Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

ANÁLISE DOS MODELOS DE REGRESSÃO ESPACIAL SAR, SEM E SAC

CAIO VIEIRA RÊGO 09/07979
MARINA GARCIA PENA 09/13383

Brasília

2012

Caio Vieira Rêgo **09/07979**
Marina Garcia Pena **09/13383**

ANÁLISE DOS MODELOS DE REGRESSÃO ESPACIAL SAR, SEM E SAC

Relatório elaborado na disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para conclusão do curso e obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Co-Orientador: Prof. Pedro Henrique Melo Albuquerque

Brasília

2012

Dedico este trabalho à minha família, pela paciência e disposição nos momentos de reclusão. Aos amigos pela distração nos momentos em que foi necessária. Aos colegas de colação pelo exemplo. E aos co-autores, a amiga Marina e o Prof° Alan Silva , pelo trabalho árduo e dedicação.

Caio Vieira Rêgo

Dedico este trabalho a meus pais, pelo amor, incentivo e dedicação que sempre me passaram, por tudo que eles representam na minha vida; a meus irmãos, por sempre estarem ao meu lado como exemplos de vida para mim; a todos os meus familiares que estiveram ao meu lado durante esse tempo; ao meu colega de trabalho e amigo, Caio, e a todos os meus amigos, que tornaram essa experiência única.

Marina Garcia Pena

Agradecimentos

A Deus, por ter nos iluminado nos momentos de dificuldade e nos dado força para concluir mais uma etapa de nossas vidas.

Ao professor orientador Alan, por estar ao nosso lado durante todo o trabalho, sempre nos auxiliando e nos motivando para a realização de um trabalho cada vez melhor e pelo exemplo de excelência profissional que ele nos deu durante todo nosso caminho.

Aos nossos pais, por estarem sempre ao nosso lado, nos apoiando no que fosse preciso, nos incentivando e, o mais importante, nos dando carinho e condições de realizarmos esse sonho.

Aos nossos amigos e colegas pelo incentivo e apoio constantes.

Aos nossos familiares pelas orações e torcida.

A todos os professores que passaram por nosso caminho, por transmitirem seus conhecimentos a nós e dividirem suas experiências conosco.

Resumo

A utilização de técnicas de estatística espacial é algo recorrente em pesquisas modernas. Uma parte importante dessa área é a regressão espacial. Três modelos amplamente disseminados são o *Spatial Autoregressive Model - SAR*, o *Spatial Error Model - SEM* e o *General Spatial Model - SAC*. Nesses modelos há coeficientes que representam a dependência espacial ou seja, neles, as informações dos "vizinhos" é utilizada para prever ou "explicar" o que está sendo estudado.

A distinção e escolha entre os modelos SAR e SEM não é simples, devido ao fato de eles possuirem uma formulação parecida. Nota-se que ao se desenvolver a estrutura do SEM, se torna um caso particular do SAR, diferindo, entretanto, na interpretação final do resultado. Foi verificado, por meio de análise empírica, que o modelo SAR, em geral, se ajusta melhor e resulta em R^2 mais altos que o SEM. Como esperado, dados com baixa ou nenhuma dependência espacial não geram modelos SAR e SEM com coeficientes significativos.

O modelo SAC foi aplicado a diferentes bancos de dados, o resultado obtido indica que esse modelo é significativo e tem bom ajuste apenas em modelos com alta dependência espacial. A matriz de proximidade binária gerou melhores resultados nos bancos em que o modelo SAC foi aplicado, sua estrutura, mais simples, não gerou coeficientes maiores que $|1|$ em nenhum caso. A matriz de distâncias, nos dados com dependência espacial elevada, exige medida corretiva. Menores distâncias de corte geram índices de dependência espacial mais altos e maiores R^2 porem influem negativamente nas variáveis não espaciais do modelo. É importante se atentar aos valores

do intercepto em cada configuração do modelo. Em alguns casos em que foram utilizadas matrizes iguais observou-se inversões de sinal no intercepto - tomando-se como base a regressão clássica -, problema que é recorrente nos casos de multicolinearidade. Essa inversão pode ser causada por se utilizar a mesma estrutura para se "retirar" a dependência da variável respostas e do erro. Se a matriz $\mathbf{W_1}$ não foi capaz de esgotar a dependência espacial do modelo é recomendado o uso de uma matriz $\mathbf{W_2}$ diferente para sanar essa dependência. Possíveis medidas corretivas em alguma matriz podem limitar o modelo: distâncias de corte muito curtas são influentes nos p-valores do teste de significância dos parâmetros do modelo, que devem ser analisados com cuidado redobrado em situações como essa.

Sumário

RESUMO **iv**

1 INTRODUÇÃO **1**

1.1 OBJETIVOS 3

2 MODELOS DE REGRESSÃO ESPACIAL **5**

2.1 INTRODUÇÃO 5

2.2 MATRIZ DE PROXIMIDADE ESPACIAL 5

2.3 ÍNDICE I DE MORAN E C DE GEARY 9

2.4 MODELO SAC 11

2.5 MODELO SAR 15

2.6 MODELO SEM 18

2.7 MODELO FAR 21

3 ANÁLISE ESTRUTURAL **23**

3.1 INTRODUÇÃO 23

3.2 COMPARAÇÃO ENTRE OS MODELOS SAR E SEM 23

3.3 ANÁLISE ESTRUTURAL DO MODELO SAC 24

3.4 MÉTODOS DE SELEÇÃO DO MODELO ESPACIAL 25

4 ANÁLISE EMPÍRICA E RESULTADOS **32**

4.1 INTRODUÇÃO 32

4.2 Simulações 32

4.3 ANÁLISE EMPÍRICA 33

4.4 RESULTADOS SAC 39

4.5 RESULTADOS DA COMPARAÇÃO ENTRE SAR E SEM 48

5 CONCLUSÃO **56**

Referências **59**

A Programação SAS para os modelos espaciais **60**

B Programação SAS as simulações dos bancos com Máxima e Mínima dependência espacial **75**

Lista de Tabelas

3.1 Exemplo Goiás - Parâmetros SAR 27

3.2 Exemplo Goiás - Parâmetros SEM 27

3.3 Exemplo Goiás - Parâmetros SAC com matrizes iguais 27

3.4 Exemplo Goiás - Parâmetros SAC com matrizes diferentes 29

3.5 Exemplo Columbus - Parâmetros SAR 30

3.6 Exemplo Columbus - Parâmetros SEM 30

3.7 Exemplo Columbus - Parâmetros SAC com matrizes iguais 30

3.8 Exemplo Columbus - Parâmetros SAC com matrizes diferentes 31

4.1 Exemplo Máxima Dependência Espacial - Ajuste 36

4.2 Exemplo Máxima Dependência Espacial - Parâmetros 36

4.3 Exemplo Mínima Dependência Espacial - Ajuste 37

4.4 Exemplo Mínima Dependência Espacial - Parâmetros 37

4.5 Exemplo Goiás - Ajuste 38

4.6 Exemplo Goiás - Parâmetros 38

4.7 Exemplo Columbus - Ajuste 39

4.8 Exemplo Columbus - Parâmetros 39

4.9 Exemplo Máxima Dependência Espacial - Medidas de Ajuste vs Distância de Corte 41
4.10 Exemplo Máxima Dependência Espacial - Medidas de Ajuste vs Distância de Corte - Parâmetros 42
4.11 Exemplo Máxima Dependência Espacial - Ajuste 44
4.12 Tabela:Exemplo Máxima Dependência Espacial - Parâmetros 44
4.13 Exemplo Máxima Dependência Espacial - Ajuste 47
4.14 Exemplo Máxima Dependência Espacial - Columbus(dist. de corte=85) - Parâmetros 47
4.15 Exemplo I de moran máximizado - Comparação do ajustamento - SAR e SEM 49
4.16 Exemplo I de Moran Maximizado - Comparação dos parâmetros - SAR e SEM 49
4.17 Exemplo I de moran minimizado - Comparação do ajustamento - SAR e SEM 49
4.18 Exemplo I de Moran Minimizado - Comparação dos parâmetros - SAR e SEM 50
4.19 Exemplo Goiás - Comparação do ajustamento - SAR e SEM 50
4.20 Exemplo Goiás - Comparação dos parâmetros - SAR e SEM 50
4.21 Exemplo Columbus - Comparação do ajustamento - SAR e SEM . . . 51
4.22 Exemplo Columbus - Comparação dos parâmetros - SAR e SEM . . . 51
4.23 Exemplo Rio de Janeiro - Comparação do ajustamento - SAR e SEM 52

4.24 Exemplo Rio de Janeiro - Comparação dos parâmetros - SAR e SEM 52

4.25 Comparação SAR e SEM - Utilização do R^2 do FAR 54

Lista de Figuras

2.1 Mapa exemplo da matriz de vizinhança 7

3.1 Esquema do Método Forward 26

3.2 Esquema do Método de Hendry 29

4.1 Matriz Binária Padronizada vs Matiz de Distância Padronizada . . . 34

4.2 Matriz Binária Padronizada vs Matiz de Distância Padronizada . . . 35

4.3 I de Moran vs Distâncias de Corte 43

4.4 Dependência Maximizada 46

Capítulo 1

INTRODUÇÃO

Os modelos de regressão são ferramentas estatísticas largamente utilizadas em todas as áreas das ciências. Por esse motivo, diferentes técnicas e diferentes modelos vêm sendo estudados ao longo do tempo. Um segmento ainda novo da estatística é a estatística espacial. Klaassen and Paelinck (1979) publicaram um trabalho que foi considerado a primeira tentativa de delinear a econometria espacial. Desde então, estudos importantes vem sendo feitos na área, e vários modelos espaciais são hoje largamente utilizados.

A estatística espacial é um ramo que leva em conta nas suas análises informações geográficas, introduzidas por meio das matrizes de contiguidade (também chamada de matriz de vizinhança). As matrizes de vizinhança informam se uma certa área geográfica é considerada ou não vizinha de outra. Pode ser considerado vizinho aquele polígono que possui um ou mais pontos em comum com o polígono analisado (vizinhanças *Queen* e *Rook*), ou ainda pode ser adotada uma matriz de distâncias para a atribuição de vizinhança. Sendo assim, ao se realizar uma análise de uma certa área, a informação dos seus vizinhos é de alguma forma incorporada nos resultados, trazendo uma maior robustes ao trabalho.

As regressões espaciais são modelos que possuem a matriz de vizinhança em algum dos seus parâmetros. Os modelos mais conhecidos são *Spatial Autoregressive Model (SAR)*, *Spatial Error Model (SEM)* e *General Spatial Model (SAC)*. No modelo SAR a variável dependente $\mathbf{y}$ é explicada por seus vizinhos e por outras covariáveis. Ou seja, a informação dos vizinhos é introduzida também como variável explicativa. Sua formulação é dada por:

$$\mathbf{y} = \rho \mathbf{W_1} \mathbf{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{1.1}$$

onde:

i - $\mathbf{y}$ é a variável dependente;

ii - ρ é parâmetro espacial responsável pela mensuração do grau de dependência espacial da variável dependente e seus respectivos vizinhos;

iii - $\mathbf{W_1}$ é a matriz de vizinhança;

iv - $\mathbf{X}$ são as variáveis independentes;

v - $\boldsymbol{\beta}$ são os coeficientes da regressão;

vi - $\boldsymbol{\epsilon}$ é o erro aleatório;

vii - σ^2 é a variância do modelo;

viii - $\mathbf{I}$ é uma matriz identidade.

Diferentemente do modelo SAR, o modelo SEM introduz a informação de vizi-

nhança apenas no erro aleatório. A formulação geral é dada por:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{X\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W_2 u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 \mathbf{I}) \end{aligned} \tag{1.2}$$

onde:

i - $\mathbf{u}$ é o erro aleatório;

ii - $\mathbf{W_2}$ é a matriz de vizinhança;

iii - λ é parâmetro espacial.

O modelo SAC é uma generalização dos dois modelos acima. A estrutura de vizinhança W aparece tanto como variável explicativa como no erro aleatório, conforme a expressão geral dada por:

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W_1 y} + \boldsymbol{X\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W_2 u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \end{aligned} \tag{1.3}$$

Conforme Anselin (1988), não existe uma distinção clara entre os modelos SAR e SEM. Além disso, segundo o autor, o modelo SAC sofre de problemas quando as matrizes $\mathbf{W_1}$ e $\mathbf{W_2}$ são iguais. Dessa forma, este trabalho irá explorar os três modelos de regressão espacial apresentados acima, buscando explicar tais problemas.

1.1 OBJETIVOS

O objetivo geral do trabalho é analisar a estrutura dos modelos de regressão espacial SAR, SEM e SAC.

Os objetivos específicos são:

- Apresentar os modelos de regressão espacial;
- Verificar as diferenças existentes entre os modelos SAR e SEM;
- Estudar o modelo SAC;
- Implementação dos algoritmos dos modelos SAR, SEM e SAC no *software* SAS 9.2.

Capítulo 2

MODELOS DE REGRESSÃO ESPACIAL

2.1 INTRODUÇÃO

Há uma larga gama de modelos de regressão espacial. Além disso, a própria variedade de características que a(s) matriz(es) de vizinhança pode(m) representar é um prenúncio da polivalência dos modelos dessa espécie.

Esse capítulo apresenta os modelos que serão trabalhados com os objetivos propostos na introdução - SAC, SAR e SEM -, bem como os estimadores de seus parâmetros mais importantes.

2.2 MATRIZ DE PROXIMIDADE ESPACIAL

Nos modelos a serem estudados neste trabalho a espacialidade é introduzida por meio da estrutura de vizinhança $\mathbf{W}$. As estruturas de vizinhança são matrizes $n \times n$ que indicam quais são os vizinhos de cada polígono i. Essa matriz é denominada matriz de proximidade espacial.

Por definição, um polígono nunca será vizinho dele mesmo, portanto a diagonal da matriz será sempre igual a zero. Usualmente atribui-se o valor 1 para indicar que a

área i é vizinha da área j, com $i \neq j$ - essa é a chamada matriz de vizinhança binária. Pode-se também normalizar a matriz de forma com que a soma dos elementos da linha seja igual a 1. Neste caso faz-se cada elemento $w_{ij} = \frac{1}{\sum_j w_{ij}}$, e a matriz passa a ser chamada de matriz normalizada ou matriz padronizada ($\mathbf{W_{Pdr}}$).

A vizinhança pode ser atribuída de várias maneiras diferentes, sendo algumas delas matrizes de forma discreta. Essas são as matrizes binárias, compostas pr 0's e 1's:

I - Na vizinhança do tipo *Rook* é considerado vizinho aquele polígono P_j que possui pelo menos um lado em comum com o polígono P_i;

II - No tipo de vizinhança *Queen* P_j é vizinho de P_i se eles possuem ao menos um ponto em comum;

III - Na vizinhança por distância determinam-se como vizinhos aqueles polígonos cujos centroides se encontrarem a uma determinada distância d_{ij} de P_i;

IV - Determinam-se os vizinhos como os k polígonos com os centroides mais próximos de P_i.

As matrizes binárias também são chamadas de matrizes de vizinhança ou matrizes de contiguidade. Essas matrizes podem também ser normalizadas, como dito anteriormente. Além das matrizes cujos elementos são discretos, temos aquelas em que os elementos w_{ij} são contínuos. Como exemplo temos as seguintes formas:

I - Os elementos da matriz $\mathbf{W}$ são funções do tamanho da fronteira l_{ij} entre P_i e P_j: $w_{ij} = \frac{l_{ij}}{l_i}$;

II - Cada elemento w_{ij} é uma função do tempo t_{ij} que se leva de uma região a outra : $w_{ij} = \frac{1}{1+t_{ij}}$;

III - Cada elemento w_{ij} é uma função da distância entre centroides d_{ij}: $w_{ij} = \frac{1}{1+d_{ij}}$;

IV - Existem outras formas como a citada acima que ao invés de se utilizar a distância utiliza-se fluxos comerciais ou fluxos migratórios por exemplo.

As matrizes de proximidade espacial mais utilizadas são as matrizes binárias. Nota-se que na prática não existe muita diferença entre a matriz binária do tipo *Queen* e a do tipo *Rook*, pois se tratando de áreas geográficas (como municípios, estados, etc.) raramente existirão casos em que um polígono toca em apenas um ponto seu vizinho (caso em que a vizinhança só é captada pela matriz *Queen*). Na Figura 2.1 temos um exemplo de como funciona a matriz de vizinhança:

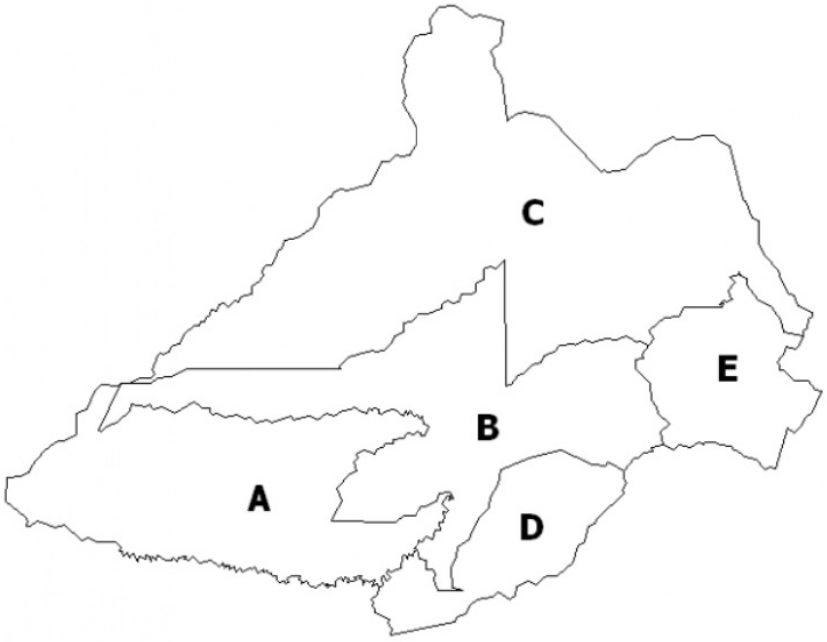

Figura 2.1: Mapa exemplo da matriz de vizinhança

$$\mathbf{W} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array} \qquad \mathbf{W}_P dr = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 0,5 & 0 & 0,5 & 0 \\ 0,25 & 0 & 0,25 & 0,25 & 0,25 \\ 0 & 0,5 & 0 & 0 & 0,5 \\ 0,5 & 0,5 & 0 & 0 & 0 \\ 0 & 0,5 & 0,5 & 0 & 0 \end{pmatrix} \end{array}$$

Como o polígono A da Figura 2.1 só apresenta dois vizinhos (B e D), os elementos da linha que representa A foi dividido por dois, resultando em um peso de $\frac{1}{2}$ para cada vizinho. Já o polígono B possui quatro vizinhos (A, C, D e E). Portanto, a linha que o representa foi dividida por quatrp, resultando em um peso de $\frac{1}{4}$. Seguindo-se o raciocínio para todos os polígonos obtem-se a matriz de vizinhança binária padronizada.

