usar um modelo de regressão espacial SAR. Dessa forma, os resultados usando o modelo CAR ficam comprometidos uma vez que a potencialidade principal desse modelo está em modelar a dependência espacial existentes nos erros.

Por fim, o terceiro método será o utilizado nesta tese. Esse método estima o valor desconhecido da variável independente no local em estudo preservando as características de dependência espacial. Isso acontece devido ao fato de que o modelo utilizado é o modelo de regressão linear clássico. Dessa forma, o erro relacionado ao valor da variável independente estimado é o mais indicado para ser usado na expressão 4.8.

4.2 Estimador intervalar do valor desconhecido x_0 da variável independente X

Baseado na metodologia de Graybill (1976), primeiramente define-se uma variável Z com distribuição normal padrão, sendo $Z = \frac{\varepsilon_{0i}}{\sqrt{Var[\varepsilon_{0i}]}}$. Para tanto, calcula-se a esperança e a variância de ε_{0i}, dadas por:

$$E\left[\varepsilon_{0i}\right] = E\left[y_{0i} - x_{0i}\hat{\beta} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0}\right)\right] = 0; \tag{4.9}$$

$$
\begin{aligned}
Var\left[\varepsilon_{0i}\right] &= Var\left[y_{0i} - x_{0i}\hat{\beta} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0}\right)\right] \\
&= Var\left[y_{0i}\right] + Var\left[x_{0i}\hat{\beta}\right] + Var\left[\hat{\lambda}(\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0})\right] - \\
&- 2Cov\left[y_{0i}, \hat{\lambda}\left(\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0}\right)\right] - 2Cov\left[y_{0i}, x_{0i}\hat{\beta}\right] + \\
&+ 2Cov\left[\hat{\lambda}\left(\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0}\right), x_{0i}\hat{\beta}\right] \\
&= Var\left[y_{0i}\right] + \hat{\lambda}^2 Var\left[\boldsymbol{w_{0c_i}u_c} + \boldsymbol{w_{00_i}u_0}\right] + x_{0i}Var[\hat{\beta}]x_{0i}^T - \\
&- 2\hat{\lambda}\left\{Cov\left[y_{0i}, \boldsymbol{w_{0c_i}u_c}\right] + Cov\left[y_{0i}, \boldsymbol{w_{00_i}u_0}\right]\right\} - 2Cov\left[y_{0i}, \hat{\beta}\right] x_{0i}^T + \\
&+ 2\hat{\lambda}\left\{Cov\left[\boldsymbol{w_{0c_i}u_c}, \hat{\beta}\right] + Cov\left[\boldsymbol{w_{00_i}u_0}, \hat{\beta}\right]\right\} x_{0i}^T \\
&= Var\left[y_{0i}\right] + \\
&+ \hat{\lambda}^2\left\{Var\left[\boldsymbol{w_{00_i}u_0}\right] + Var\left[\boldsymbol{w_{0c_i}u_c}\right] + 2Cov\left[\boldsymbol{w_{00_i}u_0}, \boldsymbol{w_{0c_i}u_c}\right]\right\} + \\
&+ x_{0i}Var[\hat{\beta}]x_{0i}^T - 2\hat{\lambda}\left\{Cov\left[y_{0i}, \boldsymbol{w_{00_i}u_0}\right] + Cov\left[y_{0i}, \boldsymbol{w_{0c_i}u_c}\right]\right\} - \\
&- 2Cov\left[y_{0i}, \hat{\beta}\right] x_{0i}^T + 2\hat{\lambda}\left\{Cov\left[\boldsymbol{w_{00_i}u_0}, \hat{\beta}\right] + Cov\left[\boldsymbol{w_{0c_i}u_c}, \hat{\beta}\right]\right\} x_{0i}^T
\end{aligned}
\tag{4.10}
$$

Os resultados de alguns termos de (4.10) serão mostrados separadamente a fim de facilitar a compreensão dos cálculos, ou seja:

- $Var\left[\boldsymbol{w_{00_i}u_0}\right] = \boldsymbol{w_{00_i}}Var\left[\boldsymbol{u_0}\right]\boldsymbol{w_{00_i}^T} = \boldsymbol{w_{00_i}}\sigma^2\boldsymbol{\Theta_{00}w_{00_i}^T}$;
- $Var\left[\boldsymbol{w_{0c_i}u_c}\right] = \boldsymbol{w_{0c_i}}Var\left[\boldsymbol{u_c}\right]\boldsymbol{w_{0c_i}^T} = \boldsymbol{w_{0c_i}}\sigma^2\boldsymbol{\Theta_{cc}w_{0c_i}^T}$;
- $Cov\left[\boldsymbol{w_{00_i}u_0}, \boldsymbol{w_{0c_i}u_c}\right] = \boldsymbol{w_{00_i}}Cov\left[\boldsymbol{u_0}, \boldsymbol{u_c}\right]\boldsymbol{w_{0c_i}^T} = \boldsymbol{w_{00_i}}\sigma^2\boldsymbol{\Theta_{0c}w_{0c_i}^T}$;
- $Cov\left[y_{0i}, \boldsymbol{w_{00_i}u_0}\right] = Cov\left[y_{0i}, \boldsymbol{u_0}\right]\boldsymbol{w_{00_i}^T} = Cov\left[y_{0i}, y_{0i} - x_{0i}\hat{\beta}\right]\boldsymbol{w_{00_i}^T} =$
  $= \left\{Cov\left[y_{0i}, y_{0i}\right] - Cov\left[y_{0i}, x_{0i}\hat{\beta}\right]\right\}\boldsymbol{w_{00_i}^T} = \left\{\sigma^2\Phi_{ii} - Cov\left[y_{0i}, \boldsymbol{y_c}\right]a^T x_{0i}^T\right\}\boldsymbol{w_{00_i}^T} =$
  $= \sigma^2\left[\boldsymbol{\Phi_{ii}} - \boldsymbol{\Phi_{0ic}}a^T x_{0i}^T\right]\boldsymbol{w_{00_i}^T}$;

- $Cov\left[y_{0i}, \boldsymbol{w_{0c_i}} \boldsymbol{u_c}\right] = Cov\left[y_{0i}, \boldsymbol{u_c}\right] \boldsymbol{w_{0c_i}^T} = Cov\left[y_{0i}, \boldsymbol{y_c} - \boldsymbol{x_c}\hat{\beta}\right] \boldsymbol{w_{0c_i}^T} =$
  $= \left\{Cov\left[y_{0i}, \boldsymbol{y_c}\right] - Cov\left[y_0, \boldsymbol{x_c}\hat{\beta}\right]\right\} \boldsymbol{w_{0c_i}^T} = \left\{\sigma^2 \boldsymbol{\Phi_{0ic}} - Cov\left[y_{0i}, \boldsymbol{y_c}\right] a^T \boldsymbol{x_c^T}\right\} \boldsymbol{w_{0c_i}^T} =$
  $= \sigma^2 \left[\boldsymbol{\Phi_{0ic}} - \boldsymbol{\Phi_{0ic}} a^T \boldsymbol{x_c^T}\right] \boldsymbol{w_{0c_i}^T};$

- $Cov\left[y_{0i}, \hat{\beta}\right] = Cov\left[y_{0i}, \boldsymbol{y_c}\right] a^T = \sigma^2 \boldsymbol{\Phi_{0ic}} a^T;$

