Relatório final

“Econometria Espacial”

Orientando: Thiago Moraes Rizzieri

Orientador: Prof. Dr. José Silvio Govone

de 20 de junho de 2022 até 20 de dezembro de 2022

Thiago Moraes Rizzieri

Sumário

Introdução

Este relatório de iniciação científica apresenta alguns tópicos iniciais da Econometria Espacial, como os principais efeitos espaciais, tipos de dados espaciais, diferentes formas de mensurar a proximidade entre os dados, análise exploratória de dados espaciais e, por fim, alguns exemplos com aplicação de modelos espaciais.

Vale ressaltar que o conteúdo presente requer um conhecimento básico de probabilidade, como entender a definição e propriedades básicas da função de probabilidade, ter conhecimento sobre esperança amostral e conhecer o conceito de distribuição de probabilidade, bem como as distribuições mais conhecidas.

E também requer um conhecimento básico de estatística, como conhecer as medidas de centro: Média, Moda, Mediana, bem como as medidas de variabilidade: Variância, Desvio padrão, desvio médio. Além disso, a variabilidade entre variáveis também será bem importante, onde se trabalha com covariância e correlação, além das medidas de forma, como a assimetria e a curtose. Representação em gráficos como o histograma também é fundamental. É importante que se saiba um pouco sobre inferência estatística também, como por exemplo, sobre a estimação, teste de hipóteses e criação de intervalos de confiança.

Capítulo 1

Definindo a Econometria Espacial

A Econometria Espacial é um ramo da Econometria que incorpora **efeitos espaciais** em seus modelos. Assim, as observações representam diferentes regiões, como cidades, estados, países, setores censitários, bairros, entre outros.

Essa presença requer uma atenção especial, pois ela invalida alguma das hipóteses do Modelo Clássico de Regressão Linear, as chamadas hipóteses de Gauss-Markov. Assim, acaba por invalidar o Teorema de Gauss-Markov que garante que o estimador pelo Método dos Mínimos Quadrados ordinários é o melhor estimador linear não viesado, também conhecido por BLUE (ALMEIDA, 2012).

À seguir, veremos alguns dos principais efeitos espaciais.

1.1 Efeitos Espaciais: Dependência Espacial

Quando trabalhamos com a Econometria Espacial, os dados deixam de ser independentes entre si. Essa afirmação vêm da primeira lei da geografia:

Definição 1.1. ***Lei de Tobler (Primeira Lei da Geografia):*** *Tudo depende de todo o restante, porém o que está mais próximo depende mais do que aquilo que está mais distante.*

Primeiramente, podemos observar que há a noção de dependência presente nesta lei, mas essa dependência depende da noção de proximidade, sendo que o mais próximo é o que influência mais. Porém, vale destacar que a ideia de proximidade pode não ser apenas física, mas também social, econômica e política.

Definição 1.2. ***Dependência Espacial:*** *A variável y_i na região i depende do valor dessa variável nas regiões próximas j, chamadas de y_j, além das variáveis explanatórias exógenas X. Ou seja,*

$$y_i = f(y_j, X), \text{ sendo } i, j = 1, \cdots, n, \text{ com } i \neq j$$

Assim, temos presente no modelo que a variável y na região i não depende apenas das variáveis X, mas também da variável y nas regiões vizinhas j, como descrito na primeira lei da geografia.

Observação: as variáveis *temporais* são **unidirecionais**, ou seja, o passado pode influenciar o futuro, mas o oposto não é verdade. Assim, a variável y em um certo momento t, também descrita como y_t, dependerá de y_{t-1}, mas o contrário não é válido.

Já a dependência espacial é **multidirecional**, podendo ter dependência de cada direção do espaço, e assim como a região A depende da região próxima B, a região B depende de A.

Existem três fontes primárias para a dependência espacial: **a interação espacial, o erro de medida dos dados espaciais e a má especificação do modelo**.

1.1.1 Interação espacial

A interação espacial possui vínculo teórico. De acordo com Odland (1988), a interação espacial é o movimento de bens, pessoas ou informação através do **espaço**.

Os eventos em um certo lugar podem afetar as condições de outro lugar se os lugares interagem entre si.

De acordo com Haining (1990), existem quatro processos espaciais na Econometria: **difusão, troca de bens e serviços, comportamento estratégico e espraiamento de um atributo.**

Difusão

Difusão é a adoção de um atributo de interesse por parte dos elementos de uma população fixa. Ex: inovações tecnológicas.

Troca de bens e serviços

Como o próprio nome diz, trata das relações comerciais entre regiões, como um sistema de insumo-produto.

Comportamento estratégico

Seja em um comportamento competitivo, como a determinação dos preços em varejo, ou cooperativo, como a disposição de lojas da mesma franquia, que deve ser a melhor possível para todas.

Espraiamento de um atributo

Em contraste com a difusão, a própria população se espalharia. Como por exemplo, em uma migração populacional como o êxodo rural.

1.1.2 Erro de medida dos dados espaciais

Anselin (1988) alerta que, como o espaço é coletado seguindo uma certa agregação, pode ocorrer uma transbordação de uma unidade espacial para outra. Temos uma boa visualização desse problema com o exemplo da Figura 1.1.

As regiões foram socio-historicamente delimitadas por 1 e 2, que poderiam ser bairros, cidades, estados, entre outros.

Já A,B e C são três eventos ocorrendo no mesmo espaço cuja divisão exata não é representada pela divisão por 1 e 2.

Pode ser que exista independência entre as áreas A, B e C, mas por conta da forma de agregação, haverá autocorrelação entre as regiões 1 e 2. Acarretando assim, numa dependência espacial via agregação.