Outra propriedade da matriz de proximidade espacial é a ordem da matriz. A matriz de primeira ordem é aquela matriz que considera como vizinhos apenas os vizinhos diretos do polígono i. Já uma vizinhança de segunda ordem considera não apenas os vizinhos diretos, mas também os vizinhos dos vizinhos, e assim por diante. Quando não é mencionado nada a respeito da ordem da vizinhança considera-se como sendo primeira ordem.

Ao se utilizar um modelo de regressão espacial e verificar-se a não significância do parâmetro de espacialidade não necessariamente existe ausência de dependência espacial. A única evidência que se tem nesse caso é de que a matriz de proximidade espacial utilizada não conseguiu capturar a dependência espacial. Essa dependência talvez possa ser capturada se utilizada uma matriz de vizinhança diferente. No

trabalho de Silva (2007) nota-se que a utilização de uma matriz binária retornou ausência de dependência espacial no modelo, entretanto ao mudar a matriz para uma matriz **W** de tempo, verificou-se a existência de dependência espacial (baixa dependência, mas ainda sim existente).

Após se estruturar a matriz de proximidade espacial desejada, é importante saber se ela será capaz de detectar a presença de autocorrelação espacial. No próximo tópico serão introduzida formas de se detectar previamente a presença dessa autocorrelação.

2.3 ÍNDICE *I* DE MORAN E *C* DE GEARY

É importante detectar a presença ou não de dependência espacial antes de se utilizar um modelo de regressão espacial para a modelagem dos dados. Quando a dependência espacial é geográfica pode-se fazer uma análise exploratória para uma prévia investigação. Essa análise exploratória - também chamada de mapa temático - é uma representação visual da variável de interesse em um mapa. Porém essa é uma análise inicial do processo, e não dá nenhuma certeza acerca da dependência espacial. Para uma análise inferencial sobre a existência de dependência espacial pode-se calcular o índice I de Moran, proposto por Moran (1950).

$$I = \frac{n \sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_i (y_i - \bar{y})^2\right) \left(\sum_i \sum_j w_{ij}\right)} \tag{2.1}$$

onde:

i - y_i é o valor da variável y na região i;

ii - y_j é o valor da variável y na região j;

iii - $\bar{y}$ é a média da variável y;

iv - w_{ij} é o elemento ij da matriz de proximidade espacial;

v - n é o número de observações.

O índice acima foi derivado a partir da mesma ideia do índice de correlação de Pearson, que é normalmente utilizado para verificar a correlação entre variáveis, em um modelo de regressão não-espacial. O índice I de Moran varia entre -1 e 1. O valor zero indica ausência de dependência espacial. Já valores próximos de 1 indicam uma autocorrelação espacial forte e positiva. O mesmo vale para valores próximos a -1, só que neste caso a autocorrelação espacial é negativa.

Outro índice que pode ser usado para se detectar a dependência espacial é o índice C de Geary, proposto por Geary (1954). Sua formulação é dada por:

$$C = \frac{n-1}{2} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\left(\sum_i (y_i - \bar{y})^2\right) \left(\sum_i \sum_j w_{ij}\right)} \tag{2.2}$$

onde os parâmetros são como na Equação 2.1.

O resultado do índice C de Geary é semelhante ao I de Moran. Entretanto, seu valor varia entre 0 e 2, sendo 0 uma forte autocorrelação espacial positiva e 2 uma forte autocorrelação espacial negativa. O valor 1 representa a ausência de autocorrelação espacial. Segundo Lembo (2005), o coeficiente G de Geary é preferido ao I de Moran quando existe uma pequena quantidade de vizinhanças.

Os índices apresentados acima são ditos índices de dependência espacial global. Eles podem servir como um teste prévio para saber se é indicada a utilização de um modelo de regressão espacial. Caso esses índices indiquem que não existe autocor-

relação espacial os parâmetros espaciais dos modelos de regressão espacial não serão significativos. Nestes casos é indicada a utilização de uma matriz de proximidade espacial diferente ou então a utilização de um modelo de regressão convencional.

Introduzido o conceito de matriz de proximidade espacial e dos índices globais de dependência espacial nos próximos tópicos serão apresentados quatro modelos de regressão espacial.

2.4 MODELO SAC

O modelo SAC, Equação 2.3, utiliza as duas matrizes de vizinhança $\mathbf{W_1}$ e $\mathbf{W_2}$ para estimar parte do valor predito e o erro, respectivamente. De outra forma, as observações vizinhas ajudam a explicar determinado valor e sua variância. Entretanto, um problema do modelo SAC pode aparecer quando as matrizes de vizinhança são iguais, fato que será discutido em capítulos seguintes.

Os modelos de regressão espacial também estão sujeitos às três suposições necessárias para a validação de um modelo de regressão:

- erros normais com média zero;
- homocedásticidade;
- erros não correlacionados.

Esse modelo é usado quando o modelo SAR (*Spatial Autoregressive Model*) apresenta evidências de erro com dependência espacial. Essa dependência deve ser investigada a partir de testes como o LM para correlação espacial do erro.

A formulação do modelo SAC é dada por:

$$\mathbf{y} = \rho\mathbf{W_1}\mathbf{y} + \boldsymbol{X\beta} + \mathbf{u}$$

$$\mathbf{u} = \lambda\mathbf{W_2}\mathbf{u} + \boldsymbol{\epsilon} \tag{2.3}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

onde os parâmetros são como na Equação 1.3.

A partir de (2.3), condicionando a função a variadas restrições, se obtém outros modelos:

I - Com $\rho = 0$ e $\lambda = 0$, o modelo resultante nada mais é que o modelo clássico de regressão linear:

$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \epsilon$

II - Com $\lambda = 0$, teremos como resultado o modelo SAR(*Spatial Autoregressive Model*)

$$\mathbf{y} = \rho\mathbf{W}_1 y + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.4}$$

III - Restringindo a Equação 2.3 de forma que $\rho = 0$, teremos como resultado o Modelo SEM(*Spatial Error Model*):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \lambda\mathbf{W}_2)^{-1}\epsilon \tag{2.5}$$

Agora, retornaremos aos estimadores dos parâmetros do Modelo Espacial Geral, o SAC. Os parâmetros foram obtidos por meio do método da máxima verossimilhança e seu cálculo está demonstrado em Anselin (1988) e Silva (2006).

Tomando as seguintes equações:

$$\begin{aligned} \boldsymbol{\epsilon} &= \mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ \mathbf{A} &= (\mathbf{I_n} - \boldsymbol{\rho}\mathbf{W_1}) \\ \mathbf{B} &= (\mathbf{I_n} - \boldsymbol{\lambda}\mathbf{W_2}) \end{aligned} \tag{2.6}$$

Os estimadores para os parâmetros do modelo em questão são da forma:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{B}'\mathbf{B}\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}'\mathbf{B}\mathbf{A}\mathbf{y} \tag{2.7}$$

O estimador da variância do modelo, σ^2:

$$\sigma^2 = ((\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{B}'\mathbf{B}(\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))/n \tag{2.8}$$

Para estimar ρ utiliza-se os seguintes passos:

1. Fazer mínimos quadrados ordinários (OLS) no modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon_0}$

2. Fazer mínimos quadrados ordinários (OLS) no modelo $\mathbf{W_1}\mathbf{y} = \mathbf{X}\boldsymbol{\beta_L} + \boldsymbol{\epsilon_L}$

3. Obter os resíduos $\mathbf{e_0} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\mathbf{0}$ e $\mathbf{e_L} = \mathbf{W_1}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$

4. Após $\mathbf{e_0}$ e $\mathbf{e_L}$ calculados, deve-se achar o ρ que maximize

   $\ln(L) = -\frac{n}{2}\ln\left(\frac{1}{n}(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L})'(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L})\right) + \ln|\mathbf{I} - \rho\mathbf{W_1}|$

A estimação de λ também depende de um algoritmo:

1. Fazer mínimos quadrados ordinários (OLS) no modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$;

2. Obter os resíduos: $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$;

3. Procurar o valor de λ que maximiza a função de verossimilhança condicionada aos valores dos $\hat{\boldsymbol{\beta}}$ encontrados

   $\ln(L) = -\frac{n}{2}\ln\left(\frac{1}{n}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\mathbf{I} - \lambda\mathbf{W_2})'(\mathbf{I} - \lambda\mathbf{W_2})(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right) + \ln|\mathbf{I} - \lambda\mathbf{W_2}|$

4. Atualizar os valores dos $\hat{\boldsymbol{\beta}}$ usando o valor de $\hat{\lambda}$ calculado. Para obter o novo valor de $\hat{\boldsymbol{\beta}}$ pode-se usar mínimos quadrados generalizados

   $\hat{\boldsymbol{\beta}} = [((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})'((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})]^{-1}((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})'((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{y})$

5. Voltar para o passo 3 até obter a convergência dos resíduos.

A partir da Matriz de Informação de Fisher (mais especificamente a inversa dela), derivam-se os erros-padrão. Os estimadores de máxima verossimilhança atingem o limite de Cramer-Rao e, portanto são eficientes. A demonstração detalhada de como são obtidos está em Silva (2006).

$$
\begin{aligned}
&-E\left(\frac{\partial^2 Ln(L)}{\partial(\sigma^2)^2}\right) = \frac{n}{2(\sigma^2)^2} \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho^2}\right) = tr((\mathbf{I_n} - \rho\mathbf{W_1})^{-1}\mathbf{W_1}(\mathbf{I_n} - \rho\mathbf{W_1})^{-1}\mathbf{W_1}) \\
&+\frac{1}{\sigma^2}tr(\mathbf{W_1}'(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2})\mathbf{W_1}(\mathbf{I_n} - \rho\mathbf{W_1})^{-1}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta})(\mathbf{I_n} - \rho\mathbf{W_1})'^{-1}) \\
&+tr((\mathbf{W_1}'(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2})\mathbf{W_1}[(\mathbf{I_n} - \lambda\mathbf{W_2})(\mathbf{I_n} - \rho\mathbf{W_1})]'^{-1} \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda^2}\right) = tr((\mathbf{I_n} - \lambda\mathbf{W_2})^{-1}\mathbf{W_2}(\mathbf{I_n} - \lambda\mathbf{W_2})^{-1}\mathbf{W_2}) \\
&+tr(\mathbf{W_2'}\mathbf{W_2}((\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2}))^{-1}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\beta^2}\right) = \frac{\mathbf{X}'(\mathbf{I_n}\lambda\mathbf{W_2})'(\mathbf{I_n}\lambda\mathbf{W_2})\mathbf{X}}{\sigma^2} \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho\partial\sigma^2}\right) = \frac{1}{\sigma^2}tr(\mathbf{W_1'}(\mathbf{I_n} - \rho\mathbf{W_1'})'^{-1}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda\partial\sigma^2}\right) = \frac{1}{\sigma^2}tr(\mathbf{W_2'}(\mathbf{I_n} - \lambda\mathbf{W_2'})'^{-1}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\boldsymbol{\beta}\partial\sigma^2}\right) = 0 \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho\partial\lambda}\right) = tr(\mathbf{W_1'}\mathbf{W_2}[(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \rho\mathbf{W_1})]'^{-1}) \\
&+tr(\mathbf{W_2'}\mathbf{I_n} - \lambda\mathbf{W_2})\mathbf{W_1}(\mathbf{I_n} - \rho\mathbf{W_1})^{-1}[(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2})]^{-1}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho\partial\boldsymbol{\beta}}\right) = \frac{1}{\sigma^2}(\mathbf{X}'(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2})) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda\partial\boldsymbol{\beta}}\right) = 0
\end{aligned}
\tag{2.9}
$$

A próxima seção iniciará um dos modelos derivados do SAC, o SAR, que apresenta a dependência espacial apenas como variável explicativa.

2.5 MODELO SAR

O modelo de regressão espacial SAR consiste em um modelo de regressão em que uma das variáveis explicativas possui uma dependência espacial com a variável a ser explicada. Ele é um caso particular do SAC onde o parâmetro espacial λ assume valor igual a zero. Diferentes aplicações podem ser feitas por meio deste modelo. As principais são na área econômica. Um exemplo de utilização do modelo é no estudo da renda de um município. Utilizando-se o modelo SAR, a variável renda (variável dependente) é explicada, além das outras covariáveis, pela renda dos vizinhos. Similarmente estudos na área de saúde e educação, por exemplo, também podem ser feitos. Portanto, a utilização modelo SAR se torna interessante quando se está realizando estudos de políticas públicas (planejamento) de uma certa área.

A formulação do modelo é dada por:

$$\mathbf{y} = \rho \mathbf{W_1 y} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{2.10}$$

onde os parâmetros são como na Equação 1.1.

O que diferencia o SAR de um modelo de regressão linear convencional é o parâmetro espacial ρ. Se esse parâmetro assumir valor zero existe ausência de dependência espacial nessa variável, ou seja, os vizinhos não exercem influência no valor da variável estudada e os resultados serão similares à regressão clássica. Esse parâmetro está presente, também, na estimação do vetor dos coeficientes da regressão $\boldsymbol{\beta}$. Esse vetor é estimado da seguinte forma:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Iy} - \rho(\mathbf{X'X})^{-1}\mathbf{X'W_1 y} \tag{2.11}$$

Nota-se que, se o parâmetro espacial for igual ou aproximadamente igual a zero, existirá pouca diferença entre o vetor $\hat{\boldsymbol{\beta}}$ estimado pelo modelo SAR e o estimado por um modelo de regressão convencional. Ou seja, temos que neste caso a utilização do modelo de regressão espacial agregou pouca informação à estimação. Entretanto, o parâmetro $\hat{\boldsymbol{\beta}}$ pode ser visto como correção do viés associado à endogeinidade do modelo espacial.

A estimação do parâmetro espacial ρ é feita por meio de um algoritmo composto por quatro passos:

I - Fazer uma regressão por mínimos quadrados no modelo $\mathbf{y} = \boldsymbol{X\beta_0} + \boldsymbol{\epsilon_0}$;

II - Realizar o mesmo procedimento de mínimos quadrados no modelo $\mathbf{W_1 y} = \boldsymbol{X\beta_L} + \boldsymbol{\epsilon_L}$;

III - Calcular os resíduos dos modelos acima: $\boldsymbol{\epsilon_0} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ e $\boldsymbol{\epsilon_L} = \mathbf{W_1 y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$;

IV - Calcular ρ que maximize a função

$\ln(L) = -\frac{n}{2}\ln\left(\frac{1}{n}(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L})'(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L})\right) + \ln|\mathbf{I} - \rho\mathbf{W_1}|$

Por fim, a estimativa de σ^2 é dada por:

$$\hat{\sigma^2} = \frac{1}{n}(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L})'(\boldsymbol{\epsilon_0} - \rho\boldsymbol{\epsilon_L}) \tag{2.12}$$

As demonstrações das equações acima podem ser encontradas em Anselin (1988) e Silva (2006).

As variâncias dos parâmetros, novamente, são estimadas a partir de equações derivadas da Matriz de Informação de Fisher (Silva, 2006). Aqui serão apresentadas:

$$-E\left(\frac{\partial^2 Ln(L)}{\partial(\sigma^2)^2}\right) = \frac{n}{2(\sigma^2)^2}$$

$$\begin{aligned}
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho^2}\right) = tr((\mathbf{I_n} - \rho\mathbf{W_1})^{-1}\mathbf{W_1}(\mathbf{I_n} - \rho\mathbf{W_1})^{-1}\mathbf{W_1}) \\
&+\frac{1}{\sigma^2}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I_n} - \rho\mathbf{W_1})'^{-1}\mathbf{W_1}'\mathbf{W_1}((\mathbf{I_n} - \rho\mathbf{W_1})^{-1}(\mathbf{X}\boldsymbol{\beta})) \\
&+tr((\mathbf{I_n} - \rho\mathbf{W_1})'(\mathbf{I_n} - \rho\mathbf{W_1})^{-1}\mathbf{W_1}'\mathbf{W_1})
\end{aligned}$$

$$-E\left(\frac{\partial^2 Ln(L)}{\partial\beta^2}\right) = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}$$

$$\begin{aligned}
&-E\left(\frac{\partial^2 Ln(L)}{\partial\rho\partial\sigma^2}\right) = \frac{1}{(\sigma^2)^2}(\mathbf{X}\boldsymbol{\beta})(\mathbf{I}_n - \rho\mathbf{W_1})'^{-1}\mathbf{W_1}'(\mathbf{I}_n - \rho\mathbf{W_1})^{-1}(\mathbf{X}\boldsymbol{\beta}) \\
&+\frac{1}{(\sigma^2)}tr(\mathbf{W}_1'((\mathbf{I}_n - \rho\mathbf{W_1})'(\mathbf{I}_n - \rho\mathbf{W_1}))^{-1}) \\
&-\frac{1}{(\sigma^2)^2}\rho(\mathbf{I}_n - \rho\mathbf{W_1})'^{-1}\mathbf{W}_1'\mathbf{W}_1(\mathbf{I}_n - \rho\mathbf{W_1})^{-1}(\mathbf{X}\boldsymbol{\beta}) \\
&-\frac{1}{(\sigma^2)}tr(((\mathbf{I}_n - \rho\mathbf{W_1})'(\mathbf{I}_n - \rho\mathbf{W_1}))^{-1})\mathbf{W}_1'\mathbf{W}_1)
\end{aligned} \tag{2.13}$$

$$-E\left(\frac{\partial^2 Ln(L)}{\partial\boldsymbol{\beta}\partial\sigma^2}\right) = 0$$

$$-E\left(\frac{\partial^2 Ln(L)}{\partial\rho\partial\boldsymbol{\beta}}\right) = \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{W}_1(\mathbf{I}_n - \rho\mathbf{W}_1)^{-1}\mathbf{X}\boldsymbol{\beta}$$

Caso não seja detectada a dependência espacial na variável explicativa é possível que ela exista no erro aleatório. Neste caso, o modelo de regressão a ser utilizado será o SEM, apresentado na seção a seguir.