- $Cov\left[\boldsymbol{w_{00_i}} \boldsymbol{u_0}, \hat{\beta}\right] = \boldsymbol{w_{00_i}} Cov\left[y_{0i} - x_{0i}\hat{\beta}, \hat{\beta}\right]$
  $= \boldsymbol{w_{00_i}} \left\{Cov\left[y_{0i}, \hat{\beta}\right] - Cov\left[x_{0i}\hat{\beta}, \hat{\beta}\right]\right\}$
  $= \boldsymbol{w_{00_i}} \left\{Cov\left[y_{0i}, \boldsymbol{y_c}\right] a^T - x_{0i} Cov\left[\hat{\beta}, \hat{\beta}\right]\right\}$
  $= \boldsymbol{w_{00_i}} \left\{\sigma^2 \boldsymbol{\Phi_{0ic}} a^T - x_{0i}\sigma^2 \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right\}$
  $= \sigma^2 \boldsymbol{w_{00_i}} \left[\boldsymbol{\Phi_{0ic}} a^T - x_{0i} \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right];$

- $Cov\left[\boldsymbol{w_{0c_i}} \boldsymbol{u_c}, \hat{\beta}\right] = \boldsymbol{w_{0c_i}} Cov\left[\boldsymbol{y_c} - \boldsymbol{x_c}\hat{\beta}, \hat{\beta}\right]$
  $= \boldsymbol{w_{0c_i}} \left\{Cov\left[\boldsymbol{y_c}, \hat{\beta}\right] - Cov\left[\boldsymbol{x_c}\hat{\beta}, \hat{\beta}\right]\right\}$
  $= \boldsymbol{w_{0c_i}} \left\{Cov\left[\boldsymbol{y_c}, \boldsymbol{y_c}\right] a^T - \boldsymbol{x_c} Cov\left[\hat{\beta}, \hat{\beta}\right]\right\}$
  $= \boldsymbol{w_{00_i}} \left\{\sigma^2 \boldsymbol{\Phi_{cc}} a^T - \boldsymbol{x_c}\sigma^2 \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right\}$
  $= \sigma^2 \boldsymbol{w_{0c_i}} \left[\boldsymbol{\Phi_{cc}} a^T - \boldsymbol{x_c} \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right];$

em que:

- $a = \left[\boldsymbol{x_c^T} (I - \lambda W)^T (I - \lambda W) \boldsymbol{x_c}\right]^{-1} \boldsymbol{x_c^T} (I - \lambda W)^T (I - \lambda W).$

Assim, obtém-se:

$$Var\left[\varepsilon_{0i}\right] = \sigma^2 \left\{\boldsymbol{\Phi_{ii}} + \hat{\lambda}^2 \left(\boldsymbol{w_{00_i}} \boldsymbol{\Theta_{00}} \boldsymbol{w_{00_i}^T} + \boldsymbol{w_{0c_i}} \boldsymbol{\Theta_{cc}} \boldsymbol{w_{0c_i}^T} + 2\boldsymbol{w_{00_i}} \boldsymbol{\Theta_{0c}} \boldsymbol{w_{0c_i}^T}\right) + \right.$$
$$+ x_{0i} \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1} x_{0i}^T - 2\hat{\lambda} \left(\left[\boldsymbol{\Phi_{ii}} - \boldsymbol{\Phi_{0ic}} a^T x_{0i}^T\right] \boldsymbol{w_{00_i}^T} + \left[\boldsymbol{\Phi_{0ic}} - \boldsymbol{\Phi_{0ic}} a^T \boldsymbol{x_c^T}\right] \boldsymbol{w_{0c_i}^T}\right)$$
$$\left. - 2\boldsymbol{\Phi_{0ic}} a^T x_{0i}^T + 2\hat{\lambda} \left(\boldsymbol{w_{00_i}} \left[\boldsymbol{\Phi_{0ic}} a^T - x_{0i} \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right] + \boldsymbol{w_{0c_i}} \left[\boldsymbol{\Phi_{cc}} a^T - \boldsymbol{x_c} \left(\boldsymbol{x_c^T \Theta_{cc}^{-1} x_c}\right)^{-1}\right]\right) x_{0i}^T \right\}.$$

Por fim, segue que

$$Var\left[\varepsilon_{0i}\right] = \sigma^2 M. \tag{4.11}$$

Com o resultado (4.11), tem-se:

$$Z = \frac{\varepsilon_{0i}}{\sqrt{Var\left[\varepsilon_{0i}\right]}} = \frac{y_{0i} - x_{0i}\hat{\beta} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}}\hat{\boldsymbol{u}}_{\boldsymbol{c}} + \boldsymbol{w_{00_i}}\hat{\boldsymbol{u}}_{\boldsymbol{0}}\right)}{\sqrt{\sigma^2 M}} \sim N\left(0,1\right) \quad (4.12)$$

e

$$U = \left(n - 2tr\left(\boldsymbol{S}\right) + tr\left(\boldsymbol{S^T S}\right)\right)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)}. \quad (4.13)$$

Como $Z \sim N\left(0,1\right)$ e $U \sim \chi^2_{n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)}$ são independentes, pode-se obter a seguinte variável:

$$T = \frac{Z}{\sqrt{\frac{U}{n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)}}}. \quad (4.14)$$

Pelo **Teorema 2**, a variável aleatória T dada em (4.14), tem distribuição t de Student com $\left(n - 2tr\left(\boldsymbol{S}\right) + tr\left(\boldsymbol{S^T S}\right)\right)$ graus de liberdade. Assim, sendo T uma quantidade pivotal,

$$P\left(-t_{\left(n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)\right)} \leq T \leq t_{\left(n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)\right)}\right) = 1 - \alpha. \quad (4.15)$$

De (4.15) tem-se que

$$T^2 \leq t^2_{\left(\alpha/2;n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)\right)}. \quad (4.16)$$

Então,

$$\frac{\left(y_{0i} - x_{0i}\hat{\beta} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}}\hat{\boldsymbol{u}}_{\boldsymbol{c}} + \boldsymbol{w_{00_i}}\hat{\boldsymbol{u}}_{\boldsymbol{0}}\right)\right)^2}{\hat{\sigma}^2 M} \leq t^2_{\left(n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)\right)}. \quad (4.17)$$

Por fim, seque que

$$\left(y_{0i} - x_{0i}\hat{\beta} - \hat{\lambda}\left(\boldsymbol{w_{0c_i}}\boldsymbol{\hat{u}_c} + \boldsymbol{w_{00_i}}\boldsymbol{\hat{u}_0}\right)\right)^2 - \hat{\sigma}^2 M t^2_{\left(n-2tr(\boldsymbol{S})+tr\left(\boldsymbol{S^T S}\right)\right)} \leq 0. \quad (4.18)$$

Dessa forma, para obter o intervalo de confiança $(1-\alpha)100\%$ de x_{0i} quando $Y = y_{0i}$ é necessário expandir o resultado (4.18), o que resulta em uma inequação quadrática, em função do valor desconhecido x_{0i} da seguinte forma: $q(x_{0i}) = ax_{0i}^2 + bx_{0i} + c \leq 0$.

De acordo com Graybill (1976) e Thonnard (2006) os valores de x_{0i} que satisfazem essa inequação, quando $a > 0$ e $b^2 - 4ac > 0$ (seção 2.1.3), formam um intervalo de confiança $(1-\alpha)100\%$ para x_{0i}.