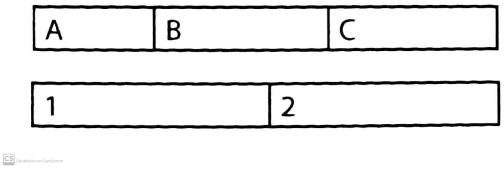

Figura 1.1: Dependência espacial pela agregação. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

1.1.3 Má especificação do modelo

Pode ocorrer de diversas formas, como por exemplo, com a omissão de uma variável relevante ou pela presença de valores discrepantes. É importante que os resíduos tenham autocorrelação espacial.

1.2 Efeitos Espaciais: Heterogeneidade Espacial

Ocorre quando não há estabilidade comportamental e, além disso, há violação da homoscedasticidade (variância não constante). Temos então, a segunda lei da geografia:

Definição 1.3. ***Segunda Lei da Geografia, por Goodchild:*** *Há diferentes respostas para y_i dependendo da localidade i ou da escala espacial β_i. Também pode ser expressa por*

$$y_i = f_i(X_i, \beta_i, \xi_i), \xi \sim (0, \Omega)$$

em que f_i denota a forma matemática funcional e ξ_i é o termo de erro. O símbolo Ω representa a matriz de variância e covariância, cuja diagonal principal não é composta por constantes. [4]

Assim, talvez a solução seja subdividir as observações por algum critério como Norte-Sul, centro-periferia ou urbano-rural.

Esse problema é extremo! Pois significa que existiria um coeficiente distinto para cada localidade.

As fontes da heterogeneidade espacial são: estrutura espacial, má especificação do modelo e erros de medida.

1.2.1 Estrutura espacial

Presente na esfera econômica, política, social, geográfica, etc. Faz os valores dos parâmetros serem diferentes dependendo de onde ocorrem. Tais caraterísticas que diferem uma região da outra podem ser valores observáveis ou não também (ALMEIDA, 2012).

1.2.2 Erros de medida dos dados espaciais

Por algum critério de agregação, assim como visto na dependência espacial.

1.2.3 Má especificação do modelo

Pode ocorrer de diversas formas, assim como visto na dependência espacial.

1.3 Efeitos Espaciais: Junção dos efeitos espaciais anteriores

Às vezes podemos ter um bom modelo com erros bem comportados, mas por conta da interdependência e interação entre regiões, pode surgir a heterocedasticidade, que consigo, leva à dependência espacial.

Um modelo ajustado como homogêneo para uma amostra de dados espaciais que exibem heterogeneidade produzirá resíduos com dependência espacial (ALMEIDA, 2012).

Temos no diagrama da Figura 1.2 como os dois efeitos espaciais se relacionam entre si. O diagrama também resume as fontes de cada um dos efeitos.

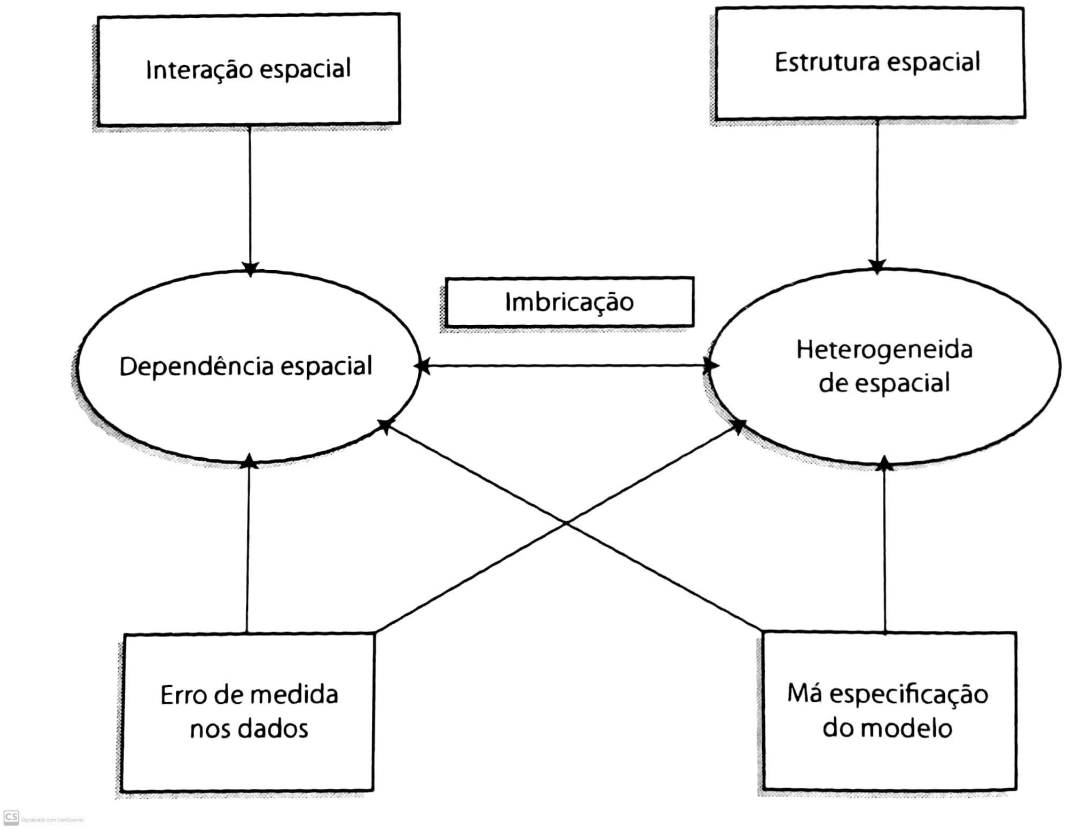

Figura 1.2: Diagrama da imbricação dos efeitos espaciais. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

1.4 Os efeitos espaciais e as hipóteses de Gauss-Markov

Veremos nesta seção como os efeitos espaciais afetam nas hipóteses de Gauss-Markov para o modelo clássico de regressão linear. Iremos citar cada uma das hipóteses e as interferências desses efeitos.