2.6 MODELO SEM

O modelo SEM, também chamado de Modelo de Autocorrelação Espacial no Erro, não possui a informação de vizinhança como variável, e sim no erro aleatório do modelo. Por esse motivo, sua compreensão se torna mais complicada. Sendo assim, o papel da informação dos vizinhos no modelo não é tão facilmente visualizado como no caso do SAR. Este modelo também pode ser considerado como um caso particular do Modelo Geral (SAC), quando $\rho = 0$.

Pelo fato de o parâmetro espacial não ter uma influência direta em nenhuma

das variáveis explicativas do modelo, gera-se uma dúvida de quando se utilizar o modelo SEM. Outra questão é sua semelhança ao SAR, e em quais casos deve-se preferi-lo ao modelo SAR. Uma primeira indicação que se pensa é a falta de variáveis explicativas, que força com que a dependência espacial seja introduzida apenas no erro aleatório. A comparação entre os dois modelos e os casos em que cada um trará um resultado mais robusto será feita no capítulo 3.

A formulação do modelo SEM é dada por:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{X\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W_2 u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 \mathbf{I}) \end{aligned} \tag{2.14}$$

onde seus parâmetros são como na Equação 1.2.

Quando comparado a um modelo de regressão linear, nota-se que a diferença entre ele e o modelo SEM é a presença de um parâmetro espacial λ no erro aleatório. É interessante perceber que o $\mathbf{u}$ que aparece na equação segue o modelo autoregressivo de primeira ordem *FAR*, que é basicamente o modelo SAR porém parametrizado de forma que $\mathbf{y}$ são desvios com relação a média.

Novamente, se o parâmetro espacial for igual a zero tem-se resultados semelhantes ao de uma regressão convencional. Apesar de não aparecer explicitamente no modelo como variável explicativa, o parâmetro espacial λ aparece na estimação da matriz $\boldsymbol{\beta}$ dos coeficientes da regressão. Portanto, esse parâmetro faz com que a estimativa da matriz $\boldsymbol{\beta}$ do modelo SEM seja diferente da matriz do modelo não espacial. Tal matriz é obtida como mostrado a seguir. Sua demonstração (assim como as dos

demais parâmetros) se encontram em Anselin (1988) e Silva (2006).

$$\hat{\boldsymbol{\beta}} = [((\mathbf{I} - \lambda\mathbf{W_2})\mathbf{X})'((\mathbf{I} - \lambda\mathbf{W_2})\mathbf{X})]^{-1}\,((\mathbf{I} - \lambda\mathbf{W_2})\mathbf{X})'((\mathbf{I} - \lambda\mathbf{W_2})\mathbf{X})\mathbf{y} \tag{2.15}$$

Os passos para a estimação do parâmetro espacial λ são:

1. Fazer mínimos quadrados ordinários no modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$;

2. Obter os resíduos do modelo de regressão acima: $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$;

3. Maximizar λ na função de verossimilhança condicionada aos valores dos $\hat{\boldsymbol{\beta}}$ encontrados

   $\ln(L) = -\frac{n}{2}\ln\left(\frac{1}{n}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\mathbf{I} - \lambda\mathbf{W_2})'(\mathbf{I} - \lambda\mathbf{W_2})(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right) + \ln|\mathbf{I} - \lambda\mathbf{W_2}|$

4. Atualizar os valores dos $\hat{\boldsymbol{\beta}}$ usando o valor de $\hat{\lambda}$ obtido. Para obter o novo valor de $\hat{\boldsymbol{\beta}}$ pode-se usar mínimos quadrados generalizados

   $\hat{\boldsymbol{\beta}} = [((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})'((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})]^{-1}((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{X})'((\mathbf{I_n} - \hat{\lambda}\mathbf{W_2})\mathbf{y})$.

5. Voltar para o passo 3. até obter a convergência dos resíduos.

Estas são as estimativas para a variâncias dos estimadores:

$$
\begin{aligned}
&-E\left(\frac{\partial^2 Ln(L)}{\partial(\sigma^2)^2}\right) = \frac{n}{2(\sigma^2)^2} \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda^2}\right) = tr((\mathbf{I_n} - \lambda\mathbf{W_2})^{-1}\mathbf{W_2}(\mathbf{I_n} - \lambda\mathbf{W_2})^{-1}\mathbf{W_2}) \\
&+tr(\mathbf{W_2'}\mathbf{W_2}((\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n} - \lambda\mathbf{W_2}))^{-1}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\beta^2}\right) = \frac{\mathbf{X}'(\mathbf{I_n}-\lambda\mathbf{W_2})'(\mathbf{I_n}-\lambda\mathbf{W_2})\mathbf{X}}{\sigma^2}) \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda\partial\sigma^2}\right) = \frac{1}{\sigma^2}[tr((\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n}-\lambda\mathbf{W_2})^{-1}\mathbf{W}_2') - \lambda tr(\mathbf{I_n} - \lambda\mathbf{W_2})'(\mathbf{I_n}-\lambda\mathbf{W_2}^{-1})\mathbf{W}_2'\mathbf{W}_2)] \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\boldsymbol{\beta}\partial\sigma^2}\right) = 0 \\
&-E\left(\frac{\partial^2 Ln(L)}{\partial\lambda\partial\boldsymbol{\beta}}\right) = 0
\end{aligned}
\tag{2.16}
$$

Por fim, a seção seguinte apresentará o último modelo de regressão espacial que será abordado neste trabalho: o FAR

2.7 MODELO FAR

O modelo de regressão espacial mais simples é o modelo FAR - *First-order spatial AR model*. Nesse modelo a variável dependente é explicada apenas por seus vizinhos, não existindo outras covariáveis. Sua formulação é dada por:

$$
\mathbf{y} = \rho\mathbf{W_1}\mathbf{y} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.17}
$$

onde:

- $\boldsymbol{\epsilon}$ é o erro aleatório;
- $\mathbf{W_1}$ é a matriz de vizinhança;
- ρ é parâmetro espacial.

Esse modelo é derivado do SAC, nos casos em que os parâmetros $\boldsymbol{X\beta} = 0$ e $\lambda\mathbf{W_2} = 0$. O FAR aplica-se quando a variável dependente y é auto explicada por sua estrutura de vizinhança. Por possuir uma estrutura simples ele é facilmente interpretado e em casos que a dependência espacial é grande não é necessário o acréscimo de variáveis adicionais para explicar a variável dependente - a estrutura espacial consegue fazer com que o modelo tenha um bom ajustamento. Nesses casos, portanto, o FAR é uma boa escolha de modelo.

Neste trabalho sua estrutura não será analisada, não cabe, portanto, um detalhamento maior do modelo e de suas estimações. Aqui, devido a sua simplicidade e ao fato de sua estrutura aparecer na formulaçao do erro aleatório no modelo SAC e SEM, o FAR servirá apenas como modelo de apoio para as comparações e análises feitas no capítulo de estudo empírico.

Apresentados os modelos que serão investigados nesse estudo, devemos avançar sobre as questões motivadoras do mesmo. O terceiro capítulo se debruça sobre análise estrutural dos três primeiros modelos introduzidos no presente capítulo. Já se fez claro que o SAR e o SEM são casos do modelo geral - o SAC. Cabe analisar o que os distingue. Será aboradada, também, a estrutura modelo SAC com diversas combinações de matrizes.

Capítulo 3

ANÁLISE ESTRUTURAL

3.1 INTRODUÇÃO

A análise simplista e recorrente aponta a diferença entre o SAR e o SEM para o local onde a dependência espacial atua. No SAR, a matriz de vizinhança tem efeito direto sobre a predição, no modelo SEM a matriz de vizinhança incorre sobre a dispersão do erro, no termo $\mathbf{u}$ da Equação 2.15. Entretanto, na prática essa diferença não é tão visível, e a utilização do SEM se torna menos intuitiva. A análise da formulação do SAC também se torna um exercício pertinente neste capítulo.

3.2 COMPARAÇÃO ENTRE OS MODELOS SAR E SEM

Como já foi mencionado, um dos problemas dos modelos de regressão espacial é a falta de uma distinção clara entre os modelos SAR e SEM. Por ter a estrutura de vizinhança presente apenas no erro aleatório, o SEM se torna um modelo de regressão menos claro do que os demais. Desenvolvendo sua fórmula e fazendo as devidas substituições podemos considerar que ele é um caso particular do SAR, o que faz com que haja uma confusão em quando ele deverá ser utilizado. Substituindo

$\mathbf{u}$ na Equação 2.15 temos que a fórmula do SEM é:

$$\begin{gathered}\mathbf{y} = \boldsymbol{X\beta} + \lambda\mathbf{W_2u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})\end{gathered} \tag{3.1}$$

Sabe-se que em um modelo de regressão a estimativa do erro aleatório (o resíduo) é igual a:

$$\hat{\mathbf{u}} = \mathbf{y} - \boldsymbol{X\beta} \tag{3.2}$$

Fazendo a substituição do valor predito de $\mathbf{u}$ da Equação 3.5 na Equação 3.4 e desenvolvendo-a temos:

$$\begin{gathered}\mathbf{y} = \boldsymbol{X\beta} + \lambda\mathbf{W_2}(\mathbf{y} - \boldsymbol{X\beta}) + \boldsymbol{\epsilon} \\ \mathbf{y} = \boldsymbol{X\beta} + \lambda\mathbf{W_2y} - \lambda\mathbf{W_2}\boldsymbol{X\beta} + \boldsymbol{\epsilon} \\ \mathbf{y} = (\mathbf{I} - \lambda\mathbf{W_2})\,\boldsymbol{X\beta} + \lambda\mathbf{W_2y} + \boldsymbol{\epsilon}\end{gathered} \tag{3.3}$$

Pela Equação 3.3 nota-se que, como no SAR, existe uma variável explicativa do modelo com o fator espacial, o que faz com o que o modelo SEM seja diferente do SAR apenas pela presença do termo $(\mathbf{I} - \lambda\mathbf{W_2})\,\boldsymbol{X\beta}$. Há, portanto, indícios de que o SEM seja um caso particular do SAR. Mais a frente neste trabalho essa semelhança será estudada.

Na seção a seguir será feita uma análise estrutural do modelo SAC.

3.3 ANÁLISE ESTRUTURAL DO MODELO SAC

Na presente seção, o modelo SAC passará por processo análogo ao imposto à SAR e SEM na anterior, para que se perceba melhor como funciona sua estrutura. Substituindo $\mathbf{u}$ na Equação 2.3 temos que a fórmula do SAC é:

$$\begin{gathered}\mathbf{y} = \rho\mathbf{W_1y} + \boldsymbol{X\beta} + \lambda\mathbf{W_2u} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})\end{gathered} \tag{3.4}$$

Sabe-se que em um modelo de regressão a estimativa do erro aleatório (o resíduo) é igual a:

$$\hat{\mathbf{u}} = \mathbf{y} - \boldsymbol{X\beta} - \rho\mathbf{W_1y} \tag{3.5}$$

Fazendo a substituição do valor predito de $\mathbf{u}$ da Equação 3.5 na Equação 3.4 e desenvolvendo-a temos:

$$\mathbf{y} = \rho\mathbf{W_1y} + \boldsymbol{X\beta} + \lambda\mathbf{W_2}(\mathbf{y} - \boldsymbol{X\beta} - \rho\mathbf{W_1y}) + \boldsymbol{\epsilon}$$

$$\begin{aligned} \mathbf{y} &= \rho\mathbf{W_1y} + \boldsymbol{X\beta} + \lambda\mathbf{W_2y} - \lambda\mathbf{W_2}\boldsymbol{X\beta} - \lambda\mathbf{W_2}\rho\mathbf{W_1y} + \boldsymbol{\epsilon} \\ \mathbf{y} &= (\mathbf{I} - \lambda\mathbf{W_2})\,\boldsymbol{X\beta} + \lambda\mathbf{W_2y} + (\mathbf{I} - \lambda\mathbf{W_2})\rho\mathbf{W_1y} + \boldsymbol{\epsilon} \\ \mathbf{y} &= (\mathbf{I} - \lambda\mathbf{W_2})\,(\boldsymbol{X\beta} + \rho\mathbf{W_1y}) + \lambda\mathbf{W_2y} + \boldsymbol{\epsilon} \end{aligned} \tag{3.6}$$

Outro problema a ser analisado entre os modelos de regressão espacial é quando as matrizes $\mathbf{W_1}$ e $\mathbf{W_2}$ são iguais no modelo SAC, como será apresentado a seguir:

$$\mathbf{y} = (\mathbf{I} - \lambda\mathbf{W})\,(\boldsymbol{X\beta} + \rho\mathbf{Wy}) + \lambda\mathbf{Wy} + \boldsymbol{\epsilon} \tag{3.7}$$

A análise estrutural do SAC não evidencia restrições ou detalhes tão facilmente. Para tal, uma abordagem matemática mais aprofundada de sua estrutura deve ser feita, o que não será objeto deste trabalho. A possibilidade de diferentes parametrizações nas matrizes $\mathbf{W_1}$ e $\mathbf{W_2}$ geram uma multiplicidade de casos na análise estrutural, por isso uma análise empírica se mostrou mais prática.

A seguir, serão apresentadas as ferramentas tradicionais para a definição do modelo espacial mais adequado.

3.4 MÉTODOS DE SELEÇÃO DO MODELO ESPACIAL

Um ponto importante é a escolha do modelo espacial mais adequado. Cada modelo tem suas peculiaridades em estimação e em interpretação. Nesta seção serão

apresentados os métodos de seleção mais comuns. O esquema a seguir representa de forma sucinta um bom método para decisão de qual modelo de regressão utilizar quando se investiga indícios de dependência espacial nos dados explorados. Esse esquema é conhecido como método clássico para especificação do modelo - chamado de método de *forward elimination*.

MÉTODO *FORWARD*

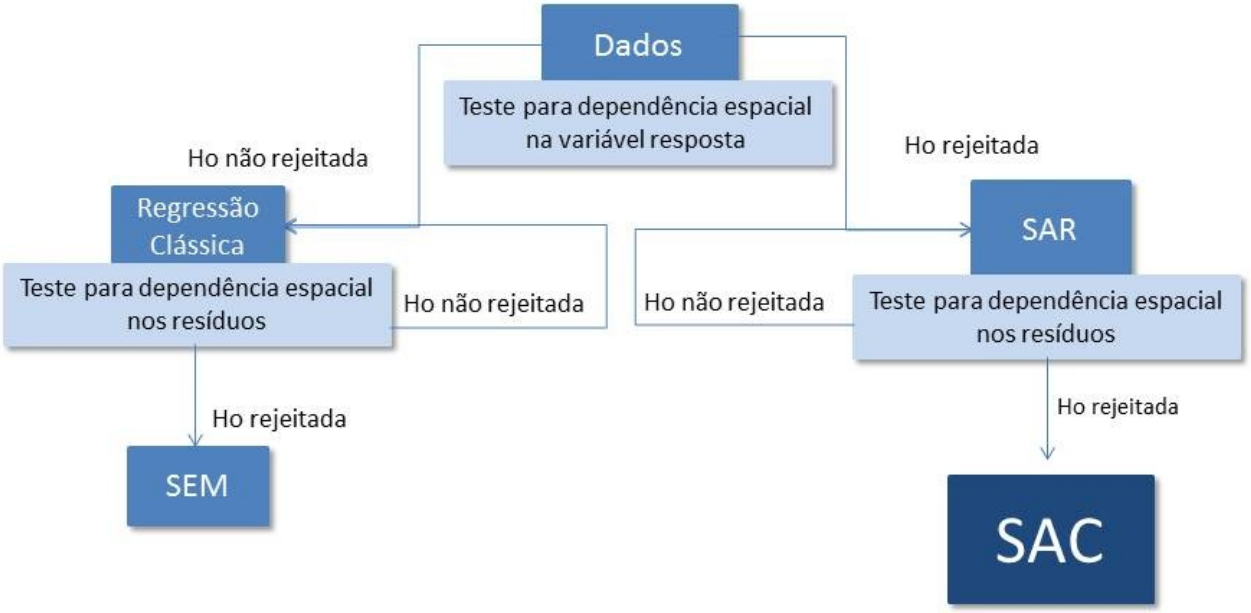

Figura 3.1: Esquema do Método Forward

O teste, cuja hipótese nula aparece como rejeitada ou não no esquema, é o teste individual para significância dos parâmetros e para uma análise como a da figura, deve ser aplicado: no SAR para o coeficiente ρ ($H0 = \rho = 0$), no SEM para $\lambda = 0$, no SAC para ρ e para λ.

Realizou-se um exemplo com os dados do estado de Goiás onde a variável dependente era a população e a explicativa o número de casas. Foram modelados, a

princípio, o SAR e o SEM para uma ilustração inicial. Os resultados estão mostrados na Tabela 3.1e Tabela 3.2.

Tabela 3.1: Exemplo Goiás - Parâmetros SAR

Estimativas modelo SAR		
Parâmetro	Coeficiente	P-Valor
ρ	-0,0058	0,5964

Tabela 3.2: Exemplo Goiás - Parâmetros SEM

Estimativas modelo SEM		
Parâmetro	Coeficiente	P-Valor
λ	-0,0464	0,6487

Pelos resultados pode-se inferir:

- A presença do coeficiente que indica dependência espacial na variável resposta, ρ é rejeitada a um p-valor=0,5964.
- A presença do coeficiente correspondente no erro aleatório, λ, é rejeitada com um p-valor=0,6487

Os dados então indicariam que o modelo regressivo adequado seria a Regressão Clássica e apontam que não há dependência espacial significativa nos dados. Porém, quando são testados ρ e λ para o modelo SAC, a inclusão do parâmetro ρ parece importante para o ajuste do modelo, como mostrado na Tabela 3.3.