Os estimadores inversos pontual e intervalar foram implementados computacionalmente usando o *software* R e foram usados na análise de dados reais.

5 APLICAÇÃO E DISCUSSÃO

Para a aplicação dos estimadores inversos pontual e intervalar propostos neste trabalho foi utilizado um conjunto de dados reais que está presente na biblioteca "spdep"do *software* R (R CORE TEAM, 2016).

Esse conjunto de dados é descrito em Anselin (1988) e inclui observações dos roubos residenciais e roubos de veículos por mil domicílios (quantidades de crimes), renda média familiar (em mil dólares) e o valor médio da habitação (em mil dólares) em cada um dos 49 bairros da cidade de Columbus, Ohio, EUA.

A distribuição espacial do número de crimes e da renda média familiar está representada geograficamente nas Figuras 6a e 6b, respectivamente. Observa-se nessa representação que há uma associação inversa entre o número de crimes e a renda média familiar. Bairros com maior número de roubos são os bairros que

apresentam a menor renda familiar.

A teoria econômica sugere que a desigualdade de renda contribui para o aumento da criminalidade (RESENDE & ANDRADE, 2011). A relação entre desigualdade de renda e criminalidade já foi objeto de análise de diversos estudos econômicos como Eberts & Schwirian (1968), Danzinger & Wheeler (1975), Fowles & Merva (1996), Demombynes & Ozner (2002), Dahlberg & Gustavsson (2005), Resende & Andrade (2011), dentre outros. Nesses estudos os autores concluiram que existe um padrão de atuação da variável desigualdade de renda sobre os indicadores de criminalidade.

Dessa forma, nota-se que a ocorrência de crimes está correlacionada com a distribuição de renda, onde essa taxa de criminalidade varia de acordo com a variação da renda média, ou seja, intuitivamente a ocorrência da criminalidade depende da distribuição de renda. Portanto, considerou-se a ocorrência de crimes como sendo a variável dependente e distribuição de renda média familiar como sendo a variável independente na análise de regressão.

Porém, conforme descreve Cordeiro (2015), observa-se que nessa situação pode ser mais complicado obter informações dos valores da renda familiar do que informações sobre número de crimes, devido à dificuldade das pessoas declararem o valor real dos rendimentos familiares. Portanto, essa situação caracteriza um problema de regressão inversa, onde tem-se o interesse em estimar o valor da renda média familiar (X) dadas as observações dos roubos residenciais e roubos de veículos por mil domicílios (Y). Incorporando-se a distribuição espacial dos dados no processo de modelagem da regressão inversa tem-se a regressão espacial inversa.

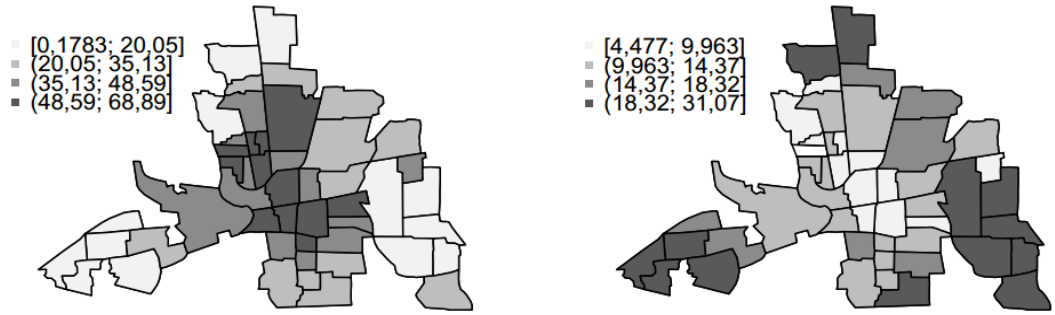

Figura 6 Mapa dos crimes relacionados a roubos residenciais e de veículos por mil domicílios (a) e o mapa da renda média mensal familiar em mil dólares (b), em Columbus, Ohio, EUA.

As quatro classes que estão nas legendas da figura acima foram obtidas a partir dos quartis que dividem a distribuição dos dados em ordem crescente em quatro partes iguais.

5.1 Ajuste do modelo CAR

Segundo Ponciano e Scalon (2010), o processo de análise espacial para dados de área inclui duas etapas: análise exploratória e modelagem. De uma forma geral, a análise exploratória permite descrever as distribuições das variáveis e os padrões de associação espacial das mesmas. A análise exploratória fornece indicadores sobre possíveis modelos que serão analisados através de procedimentos de estimação e validação.

Inicialmente realizou-se uma análise exploratória das variáveis envolvidas na análise. As estatísticas descritivas clássicas e o índice de Moran dessas variáveis estão na Tabela 3.

Tabela 3 Estatísticas descritivas das variáveis: números de crimes (CRIME), valor da renda média familiar (RENDA) e índice de Moran.

Estatística	CRIME	RENDA
Mínimo	0,17	4,48
Mediana	34,00	13,38
Média	35,13	14,37
Máximo	69,89	31,07
Desvio padrão	16,73	5,70
Moran	$0,49^{*}$	$0,42^{**}$

$^{*}(valor - p = 4,5 \times 10^{-8})$

$^{**}(valor - p = 1,3 \times 10^{-6})$

O índice de Moran foi utilizado para descrever a estrutura de dependência espacial das variáveis envolvidas medindo o nível de autocorrelação espacial entre as unidades (bairros). Conforme a Tabela 3, pode-se verificar que os dados são autocorrelacionados espacialmente tanto para a variável crime como para a variável renda com valores significativos desse índice iguais a 0,486 e 0,417, respectivamente, com valor-p <0,0001.

Testou-se a suposição de independência dos resíduos utilizando o índice de Moran obtendo o valor I = 0,16, o que revelou a presença de autocorrelação espacial dos resíduos (valor-p <0,05), indicando que os resíduos apresentam dependência espacial, o que normalmente acontece quando as variáveis utilizadas no modelo são espacialmente dependentes.

Baseado nos índices de Moran calculados para as variáveis e para os resíduos, pôde-se concluir que os bairros mais próximos são mais semelhantes entre si, ou seja, existe um componente espacial envolvido nas variáveis analisadas. O cálculo do índice de Moran foi realizado com a matriz de peso espacial W construída de acordo com o critério que se A_i compartilha um lado comum com A_j então $w_{ij} = 1$, caso contrário $w_{ij} = 0$.

Depois da análise exploratória inicial, baseado no valor do índice de Mo-

ran buscou-se capturar a estrutura de autocorrelação espacial ajustando o modelo de regressão espacial autorregressivo condicional (modelo CAR). Após estimar o parâmetro autorregressivo espacial λ desse modelo, pode-se testar a significância desse parâmetro, utilizando-se o teste de Wald, o teste da razão de verossimilhança ou o teste dos multiplicadores de Lagrange. Segundo Ywata e Albuquerque (2011), testando-se a significância do parâmetro λ testa-se implicitamente a presença de dependência espacial, pois a ideia básica no modelo CAR é incorporar a autocorrelação espacial como componente do modelo. Caso se observe a ausência de autocorrelação espacial ($\lambda = 0$), o modelo CAR (2.56 e 2.57) é o próprio modelo de Guass-Markov geral, ou seja, o modelo de regressão linear clássico.