1.4.1 Linearidade dos parâmetros

Os parâmetros estarem lineares no modelo ainda é válido quando temos dependência espacial, mas nem sempre ocorre quando temos heterogeneidade espacial.

1.4.2 Não colinearidade perfeita

Rigorosamente, continua válido. Mas se o econometrista estiver trabalhando com muitas regiões, situação que dificilmente ocorre, esse problema pode ser atenuado pela grande variação nos dados.

1.4.3 Média condicional zero

É violado de diversas formas. Suponha por exemplo, que o desempenho acadêmico dos alunos dependa apenas de sua inteligência e o padrão socioeconômico de seus pais.

Contudo, considerando dados agrupados, o desempenho também varia pelo tipo de escola frequentada pelo aluno, como por exemplo, se é pública ou privada, do seu estado localizado, do seu bairro, entre outros fatores espaciais.

1.4.4 Homoscedasticidades

Diz respeito à variância ser constante, mas que para dados espaciais, é comum que não seja sempre constante por conta da dependência espacial e da heterogeneidade.

1.4.5 Independência dos erros

Precisamos de uma amostra aleatória das regiões, que dificilmente ocorre com unidades geográficas, pois caso contrário, muito provavelmente será violado. Ou seja, o erro de uma região i pode ter correlação linear com o erro da região j.

1.4.6 Normalidade dos erros

Apesar de não ser uma das hipóteses de Gauss-Markov, a normalidade dos erros é bastante verificada pela sua importância na inferência estatística. Se a hipótese de média condicional zero e/ou a hipótese de independência dos erros for violada, essa hipótese também acaba sendo violada.

Capítulo 2

O que há de especial nos dados espaciais?

O modelo espacial apresenta não apenas quanto um certo fenômeno é afetado, mas **onde** ele é mais afetado também. Iremos ver nesse capítulo, como trabalhar formalmente com os dados espaciais e os seus principais tipos.

2.1 Processo Estocástico Espacial

A ideia por trás do Processo Estocástico espacial é de que, quando vemos dados em um mapa, estamos vendo uma única realização de um processo estocástico dentre inúmeras possíveis.

Por exemplo, o PIB em um certo páis em 2020 poderia ter tido valores diferentes caso as variáveis tivessem sido um pouco diferentes. Dificilmente o PIB seria 0, com uma probabilidade quase inexistente, mas seria bem possível que o valor fosse próximo do que realmente foi.

Definição 2.1. ***(Processo Estocástico)*** *Seja T um conjunto qualquer. Um processo estocástico é uma família $\{X(t), t \in T\}$ tal que para cada $t \in T$, $X(t)$ é uma variável aleatória.*

Se T é enumerável dizemos que $\{X(t), t \in T\}$ é um ***processo estocástico discreto****, e se T não for enumerável dizemos que $\{X(t), t \in T\}$ é um* ***processo estocástico contínuo****.*

Por exemplo, T poderia ser o conjunto de meses do ano e $X(t)$ a quantidade de bicicletas vendidas por mês.

Definição 2.2. ***(Processo Estocástico Espacial)*** *Se D representa um conjunto de unidades espaciais, sendo enumerável ou não, então $\{X_i, i \in D\}$ é um* ***processo estocástico espacial****.*

Os dados espaciais são divididos entre: dados pontuais (*point pattern data*), dados agrupados em área (*lattice data*) e dados contínuos (Geoestatística).

2.2 Dados pontuais (*point pattern data*)

Neste caso, cada região i representa um ponto no espaço, usualmente chamado de **grafo**. Muito utilizado na epidemiologia e criminologia, além da Economia.

Um meio de trabalhar seria transformar em um mapa de polígonos utilizando seus centroides, através do Diagrama de Voronoy (ou polígonos de Thiessen), em que qualquer ponto do polígono é mais próximo do seu centroide do que outro centroide de outro grupo.

2.3 Dados em área (*lattice data*)

Cada região i representando um certo polígono, que podem ser regulares ou não, sendo que geralmente costumam não ser.

Dentro de cada polígono há um valor constante que se altera ao passar pela fronteira. Como por exemplo, os índices econômicos de cada estado.

2.4 Dados contínuos (Geoestatística)

Neste caso, i representa uma variável contínua. A Geoestatística é modelada usando o **variograma**.

Definição 2.3. ***Variograma:*** *é variância da diferença entre uma variável observada em duas regiões diferentes. Ou seja, revela a força de associação entre os pares de localidades como uma função contínua da distância que os separa. Observemos que o variograma é uma função crescente.*

Geralmente queremos inferir informações sobre os dados não observados (Darmofal, 2006). Por isso, utiliza-se a **Krigagem**, técnica que usa combinação linear de pesos em lugares conhecidos para prever valores nas regiões desconhecidas gerando uma superfície de previsão contínua.

A Geoestatística é útil para trabalhar com fenômenos físicos como a temperatura, porém não é tão utilizado na Economia quanto os outros dois acima.

2.5 Dados espaciais e inferência estatística

O que garante que um mapa (com informações estatísticas) seja bem representativo dentro de uma população de mapas possíveis? O que é necessário assumir?