Tabela 3.3: Exemplo Goiás - Parâmetros SAC com matrizes iguais

Estimativas modelo SAC - Matrizes Iguais		
Parâmetro	Coeficiente	P-Valor
ρ	-0,0705	0,0005
λ	-0,0076	0,9405

Essa interpretação dúbia do efeito da dependência espacial no modelo é referente ao método de estimação dos coeficientes de regressão no modelo SAC. A estimação multivariada resulta em p-valores indicativos da necessidade de adição ou subtração de determinado coeficiente no modelo que se modificam em função dos coeficientes que já se encontram na regressão. Em outras palavras, é possível que o modelo indicado como mais adequado mude de acordo com o critério de seleção iterativo (*forward, backward ...*) ou grau de complexidade desejado.

MÉTODO DE HENDRY

O método forward previamente citado, de acordo com Maddala (1992), é baseado em "excessiva pré-simplificação com testes diagnósticos inadequados". Florax et al. (2003) apresentam um estudo de eficiência que indica um processo semelhante para escolha do modelo mais adequado aos dados, porém que se inicia com o modelo saturado e se testa a significância dos coeficientes.

Se considerarmos a abordagem clássica como um procedimento *stepwise de forward elimination* (inicia-se com o modelo mais simples e se adiciona os coeficientes para então testar sua significância a cada passo), veremos a abordagem proposta como *stepwise de backward elimination* - conhecida como metodologia de Hendry (1979). A Figura 3.2 esquematiza como seria tal abordagem.

Quando são utilizadas matrizes diferentes para a estimação dos parâmetros do modelo SAC, encontra-se resultados mostrados na Tabela 3.4.

Nota-se que mesmo com matrizes $\mathbf{W_1}$ e $\mathbf{W_2}$ distintas o problema permanece. Anselin (1988) não explicita em sua obra o porquê existe a restrição de matrizes

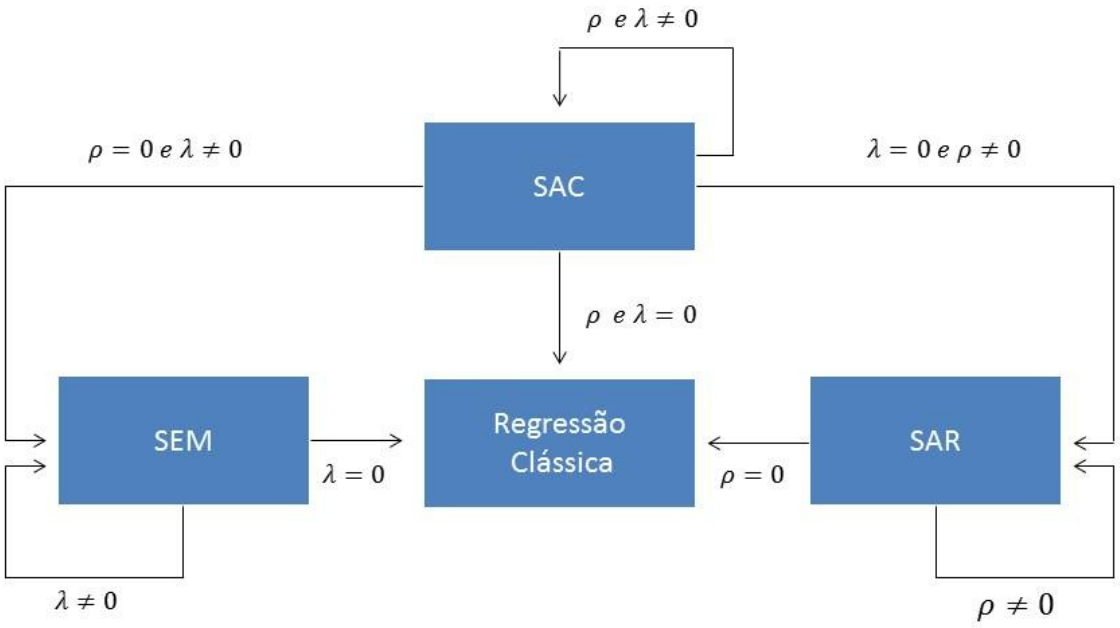

Figura 3.2: Esquema do Método de Hendry

Tabela 3.4: Exemplo Goiás - Parâmetros SAC com matrizes diferentes

Estimativas modelo SAC - Matrizes Diferentes		
Parâmetro	Coeficiente	P-Valor
ρ	-0,0669	<0,0001
λ	-0,00503	0,9828

iguais. Sendo assim, um dos objetivos desse trabalho é tentar explicitar esse problema, de forma empírica ou metodológica. Para uma corroboração da análise, é conveniente a investigação de outros exemplos a fim de se estudar os modelos - como o caso de Goiás foi selecionado como exemplo por conveniência, pode-se indagar se que o resultado foi fruto de coincidência e não de um padrão.

Um corrente exemplo utilizado em estatística espacial são os dados de 1980 de Columbus, capital do estado de Ohio nos Estados Unidos. Essa base foi escolhida para a aplicação de um exercício semelhante ao caso da pauta anterior (dos dados

de Goiás). O resultado obtido na regressão em que a variável crime é a variável dependente e a variável renda é a explicativa é mostrado nas Tabelas 3.5, 3.6, 3.7 e 3.8,

Tabela 3.5: Exemplo Columbus - Parâmetros SAR

Estimativas modelo SAR		
Parâmetro	Coeficiente	P-Valor
ρ	0,4229	0,00112

Tabela 3.6: Exemplo Columbus - Parâmetros SEM

Estimativas modelo SEM		
Parâmetro	Coeficiente	P-Valor
λ	2,4056	<0,0001

Tabela 3.7: Exemplo Columbus - Parâmetros SAC com matrizes iguais

Estimativas modelo SAC - Matrizes Iguais		
Parâmetro	Coeficiente	P-Valor
ρ	0,0562	0,9017
λ	0,4014	0,3376

No caso de Columbus, nota-se que no modelo SAR o parâmetro ρ é significativo, da mesma forma que λ no modelo SEM. Entretanto, ao se estimar os parâmetros por meio do modelo SAC com matrizes iguais ambos os parâmetros tornam-se não significativos. Alterando-se a matriz $\mathbf{W_2}$ apenas λ torna-se significativo. Vale ressaltar que tanto no exemplo de Goiás quanto no exemplo de Columbus foi utilizado a matriz binária padronizada como matriz $\mathbf{W_1}$ e a matriz de distâncias como matriz $\mathbf{W_2}$.

A partir dos exemplos, nota-se que a simples substituição da matriz $\mathbf{W_2}$ por uma matriz diferente de $\mathbf{W_1}$ gera uma alteração na definição do modelo definido

Tabela 3.8: Exemplo Columbus - Parâmetros SAC com matrizes diferentes

Estimativas modelo SAC - Matrizes Diferentes		
Parâmetro	Coeficiente	P-Valor
ρ	0,1258	0,4871
λ	1,7519	0,0044

como mais adequado. A escolha do modelo, então, também está relacionada a parametrização da matriz de distâncias.

No capítulo dedicado aos resultados será demonstrada uma análise empírica que explora os desdobramentos da escolha de um ou outro modelo. O capítulo seguinte irá abordar as simulações e análises empíricas do SAR, do SEM e do SAC.

Capítulo 4

ANÁLISE EMPÍRICA E RESULTADOS

4.1 INTRODUÇÃO

Na seção 3.4 foram apresentados 2 métodos iterativos para a seleção de modelos de regressão espacial. Embora esses métodos sejam análogos aos métodos clássicos de eficácia conhecida, no caso da regressão espacial os resultados nem sempre são satisfatórios. No capítulo 4, o assunto será estudado mais afundo e novas características serão investigadas: a relação entre a adesão aos modelos e força da dependência espacial, efeitos de diferentes parametrizações das matrizes de vizinhança e implicações da estimação por matrizes diferentes no modelo SAC. Será também analisada empiricamente a diferença entre o SAR e o SEM.

4.2 Simulações

Com fins de facilitar a replicabilidade dos resultados obtidos, estão dispostos no Apêndice B e na presente seção (resumidamente) a forma como os bancos de dados simulados foram contruídos;

O banco referenciado como de Mínima Dependência Espacial foi gerado a partir

da adição de uma nova variável ao banco de dados Goiás preexistente. Essa variável corresponde a observações aleatórias de uma distribuição normal (semente=2). As demais alterações na dist. normal em questão (como pode ser visto no apêndice B) são referentes à aproximação da grandeza da soma de quadrados total dos dados com a soma de quadrados referente ao banco com Máxima Dependência Espacial.

Quando da utilização de matrizes de distância, os índices de dependência espacial medem a associação entre determinada variável e a distância entre os centróides das regiões observadas.Para geras bancos com dependência máxima, foram gerados valores para a variável MAX, que são função direta das distâncias entre o centróide de uma região predeterminada e as demais regiões do mapa.

As simulações resultaram em bancos de dados com características satisfatórias para a análise empírica

4.3 ANÁLISE EMPÍRICA

Para aplicar os modelos em bancos de dados com diferentes dependências espaciais utilizou-se o I de Moran como base. Já havia sido constatado dependência espacial fraca no banco referente aos dados de Goiás. Dois bancos de dados foram simulados: um prezando pela máxima dependência espacial e um com dependência estatisticamente nula (p-valor para $(H_0 : I = 0) = 0{,}492008$).

A Equação 2.1 da seção 2.3 deste trabalho mostra que a matriz de proximidade espacial é utilizada no cálculo do índice I de Moran. Até agora, estava sendo uti-

lizado para caracterizar cada banco de dados, os índices provenientes da utilização da matriz binária padronizada, porém a matriz de distância padronizada também aparece para gerar os modelos presentes nas seções e capítulos seguintes. Portanto é importante que se calcule novamente os I's através desta última. Os mapas dos bancos e índices estão dispostos na Figura 4.1.

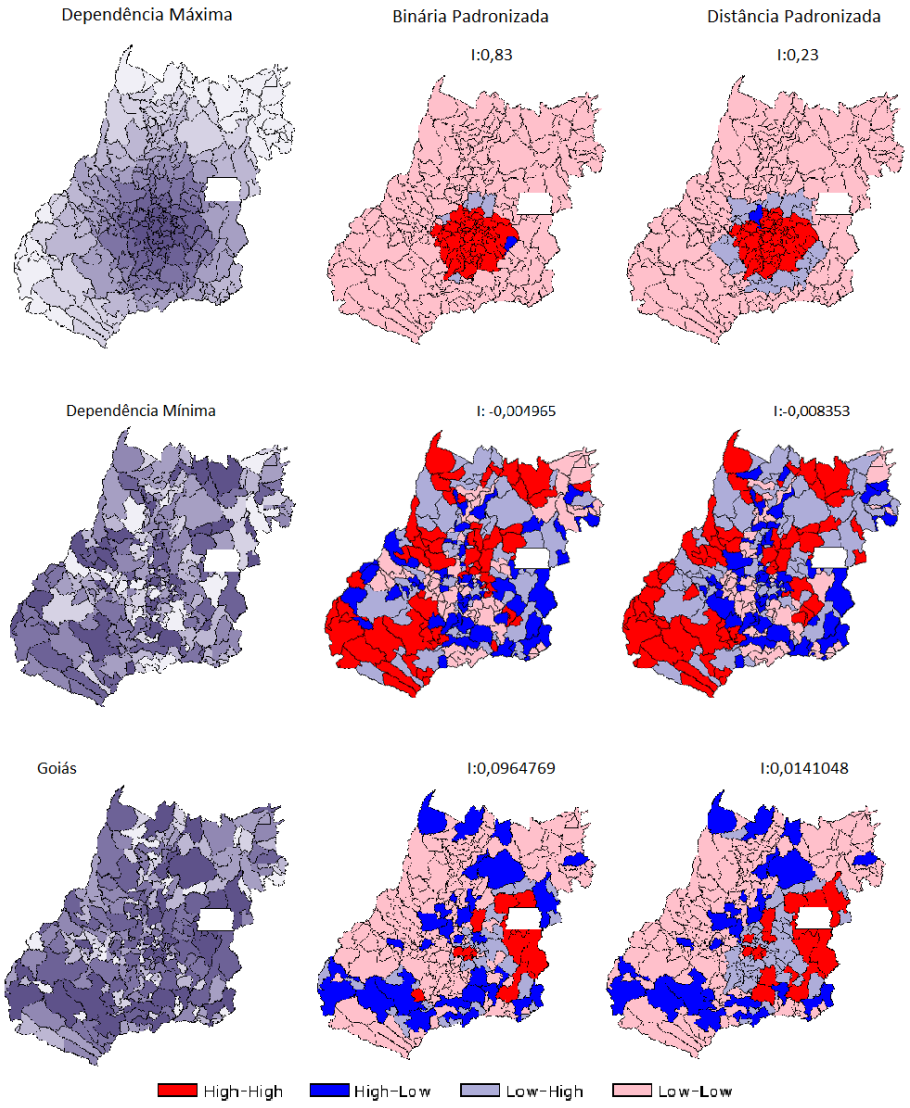

Figura 4.1: Matriz Binária Padronizada vs Matiz de Distância Padronizada

Parte da diferença entre a as medidas de ajuste e o R^2 dos modelos iguais com matrizes de vizinhança diferentes (SAR com matriz de proximidade binária ou de

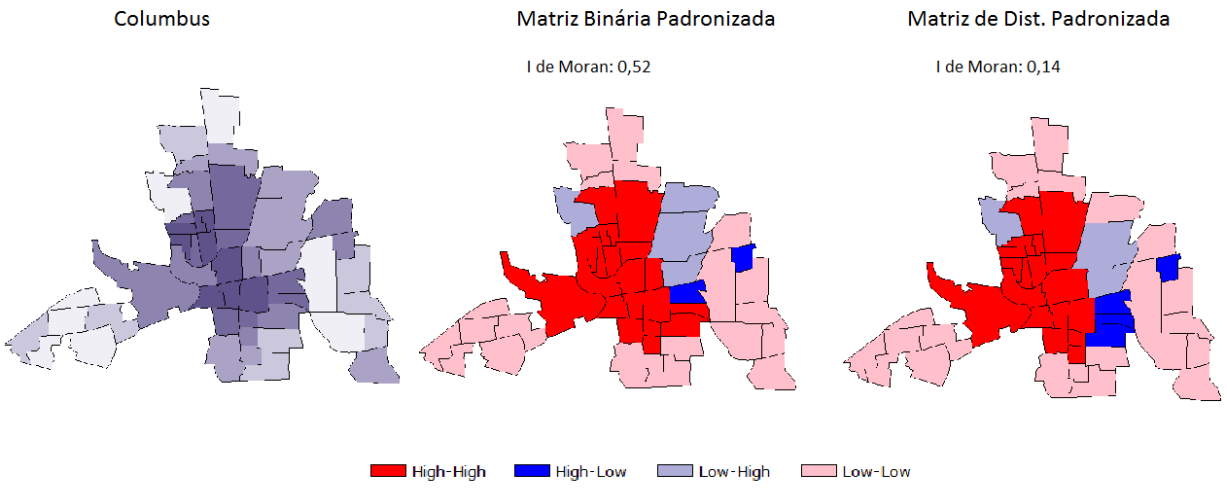

Figura 4.2: Matriz Binária Padronizada vs Matiz de Distância Padronizada

distância por exemplo) podem ser alocadas na mudança da dependência espacial, ocasionada pela matriz utilizada na estimação (em um modelo espacial, espera-se R^2 maior e melhor ajuste em um banco de dados com maior índice de dependência espacial).

Foram montadas tabelas com os valores dos coeficientes e medidas de ajuste para os modelos SAR, SEM e SAC, e para as matrizes de distância da forma binária padronizada e de distância padronizadas (padronizada, nesse ponto, indica que suas linhas somam 1 para garantir os limites superior e inferior dos parâmetros espaciais entre 1 e -1). [1]

Primeiramente é necessário se analisar os p-valores dos coeficientes do modelo. Como esperado, os bancos com I de moran próximos a 0, mostrados na Tabela 4.4, rejeitam os modelos de regressão espacial. No banco que simula os dados com

[1] O modelo SAC é estruturalmente semelhante ao modelo SAR com exceção do erro espacialmente dependente. Sua estrutura de erro tem estimação da forma autoregressiva de primeira ordem. Para comparação, aparece na tabela o coeficiente λ - na área referente ao SAR - calculado a partir do resíduo do SAR (utiliza-se o modelo FAR no resíduo).

Tabela 4.1: Exemplo Máxima Dependência Espacial - Ajuste

I de Moran=0,83.P-valor=0				
Modelo	Matriz	MSE	R^2	AIC
SAR	Bin.	741.855,78	0,88526	3275,08
SAR	Dist.	1.798.936,30	0,72177	3489,45
SEM	Bin.	741.855,78	0,88941	4170,90
SEM	Dist.	1.784.069,08	0,71203	3799,49
FAR	Bin.	757.247,34	0,88200	-
FAR	Dist.	1.784.069,08	0,72177	-
SAC	Bin.	763.137,67	0,88098	3283,94
SAC	Dist.	1.322.983,69	0,79368	3417,09
SAC	B. e D.	683.743,27	0,89337	3257,35
SAC	D. e B.	790.907,36	0,87666	3292,59
Reg.Clás.	-	6.003.493,00	0,0715	3779,09

Tabela 4.2: Exemplo Máxima Dependência Espacial - Parâmetros

I de Moran=0,83.P-valor=0								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	39,28	0,55	0,0020	0,007	0,953	0	0,110	0,26
SAR - Dist.	-5541,20	0,00	0,0031	0,006	3,083	0	3,891	0,00
SEM - Bin.	69,72	0,00	0,00038	0,569	-	-	1,003	0,00
SEM - Dist.	2596,10	0,00	0,0030	0,011	-	-	2,705	0,00
FAR - Bin.	-	-	-	-	0,960	0	-	-
FAR - Dist.	-	-	-	-	2,528	0	-	-
SAC - Bin.	177,70	0,00	0,0014	0,059	0,890	0	0,346	0,00
SAC - Dist.	-3547,02	0,00	0,0024	0,016	2,325	0	2,690	0,00
SAC - B. e D.	-59,27	0,52	0,0015	0,037	1,004	0	0,403	0,38
SAC - D. e B.	-2904,74	0,00	0,0006	0,428	1,945	0	0,888	0,00
Reg.Clás.	1957,58	0,00	0,0088	0,000	-	0	-	0,00

máxima dependência espacial, os coeficientes λ e ρ aparecem na Tabela 4.2 como significativos.