De acordo com a Tabela 4 , o parâmetro espacial autorregressivo possui valor igual a 0,983, estatisticamente significativo pelo teste de Wald, confirmando que os bairros mais próximos são mais semelhantes entre si e que a dependência espacial foi modelada adequadamente. Ainda, para efeito de comparação, os modelos de regressão linear simples e espacial autorregressivo condicional (CAR) foram comparados através dos seguintes critérios descritos em Draper e Smith (1998): critério de Akaike e erro quadrático médio. Pode-se observar na Tabela 4 que o modelo CAR obteve melhor desempenho uma vez que apresenta menores valores nos dois critérios.

Tabela 4 Camparação dos modelo de regressão linear simples e modelo CAR.

Modelos	EQM	AIC	Componente aut. espacial ($\hat{\lambda}$)
Linear simples	144,5	294,64	$--$
CAR	1,2	172,31	0,983*

*significativo a 1%.

5.2 Estimação do valor x_0 desconhecido da variável X

Para aplicar os estimadores propostos objetivando estimar uma observação não pertencente a uma amostra particionou-se a região de estudo em dois tipos de unidades espaciais. Primeiramente, tem-se unidades em que os valores das variáveis independente e dependente são conhecidas (X_c, Y_c), ou seja, essas unidades pertencem à amostra selecionada e são necessárias na realização da primeira etapa do processo de calibração (ajuste do modelo). O outro tipo de unidades espaciais são aquelas onde os valores da variável independente são desconhecidos e os valores da variável dependente são conhecidos (X_0, Y_0), ou seja, é onde deseja-se estimar um valor x_0 desconhecido da variável independente X.

Baseado no que foi dito no parágrafo anterior, algumas situações hipotéticas foram usadas para o efetivo uso dos estimadores. A seguir apresenta-se a configuração de cada uma dessas situações bem como o resultado da estimação espacial inversa por meio dos estimadores propostos em (4.8) e (4.18).

Situação 1

Nessa primeira situação foram selecionados ao acaso 48 bairros onde os valores das variáveis X (renda média familiar) e Y (crimes) são conhecidos, e um bairro onde considera-se conhecido somente o valor y_0 da variável dependente e onde deseja-se estimar o valor desconhecido x_0 da variável independente. A configuração espacial deste cenário está representada na Figura 7.

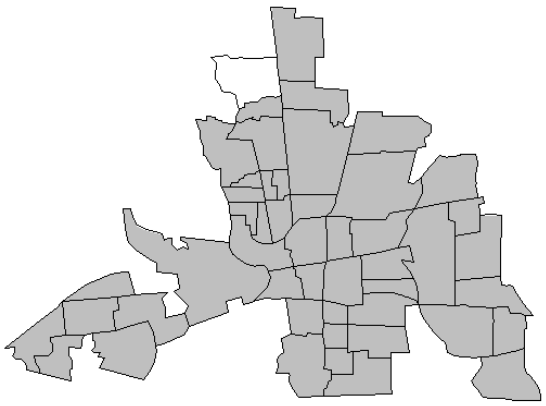

Figura 7 As áreas sombreadas evidenciam os 48 bairros usados na primeira etapa da calibração espacial, ou seja, usadas no ajuste do modelo.

Inicialmente construiu-se a matriz de vizinhança desses 48 bairros, em seguida ajustou-se o modelo CAR por meio da maxima verossimilhança, considerando como variável independente "renda média familiar"e como variável dependente "crimes". Os parâmetros do modelo podem ser observados na Tabela 5.

Tabela 5 Estimativas dos parâmetros do modelo CAR ajustado utilizando 48 unidades espaciais (ou bairros) selecionadas ao acaso.

Parâmetros	Estimativas	Erro-padrão	z calculado	Valor-p
Constante	34,930	2,933	11,908	$2,2 \times 10^{-16}$
Renda média familiar	-1,590	0,339	-4,679	$2,8 \times 10^{-6}$
Componente espacial aut. (λ)	0,446	0,157	2,835	$4,0 \times 10^{-3}$

De acordo com a Tabela 5, a variável renda média familiar possui um coeficiente negativo, o que indica que quanto maior a renda, menor será o número de crimes. O coeficiente espacial autorregressivo apresentou um valor estatisticamente significativo a 1%, o que evidencia a dependência espacial nos resíduos.

Na segunda etada do processo, estimou-se o valor desconhecido da variável independente de forma pontual e intervalar. Essas estimativas podem ser observadas na Tabela 6.

Tabela 6 Estimativa pontual e intervalar para a renda média familiar para o bairro onde não se conhece o valor dessa variável.

Crimes	Renda média familiar (valor estimado)	Estimativa intervalar	Renda média familiar (valor observado)
18,80	21,72	[6,94; 43,65]	21,23

Pode-se observar que o valor real está próximo do valor pontual estimado e está dentro do intervalo de confiança obtido. Este resultado evidencia que os estimadores funcionaram de maneira satisfatória.

Para validar o modelo foi feita a validação cruzada *leave-one-out*. Segundo Snee (1977) a validação cruzada é um método eficaz de avaliação de um modelo de regressão. O gráfico da validação cruzada, que apresenta os valores preditos pelo modelo versus os valores reais, é apresentado na Figura 8.

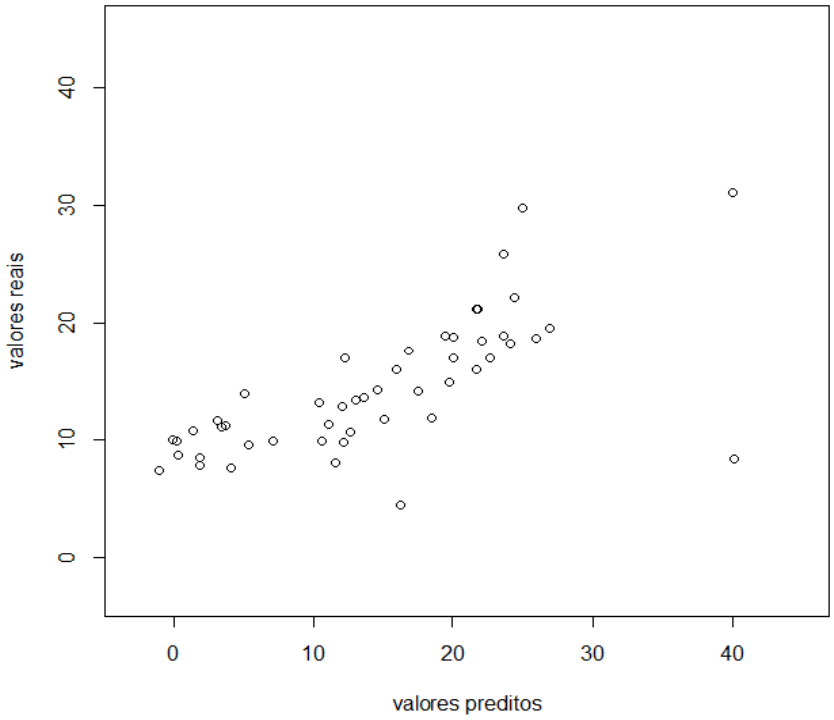

Figura 8 Gráfico de dispersão dos valores reais e dos valores preditos da variável X.