Precisamos impor algumas condições de estabilidade do modelo, restringindo o grau de dependência e heterogeneidade do processo estocástico, que chamamos de **regularidade**.

Na ausência de regularidade, o mapa, única realização do processo estocástico espacial, não seria representativo. Uma análise confirmatória é feita à posteriori.

2.5.1 Regularidade por estacionariedade

Uma forma de regular o processo estocástico é pela chamada **estacionariedade**.

Definição 2.4. *Um processo estocástico é chamado de* ***processo estocástico estacionário*** *se satisfazer os seguintes itens:*

1. *Média constante:* $E(y_i) = \mu$;
2. *Variância constante:* $Var(y_i) = \sigma^2$;
3. *Covariância depende da distância,* d_{ij}*, entre os valores. Temos então:* $cov(y_i, y_j) = \sigma^2 c(d_{ij})$.

No terceiro item, a direção entre duas regiões, que é representada por uma angulação, não é considerada na função c, pois teríamos diferentes autocorrelações para uma mesma distância separadora de regiões. Assim, não importa se as regiões A e B estão há $100km$ de distância uma

da outra estão conectadas de cima pra baixo ou da esquerda pra direita, as regiões terão a mesma covariância em ambos os casos.

Essa propriedade que assumimos onde a direção na qual ocorre o fenômeno não possui importância é chamada de **isotropia**.

2.6 Problemas com dados espaciais

Os dados espaciais possuem alguns problemas especiais e particulares. No capítulo anterior comentamos dos efeitos espaciais, como a dependência espacial e a heterogeneidade, mas ainda há alguns outros problemas que iremos comentar.

2.6.1 Falácia ecológica

Esse problema ressalta um importante cuidado com os resultados das análises. O comportamento individual deve ser analisado por dados individuais e o comportamento agregado deve ser analisado pelos dados agregados.

Assim, a falácia ecológica consiste em usar estimativas obtidas por dados agrupados para inferir no nível individual. Mas por que essa inferência pode existir?

Primeiro, alguns comportamento individuais podem ser influenciados por dados agregados. Além disso, pode faltar dados em nível individual.

Existem alternativas para solucionar este problema, mas não iremos abordar aqui.

2.6.2 Problema da Unidade de Área Modificável

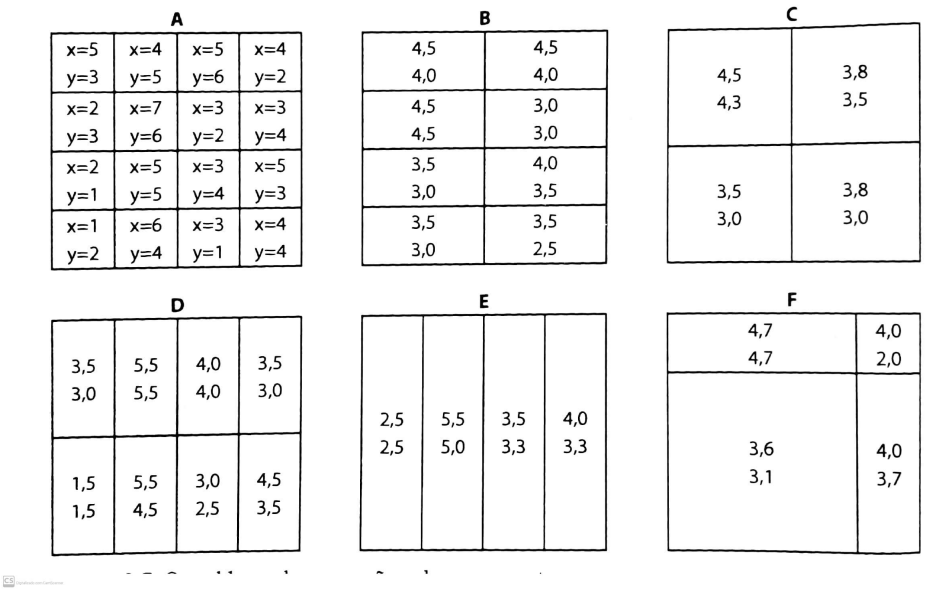

Figura 2.1: Exemplo de problema de unidade de área modificável. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

Os resultados de uma análise de dados agregados dependem da definição do critério usado para agregação. Mais especificamente, podemos destrinchar esse problema em dois.

Primeiramente, a **escala** da agregação pode afetar na variância e na correlação das variáveis. Além da escala, há várias formas de se combinar os polígonos do agregamento, como divisão por bairros, ou por outros critérios. Essas combinações são chamadas de **zoneamentos**.

Podemos visualizar melhor este problema com o exemplo na Figura 2.1. Na divisão A, cada região possui um certo valor da variável x e da variável y. Quando alguma parcela é agregada, como nas situações B, C, D, E e F, as variáveis são tomadas pelas médias dos valores x e y das parcelas envolvidas.

Na Figura 2.2 temos uma tabela com mais informações sobre esse exemplo, como a média geral de x e y, as variâncias de x e y e a correlação entre x e y, para cada zoneamento. A tabela ainda é dividida entre **Sem ponderação**, em que todas as regiões possuem o mesmo peso, e **Com ponderação**, em que maiores regiões possuem peso maior, situação que apenas ocorre no zoneamento F.

Assim, apesar das médias serem iguais, considerando a ponderação no zoneamento F, as variâncias e correlações podem variar.

Tabela 2.1. Estatísticas do impacto dos problemas de agregação e de zoneamento.