Um ponto interessante é a estimação dos coeficientes por meio das diferentes matrizes de distância. A matriz de distância gerou coeficientes de maior magnitude em todos os modelos, ultrapassando, em algumas vezes, o limite $|1|$. A tentativa de sanar esse problema é tratada logo a frente neste capítulo.

O banco referente aos dados do distrito de Columbus, cujos resultados de es-

Tabela 4.3: Exemplo Mínima Dependência Espacial - Ajuste

I de Moran=-0,004965.P-valor=0,492008				
Modelo	Matriz	MSE	R^2	AIC
SAR	Bin.	9.671.194,20	0,00036	3896,48
SAR	Dist.	9.671.194,20	0,00036	3896,48
SEM	Bin.	9.591.267,01	0,00039	3896,51
SEM	Dist.	9.509.096,70	0,00888	3896,53
FAR	Bin.	9.593.861,36	0,00008	-
FAR	Dist.	9.518.123,90	0,00798	-
SAC	Bin.	9.144.184,81	0,04695	3884,93
SAC	Dist.	9.505.101,36	0,00934	3894,30
SAC	B. e D.	9.485.918,89	0,01134	3893,81
SAC	D. e B.	9.486.977,37	0,01122	3893,83
Reg.Clás.	-	9.672.188,00	0,00025261	3894,50

Tabela 4.4: Exemplo Mínima Dependência Espacial - Parâmetros

I de Moran=-0,004965.P-valor=0,492008								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	8303,94	0,00	0,0007	0,800	-0,013	0,91	-0,002	0,99
SAR - Dist.	13130,74	0,00	0,0006	0,802	-0,607	0,00	-0,217	0,63
SEM - Bin.	8195,35	0,00	0,0007	0,790	-	-	-0,015	0,88
SEM - Dist.	8165,88	0,00	0,0006	0,809	-	-	-0,605	0,21
FAR - Bin.	-	-	-	-	-0,012	0,91	-	-
FAR - Dist.	-	-	-	-	-0,571	0,23	-	-
SAC - Bin.	11348,39	0,00	0,0001	0,982	-0,382	0,20	0,330	0,17
SAC - Dist.	11189,48	0,36	0,0006	0,807	-0,371	0,80	-0,375	0,81
SAC - B. e D.	7815,63	0,00	0,0006	0,826	0,043	0,70	-0,723	0,19
SAC - D. e B.	1401,56	0,00	0,0005	0,841	-0,715	0,19	0,042	0,72
Reg.Clás	8198,49	0,00	0,0006	0,806	-	-	-	-

timação dos parâmetros estão na Tabela 4.8, expõe dados com dependência espacial moderada (I de moran= 0.5002) indica que são significativos os parâmetros espaciais dos modelos SAR e SEM tanto para matriz de distância padronizada e matriz binária de vizinhança padronizada. O critério de Akaike, presente na Tabela 4.7, aponta que o modelo de melhor ajuste é o SAR, que, quando estimado pela matriz de distância padronizada, possui um ajuste ligeiramente melhor (236,89 contra 237,46) que o caso complementar. A comparação entre os modelos SAR e SEM e

Tabela 4.5: Exemplo Goiás - Ajuste

I de Moran : 0,0964769.P-valor=0,0066997				
Modelo	Matriz	MSE	R^2	AIC
SAR	Bin.	15.787.762,00	0,99735	4015,08
SAR	Dist.	15.748.571,00	0,99736	4014,48
SEM	Bin.	15.657.284,17	0,99735	4015,22
SEM	Dist.	15.618.417,45	0,99735	4015,21
FAR	Bin.	321.777.218,00	0,02565	-
FAR	Dist.	325.538.032,00	0,01426	-
SAC	Bin.	16.582.624,39	0,99719	4028,98
SAC	Dist.	15.645.741,81	0,99735	4014,90
SAC	B. e D.	16.912.298,70	0,99714	4033,74
SAC	D. e B.	15.930.323,11	0,99730	4019,26
Reg.Clás.	-	15.796.476,94	0,99735	4013,21

Tabela 4.6: Exemplo Goiás - Parâmetros

I de Moran : 0,0964769.P-valor=0,0066997								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	-2033,97	0,00	4,2253	0,0	-0,002	0,63	-0,024	0,81
SAR - Dist.	-1321,85	0,00	4,2265	0,0	-0,035	0,00	-0,009	0,98
SEM - Bin.	-2106,21	0,00	4,2245	0,0	-	-	-0,027	0,79
SEM - Dist.	-2086,71	0,00	4,2245	0,0	-	-	0,067	0,87
FAR - Bin.	-	-	-	-	0,192	0,04	-	-
FAR - Dist.	-	-	-	-	0,576	0,02	-	-
SAC - Bin.	-1466,37	0,00	4,2350	0,00	-0,025	0,00	-0,002	0,98
SAC - Dist.	-1892,35	0,04	4,2249	0,00	-0,009	0,82	-0,035	0,93
SAC - B. e D.	-1372,19	0,00	4,2367	0,00	-0,029	0,00	-0,034	0,94
SAC - D. e B.	-3294,54	0,00	4,2211	0,00	0,055	0,17	-0002	0,98
Reg.Clás.	-2088,96	0,00	4,2244	0,00	-	-	-	-

suas diferenças também serão discutidas mais a frente nesse capítulo.

A diferença nos valores do modelo SAC e do modelo SAR com λ estimado a partir do modelo FAR é referente ao primeiro estimar os λ e ρ de forma simultânea, enquanto o segundo não o faz. Os coeficientes do SAC parecem equilibrar a divisão da dependência espacial - indicando que a estimação simultânea causa maior paridade entre os coeficientes espaciais de estimação e erro. Há indícios de que a magnitude desses coeficientes esteja também relacionada com a matriz que os estima

Tabela 4.7: Exemplo Columbus - Ajuste

I de Moran : 0,5008.P-valor=0				
Modelo	Matriz	MSE	R^2	AIC
SAR	Bin.	112,58	0,58949	237,46
SAR	Dist.	111,29	0,59421	236,89
SEM	Bin.	112,58	0,57016	251,64
SEM	Dist.	111,29	0,58072	253,49
FAR	Bin.	3151,00	0,44941	-
FAR	Dist.	187,88	0,31493	-
SAC	Bin.	119,25	0,56519	240,28
SAC	Dist.	114,69	0,58181	238,37
SAC	B. e D.	114,41	0,58284	238,25
SAC	D. e B.	112,83	0,58859	237,57
Reg.Clás.	-	147,58	0,48380	246,68

Tabela 4.8: Exemplo Columbus - Parâmetros

I de Moran : 0,5008.P-valor=0								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	42,32	0,000	-1,456	0,000	0,4229	0,001	0,0334	0,865
SAR - Dist.	37,37	0,000	-1,411	0,000	0,4956	0,000	0,1321	0,603
SEM - Bin.	56,62	0,000	-1,518	0,000	-	-	0,4890	0,001
SEM - Dist.	54,01	0,000	-1,412	0,000	-	-	0,6147	0,000
FAR - Bin.	-	-	-	-	0,6642	0,000	-	-
FAR - Dist.	-	-	-	-	0,4948	0,000	-	-
SAC - Bin.	55,66	0,002	-1,580	0,000	0,0562	0,902	0,4014	0,338
SAC - Dist.	49,29	0,043	-1,475	0,000	0,1767	0,777	0,4536	0,394
SAC - B. e D.	49,75	0,000	-1,466	0,000	0,1682	0,539	0,4516	0,138
SAC - D. e B.	47,51	0,001	-1,499	0,000	0,2489	0,443	0,3552	0,188
Reg.Clás.	64,46	0,000	-2,041	0,000	-	-	-	-

(comparando-se, por exemplo, com a estimação dos modelos SAC por matrizes distintas). O coeficiente relacionado a matriz de distância se apresenta maior quando comparado ao mesmo coeficiente calculado com a matriz binária.

4.4 RESULTADOS SAC

Uma rápida análise das tabelas anteriores revela uma grave incoerência aos modelos autoregressivos: coeficientes maiores que $|1|$ (encontrado, principalmente, quando

os coeficientes são estimados a partir da matriz de distância). Parece provável que o tamanho da matriz aliado ao grande número de casas decimais seja responsável por esse problema. Nos bancos com dependência máxima, mínima e para os dados de Goiás são utilizados cerca de 58000 elementos na matriz de distância, nesse nível, qualquer tipo de aproximação nas contas que levaram a construção da matriz de distância podem influenciar os coeficientes.

Uma provável solução é diminuir o número de elementos influentes dessa matriz - ou seja, diminuir o número de elementos não nulos. É de se esperar que dados de um polígono muito distante de outro tenha pouca, ou nenhhuma, influência no local analisado. Portanto fez-se a tentativa de colocar uma distância limite para considerar que um polígono(no caso município) seja influente na matriz de distância; distâncias acima da definida geram entradas de valor 0 na matriz de distância.

Os resultados foram satisfatórios como pode ser visto na Tabela 4.9. Nela foram simulados SAR, SEM e SAC utilizando-se diferentes distâncias como corte, com o intuito de decidir qual é a mais eficaz sem grande perda de qualidade de ajuste. Foram, também, calculados os novos I's de Moran para cada distância - apresentados na Figura 4.3. A Tabela 4.10 mostra os parâmetros estimados para cada caso.

Uma premissa importante para a estimação dos coeficientes β e ρ é a padronização da matriz $\mathbf{W}$, a nova matriz de distâncias foi calculada de forma que a linhas continuassem somando 1. Portanto, a influência das observações próximas aumentou quando se cortou a influência das observações mais distantes. É por isso que a dependência espacial aumenta para menores distâncias de corte: os municípios

Tabela 4.9: Exemplo Máxima Dependência Espacial - Medidas de Ajuste vs Distância de Corte

Medidas de Ajuste vs Distância de Corte					
Distância	Modelo	MSE	R^2	AIC	I de Moran
30	SAR	829.997,18	0,87056	3304,26	0.91
	SEM	829.997,18	0,90598	3851,90	
	SAC	641.316,35	0,89999	3241,85	
50	SAR	1.277.606,61	0,80075	3408,64	0,76
	SEM	1.277.606,61	0,82403	3857,77	
	SAC	1.699.281,68	0,73499	3477,66	
100	SAR	2.162.581,88	0,66274	3536,01	0,58
	SEM	2.162.581,88	0,66927	4885,59	
	SAC	1.809.162,53	0,71786	3492,83	
150	SAR	2.483.124,12	0,61275	3569,46	0,46
	SEM	2.483.124,12	0,58838	4396,80	
	SAC	2.671.771,98	0,58333	3587,18	
200	SAR	2.855.925,19	0,55461	3603,31	0,37
	SEM	2.855.925,19	0,54441	3853,05	
	SAC	1.925.130,21	0,69977	3507,86	

vizinhos têm um peso maior na estimação de cada ponto pois acumulam o peso daqueles elementos que agora aparecem como 0 na matriz de distância.

O resultado dessa mudança foi satisfatório, embora as medidas de diagnóstico indiquem que as distâncias de 30 e 50 tenham gerado melhores modelos (com estas distâncias a dependência é tão alta que o coeficiente β deixa de ser significativo), é mais adequado utilizar a distância corte em 100 para que a comparação com os modelos previamente calculados seja tão honesta quanto possível. O resultado do modelo com a distância de corte igual a 100 é mostrado nas Tabelas 4.11 e 4.12.

Para a decisão de qual modelo SAC utilizar (nessa seção excluímos a possibilidade de se utilizar SAR ou SEM) em cada banco de dados, utilizaremos o R^2, o critério de Akaike (AIC) e os p-valores dos testes com H_0 :coeficiente=0.

Pelo banco referente à simulação que maximiza o I de Moran, ao observar o R^2,

Tabela 4.10: Exemplo Máxima Dependência Espacial - Medidas de Ajuste vs Distância de Corte - Parâmetros

Medidas de Ajuste vs Distância de Corte									
Dist.	Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
30	SAR	429,40	0,000	0,0019	0,012	0,836	0,00	0,429	0
	SEM	888,53	0,000	-1,0974	0,999	-	-	1,009	0
	SAC	814,96	0,000	0,0004	0,542	0,678	0,00	0,639	0
50	SAR	188,23	0,030	0,0031	0,001	0,872	0,00	0,482	0
	SEM	738,14	0,013	0,0013	0,138	-	-	1,009	0
	SAC	1062,85	0,003	0,0016	0,147	-1,044	0,00	1,058	0
100	SAR	-81,94	0,407	0,0048	0,000	0,929	0,00	0,679	0
	SEM	25836,69	0,094	0,0038	0,002	-	-	1,006	0
	SAC	-363,18	0,212	0,0036	0,002	1,034	0,00	0,675	0
150	SAR	-1093,96	0,000	0,0043	0,001	1,336	0,00	2,248	0
	SEM	10395,55	0,011	0,0047	0,001	-	-	1,025	0
	SAC	-13948,21	0,000	0,0023	0,094	4,726	0,00	1,121	0
200	SAR	-1296,23	0,000	0,0048	0,001	1,373	0,00	2,609	0
	SEM	3443,24	0,000	0,0047	0,001	-	-	1,2538	0
	SAC	-7315,49	0,000	0,0028	0,018	3,206	0,00	1,209	0

são candidatos: modelo SAC, com matrizes binárias e $\mathbf{W_1} = \mathbf{W_2}$; modelo SAC com matrizes $\mathbf{W_1}$ = binária e $\mathbf{W_2}$ = distância; e SAC com matrizes $\mathbf{W_1}$ = distância e $\mathbf{W_2}$ = binária. Por conveniência os chamaremos respectivamente por A, B e C.

A diferença entre o R^2 dos modelos aparece na terceira casa decimal e, portanto, não é determinante para a definição da melhor configuração entre as matrizes. O critério AIC também não aponta diferenças determinantes entre os 3 modelos, porém indica que o modelo B é o melhor.

A análise dos p-valores exclui o modelo C, pois indica que apenas seu coeficiente λ é significativo. O coeficiente β em A e o intercepto em B se apresentam sensíveis ao nível de significância, adotou-se o nível de significância $\alpha = 0,05$ e os p-valores de seus testes de significância estão por volta de 0,06; portanto, não há grandes perdas em considerá-los significativos.

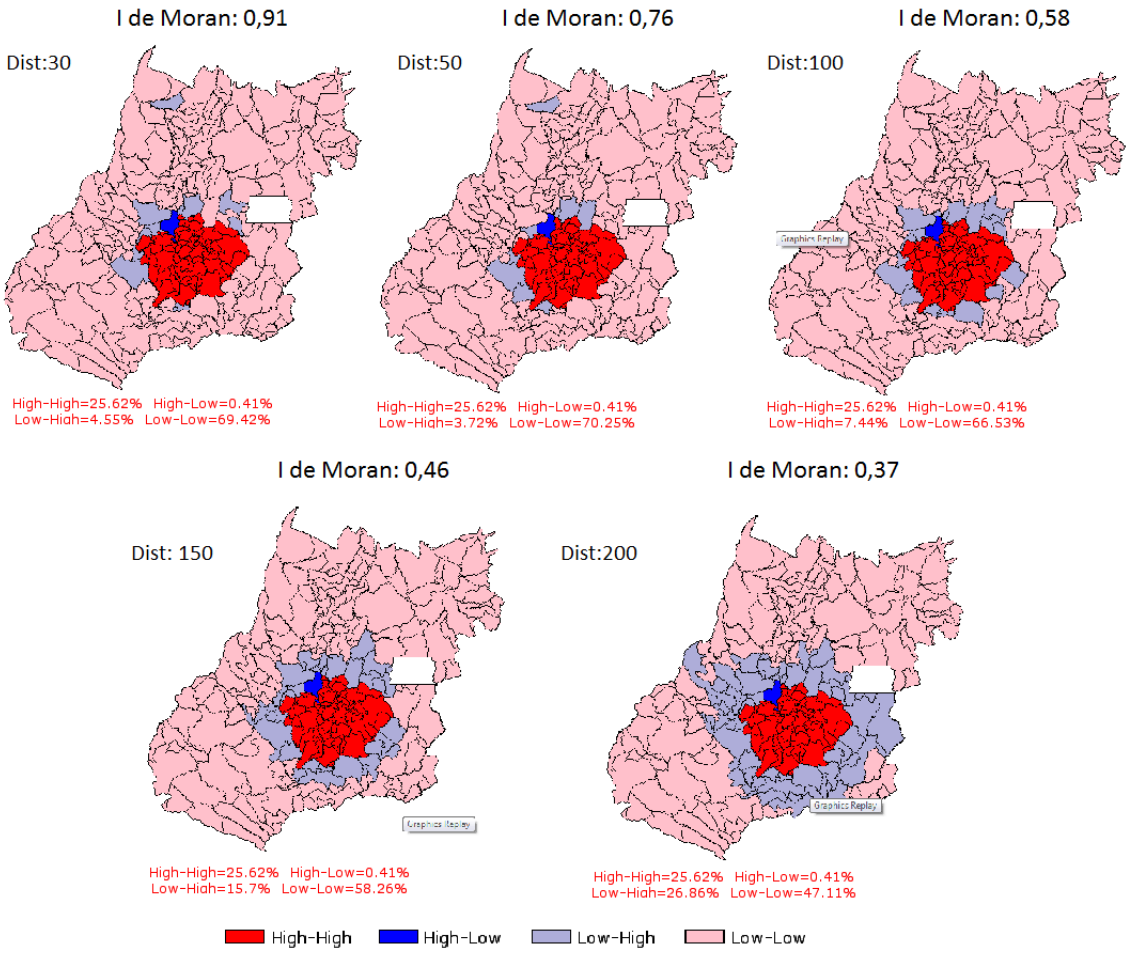

Figura 4.3: I de Moran vs Distâncias de Corte

Nesse ponto parece razoável que a decisão seja tomada com base no AIC, então se opta pelo modelo B (AIC=3270,63 ante 3283,93 do modelo A). Por outro lado, a escolha do modelo A também parece razoável por não exigir medida corretiva na matriz de distância utilizada (que poderia causar perda de informação no modelo). Portanto, o veredicto é que tanto o modelo SAC com matrizes binárias e $\mathbf{W_1} = \mathbf{W_2}$ quanto o modelo SAC com matrizes $\mathbf{W_1}$ = binária e $\mathbf{W_2}$ = distância pode ser escolhido sem grande perdas de capacidade de previsão e ajuste.[2]

[2]Se adotadas as distâncias de corte menores que 100, perceberíamos que AIC e R^2 apontariam os modelos baseados na matriz de distâncias padronizados como preferíveis em relação ao modelo que se utiliza apenas da matriz binária de vizinhança e uma análise mais cuidadosa do comportamento dos testes de significância para os coeficientes de regressão seria necessária.