Observa-se no gráfico uma relação linear entre os valores reais e os preditos com um coeficiente de correlação igual a 0,693, o que comprova que o modelo é preciso em se tratando de predição. Existe uma observação (observação 7) que apresentou um comportamento diferente, com seu valor afastado dos demais. Trata-se de uma região de fronteira da cidade de Columbus que possivelmente sofre influencia de regiões que não pertencem a área de estudo e que exerce efeito no ajuste do modelo.

Situação 2

Nesse segundo momento, para a primeira etapa do processo de calibração espacial foram selecionadas ao acaso 45 unidades (bairros) conforme a Figura 9.

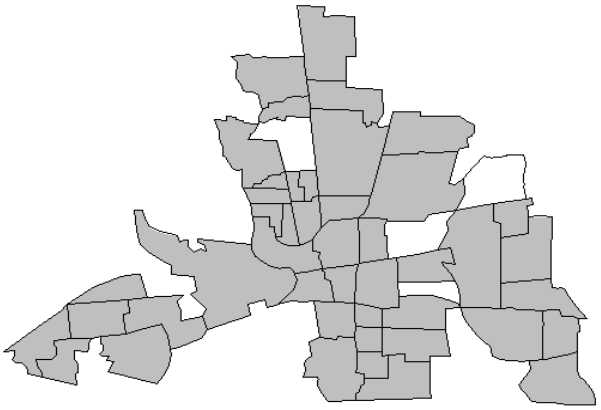

Figura 9 As áreas sombreadas são os 45 bairros utilizados na fase de ajuste do modelo.

Depois de construir a matriz de vizinhança desses bairros, novamente ajustou-se o modelo CAR e os valores estimados dos parâmetros do modelo podem ser observados na Tabela 7.

Tabela 7 Estimativas dos parâmetros do modelo CAR ajustado considerando as 45 áreas selecionadas ao acaso.

Parâmetros	Estimativas	Erro-padrão	z calculado	Valor-p
Constante	34,482	2,926	11,781	$2,2 \times 10^{-16}$
Renda familiar	-1,632	0,341	-4,786	$1,6 \times 10^{-6}$
Componente espacial aut. (λ)	0,405	0,162	2,486	$1,2 \times 10^{-2}$

O coeficiente espacial autorregressivo estimado $\hat{\lambda}$ apresentou um valor significativo diferente de zero, o que mostra que os resíduos apresentam dependência espacial e que o modelo CAR conseguiu modelar a dependência espacial na área em estudo, pois conforme salienta Câmara et al. (2004), a dependência espacial nos resíduos pode ser reflexo da autocorrelação presente nos dados, que pode se manifestar por diferenças regionais sistemáticas nas relações do modelo, ou ainda,

por uma tendência espacial contínua.

O interesse na segunda etapa da calibração espacial é obter os valores da variável independente em função dos valores da variável dependente. Dessa forma, considerou-se que nos quatro bairros restantes tem-se apenas os valores y_0 da variável crimes e os valores x_0 da variável renda média familiar são desconhecidos.

Esses valores da variável independente foram estimados e estão apresentados na Tabela 8.

Tabela 8 Estimativa pontual e intervalar para a renda média familiar nos bairros onde não se conhece o valor real dessa variável.

Crimes	Renda média familiar (valor estimado)	Estimativa intervalar	Renda média familiar (valor observado)
33,70	15,19	[-1,78; 33,19]	11,70
34,01	14,22	[-2,87; 32,18]	13,59
38,42	11,16	[-7,19; 27,38]	11,33
41,96	10,74	[-8,48; 26,89]	9,91

Como pode ser observado na Tabela 8, os limites inferiores possuem valores negativos, o que deve ser admitido teoricamente. Porém, do ponto de vista prático não é útil, visto que trata-se da variável renda média familiar.

À medida que a quantidade de crimes aumenta observa-se que os valores estimados para a renda média familiar diminui, o que está em consonância com o que o modelo ajustado propõe.

Além disso, novamente obsrva-se que os valores reais estão próximos dos valores pontuais estimados e estão dentro dos intervalos de confiança obtidos. **Situação 3**

Nesse terceiro cenário de estudo, foram selecionadas ao acaso 40 unidades (bairros) para a primeira etapa da calibração espacial, conforme pode ser observado na Figura 10.

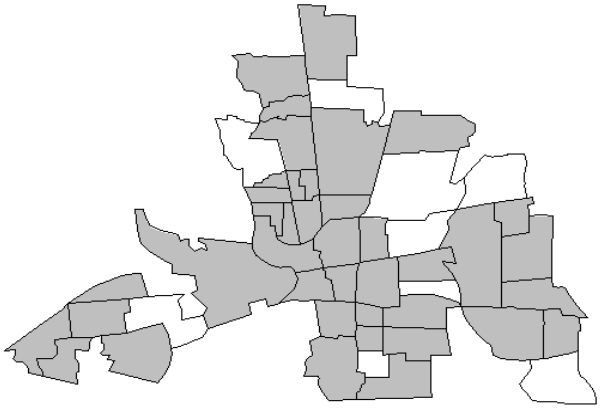

Figura 10 As áreas sombreadas são os 40 bairros utilizadas no ajuste do modelo.

Novamente, depois de obter a matriz de vizinhança espacial dessas 40 áreas, ajustou-se o modelo CAR via máximaverossomilhança para a modelagem espacial. Os valores estimados dos parâmetros do modelo estão apresentados na Tabela 9.

Tabela 9 Estimativas dos parâmetros do modelo CAR ajustado considerando as 40 áreas selecionadas ao acaso.

Parâmetros	Estimativas	Erro-padrão	z calculado	Valor-p
Constante	36,143	2,994	12,069	$2,2 \times 10^{-16}$
Renda familiar	-1,648	0,281	-5,856	$4,7 \times 10^{-9}$
Componente espacial aut. (λ)	0,522	0,143	3,643	$2,0 \times 10^{-4}$

O coeficiente estimado $\hat{\lambda}$ apresentou um valor positivo estatisticamente diferente de zero, evidenciando a dependência dos resíduos e que os valores em bairros que fazem fronteira entre si são similares, ou seja, há uma dependência espacial modelada nesse ajuste considerando os 40 bairros selecionados .

Dando sequência ao processo de calibração espacial, estimou-se os valores

x_0 da variável independente renda média familiar em cada um dos nove bairros que não foram considerados no ajuste do modelo, e onde em tese existe a informação somente dos valores y_0 da variável dependente crimes. Os valores estimados considerando o valor y_0 em cada um desses bairros podem ser observados na Tabela 10.

Tabela 10 Estimativa pontual e intervalar para a renda média familiar nos bairros não se conhece o valor real dessa variável.

Crimes	Renda média familiar (valor estimado)	Estimativa intervalar	Renda média familiar (valor observado)
0,17	37,29	[25,34; 62,36]	8,43
23,97	20,17	[9,04; 34,73]	14,94
26,64	18,70	[7,31; 32,61]	11,81
27,82	20,35	[9,85; 35,62]	18,95
30,51	17,48	[5,85; 30,61]	17,58
30,62	16,49	[4,69; 29,40]	15,95
33,70	15,80	[3,83; 28,72]	11,70
34,00	14,91	[2,92; 27,93]	13,59
41,96	11,54	[-1,63; 23,31]	9,91

Percebe-se novamente a associação inversa entre as variáveis "crimes"e "renda média familiar"e também que os valores reais estão dentro dos intervalos obtidos. Exceto pelo primeiro intervalo que trata-se da observação 7 já mencionada anteriormente, nota-se que a amplitude dos intervalos de confiança diminuiu em relação às outras duas situações estudadas (Tabela 6 e Tabela 8), gerando uma maior precisão. Esse menor valor de amplitude está associado ao fato de que, conforme pode ser observado na expressão (4.10), o vetor de proximidade entre as áreas onde não se conhece os valores x_0 da variável independente é usado na compilação da estimação intervalar de x_0. Uma vez que mais áreas fazem parte do segunda etapa da calibração espacial, então a estimação pontual e intervalar da variável "renda média familiar", x_0, em uma determinada localidade, sofre in-

fluência direta das outras regiões onde não se conhece os valores da "renda média familiar", mas que fazem fronteira com a região onde se está estimando.