	Sem ponderação						Ponderado				
	n	$\bar{x}$	$\bar{y}$	s_x^2	s_y^2	r_{xy}	$\bar{x}$	$\bar{y}$	s_x^2	s_y^2	r_{xy}
A	16	3,88	3,44	2,36	2,37	0,66	3,88	3,44	2,36	2,37	0,66
B	8	3,88	3,44	0,30	0,40	0,88	3,88	3,44	0,30	0,40	0,88
C	4	3,88	3,44	0,14	0,26	0,94	3,88	3,44	0,14	0,26	0,94
D	8	3,88	3,44	1,55	1,34	0,95	3,88	3,44	1,55	1,34	0,95
E	4	3,88	3,44	1,17	0,98	0,98	3,88	3,44	1,17	0,98	0,98
F	4	4,06	3,36	0,16	0,93	0,64	3,88	3,44	0,18	0,48	0,80

Fonte: Adaptado de Waller e Gotway (2004, p. 106).

Figura 2.2: Tabela com estatísticas de cada zoneamento. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

Esses dois problemas, nível da escala e arranjo das zonas, compõem o Problema da Unidade de Área Modificável.

2.6.3 Efeito da beirada (Edge Effect)

Dados próximos de fronteiras podem ter correlações com regiões de fora da área de estudo. Como por exemplo, nas fronteiras de um país com outro. Além disso, regiões que estão na beirada de estudo costumam ter menos vizinhos que as que estão no interior, fornecendo menos informação para a defasagem espacial.

Na literatura existem tentativas para lidar com esse problema, mas não serão abordadas aqui.

2.6.4 Outliers e pontos de alavancagem

É comum, principalmente devido à desigualdade socioeconômica em nosso país, haver *outliers* e pontos de alavancagem, como por exemplo, com condomínios residenciais próximos de zonas periféricas.

Mas deve-se estar atento, pois por mais que sejam valores discrepantes, ainda podem trazer informações relevantes sobre a situação local.

Para visualizar esses *outliers*, além do conhecido gráfico BoxPlot, podemos utilizar o Box Map, onde cada região apresenta qual quartil está mais próximo.

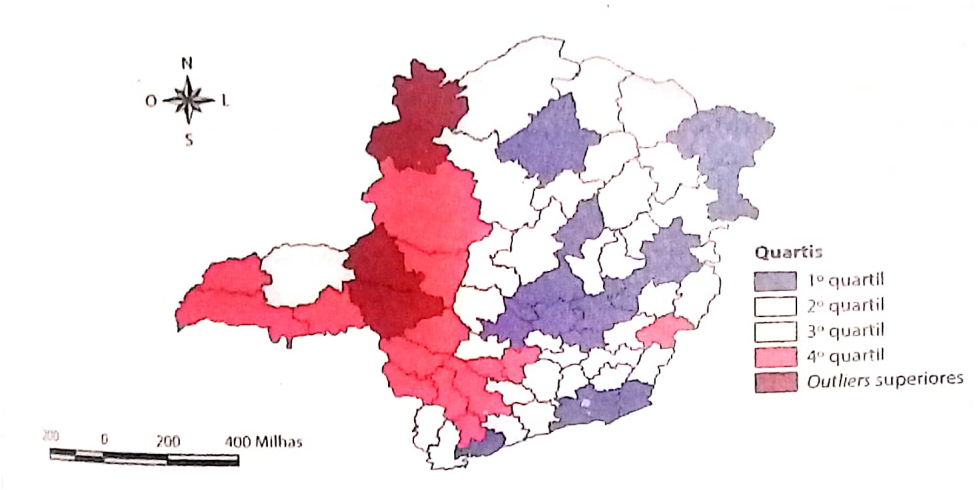

Figura 2.3: *Box map* da área colhida *per capita* em Minas Gerais. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

Capítulo 3

Matrizes de Ponderação Espacial W

Sabemos que a autocorrelação espacial implica na influência de valores vizinhos. Mas como podemos medir essa influência?

Existem duas formas para mensurarmos a dependência espacial. Uma delas é a abordagem Geoestatística, em que todos os pares de dados são classificados pela distância que os separa, usando o **variograma**, como mencionado anteriormente.

Na outra forma, precisamos impor um arranjo para as interações com uma **matriz de ponderação espacial** W. As regiões mais conectadas entre si interagem mais do que as menos conectadas, sendo que cada conexão é representada em W. Observemos que, como dito anteriormente, as conexões não precisam necessariamente serem geográficas, podendo serem culturais e institucionais também (Plaigin, 2009). Porém, há dois problemas na escolha de W:

1. Muitas vezes a escolha é arbitrária, visto que não há um teste formal para definí-la;
2. Problemas com sensibilidade dos dados à escolha da matriz.

Assim, torna-se bem importante e delicado a escolha da matriz de ponderação espacial.

3.1 Tipologia das matrizes W

Uma matriz de ponderação espacial é uma matriz quadrada de n por n, representando como as n regiões interagem entre si. Cada componente w_{ij} representa o grau de conexão entre duas regiões, ou seja, como a região i influencia na região j.

3.1.1 Proximidade geográfica

Ao trabalharmos com proximidades geográficas, podemos utilizar matrizes W que trabalhem com contiguidade entre regiões ou distância entre regiões.

Contiguidade

Neste caso, estaremos considerando as conexões das fronteiras das regiões. Como por exemplo, temos os estados que fazem fronteira com o estado de São Paulo e os que não fazem. Assim, temos uma matriz com valores binários, onde 1 representa uma conexão e 0 representa que não

há conexão. Os vizinhos de São Paulo são Minas Gerais, Rio de Janeiro, Paraná e Mato Grosso do Sul.