Tabela 4.11: Exemplo Máxima Dependência Espacial - Ajuste

I de Moran=0,83.P-valor=0				
Modelo	Matriz	MSE	R^2	AIC
SAR	Bin.	741.855,78	0,88526	3275,08
SAR	Dist.	2.162.581,88	0,66274	3536,01
SEM	Bin.	741.855,78	0,88941	4170,90
SEM	Dist.	2.162.581,88	0,66927	4885,59
FAR	Bin.	757.247,34	0,88200	-
FAR	Dist.	2.161.269,53	0,66295	-
SAC	Bin.	763.137,67	0,88098	3283,94
SAC	Dist.	1.809.162,53	0,71786	3492,83
SAC	B. e D.	722.306,75	0,88735	3270,63
SAC	D. e B.	736.840,01	0,88509	3275,45
Reg.Clás.	-	6.003.493,00	0,0715	3779,09

Tabela 4.12: Tabela:Exemplo Máxima Dependência Espacial - Parâmetros

I de Moran=0,83.P-valor=0								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	39,28	0,55	0,0020	0,007	0,953	0	0,110	0,26
SAR - Dist.	-81,94	0,407	0,0048	0,000	0,929	0,00	0,679	0,00
SEM - Bin.	69,72	0,00	0,0030	0,011	-	-	2,705	0,00
SEM - Dist.	25836,69	0,094	0,0038	0,002	-	-	1,006	0,00
FAR - Bin.	-	-	-	-	0,960	0	-	-
FAR - Dist.	-	-	-	-	1,140	0	-	-
SAC - Bin.	177,70	0,00	0,0014	0,059	0,890	0	0,346	0,00
SAC - Dist.	-363,18	0,212	0,0036	0,002	1,034	0,00	0,675	0,00
SAC - B. e D.	76,39	0,060	0,0017	0,021	0,935	0,000	0,4334	0,02
SAC - D. e B.	818,91	0,142	0,0004	0,540	0,471	0,133	0,9255	0,00
Reg.Clás.	1957,58	0,00	0,0088	0,000	-	0	-	0,00

Um ponto a favor dos modelos estimados com matrizes distintas é o fato de não se ter observado nele sérias mudanças de sinal nos parâmetros estimados em relação aos outros modelos estimados (quando ocorreram, em módulo, o valor não parece absurdo) - principalmente em relação ao modelo de regressão clássica, tomado como base. A inversão de sinal gerada é característica de regressões onde aparece multicolinearidade, como nesses casos é comum que os p-valores não indiquem a realidade, a inversão no sinal no intercepto pode ser um indicativo de problema

mais grave. É necessário, portanto, que se adote uma abordagem diferente para sanar a questão.

A mudança de sinal pode ter sido causada pela utilização da mesma estrutura de proximidade espacial em $\mathbf{W_1}$ e $\mathbf{W_2}$. A princípio "explica-se" a dependência espacial via matriz de vizinhanca ou distância; é de se esperar que toda a contribuição dessa parametrização da proximidade espacial seja exaurida e, então, seria natural se utilizar de alguma informação diferente para esgotar a dependência restante (como uma parametrização distinta). A mesma estrutura, porém, foi utilizada, o que não teria sentido quando se analisa dessa forma - se a estrutura não foi possivel de "retirar" toda a dependência espacial da primeira vez não teria porque utiliza-la de novo,é válido imaginar que outra estrutura de matriz de proximidade espacial capture a dependência restante; uma diferente abordagem de escolha da matriz $\mathbf{W_2}$ seria recomendada. Portanto, apesar de um bom ajustamento dos modelos com matrizes iguais eles não necessariamente serão a melhor escolha. É importante que se observem os parâmetros e as diferenças obtidas comparando-se com outros modelos.

O modelo cuja dependência espacial é mínima ($I = -0,004965$) não exige avaliação profunda. Obviamente o ajuste do modelo SAC é ruim para qualquer configuração matricial, o "grau de explicação"(R^2) do modelo não é razoável e os p-valores rejeitam qualquer modelo SAC. Portanto deve-se procurar outra técnica para a tentativa de modelar os dados.

Para os dados de Goiás temos um caso interessante: percebe-se que a Regressão Clássica se adequa de forma excelente aos dados, o que implica que a parte espacial

da modelagem poderia agregar informação, mas dificilmente o aumento na complexidade do modelo "se pagaria". Se ainda assim o modelo SAC fosse escolhido, os p-valores acerca dos testes $H_0 : \lambda = 0$ e $H_0 : \rho = 0$ afastariam essa possibilidade.

Columbus é um exemplo sólido em estatística espacial, mas, assim como os dois citados por último, rejeitou-se a dependência espacial no erro ($\lambda = 0$).

A conclusão é que, em geral, o modelo SAC se adequa apenas a bancos de dado com elevada dependência espacial. Para confirmar essa hipótese, foi gerado um banco de dados que buscou maximizar a dependência espacial no exemplo de Columbus, a partir da criação de uma variável que é função da distância. O resultado obtido está apresentado nas Tabelas 4.13 e 4.14 e na figura 4.4.

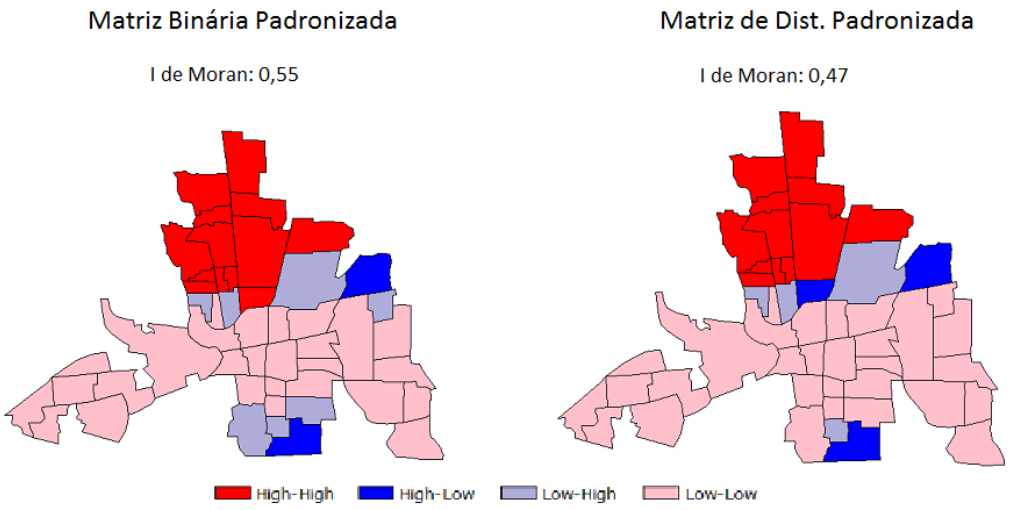

Figura 4.4: Dependência Maximizada

O valor do índice obtido (0,55 para matriz binária e 0,45 para matriz de distância) não parece indicar dependência tão alta. Porém, o valor máximo que o índice pode alcançar é limitado pelo número de polígonos (no caso, municípios) utilizados em sua estimação. Mapas com menos polígonos dão origem a I's de Moran limitados

por valores mais baixos.

Tabela 4.13: Exemplo Máxima Dependência Espacial - Ajuste

Columbus				
Modelo	Matriz	MSE	R^2	AIC
SAC	Bin.	11.024.669,26	0,76266	800,57
SAC	Dist.	18.610.126,29	0,59936	826,22
SAC	B. e D.	21.091.218,79	0,54595	832,35
SAC	D. e B.	20.340.684,86	0,56210	830,58
Reg.Clás.	-	48.426.675	0,0000	869,4

Tabela 4.14: Exemplo Máxima Dependência Espacial - Columbus(dist. de corte=85) - Parâmetros

Columbus								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAC - Bin.	-1609,60	0,055	66,862	0,22	1,07	0,00	-1,01	0,001
SAC - Dist.	-935,52	0,463	79,847	0,35	1,00	0,00	-0,68	0,007
SAC - B. e D.	1682,62	0,626	37,221	0,69	0,81	0,03	-0,36	0,724
SAC - D. e B.	-1576,17	0,301	127,428	0,20	1,00	0,00	-0,34	0,103
Reg.Clás.	10797	0,00	4,872	0,98	-	-	-	-

No caso em pauta, o único modelo que rejeitou um dos parâmetros espaciais, a qualquer nível de significância razoável, foi o SAC(w1=bin,w2=dist) que rejeitou a presença de λ no modelo a um p-valor 0,724. O modelo SAC(w1=dist,w2=bin) indicou que λ também não faz parte do modelo, porém com p-valor=0,103. Os resultados, então, dão mais indícios que a validade do modelo SAC está relacionada à altos valores de dependência espacial. Em tempo, o modelo SAC(w1=w2)[bin] foi que melhor se ajustou aos dados : $R^2 = 0,76$ e $AIC = 800,6$. O problema da inversão de sinal está novamente presente. No caso, apenas o modelo que estima ρ pela matriz binária padronizada e λ pela matriz de distâncias padronizadas tem o intercepto com mesmo sinal que a regressão clássica.

É razoável que se utilize a estrutura de matriz de proximidade espacial mais

simples para descrever y, ou seja, no exemplo,equivale à matriz de vizinhança padronizada ser utilizada para estimar ρ, e se utilize da matriz mais complexa para caracterizar o erro. É claro que em exemplos aplicados, a escolha de qual matriz deve estimar qual segmento do modelo é mais natural pois o contexto do problema dá dicas de qual parametrização faz mais sentido no problema. Define-se a estrutura referente a y e a restante caracteriza o erro.

4.5 RESULTADOS DA COMPARAÇÃO ENTRE SAR E SEM

Os modelos de regressão espacial SAR e SEM são muito parecidos, como mostrado na seção 3.2 deste trabalho. Se analisada, puramente, a fórmula de ambos os modelos, uma diferenciação entre eles se torna complicada, visto que o termo $(\mathbf{I} - \lambda \mathbf{W_2})\, \boldsymbol{X\beta}$ apresentado na Equação 3.3 a princípio não possui nenhum significado na interpretação do modelo (uma análise mais a fundo desse significado não será abordada neste trabalho). Será feita, portanto, uma estudo empírico dos dois modelos com a finalidade de se observar a existência de possíveis diferenças entre ambos os modelos.

Primeiramente serão analisados os casos mostrados na seção anterior, com foco apenas nos dois modelos. Em todos os exemplos dessa seção as matrizes utilizadas foram padronizadas.

No exemplo em que o I de Moran é maximizado a Tabela 4.15 mostra as medidas de ajustamento em ambos os modelos. Pela tabela, nota-se que pelo critério AIC - em que menores valores representam melhores modelos - o SAR será um modelo melhor

Tabela 4.15: Exemplo I de moran máximizado - Comparação do ajustamento - SAR e SEM

Comparação SAR e SEM				
Modelo	Matriz	MSE	R^2	AIC
SAR	Binária	741.8555,78	0,885	3275,08
SEM	Binária	741.8555,78	0,889	4170,90
SAR	Distância	1.798.936,30	0,721	3489,45
SEM	Distância	1.784.069,08	0,712	3799,49

Tabela 4.16: Exemplo I de Moran Maximizado - Comparação dos parâmetros - SAR e SEM

Comparação SAR e SEM								
Modelo	Intercepto	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	39,28	0,5	0,0020	0,007	0,953	0	-	-
SEM - Bin.	69,72	0	0,0030	0,011	-	-	2,7	0
SAR - Dist.	-81,94	0,4	0,0048	0	0,929	0	-	-
SEM - Dist.	25836,69	0,09	0,0038	0,002	-	-	1	0

que o SEM (utilizando-se tanto a matriz binária quanto a matriz de distâncias). Pelo critério do R^2 - em que quanto maior melhor - o SAR será melhor que o SEM no caso em que se utiliza a matriz de distância e será aproximadamente igual no caso de matrizes binárias(as diferenças são mínimas em termos de R^2 de forma que nesse caso, o critério não é determinante na escolha do modelo). Após uma análise conjunta dos critérios chega-se a conclusão de que o SAR é o modelo que melhor se ajusta. Complementarmente, a Tabela 4.16 mostra que os modelos espaciais são adequados ao problema.

Tabela 4.17: Exemplo I de moran minimizado - Comparação do ajustamento - SAR e SEM

Comparação SAR e SEM				
Modelo	Matriz	MSE	R^2	AIC
SAR	Binária	9.671.194,20	0,00036	3896,48
SEM	Binária	9.591.267,01	0,00039	3896,51
SAR	Distância	9.588.339,20	0,00892	3894,40
SEM	Distância	9.509.096,70	0,00888	3896,53

Tabela 4.18: Exemplo I de Moran Minimizado - Comparação dos parâmetros - SAR e SEM

Comparação SAR e SEM								
Modelo	Intercepto	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	8303,94	0	0,0007	0,80	-0,01	0,9	-	-
SEM - Bin.	8195,35	0	0,0007	0,79	-	-	-0,015	0,9
SAR - Dist.	13130,74	0	0,0006	0,80	-0,61	0,0	-	-
SEM - Dist.	8165,88	0	0,0006	0,80	-	-	-0,605	0,2

Para o caso em que o I de Moran é estatisticamente igual a zero tanto os valores de AIC do SAR e do SEM quanto os valores do R^2 desses modelos, mostrados na Tabela 4.17, são aproximadamente iguais. Para ambos os casos o AIC é extremamente alto e o R^2 é extremamente baixo, o que já era esperado, se conderado que a base de dados foi gerada para que não houvesse dependência espacial entre as variáveis, como mostrado na Tabela 4.18. Portanto, os modelos de regressão espacial não se adequam bem aos dados.

Tabela 4.19: Exemplo Goiás - Comparação do ajustamento - SAR e SEM

Comparação SAR e SEM				
Modelo	Matriz	MSE	R^2	AIC
SAR	Binária	15.787.762,000,997		4015,08
SEM	Binária	15.657.284,170,997		4014,48
SAR	Distância	15.748.571,000,997		4015,22
SEM	Distância	15.618.417,450,997		4015,21

Tabela 4.20: Exemplo Goiás - Comparação dos parâmetros - SAR e SEM

Comparação SAR e SEM								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	-2033,97	0	4,2253	0	-0,002	0,6	-	-
SEM - Bin.	-2106,21	0	4,2245	0	-	-	-0,027	0,79
SAR - Dist.	-1321,85	0	4,2265	0	-0,035	0	-	-
SEM - Dist.	-2086,71	0	4,2245	0	-	-	0,067	0,87

No caso de Goiás os resultados também foram bem próximos. Nota-se, entre-

tanto, através da análise da Tabela 4.20, que, apesar de significativos, os coeficientes estimados de ρ para o SAR e de λ para o SEM são muito baixos (próximos a zero), o que corrobora com o I de Moran de 0,09 (ou seja, uma dependência espacial baixa). Portanto, tem-se um caso semelhante ao que a dependência espacial é estatisticamente igual a zero. O R^2 alto mostrado na Tabela 4.19 é referente à boa adequabilidade do coeficiente relacionado a parte não-espacial do modelo (o β da Tabela 4.19). A regressão clássica por si só causou um O R^2 de 0,99730 .

Tabela 4.21: Exemplo Columbus - Comparação do ajustamento - SAR e SEM

Comparação SAR e SEM				
Modelo	Matriz	MSE	R^2	AIC
SAR	Binária	112,58	0,589	237,46
SEM	Binária	112,58	0,570	251,64
SAR	Distância	111,29	0,594	236,89
SEM	Distância	111,29	0,580	253,49

Tabela 4.22: Exemplo Columbus - Comparação dos parâmetros - SAR e SEM

Comparação SAR e SEM								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	42,32	0	-1,456	0	0,422	0,001	-	-
SEM - Bin.	56,62	0	-1,518	0	-	-	0,489	0,001
SAR - Dist.	37,37	0	-1,411	0	0,495	0,000	-	-
SEM - Dist.	54,01	0	-1,412	0	-	-	0,614	0,000

Com os dados de Columbus conclui-se por meio das medias de ajustes mostradas na Tabela 4.21 que o SAR explica melhor os dados que o SEM. Essa conclusão vem do fato de que a utilização tanto da matriz binária na estimação quanto da matriz de distâncias resulta em um valor do AIC no SAR menor que no SEM (SAR - Binária: 237,46 x SEM - Binária: 251,64 e SAR - Distância: 236,89 x SEM - Distância: 253,49). Conclusão essa reforçada pelo critério R^2, para os dois tipos

de parametrização matricial, é maior para o SAR (SAR - Binária: 0,589 x SEM - Binária: 0,570 e SAR - Distância: 0,594 x SEM - Distância: 0,580). Como visto na Tabela 4.22, os modelos espaciais se aplicam ao banco de dados, dado que os coeficientes espaciais são significativos. Dessa forma, para este exemplo o SAR seria o modelo mais adequado.