Outro fator que influencia a amplitude do intervalo que deve ser levado em consideração é a variância presente nos dados usados na primeira etapa de estimação do modelo, isto é, se para o conjunto de 40 observações selecionadas ao acaso houve um menor valor da variância (S^2), do que quando foram selecionadas as 45 e 48 observações, então esse menor valor de variância influencia diretamente o tamanho da amplitude do intervalo.

6 CONSIDERAÇÕES FINAIS

A abordagem da calibração espacial ou regressão espacial inversa usando o modelo condicional autorregressivo CAR mostrou ser apropriada na análise dados de área com dependência espacial. Porém, mais estudos precisam ser realizados, tendo analisando-se o tipo da estrutura espacial. A forma como as matrizes de dependência espacial são determinadas, levando-se em consideração a irregularidade dos mapas de municípios e de setores censitários, pode afetar o desempenho da regressão espacial inversa, uma vez que este fator pode influenciar o desempenho dos modelos autorregressivos.

O modelo proposto de calibração espacial é vantajoso no sentido de que ao se estimar os parâmetros de um modelo considerando a informação espacial, no caso de modelos autorregressivos, obtém-se um ganho de eficiência dos estimadores. Na literatura quase todos os modelos de regressão inversa existentes não consideram a informação espacial, existindo apenas o modelo proposto por Cordeiro (2015) com tal particularidade. Dessa forma, este trabalho vem a enriquecer a análise de dados de área, fornecendo um ferramenta útil para casos que configurem a necessidade de se obter o valor de uma variável independente conhecendo-se o valor da variável dependente.

Por fim, um grande potencial que esse modelo de regressão espacial inversa tem está no fato que ele pode ser um método eficiente de imputação de dados faltantes na análise de dados de área. Mas para essa utilização é necessário que se conheça os valores observados de uma variável dependente. Constantemente, um problema comum que surge em investigações científicas é a ocorrência de dados faltantes (*missing data*), muitas vezes a "solução"para esses problemas é a decisão do pesquisador em desconsiderar o local onde não se conhece o valor de alguma das variáveis, visto que a maioria das técnicas estatísticas foram desen-

volvidas para a análise de dados completos. Porém, desconsiderar esses locais do estudo pode gerar inferências que não são válidas, dado o fato de que na análise de dados de área a informação de um local é influenciada pela informação de locais vizinhos.

REFERÊNCIAS

ANSELIN, L. **Spatial econometrics: methods and models**. Dordrecht: Kluwer Academic, 1988. 189 p.

_____. Under the hood: issues in the specification and interpretation of spatial regression models. **Agricultural Economics**, Amsterdam, v. 27, n. 3, p. 247-267, Nov. 2002.

ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: Associação Brasileira de Estatística, 2001. 131 p.

BAILEY, T. C.; GATRELL, A.C. **Interactive Spatial Data Analysis**. Essex: Longman, 1995. 71 p.

BREUSCH, T.; PAGAN, A. Teste simples para heterocedasticidade e coeficiente de variação aleatória Econométrica. **Sociedade Econométrica**, Rio de Janeiro, v. 47, p. 1287-1294, 1979.

BROWN, P. J. Multivariate calibration. **Journal of the Royal Statistical Society Series B-Methodological**, London, v. 44, n. 3, p. 287-321, Feb. 1982.

CÂMARA, G. et al. Análise espacial de áreas. In: DRUCK, S. et al. (Ed.). **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2002. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap1-intro.pdf>. Acesso em: 13 mar. 2017.

CASELLA, G.; BERGER, R. **Inferência estatística**. 2^{nd} ed. São Paulo: C. Learning, 2001. 588 p.

CHARNET, R. et al. **Análise de modelos de regressão linear**. Campinas: Unicamp, 2008. 357 p.

CLIFF, A. D.; ORD, J. K. **Spatial processes: models and applications**. London:

Pion, 1981. 266 p.

COLLINS, K.; BABYAK, C.; MOLONEY, J. Treatment of spatial autocorrelation in Geocoded crime data. In:____. **American Statistical Association, section on servey research methods**. Ottawa: Statistics Canada, 2006. p. 2864 - 2871.

CORDEIRO, L. L. **Estimação em regressão espacial inversa**. 2015. 89 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2015.

CRESSIE, N. A. C. **Statistics for spatial data**. Chichester: J. Wiley, 1991. 900 p.

DAHLBERG, M.; GUSTAVSSON, M. Inequality and crime: separating the effects of permanent and transitory income. **Oxford Bulletin of Economics and Statistics**, v. 70, n. 2, p. 129-153, Mar. 2008.

DANZIGER, S., WHEELER, D. The economics of crime: punishment or income redistribution. **Review of Social Economy**. v. 33, n.2, p. 113-131, Oct. 1975.

DEMOMBYNES, G.; ÖLER, B. Crime and local inequality in South Africa. **Journal of Development Economics**, v. 76, n. 2, p. 265-292, Nov. 2002.

DINIZ FILHO, J. A. F.; BINI, L. M.; HAWKINS, B. A. Spatial autocorrelation and red herrings in geographical ecology. **Global Ecology and Biogeography**, v. 12, n. 1, p. 53-64, Jan. 2003.

DOMINGOS FILHO, M.; SAMOHYL, R. W. Chart control of calibration functional with assumption of variance of measurement errors known. **Revista Brasileira de Biometria**, Lavras, v.34, n.1, p.183-209, Jun. 2016.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3^{rd} ed. New York: J. Wiley, 1998. 706 p.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA,

2004. 209 p.

EBERTS, P., SCHWIRIAN, K. Metropolitan crime rates and relative deprivation. **Criminologica** v.5, n. 4, p. 43-52 , Feb. 1968.

EISENHART, C. The interpretation of certain regression methods, and their use in biological and industrial research. **Annals of Mathematical Statistics**, Ann Arbor, v. 10, p. 162-186, Jun. 1939.

FIELLER, E. C. Some problems in interval estimation. **Journal of the Royal Statistical Society**, London, v. 16, p. 175-185, Jan. 1954.

FOWLES, R., MERVA, M. Wage inequality and criminal activity: an extreme bounds analysis for the United States, 1975-90. **Criminology**, v. 34, n. 2, 163-182, May 1996.

GETIS, A.; ORD, J.K., The Analysis of Spatial Association by Use of Distance Statistics. In: **Geographical Analysis**, v. 24, n. 3, p. 190-206, Jul. 1992.

GRAYBILL, F. A. **Theory and application of linear model**. North Situate: Duxbury, 1976. 740 p.

GUIMARÃES, E. C. **Geoestatística Básica e Aplicada**. Uberlândia: Universidade Federal de Uberlândia, Faculdade de Matemática, 2004. 74 p. Apostila.