Podemos ter ainda uma matriz de vizinhança de segunda ordem, onde colocamos os vizinhos dos vizinhos como 1 e 0 caso contrário. Os vizinhos de segunda ordem do estado de São Paulo são Santa Catarina (vizinho de Paraná), Bahia (vizinho de Minas Gerais), Espírito Santo (vizinho de Minas Gerais e Rio de Janeiro) e Goiás (vizinho de Minas Gerais). Porém, Mato Grosso do Sul, que é vizinho de Minas Gerais, já é vizinho de São Paulo, e por isso não é colocado na matriz de vizinhança de segunda ordem. É importante que as conexões não sejam redundantes como nesse caso anterior.

Distância geográfica

Aqui utilizamos a distância entre regiões para calcular a matriz de proximidade. Uma forma bem utilizada de se calcular, é através do método dos **k vizinhos mais próximos.**

É criada uma matriz W no qual:

$$W_{ij}(k) = \begin{cases} 1, & d_{ij} \leq d_i(k) \\ 0, & d_{ij} > d_i(k) \end{cases}$$

ou seja, dado a quantidade de vizinhos próximos que queremos colocar na nossa matriz, como por exemplo limitar os dois vizinhos mais próximos, encontramos a distância de corte $d_i(k)$ que obtém os k vizinhos mais próximos. Assim, qualquer região que estiver mais próximo da distância de corte recebe 1 na matriz, e se estiver mais distante recebe 0. Observemos que a distância de corte $d_i(k)$ varia de cada região i.

Neste caso, garantimos que todas as regiões tenham o mesmo número de conexões, além de evitar problemas como casos de regiões com poucas (ou nenhuma) fronteira.

Ainda podemos calcular uma matriz de k vizinhos mais próximos com valores não binários também. Utilizamos os inversos das distâncias, ou seja:

$$W_{ij}(k) = \begin{cases} \frac{1}{d_{ij}}, & d_{ij} \leq d_i(k) \\ 0, & d_{ij} > d_i(k) \end{cases}$$

garantindo que quanto mais próximo a região estiver (distância menor), maior o valor será.

Mas quantos k vizinhos devemos considerar?

Baumont, propõe o seguinte procedimento para escolhermos:

1. Roda-se uma regressão linear por mínimos quadrados ordinários.
2. Testam-se os resíduos para autocorrelação espacial por intermédio do valor da estatística I de Moran, usando L matrizes de k vizinhos mais próximos. Variando L de $k = 1$ à $k = 20$.
3. Define-se o valor k que tenha gerado o maior valor I de Moran, que seja significativo estatisticamente.

Métricas

Para as distâncias geográficas, precisamos também definir a métrica na qual iremos calcular estas distâncias.

A mais comum é a **distância euclidiana** em que tomamos a distância entre dois pontos através do segmento que liga ambos. Pode ser bem útil quando trabalhamos com distâncias menores, mas nem sempre ele será a melhor métrica.

Lembremos que a menor distância entre dois pontos na esfera é dada pela distância entre os pontos dentro do chamado círculo maior, em que o centro do círculo coincide com o centro da esfera.

Quando trabalhamos com distâncias cada vez maiores, pode ser preferível a distância na esfera, afinal, a Terra não é plana.

3.1.2 Proximidade socioeconômica

A interação espacial pode ser melhor representada por forças socioeconômicas ao invés das espaciais.

Há três naturezas de matrizes econômicas: similaridade, dissimilaridade e por fluxos.

Ambas similaridade e dissimilaridade trabalham com as diferenças entre os índices sociais ou econômicos entre as regiões, como renda per capita, taxa de desemprego, IDH, entre outros.

Geralmente tomamos como o módulo da diferença entre os valores das duas regiões para a calcular a dissimilaridade e o inverso desse mesmo módulo da diferença para calcular a similiaridade entre as regiões.

Por fim, temos a proximidade por fluxos, ou seja, quanto há interatividade comercial entre as regiões, como acordos comerciais. Uma forma de medir seria colocando como 1 se a intensidade das relações comerciais for maior que a média e 0 se for menor.

3.2 Operador de defasagem espacial

Ao lidarmos com séries temporais, buscamos um operador de defasagem espacial B de modo que $By_t = y_{t-1}$. Ou seja, um operador que faça y retornar seu valor no tempo.

Já no contexto espacial, fica mais abstrato a ideia por onde a "defasagem pode ir".

Neste caso, temos o operador de defasagem pela multiplicação de W por y, denotada por Wy.

3.3 Qual matriz W usar?

Não há um teste formal para isso ainda, mas há uma abordagem proposta por Baumont. Mesma abordagem utilizada para encontrar o valor de k para o método dos k vizinhos mais próximos.

1. Roda-se uma regressão linear por mínimos quadrados ordinários.
2. Testam-se os resíduos para autocorrelação espacial por intermédio do valor da estatística I de Moran, usando L matrizes de k vizinhos mais próximos. Variando L de $k = 1$ à $k = 20$.
3. Define-se o valor k que tenha gerado o maior valor I de Moran, que seja significativo estatisticamente.

Capítulo 4

Análise Exploratória de Dados Espaciais

Antes de qualquer análise sofisticada, uma análise exploratória pode ser fundamental, como por exemplo, para identificar *outliers*, *clusters* e para termos uma boa visualização dos dados.

Nessa análise, podemos testar a aleatoriedade espacial do conjunto de dados, ou seja, se um valor depende mesmo do que acontece em sua vizinhança.