Tabela 4.23: Exemplo Rio de Janeiro - Comparação do ajustamento - SAR e SEM

Comparação SAR e SEM				
Modelo	Matriz	MSE	R^2	AIC
SAR	Binária	1,161	0,468	19,587
SEM	Binária	1,161	0,423	39,728
SAR	Distância	1,255	0,425	26,683
SEM	Distância	1,255	0,415	37,075

Tabela 4.24: Exemplo Rio de Janeiro - Comparação dos parâmetros - SAR e SEM

Comparação SAR e SEM								
Modelo	Interc.	P-valor	β	P-valor	ρ	P-valor	λ	P-valor
SAR - Bin.	9,38	1,64E-7	1,27E-6	2,09E-9	0,39	2,14E-4	-	-
SEM - Bin.	15,79	0	1,20E-6	3,59E-8	-	-	0,37	3,36E-3
SAR - Dist.	6,64	4,67E-2	1,40E-6	2,01E-10	0,57	5,96E-3	-	-
SEM - Dist.	15,91	0	1,36E-6	5,36E-5	-	-	0,62	3,58E-3

Para corroborar a análise outro exemplo foi feito a fim de se verificarem as diferenças entre ambos os modelos. O exemplo utilizado é com os dados do estado do Rio de Janeiro. A variável dependente escolhida foi o logaritmo da renda do município (ln(renda)) e a variável explicativa foi a população do município. Nota-se que nesse exemplo o parâmetro espacial é significativo tanto para o modelo SAR quando para o modelo SEM, como mostrado na Tabela 4.24. Além disso, esses parâmetros são próximos ou maiores que 0,5, oque mostra que essa dependência é moderada. A análise da Tabela 4.23 com o ajustamento de ambos os modelos,

permite verificar, novamente, que o SAR é o modelo que melhor se ajusta segundo as medidas AIC e R^2. Para ambas as matrizes de vizinhança - binária e de distância - o R^2 do SAR foi maior do que do SEM, indicando que o primeiro explica melhor a variável dependente. Os valores obtidos para o critério de Akaike (AIC) com o modelo SAR foram menores que os valores obtidos para o modelo SEM (para matrizes binária e padronizada), confirmando que o SAR se ajusta melhor.

Por meio da análise empírica conclui-se que, quando existe dependência espacial, em geral o SAR será um modelo que melhor se ajusta aos dados (quando comparado com o SEM). Quando a dependência espacial é inexistente ou muito baixa não cabe comparar um modelo com o outro, já que nesse caso não é indicada a utilização de um modelo de regressão espacial e sim de outras metodologias que melhor se adequem a natureza dos dados.

Para corroborar a análise, o seguinte estudo será feito: primeiro será rodado um FAR na variável dependente e analisado, como medida de ajuste desse modelo, o R^2; depois será ajustada uma regressão clássica, e, nos resíduos gerados por ela, será ajustado um modelo FAR, com um respectivo R^2. Como o modelo SAR, em geral, se mostrou melhor que o SEM no ajustamento espera-se, com esse exercício, que o o FAR na variável dependente seja mais bem ajustado do que o FAR nos resíduos da regressão clássica - dado que esse último é semelhante ao modelo SEM.

Os resultados mostrados na Tabela 4.25 foram gerados apenas para os bancos de dados onde era plausível o uso de modelos de regressão espacial. Essa tabela mostra que, como esperado, a dependência espacial na variável y (variável dependente) gera,

Tabela 4.25: Comparação SAR e SEM - Utilização do R^2 do FAR

Modelo FAR		
Base	Variável	R^2
I de Moran Máximo	Dependente	0,882
I de Moran Máximo	Resíduo	0,788
Columbus	Dependente	0,449
Columbus	Resíduo	0,103
Rio de Janeiro	Dependente	0,215
Rio de Janeiro	Resíduo	0,070

em geral, um melhor ajustamento do modelo. Esse fato, portanto, confirma o que foi analisado nessa seção por meio da comparação dos resultados de ambos os modelos.

O modelo SAR é, portanto, geralmente o melhor modelo espacial a se utilizar, quando comparado com o SEM. Além de, em geral, se ajustar melhor aos dados esse é um modelo mais intuitivo de ser aplicado e mais fácil de ser interpretado. Como apontado na seção 2.3 o fato da dependência espacial entrar no modelo como uma variável explicativa é algo simples de se entender e seus parâmetros também são de fácil interpretação, mesmo para uma pessoa que não é da área. Com o modelo SAR pode-se "traduzir em palavras" o modelo - como, por exemplo, a escolaridade do meu município é explicada pela escolaridade dos municípios vizinhos e por outras variáveis. Adicionar a dependência espacial apenas no erro aleatório torna o resultado final menos claro, principalmente quando o interessado não conhece bem técnicas estatísticas.

Apesar de, no geral, o modelo SAR ser indicado quando analisados os critérios AIC e R^2 não se pode descartar o uso do modelo SEM sempre. Quando da opção por um modelo de regressão adequado é importante sim analisar critérios como os mencionados, mas a utilização do modelo melhor avaliado não é compulsória.

É possível que se opte pela utilização do SEM mesmo que seu desempenho nas medidas de ajuste indique desempenho aquém do SAR. É importante levar em conta os conhecimentos subjetivos do pesquisador e os objetivos do estudo, é possível que o modelo com ajuste mais pobre aos dados possua interpretação mais natural ou preencha alguma outra premissa da pesquisa.

Caso o interessado na modelagem saiba que não existe influência espacial como variável explicativa - mais especificamente no termo $\rho\mathbf{W_1y}$ - o modelo a ser utilizado pode ser o SEM. Existe ainda a hipótese de o pesquisador não desejar que a influência da espacialidade apareça apenas na variável explicativa y e sim em um conjunto maior de variáveis. Por exemplo: em um estudo que se tem a renda como variável dependente, pode ser que o pesquisador não queira que apenas a variável renda dos vizinhos entre no modelo com uma informação espacial, mas sim que essa influência seja mais geral, de outras variáveis. Quando se utiliza o SAR a dependência espacial fica restrita a uma variável só, presente no termo $\rho\mathbf{W_1y}$. Ao deixar essa dependência no erro aleatório essa dependência também aparecerá no termo $(\mathbf{I} - \lambda\mathbf{W_2})\,\boldsymbol{X\beta}$, como foi mostrado na Equação 3.3.

Capítulo 5

CONCLUSÃO

Este trabalho abordou dois estudos: o primeiro era verificar as diferenças existentes entre os modelos SAR e SEM, visto que aparentemente eles eram muito próximos um do outro; o segundo era verificar as diferenças existentes entre os modelos SAR e SEM, visto que aparentemente eles eram muito próximos um do outro. A princípio, a análise estrutural de todos os modelos foi utilizada para se chegar ao resultado desejado. Essa metodologia, entretanto, não se mostrou como suficiente, dado o nível de aprofundamento deste trabalho. Uma análise empírica, portanto, foi realizada para detalhamento maior do estudo.

Comparando-se a estrutura dos modelos SAR e SEM, chegou-se a conclusão de que o SEM é um caso particular do SAR. A estrutura de dependência espacial característica deste modelo ($\rho\mathbf{Wy}$) também aparece no modelo SEM. Na comparação empírica buscou-se verificar qual dos dois modelos melhor se ajustava, ou seja, qual dos dois melhor explicava a variável dependente. Com base nos resultados, chegou-se a conclusão de que, em geral, o modelo que melhor se ajusta quando existe dependência espacial é o SAR. Entretanto, isso não elimina o uso do SEM. Por apresentar mais que um termo com o parâmetro que representa a dependência es-

pacial o pesquisador pode desejar, baseado em seus conhecimentos prévios e em seu julgamento subjetivo, utilizar o SEM na modelagem.

Já para o modelo SAC, a análise estrutural foi inconclusiva. Devido a sua maior complexidade uma abordagem matemática mais aprofundada deve ser feita, o que não cabe a este trabalho. A análise empírica se tornou, portanto, a mais adequada para o estudo do modelo. Um problema apontado em alguns resultados da análise empírica, foi o aparecimento de coeficientes para ρ e(ou) λ maiores que —1—, o que não é possível já que esses são coeficientes no intervalo de -1 e 1. Esses problemas ocorreram quando as matrizes de vizinhança foram excessivamente grandes e com grande parte dos elementos não nulos. A solução adotada para resolver o problema foi aumentar o número de zeros na matriz (através da estipulação de distâncias de corte). Outro problema apontado durante a análise dos dados é a grande variação dos valores estimados com diferentes padronizações da matriz de vizinhança e diferentes configurações dessas matrizes $\mathbf{W_1}$ e $\mathbf{W_2}$ (quando elas são iguais ou distintas).

A análise empírica evidenciou que o SAC só apresenta um bom ajustamento quando há uma forte dependência espacial nos dados e, em geral, melhores resultados de ajustamento foram obtidos com as matrizes binárias padronizadas (comparando-as com as matrizes de distância padronizadas). A matriz de distância padronizada teve que ser ajustada nos casos que a dependência espacial era alta para que os coeficientes não ultrapassassem seus limites (problema apontado anteriormente). É importante ressaltar que a distância de corte possui uma influência no resultado final também. Menores distâncias de corte geram resultados com maior dependência espa-

cial e maior R^2. Esse resultado, entretanto, influencia negativamente as variáveis do modelo que não possuem dependência espacial. Portanto, a escolha do valor de corte deve ser cuidadosamente analisada, com base em medidas de ajuste e significância dos parâmetros. Da mesma forma, as matrizes de vizinhança devem ser escolhidas considerando tudo que foi dito, bem como se $\mathbf{W_1}$ e $\mathbf{W_2}$ serão iguais ou não. No caso de medidas de ajuste próximas, é importante frisar dois pontos: a análise por matrizes iguais, no caso de apenas uma delas exigir medidas corretivas, pode ser mais fiel (utiliza-se a matriz que não exigiu correção); e o segundo ponto, deve-se atentar para o sinal do intercepto, no caso de matrizes distintas não se observa inversões de sinal tão frequentes como no caso complementar, variações essas que podem indicar problema estutural no modelo, semelhantes ao caso de multicolinearidade.

Referências Bibliográficas

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science.* Volume 89, number 1.

Florax, R. J., Folmer, H., & Rey, S. J. (2003). Specification searches in spatial econometrics: the relevance of hendry's methodology. *Regional Science and Urban Economics.*

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician.* 5 115-145.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macroeconomics: The transactions demand for money. *In P.Ormerod (Ed.), Economic Modelling.* pp. 217-242.

Klaassen, L. & Paelinck, J. (1979). *Spatial Econometrics.* Saxon House.

Lembo, A. J. (2005). Spatial autocorrelation. *Department of Crop and Soil Sciences.*

LeSage, J. P. (1999). *The Theory and Practice of Spatial Econometrics.* University of Toledo.

Maddala, G. S. (1992). *Introduction to Econometrics.* Macmillan.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika.*

Silva, A. R. (2006). Avaliação de modelos de regressão espacial para análise do cenário de transporte rodoviário de carga.

Silva, A. R. (2007). Análise da matriz de proximidades espacial para problemas de transporte. *XXI Congresso da ANPET - Associação Nacional de Pesquisa e Ensino em Transportes.* Rio de Janeiro.