HAINING, R. P. **Spatial Data Analysis in the Social and Environmental Sciences**. Cambridge: Cambridge University Press, 1990. 432 P.

KYUNG, M.; GHOSH, S. K. Maximum likelihood estimation for directional conditionally autoregressive models. **Journal of Statistical Planning and Inference**, v.140, n. 11, p. 3160-3179, Nov. 2010.

KRUTCHKOFF, R. G. Classical and inverse regression methods of calibration. **Technometrics**, Washington, v. 9, n. 3, p. 425-439, Aug. 1967.

_____. Classical and inverse regression methods of calibration in extrapolation. **Technometrics**, Washington, v. 11, n. 3, p. 605-608, Aug. 1969.

LESAGE, J.; PACE, R. K. **Introduction to spatial econometrics**. Boca Raton: CRC Press, 2009. 321 p.

LIEBERMAN, G. J.; MILLER, R. G.; HAMILTON, M. A. Unlimited simultaneous discrimination intervals in regression. **Biometrika**, London, n. 54, p. 133-145, Jun. 1967.

MANSKI, C. Identification of endogenous social effects: the reflection problem. **Rev. Econ. Stud.**, Oxford, v.60, n.3, p.531-542, Jul. 1993.

MARTINELLI, S.; KRUTCHKOFF, R. G. On the Choice of Regression in Linear Calibration. Comments on a Paper by R. G. Krutchkoff. **Technometrics**, Vol. 12, n. 1, p. 157-161, Feb. 1970.

MATHEW, T.; KASALA, S. An exact confidence region in multivariate calibration. **Annals of Statistics**, Hayward, v. 22, n. 1, p. 94-105, Mar. 1994.

MILTINO, A.F.; UGARTE, M.D.; REINALDOS, L. G. Alternative models for describing spatial dependence among dwelling selling prices. **Journal of Real Estate Finance and Economics**, v. 29, n.2, p. 193-209, Set. 2004.

MORAN, P. A. P. Notes on continuous stochastic phenomena. **Biometrika**, London, v. 37, p. 17-23, Jun. 1950.

OLIVEIRA, E. C.; AGUIAR, P. F. Validação da metodologia da avaliação de incerteza em curvas de calibração melhor ajustadas por polinômios de segundo grau. **Química Nova**, São Paulo, v. 32, n. 6, p. 1571-1575, Jan. 2009.

ORD, J. K. Estimation methods for models of spatial interaction. **Journal of the American Statistical Association**, New York, v. 70, n. 3, p. 120-126, Mar. 1975.

PLANT, R. E. **Spatial data analysis in ecology and agriculture in R**. New York: CRC, 2012. 617 p.

PONCIANO, P. F.; SCALON, J. D. Análise espacial da produção leiteira usando um modelo autoregressivo condicional. **Semina: Ciências Agrárias**, Londrina, v. 31, n. 2, p. 487-496, Jun. 2010.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2016. Disponível em: <http://www.r-project.org>. Acesso em: 12 dez. 2016.

RESENDE, J. P.; ANDRADE, M. V. Crime social, castigo social: desigualdade de renda e taxas de criminalidade nos grandes municípios brasileiros. **Estud. Econ.**, São Paulo , v. 41, n. 1, p. 173-195, Mar. 2011 .

ROSENBERG, M. S.; SOKAL, R. R.; ODEN, N. L.; DIGIOVANII, D. Spatial autocorrelation of cancer in Western Europe. **European Journal of Epidemiology**, v. 15, n. 1, p. 15-22, Jan. 1999.

SHUKLA, G. K. On the problem of calibration. **Technometrics**, Washington, v. 14, n. 3, p. 547-553, Aug. 1972.

SILVA, N. C. et al. Análise de dados de área aplicada a dois indicadores econômicos de mesorregiões do estado de Minas Gerais. **Revista Brasileira de Biometria**, São Paulo, v.29, n.3, p.369-395, Jul./Set. 2011.

SNEE, R. D. Validation of regression models: methods and examples. **Technometrics**, Washington, v. 19, n. 4, p. 415-428, Nov. 1977.

THONNARD, M. **Confidence Intervals in Inverse Regression**. 2006. 78p. Dissertation (Master in Mathematics and Computer Science) - Technische Universiteit Eindhoven, Eindhoven, 2006.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. New York: J. Wiley, 2004. 518 p.

WILLIAMS, E. J. A note on regression methods in calibration. **Technometrics**, Washington, v. 11, n. 1, p. 189-192, Feb. 1969.

YWATA, A. X. C.; ALBUQUERQUE, P. H. M. Métodos e modelos em econometria espacial: uma revisão. **Revista Brasileira de Biometria**, São Paulo, v. 29, n. 2, p. 273-306, Jul. 2011.

APÊNDICES

APÊNDICE A: Procedimento Newton-Raphson para a determinação de $\hat{\lambda}$

A estimativa de máxima verossimilhança $\hat{\lambda}$ é o valor de λ que minimiza

$$f(\lambda) = -\frac{2}{n}\sum_{i=1}^{n}(1-\lambda\gamma_i) + \ln\left(s^2\right), \tag{A1}$$

onde $s^2 \equiv s^2(\lambda) = u^T u - 2\lambda u^T u_L + \lambda^2 u_L^T u_L$, e $f(\lambda)$ é o logaritmo da expressão 2.69.

Fazendo a primeira e segunda derivadas de A1 em relação a λ tem-se, respectivamente:

$$f_\lambda(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\frac{\gamma_i}{(1-\lambda\gamma_i)} + \frac{2\left(\lambda u_L^T u_L - u^T u_L\right)}{s^2} \tag{A2}$$

e

$$f_{\lambda\lambda}(\lambda) = \frac{2}{n}\sum_{i=1}^{n}\frac{(\gamma_i)^2}{(1-\lambda\gamma_i)^2} + \frac{2u_L^T u_L}{s^2} - \frac{4\left(\lambda u_L^T u_L - u^T u_L\right)^2}{s^4}. \tag{A3}$$

Em seguida, $\hat{\lambda}$ por ser obtida interativamente através da expressão:

$$\lambda_{r+1} = \lambda_r - \frac{f_\lambda(\lambda_r)}{f_{\lambda\lambda}(\lambda_r)}. \tag{A4}$$

Segundo Ord (1975), um valor inicial útil é $\lambda_0 = u^T u_L / u^T u$.