De acordo com Odland [7], um mapa oferece duas informações: os dados de cada região e como essas regiões estão agrupadas.

4.1 I de Moran

I de Moran é um coeficiente de autocorrelação espacial global formulado por:

$$I = \frac{n \sum \sum w_{ij} z_i z_j}{S_0 \sum z_i^2}$$

em que w_{ij} é um elemento da matriz de proximidade W, S_0 é a soma de todos os elementos de W e z representa a variável de interesse padronizada.

4.1.1 Como verificar a significância do índice

Há duas formas de se verificar. Na primeira delas, precisamos assumir que a variável padronizada $Z(I)$, dada por:

$$Z(I) = \frac{I - E(I)}{DP(I)}$$

possui distribuição normal com média zero e variância unitária. Assim, podemos verificar sua significância através da tabela normal padronizada.

A segunda é feita através de uma **permutação aleatória**. Esse procedimento é feito da seguinte forma:

1. Todos os valores da variável são permutados (embaralhados) aleatoriamente entre as regiões;
2. A estatística I é calculada para cada permutação, gerando uma distribuição empírica;
3. Podemos verificar se o nosso I original está em uma região crítica nessa distribuição.

Dessa forma, verificamos se o valor é distribuído de forma aleatória, como ao permutarmos, ou não, apresentando autocorrelação espacial significativa.

4.1.2 Diagrama de dispersão de Moran

Podemos visualizar o valor I geometricamente. Podemos realizar um gráfico de dispersão com a defasagem espacial Wz em função da variável padronizada z, como na Figura 4.1. Assim, I representa o coeficiente angular da reta de regressão desse diagrama.

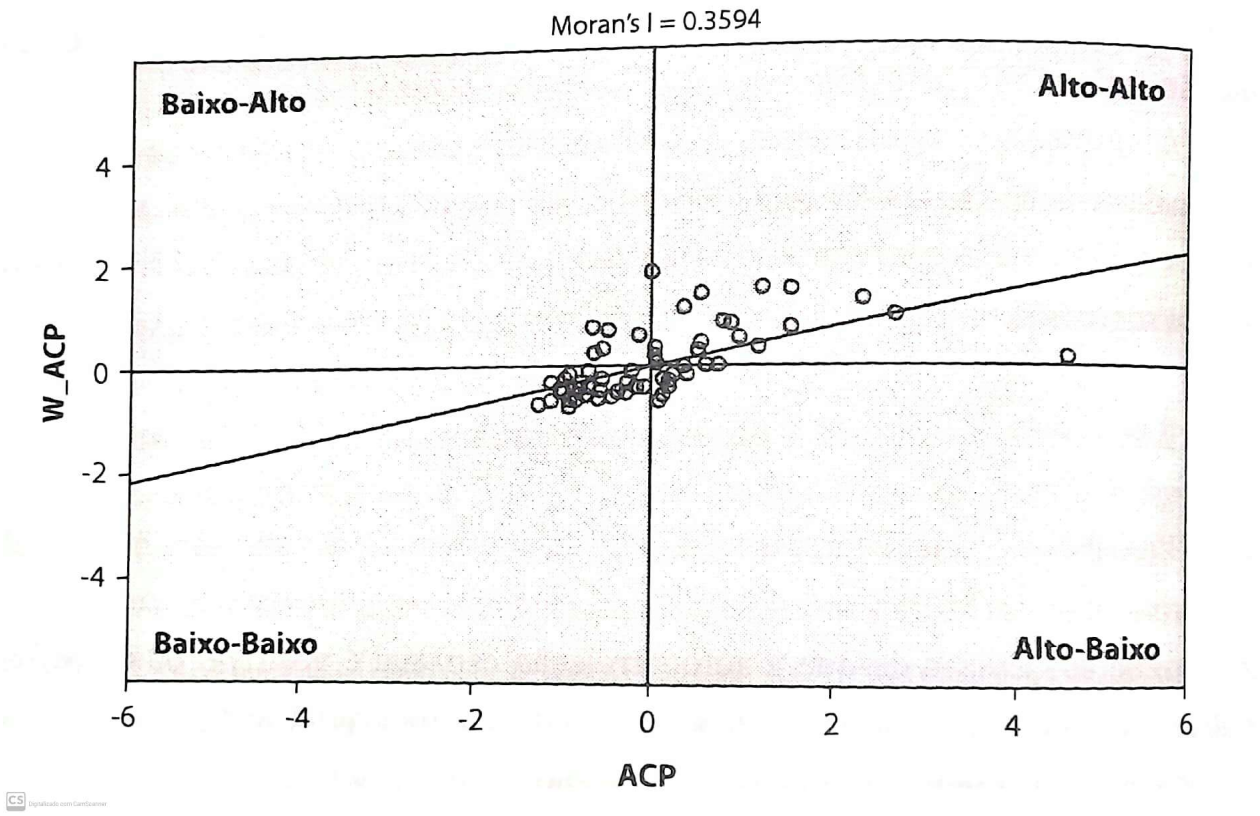

Figura 4.1: Diagrama de dispersão de Moran. Fonte: ALMEIDA, Eduardo. *Econometria Espacial Aplicada* [1].

Em cada quadrante do gráfico, podemos obter interpretações importantes. Como por exemplo, os valores que estão situados no quadrante Alto-Alto são regiões que possuem valores altos da variável y, ou seja acima da média, e são rodeadas por regiões que também possuem valores altos (ALMEIDA, 2012).