Apêndice A

Programação SAS para os modelos espaciais

```
%macro far(baseY,varY,matrizw1);
/******************* far model ********************/

proc iml;
use &baseY. var {&varY.};
read all into y;
use &matrizw1.;
read all into w;
close &matrizw1.;
n=nrow(y);
I=I(n);
y=y-y[:];
start max_like(p) global (y, w, n, I);
lnl=-(n/2)*log((1/n)*(y-p*w*y)`*(y-p*w*y))+log(abs(det(I-p*w)));
return(lnl);
finish max_like;
p=0.01;
optn={1};
call nlpnms(rc,xr,"max_like",p,optn);
p=xr;print p;
sigma2=(1/n)*(y-p*w*y)`*(y-p*w*y);
detval=det(I-p*w);
loglik=-(n/2)*log(2*3.14159)-(1/2*sigma2)*((y-p*w*y)`*
(y-p*w*y))-(n/2)*log(sigma2)
+log(detval);
```

```
yhat=p*w*y;
res=y-yhat;
rsqr1=res`*res;
rsqr2=y`*y;
rsqr=1-rsqr1/rsqr2;
varp=j(2,2,0);
b=I-p*W;
wb=W*inv(b);
term1=trace(inv(b`*b)*(W`*W));
varp[1,1]=term1+trace(w*inv(b)*w*inv(b));
varp[2,2]=n/(2*sigma2*sigma2);
varp[2,1]=-(1/sigma2)*(p*term1-trace(inv(b`*b)*w));
varp[1,2]=varp[2,1];
varp=inv(varp);
tstat=p/sqrt(varp[1,1]);
probt=2*(1-probt(abs(tstat),n-2));
print 'Rho' p tstat probt,,
sigma2 loglik;
create inf_far var{rsqr tstat probt p sigma2 loglik};
append;
create pred_res_far var{yhat res y};
append;
quit;
%mend;

/********** sar model ********************/
%macro SAR(baseY,varY,baseX,varX,matrizw1);

proc iml;
use &baseY. var{&varY.};
read all into y;
use &baseX. var {&varX.};
read all into x;
x=choose(x=.,0,x);
use &matrizw1.;
read all into w;
n=nrow(y);
eig=eigval(x`*x);
maxa=max(eig);
mina=min(eig);
```

```
IC=sqrt(maxa/mina);
l=j(n,1,1);
x=l||x;
B0=inv(x`*x)*x`*y;
e0=y-X*B0;
Bl=inv(x`*x)*x`*W*y;
el=W*y-X*Bl;
I=I(n);
free l;
start max_like(p) global (y,W,n,x,e0,el,I);
lnl=-(n/2)*log((1/n)*(e0-p*el)`*(e0-p*el))+log(abs(det(I-p*w)));
return(lnl);
finish max_like;
p=0.01;
optn={1};
call nlpnms(rc,xr,"max_like",p,optn);
p=xr;print p;
B=(B0-p*Bl);
sigma2=(1/n)*(e0-p*el)`*(e0-p*el);
yhat=p*w*y+x*B;
nvar=ncol(x);
res=y-yhat;
rsqr1=res`*res;
ym=y-y[:];
rsqr2=ym`*ym;
rsqr=1-rsqr1/rsqr2;print rsqr;
rsqradj=1-((n-1)/(n-nvar))*(1-rsqr);
fvar={'Modelo','Erro','Total'};
gl=j(3,1,0);
SQ=j(3,1,0);
QM=j(2,1,0);
gl[1]=nvar-1;
gl[3]=n-1;
gl[2]=gl[3]-gl[1];
SQ[3]=sum(ym#ym);
SQ[2]=sum(res#res);
SQ[1]=SQ[3]-SQ[2];
QM[1]=SQ[1]/gl[1];
QM[2]=SQ[2]/gl[2];
testeF=QM[1]/QM[2];
```

```
probf=1-probf(testef,gl[1],gl[2]);
print "Anova",,fvar gl SQ QM testeF probf;
create anova var{fvar gl SQ QM testeF probf};
append;
t=I-p*w;
ti=inv(t);
pterm = trace(W*ti*W*ti + W*ti*(W*ti)`);
xpx = j(nvar+2,nvar+2,0);
xpx[1:nvar,1:nvar] = (1/sigma2)*(x`*x);
xpx[1:nvar,nvar+1] = (1/sigma2)*x`*W*ti*x*B;
xpx[nvar+1,1:nvar] = xpx[1:nvar,nvar+1]`;
xpx[nvar+1,nvar+1] = (1/sigma2)*B`*x`*ti`*W`*W*ti*x*B + pterm;
xpx[nvar+2,nvar+2] = n/(2*sigma2*sigma2);
xpx[nvar+1,nvar+2] = (1/sigma2)*trace(W*ti*ti`)+
(1/sigma2*sigma2)*(B`*x`*ti`*w`*ti*x*B)
-(1/sigma2*sigma2)*(p*B`*x`*ti`*w`*w*ti*x*B)-
(1/sigma2)*(p*trace(inv(t`*t)*w`*w))
-(1/sigma2*sigma2)*(B`*x`*w*ti*x*B);
xpx[nvar+2,nvar+1] = xpx[nvar+1,nvar+2];
xpxi = inv(xpx)*I(nvar+2);
tmp = vecdiag(xpxi[1:nvar+1,1:nvar+1]);
bvec=B//p;
do i=1 to nvar+1;
if tmp[i]<0 then do;
tmp[i]=tmp[i]*(-1);
end;
end;
tstat = bvec/(sqrt(tmp));
probt=2*(1-probt(abs(tstat),n-2));
detval = det(I-p*w);
eD = (I-p*w)*y-x*B;
tmp2 = 1/(2*sigma2);
epe = eD`*eD;
llike = -(n/2)*log(2*3.14159)-(n/2)*log(sigma2)-n/2+detval ;
tmp1=sqrt(tmp);
*AIC=-2*llike+2*(nvar+1);
AIC=n*log(sum(res#res)/n)+2*(nvar+1);
BIC=-2*llike+log(n)*(nvar+1);
/* testes heterocedasticidade */;
pi=res#res/sigma2;
```

```
B0i=inv(x`*x)*x`*pi;
e0i=pi-X*B0i;
Bli=inv(x`*x)*x`*W*pi;
eli=W*pi-X*Bli;
I=I(n);
start max_like(p_i) global (W,n,x,e0i,eli,I);
lnl=-(n/2)*log((1/n)*(e0i-p_i*eli)`*(e0i-p_i*eli))+log(abs(det(I-p_i*w)));
return(lnl);
finish max_like;
p_i=0.01;
optn={1};
call nlpnms(rc,xr,"max_like",p_i,optn);
p_i=xr;
Bi=(B0i-p_i*Bli);
pihat=p_i*w*pi+x*Bi;
meanpi=sum(pi)/n;
SQEi=(pihat-meanpi)`*(pihat-meanpi);
brus=SQEi/2;
pbrus=1-probchi(brus,nvar-1);
create brus var{brus nvar pbrus};
append;
use &baseX. var {&varX.};
read all into x;
x0=x#x;
nn=ncol(x);
if nn>1 then x1=j(nrow(x),comb(nn,2),0);
do i=1 to nn-1;
do j=i+1 to nn;
if nn<=3 then do;
x1[,i+j-2]=x[,i]#x[,j];
end;
if nn=4 & i=1  then do;
x1[,i+j-2]=x[,i]#x[,j];
end;
if nn=4 & i>1 & i<=3  then do;
x1[,i+j-1]=x[,i]#x[,j];
end;
if nn=5 & i=1 then do;
x1[,i+j-2]=x[,i]#x[,j];
end;
```

```
if nn=5 & i=2 then do;
x1[,i+j]=x[,i]#x[,j];
end;
if nn=5 & i>2 & i<=4  then do;
x1[,i+j+1]=x[,i]#x[,j];
end;
end;
end;
l=j(n,1,1);
if nn=1 then x=l||x||x0;
else x=l||x||x0||x1;
B0r=inv(x`*x)*x`*res;
e0r=res-x*B0r;
Blr=inv(x`*x)*x`*W*res;
elr=W*res-x*Blr;
I=I(n);
start max_like(pr) global (W,n,e0r,elr,I);
lnl=-(n/2)*log((1/n)*(e0r-pr*elr)`*(e0r-pr*elr))+log(abs(det(I-pr*w)));
return(lnl);
finish max_like;
pr=0.01;
optn={1};
call nlpnms(rc,xr,"max_like",pr,optn);
pr=xr;
Br=(B0r-pr*Blr);
reshat=pr*w*res+x*Br;
resr=res-reshat;
rsqr1r=resr`*resr;
resm=res-(sum(res)/n);
rsqr2r=resm`*resm;
rsqrr=1-rsqr1r/rsqr2r;
white=n*rsqrr;
param=(nvar*(nvar+1))/2;
probw=1-probchi(abs(white),param);
print "Parametros Estimados",,
{"Interepto",&varX.,"p"} bvec tmp1 tstat probt;
create par_reg_sar var{bvec tmp1 tstat probt};
append;
create inf_sar var{rsqr rsqradj sigma2 llike AIC BIC white param probw};
append;
```

```
create pred_res_sar var{yhat res y};
append;
quit;

%mend;




/**************** sem model **********************/

%macro sem(baseX,varX,baseY,varY,matrizw1);
proc iml;
use &baseY. var{&varY.};
read all into y;
use &baseX. var {&varX.};
read all into x;
x=choose(x=.,0,x);
use &matrizw1.;
read all into w;
n=nrow(y);
I=I(n);
l=j(n,1,1);
x=l||x;
nvar=ncol(x);
* Estatística I de Moran;
b = inv(x`*x)*x`*y;
e = y - x*b;
epe = e`*e;
mi = (e`*W*e)/epe;
M = I - x*(inv(x`*x))*x`;
tmw = trace(M*W);
meani = tmw/(n-nvar);
vari =  trace((M*W)*(M*W`)) + trace((M*W)*(M*W)) + tmw*tmw;
vari = vari/((n-nvar)*(n-nvar+2));
vari = vari - meani*meani;
mis = (mi-meani)/sqrt(vari);
probm = 2*(1-probnorm(abs(mis)));
/* Estatística LM */;
b = inv(x`*x)*x`*y;
```

```
e = y-x*b;
sigma2 = (e`*e)/n;
t1 = trace((W+W`)*W);
lm1 = (e`*W*e)/sigma2;
lmerr = (lm1*lm1)*(1/t1);
problm = 1-probchi(lmerr,1);
/* Estatística LR */;
b = inv(x`*x)*x`*y;
e0=y - x*b;
ed = y - x*b;
econverge = eD;
criteria = 0.001;
converge = 1;
iter = 1;
itermax = 100;
p=0.01;
do while (converge > criteria & iter < itermax);
start max_like(p) global (ed, w, I, n);
lnl=-(n/2)*log((1/n)*ed`*(I-p*w)`*(I-p*w)*ed)+log(abs(det(I-p*w)));
return(lnl);
finish max_like;
optn={1};
call nlpnra(rc,xr,"max_like",p,optn);
p=xr;
xs = x - p*W*x;
ys = y - p*W*y;
begls = inv(xs`*xs)*(xs`*ys);
eD = y - x*begls;
converge = max(abs(eD - econverge));
econverge = eD;
iter = iter + 1;
end;
l=xr;

xs = x - l*W*x;
ys = y - l*W*y;
begls = inv(xs`*xs)*(xs`*ys);
eD = y - x*begls;
Be = (I - l*W)*eD;
epe = Be`*Be;
```

```
sig1 =(1/n)*epe;
epe0 = e0`*e0;
sig0 = epe0/n;

*Estatística LM error para correlação espacial dos resíduos em um Modelo SAR;
B0=inv(x`*x)*x`*y;
e0=y-X*B0;
Bl=inv(x`*x)*x`*W*y;
el=W*y-X*Bl;
I=I(n);
start max_like(p) global (y,W,n,x,e0,el,I);
lnl=-(n/2)*log((1/n)*(e0-p*el)`*(e0-p*el))+log(abs(det(I-p*w)));
return(lnl);
finish max_like;
p=0.01;
optn={1};
call nlpnms(rc,xr,"max_like",p,optn);
p=xr;
B=(B0-p*Bl);
sigma2=(1/n)*(e0-p*el)`*(e0-p*el);
yhat=p*w*y+x*B;
e=y-yhat;
t=I-p*w;
ti=inv(t);
pterm = trace(W*ti*W*ti + W*ti*(W*ti)`);
xpx = j(nvar+2,nvar+2,0);
xpx[1:nvar,1:nvar] = (1/sigma2)*(x`*x);
xpx[1:nvar,nvar+1] = (1/sigma2)*x`*W*ti*x*B;
xpx[nvar+1,1:nvar] = xpx[1:nvar,nvar+1]`;
xpx[nvar+1,nvar+1] = (1/sigma2)*B`*x`*ti`*W`*W*ti*x*B + pterm;
xpx[nvar+2,nvar+2] = n/(2*sigma2*sigma2);
xpx[nvar+1,nvar+2] = (1/sigma2)*trace(W*ti*ti`)+(1/sigma2*sigma2)*
(B`*x`*ti`*w`*ti*x*B)
-(1/sigma2*sigma2)*(p*B`*x`*ti`*w`*w*ti*x*B)-(1/sigma2)*
(p*trace(inv(t`*t)*w`*w))
-(1/sigma2*sigma2)*(B`*x`*w*ti*x*B);
xpx[nvar+2,nvar+1] = xpx[nvar+1,nvar+2];
xpxi = inv(xpx)*I(nvar+2);
tmp = vecdiag(xpxi[1:nvar+1,1:nvar+1]);
rhot = tmp[nvar+1,1];
```

```
varr = rhot*rhot;
A = I-p*W;
AI = inv(A);
W2 =W;
T22 = trace(W2*W2 + W2`*W2);
T21 = trace(W2*W*AI + W2`*W*AI);
lm1 = (e`*W2*e)/sigma2;
Tterm = (T22 - T21*T21*varr);
TI = inv(Tterm);
lratio = lm1*lm1*TI;
problr = 1-probchi(lratio,1);

* Estatística Wald;
z = I-W;
z = I-l*W;
zi = inv(z);
t1 = trace(W*z);
t2 = trace(W*z)**2;
t3 = trace((W*z)`*(W*z));
walds = (l**2) *(t2 + t3 - (1/n)*(t1*t1));
probw = 1-probchi(walds,1);
create testes_espac var{mi mis probm lmerr problm lratio problr walds probw };
append;

* Estimando o Modelo;
b = inv(x`*x)*x`*y;
ed = y - x*b;
econverge = eD;
criteria = 0.001;
converge = 1;
iter = 1;
itermax = 100;
p=0.01;
do while (converge > criteria & iter < itermax);
start max_like(p) global (ed, w, I, n);
lnl=-(n/2)*log((1/n)*ed`*(I-p*w)`*(I-p*w)*ed)+log(abs(det(I-p*w)));
return(lnl);
finish max_like;
optn={1};
call nlpnra(rc,xr,"max_like",p,optn);
```

```
p=xr;
xs = x - p*W*x;
ys = y - p*W*y;
begls = inv(xs`*xs)*(xs`*ys);
eD = y - x*begls;
converge = max(abs(eD - econverge));
econverge = eD;
iter = iter + 1;
end;
p=xr;
xs = x - p*W*x;
ys = y - p*W*y;
begls = inv(xs`*xs)*(xs`*ys);
eD = y - x*begls;
yhat=x*begls;
res=y-yhat;
ym = y - y[:];
rsqr1 = epe;
rsqr2 = ym`*ym;
rsqr = 1 - rsqr1/rsqr2;
rsqr1 = rsqr1/(n-nvar);
rsqr2 = rsqr2/(n-1);
rsqradj = 1 - (rsqr1/rsqr2);
fvar={'Modelo','Erro','Total'};
gl=j(3,1,0);
SQ=j(3,1,0);
QM=j(2,1,0);
gl[1]=nvar-1;
gl[3]=n-1;
gl[2]=gl[3]-gl[1];
SQ[3]=sum(ym#ym);
SQ[2]=sum(res#res);
SQ[1]=SQ[3]-SQ[2];
QM[1]=SQ[1]/gl[1];
QM[2]=SQ[2]/gl[2];
testeF=QM[1]/QM[2];
probf=1-probf(testef,gl[1],gl[2]);
create anova var{fvar gl SQ QM testeF probf};
append;
Be = (I - p*W)*eD;
```

```
epe = Be`*Be;
sigma2_ =(1/n)*epe;
B = (I - p*W);
BI = inv(B); WB = W*BI;
pterm = trace(WB`*WB);
nvar=ncol(x);
xpx = j(nvar+2,nvar+2,0);
xpx[1:nvar,1:nvar] = (1/sigma2_)*x`*B`*B*x;
xpx[nvar+1,nvar+1] = trace(WB*WB) + pterm;
xpx[nvar+2,nvar+2] = n/(2*sigma2_*sigma2_);
xpx[nvar+1,nvar+2] = -(1/sigma2_)*(p*trace(WB`*WB) - trace(BI`*WB));
xpx[nvar+2,nvar+1] = xpx[nvar+1,nvar+2];
tmp = vecdiag(inv(xpx));
bvec = begls//p;
tmp=remove(tmp,nvar+2);
tmp1=tmp`;
do i=1 to nvar+1;
if tmp1[i]<0 then do;
tmp1[i]=tmp1[i]*(-1);
end;
end;
tstat = bvec/(sqrt(tmp1));
tmp1=sqrt(tmp1);
probt=2*(1-probt(abs(tstat),n-2));
g=abs(det(I-p*W));
if g=0 then g=10**(-40);
llike = -(n/2)*log(2*3.14159)-(n/2)*log(sigma2_)-n/2+log(g);
AIC=n*log(sum(res#res)/n)+2*(nvar+1);
BIC=-2*llike+log(n)*(nvar+1);

create par_reg_sem var{bvec tmp1 tstat probt};
append;
create inf_sem var{rsqr rsqradj sigma2 llike AIC BIC};
append;
create pred_res_sem var{yhat res y};
append;
quit;
%mend;
```

```
/*********** modelo geral *****************/
%macro sac(baseX, varX, baseY, varY, matrizw1, matrizw2);

proc iml;
use &baseY. var{&varY.};
read all into y;
use &baseX. var {&varX.};
read all into x;
x=choose(x=.,0,x);
use &matrizw1.;
read all into w1;
use &matrizw2.;
read all into w2;
n=nrow(y);
I=I(n);
l=j(n,1,1);
x=l||x;
nvar=ncol(x);
start max_like(parm) global(I,n,y,x,W1,W2);
z1=(I-parm[1,1]*W1);
z2=(I-parm[1,2]*W2);
b=inv(x`*z1`*z1*x)*(x`*z1`*z1*z2*y);
ed=z2*y-x*b;
epe =ed`*z1`*z1*ed;
lnl = -(n/2)*log(epe/n) + log(abs(det(I-parm[1,1]*w1)))
+ log(abs(det(I-parm[1,2]*w2)));
return(lnl);
finish max_like;
parm={0.01 0.01};
optn={1};
call nlpnra(rc,xr,"max_like",parm,optn);
p=xr[1,1];
l=xr[1,2];
A = I-p*W1;
B = I-l*W2;
b0= inv(x`*B`*B*x)*(x`*B`*B*A*y);
e = B*A*y-B*x*b0;
yhat =y-e;
res=y-yhat;
sigu = e`*e;
```

```
sigma2 = sigu/n;
ym=y-y[:];
rsqr1=sigu;
rsqr2=ym`*ym;
rsqr=1-rsqr1/rsqr2;
rsqr1=rsqr1/(n-nvar);
rsqr2=rsqr2/(n-1);
rsqradj=1-(rsqr1/rsqr2);
fvar={'Modelo','Erro','Total'};
gl=j(3,1,0);
SQ=j(3,1,0);
QM=j(2,1,0);
gl[1]=nvar-1;
gl[3]=n-1;
gl[2]=gl[3]-gl[1];
SQ[3]=sum(ym#ym);
SQ[2]=sum(res#res);
SQ[1]=SQ[3]-SQ[2];
QM[1]=SQ[1]/gl[1];
QM[2]=SQ[2]/gl[2];
testeF=QM[1]/QM[2];
probf=1-probf(testef,gl[1],gl[2]);
create anova var{fvar gl SQ QM testeF probf};
append;
xpx = j(nvar+3,nvar+3,0);
BI = inv(B); AI = inv(A); WB = W2*BI; WA = W1*AI;
xpx[1:nvar,1:nvar] = (1/sigma2)*(x`*B`*B*x);
term1 = trace(WA*WA);
term2 = (1/sigma2)*trace(W1`*B`*B*WA*(x*b0)*(x*b0)`*AI`);
term3 = trace(W1`*B`*B*W1*inv(B*A)`);
xpx[nvar+1,nvar+1] = term1+term2+term3;
term1 = trace(WB*WB);
term2 = trace(WB`*WB);
xpx[nvar+2,nvar+2] = term1+term2;
xpx[nvar+3,nvar+3] = n/(2*sigma2**2);
xpx[1:nvar,nvar+1] = (1/sigma2)*(x`*B`*B*WA*x*b0);
xpx[nvar+1,1:nvar] = xpx[1:nvar,nvar+1]`;
xpx[nvar+2,1:nvar] = j(1,nvar,0);
xpx[1:nvar,nvar+2] = xpx[nvar+2,1:nvar]`;
xpx[nvar+3,nvar+1] = (1/sigma2)*trace(W1*AI);
```

```
xpx[nvar+1,nvar+3] = xpx[nvar+3,nvar+1];
xpx[nvar+3,nvar+2] = (1/sigma2)*trace(W2*BI);
xpx[nvar+2,nvar+3] = xpx[nvar+3,nvar+2];
term1 = trace(W1`*W2*inv(B*A)`);
term2 = trace(W2`*B*WA*inv(B`*B));
xpx[nvar+1,nvar+2] = term1+term2;
xpx[nvar+2,nvar+1] = xpx[nvar+1,nvar+2];
tmp = vecdiag(inv(xpx));
tmp=remove(tmp,nvar+3);
tmp1=tmp`;
bvec=b0//p//l;
do i=1 to nvar+2;
if tmp1[i]<0 then do;
tmp1[i]=tmp1[i]*(-1);
end;
end;
tstat = bvec/sqrt(tmp1);
tmp1=sqrt(tmp1);
probt=2*(1-probt(abs(tstat),n-2));
llike = -(n/2)*(1+log(2*3.14159))-(n/2)*log(sigma2)+log(abs(det(I-p*W1)))
+log(abs(det(I-l*W2)));
AIC=n*log(sum(res#res)/n)+2*(nvar+1);
BIC=-2*llike+log(n)*(nvar+2);
create par_reg_sac var{bvec tmp1 tstat probt};
append;
create inf_sac var{rsqr rsqradj sigma2 llike AIC BIC};
append;
create pred_res_sac var{yhat res y};
append;
quit;

%mend;
```

Apêndice B

Programação SAS as simulações dos bancos com Máxima e Mínima dependência espacial

```
data Populacao_goias;merge mapa.Populacao_goias(in=a) mapa.coordenadas(where=(UF='GO'))
proc iml;
use Populacao_goias;
read all var{x} into x;
read all var{y} into y;
COORD=x || y;
n=nrow(coord[,1]);
d=j(1,3,0);
nome={"idi" "idj" "d"};
create _dist_ from d[colname=nome];
do i=1 to n;
do j=i+1 to n;
d[1]=i;
d[2]=j;
d[3]=sqrt((COORD[i,1]-COORD[j,1])**2+(COORD[i,2]-COORD[j,2])**2);
append from d;
end;
end;
quit;
proc sort data=_dist_;by d;run;
/* Banco com dependencia mínima*/
data Populacao_goiasaleat;set Populacao_goias;
idi=_n_;
```

```
aleatorio=((rannor(2)+3)*2966)-600;/*alterações para aproximar o valor máximo e mínimo
última, no banco subsequente */
run;
/* Banco com dependencia máxima*/
data _dist_2; set _dist_;
max=(d*500000)/4.6103822717;
run;
data um;
input idi idj;
cards;
1 1
;
run;
data _dist_3;
set _dist_2 um;
if idi^=1 then delete;
if idi=1 and idj=1 then max=0;
run;
proc sql noprint;select sum(1/d) into: sum from _dist_3;quit;%put &sum;
data _dist_3;set _dist_3;
max=500000*((1/d)/&sum);
if idi=1 and idj=1 then max=17800; /*17800 como valor máximo da variável max (para dimi
run;
proc sql;select sum(max) from _dist_3;quit;
proc  sort data=_dist_3;
by idj;
run;
data moranmax;
merge _dist_3 populacao_goias;
keep max d nome populacao casa codigo x y;
run;
goptions reset=all reset=global;
proc gmap data=moranmax map=mapa.goias all;
id codigo;
choro max;
run;
quit;
/*simulaçao dependencia máxima columbus*/
data mapa.columbus;
set mapa.columbus;
```

```
id=code;
run;
data um;
input idi idj d;
cards;
1 1 0
;
run;
data dist2;
set dist um;
if idi^=1 then delete;
if idj=1 and idj=1 then max=0;
run;
proc sql noprint;select sum(1/d) into: sum from dist2;quit;%put &sum;
data dist2;set dist2;
max=500000*((1/d)/&sum);
if idi=1 and idj=1 then max=32500;
run;
data columbusmax;
merge dist2  mapa.Columbus_base;
keep max d crime inc x y code;
run;
data mapa.columbus_base;
set  mapa.columbus_base;
drop maxy miny maxwy minwy xxd xxm;
run;
proc gmap data=columbusmax map=mapa.columbus all;
id code;
choro max;
run;
quit;
```