ANEXOS

ANEXO A: Código do *software* R para a estimação inversa pontual e intervalar em regressão inversa linear simples.

```
#Fazendo estimação inversa pontual e intervalar usando a regressão clássica
y=c()   #vetor com os valores da variável dependente usados na estimação
x=c()   #vetor com os valores da variável independente usados na estimação
alpa=0.05    ## nível de significância
y0=c()  ## vetor com os k valores da variável dependente
y0i=mean(y0)
#função para calcular o intervalo usando a regressão inversa simples
invregsimples<-function(x, y, y0, alfa)
{
n =length(x) #quantidade de observações utilizadas no ajuste
xbar=mean(x) #média de x
k=length(y0) #tamanho de y0
ybar=mean(y) #média de y
x1=x-xbar
y0bar=mean(y0) #média dos valores de y0
#Estimação do modelo centrado
reg <- lm(y~x1)
#valor estimado x0 dado uma nova observação y0
x0=xbar+(y0bar-ybar)/coef(reg)[2]
# Valores para gerar o intervalo de confiança
tc=qt((1-alfa/2),reg$df.residual)
s2yx=sum(reg$residual^2)/reg$df.residual
a= (coef(reg)[2]^2)-((s2yx*(tc^2))/(sum((x-xbar)^2)));a
m1=xbar+(coef(reg)[2]*(y0bar-ybar)/a);m1
h1=((tc)*sqrt(s2yx)/a)*sqrt(a*((1/n)+(1/k))+((y0bar-ybar)^2)/(sum((x-xbar)^2)));h1
# Gerando o intervalo de confiança x0.
li1= m1-h1;li1
ls1= m1+h1;ls1
list(x0,c(li1,ls1))
```

```
}
invregsimples(x,y,y0i,alfa=0.05)
```

ANEXO B: Código do *software* R para a estimação espacial inversa pontual e intervalar utilizando o modelo CAR.

```
###### Bibliotecas necessárias ########
rm(list=ls())
library(agricolae)
library(geoR)
library(MASS)
require(GeoXp)
library(car)
library(spdep)
library(rgdal)

################ Conjunto de dados #######
data(columbus)
columbus
Y=columbus[,9]   ## variável dependente:CRIME  ##
X=columbus[,8]   ## Variável independente: RENDA MÉDIA ##
Xm=columbus[,8]-mean(columbus[,8]) #valores centrados na média

############### Matriz de vizinhança espacial ###############
COL.listw=nb2listw(col.gal.nb, style="W")#lista de vizinhaça
COL.listw
w=listw2mat(COL.listw);w  #matriz de vizinhança padronizada
I = diag(length(columbus$INC))
I

############## Ajuste do modelo CAR ######################
ajust=errorsarlm(Y~Xm,data=columbus,nb2listw(col.gal.nb, style="W"),
etype="error", method="eigen", quiet=FALSE, tol.solve=1.0e-10)
summary(ajust)
```

```
#### Função para obter a estimativa pontual e intervalar ###
### INTERVALO DE CONFIANÇA/PREDIÇÃO PARA O MODELO CAR  ##########
#######separando o conjunto de dados em duas partes######
#r é um vetor que contém as unidades a serem consideradas na
#primeira etapa da calibração
#h é um vetor que contém as unidades utilizadas na
#segunda etapa da calibração

h=c()    ## vetor que explicita quais unidades não estão na amostra
r=c()      ## vetor que exlicita quais unidades estão na amostra
alfa=0.05    ## nível de significância para o intervalo
espacial_inv=function(h,r,i,alfa){
tcolumbus<-readOGR(dsn = system.file("etc/shapes", package="spdep"),
layer="columbus")

# Separa a região da primeira etapa
spols<-polygons(tcolumbus)[r]
# Cria lista de vizinhos a partir do objeto "Spatial Polygon"
polgal<-poly2nb(spols)
# Cria uma lista de pesos a partir da lista de vizinhos
W_polgal<-nb2listw(polgal)

Xc1=X[r]-mean(X[r]) #valores de XC centrados na média
nc=length(Xc1)
xnc=cbind(rep(1,nc),Xc1)
#####estimando o modelo com as r observações ####
ajustc=errorsarlm(columbus$CRIME[r]~Xc1,data=columbus,
nb2listw(polgal,style="W"),etype="error", method="eigen",
quiet=FALSE, tol.solve=1.0e-10)

lambdac=ajustc$lambda

###matriz de vizinhaça das regiões da primeira etapa #####
COL.listw1=nb2listw(polgal, style="W")
wc=listw2mat(COL.listw1)

### matriz indentidade e coeficientes
```

```
Ic = diag(length(columbus$INC[r]))
Bc=cbind(coef(ajustc)[2],coef(ajustc)[3])

###montando a matriz S para obter o sigma2
matrizSc=(lambdac*wc+xnc%*%solve(t(xnc)%*%xnc)%*%t(xnc)%*%(Ic-lambdac*wc))
tracomatrizSc=sum(diag(matrizSc))      #tr(S)
tracomatrizSSc=sum(diag(t(matrizSc)%*%matrizSc))  #tr(S'S)
SSEc=deviance(ajustc);  #SQResiduo
s2c=SSEc/(nc-2*tracomatrizSc+tracomatrizSSc);s2c  #variância

###regiões em que só se conhece os valores observados de y###
Y0=Y[h]
###Estimativas de componentes individuais###
##UESTIMADO é o vetor dos erros estimados conforme o final da seção (4.1)
x0i=mean(X[r])+(Y0[i]-lambdac*(w[h[[i]],h]%*%UESTIMADO+w[h[[i]],r]%*%
residuals(ajustc))-coef(ajustc)[2])/(coef(ajustc)[3])

tcc=qt((1-alfa/2),(nc-2*tracomatrizSc+tracomatrizSSc)) ## quantil  t
kc=s2c%*%(tcc^2)
at=t((solve(t(xnc)%*%t((Ic-lambdac*wc))%*%(Ic-lambdac*wc)%*%
(xnc)))%*%t(xnc)%*%t((Ic-lambdac*wc))%*%(Ic-lambdac*wc))

OO=(solve(I-lambdac*(w)))%*%(solve((I-lambdac*t(w))))
PHI=(solve(I-lambdac*(w)))%*%(solve((I-lambdac*t(w))))
M=solve(t(xnc)%*%solve(OO[r,r])%*%xnc)

## esses objetos são oriundos da expansão do expressão (4.18)
i=PHI[h[[i]],h[[i]]]+(lambdac^2)*(w[h[[i]],h]%*%OO[h,h]%*%
cbind(w[h[[i]],h])+w[h[[i]],r]%*%OO[r,r]%*%w[h[[i]],r]+
2*w[h[[i]],h]%*%OO[h,r]%*%cbind(w[h[[i]],r]))

i_i=-2*lambdac*PHI[h[[i]],h]%*% cbind(w[h[[i]],h])

i_i_i=-2*lambdac*(PHI[h[[i]],r]-PHI[h[[i]],r]%*%at%*%t(xnc))%*%
cbind(w[h[[i]],r])

i_v=-2*(PHI[h[[i]],r]-lambdac*(w[h[[i]],h]\%*\%PHI[h,r]+w[h[[i]],r]%*%
PHI[r,r]))%*%at
```

```
v=-2*lambdac*w[h[[i]],r]%*%xnc%*%M
cc=i+i_i+i_i_i

####calculando o intervalo###
l1m=Y0[i]-(lambdac)*(w[h[[i]],h]\%*\%UESTIMADO+w[h[[i]],r]%*%
residuals(ajustc))

## coeficientes do polinômio
a=Bc[,2]^2-(M[2,2]+2*lambdac*M[2,2])*kc
b=-(2*(l1m-Bc[,1])*Bc[,2])-kc*(2*M[1,2]+i_v[1,2]-v[1,2])
c=(l1m-Bc[,1])^2-kc*(M[1,1]+i_v[1,1]+v[1,1]+cc)
deltam=(b^2)-(4*a*c)
xUm=mean(X[r])+(-b+sqrt(deltam))/(2*a)
xLm=mean(X[r])+(-b-sqrt(deltam))/(2*a)
list(x0i,c(xLm,xUm))
}
espacial_inv(h,r,i=,alfa) ## i é uma posição do vetor h
```