Da mesma forma, temos os quadrantes Baixo-Alto, cujo as regiões possuem valores baixos, isto é abaixo da média, próximo de regiões com valores altos, Baixo-Baixo, que são regiões com valores baixos próximas de regiões com valores baixos, e Alto-Baixo, que são regiões com valores altos próximas de regiões com valores baixos (ALMEIDA, 2012).

O diagrama é bastante útil para encontrar os *outliers* espaciais, valores que não seguem o mesmo padrão de comportamento espacial, e pontos de alavancagem, valores que seguem o mesmo padrão de comportamento espacial, mas que são muito discrepantes.

Por exemplo, na Figura 4.1 temos uma reta de regressão que possui inclinação positiva. Assim, valores extremos nos quadrantes Baixo-Alto e Alto-Baixo são considerados *outliers* espaciais, enquanto que valores extremos nos quadrantes Baixo-Baixo e Alto-Alto são considerados pontos de alavancagem.

Mas isso não deve ser feito apenas visualizando o diagrama. Estatísticas como a distância de Cook devem ser utilizadas para detectar os possíveis valores discrepantes.

4.2 C de Geary

Outra medida de autocorrelação espacial global bastante conhecida é a chamada C de Geary, dada por:

$$C = \frac{(n-1)\sum\sum w_{ij}(y_i - y_j)^2}{2\sum\sum w_{ij}\sum(y_i - \overline{y})^2}$$

Sua significância pode ser verificada das mesmas duas formas que o I de Moran descritas acima.

4.3 Joint Count

É um teste de autocorrelação para variáveis qualitativas binárias. Ela faz uma contagem das fronteiras (junções) entre as regiões.

Capítulo 5

Aplicando os modelos

Neste capítulo serão trabalhados os seguintes temas: Especificação do modelo, estimação e avaliação do modelo, através de exemplos de aplicações.

5.1 Função de produção agropecuária espacial

Em 2005, foi elaborada uma função de produção espacial agropecuária para o Estado de Minas Gerais em nível de microrregião utilizando variáveis de infraestrutura de transporte e o controle para dependência espacial, por Eduardo Almeida, autor da referência Econometria Espacial Aplicada [1].

A desagregação por microrregiões foi realizada devido ao fato de que os dados de infraestrutura de transporte estarem neste nível geográfico.

As variáveis do modelo são todas mensuradas **per capita**. Dentre elas, a variável dependente é o valor da produção agropecuária em 1996. As variáveis independentes são trabalho, capital e o estoque de infraestrutura de transportes (densidade rodoviária e densidade ferroviária).

5.1.1 Especificação e Estimação

A modelagem foi especificada como uma função de produção Cobb-Douglas. Foram estimados alguns modelos pelo MQO, através da linearização da função Cobb-Douglas, tomando o logaritmo de todas as variáveis incorporadas.

Além desses modelos, foram especificados dois modelos bem conhecidos quando buscamos modelar a dependência espacial: o modelo de defasagem espacial (SAR) e o modelo de erro autoregressivo (SEM). Descreveremos brevemente sobre cada um dos modelos.

Modelo de defasagem espacial ou modelo SAR

Neste modelo, consideramos que as variáveis dependentes, y_i, apresentam algum tipo de interação entre si, por influenciarem umas às outras. Assim, esse modelo pode ser expresso em sua forma *pura* da seguinte forma:

$$y = \rho W y + \epsilon$$

em que y é um vetor n por 1 de observações da variável dependente (considerando n regiões no modelo), ρ é o coeficiente autorregressivo espacial, de modo que $-1 < \rho < 1$, Wy é um vetor n

por 1 de defasagens espaciais e ϵ é um vetor n por 1 de termos de erro aleatório, com média zero e variância constante.

O sinal do valor ρ também nos diz muito sobre a variável dependente y_i. Se ρ é positivo, então existe autocorrelação global positiva. Ou seja, se há um valor alto (baixo) de y nas regiões vizinhas de y_i, então seu valor é aumentado (diminuido) também. Enquanto que se ρ for negativo, então há autocorrelação global negativa. Ou seja, um valor baixo (alto) de y nas regiões vizinhas, aumenta (diminui) o valor de y_i.

Caso o parâmetro espacial ρ não for estatisticamente significativo, pode-se considerar que o coeficiente é zero, não existindo evidências de que exista autocorrelação espacial.

Se incluímos as variáveis explicativas X, temos a versão mista do modelo SAR:

$$y = \rho W y + \beta X + \epsilon$$

em que X é uma matriz n por k de variáveis explicativas (considerando n regiões e k variáveis explicativas) com um vetor associado k por 1 de coeficientes de regressão β.

Modelo de erro autoregressivo espacial ou modelo SEM

Neste modelo, a dependência espacial é residual. O significado intuitivo desse modelo é de que o padrão espacial manifestado no termo de erro é dado por efeitos não modelados por conta da falta da qualidade de medida, que, por sua vez, não são distribuidos aleatoriamente no espaço, mas, ao contrário, estão espacialmente correlacionados (ALMEIDA, 2012).

Vale ressaltar que esses efeitos não modelados não podem ter correlação com alguma variável explicativa do modelo.

Podemos expressar o modelo SEM, contendo erro espacial autoregressivo de primeira ordem como abaixo:

$$\begin{cases} y = X\beta + \xi \\ \xi = \lambda W \xi + \epsilon \end{cases}$$

no qual o coeficiente λ é o parâmetro do erro autorregressivo espacial que acompanha a defasagem $W\xi$. Assim, nesse modelo, os erros associados com qualquer observação são uma média dos erros nas regiões vizinhas mais um componente de erro aleatório.

5.1.2 Avaliação

Quando há erros autocorrelacionados e/ou heterocedásticos, o coeficiente de determinação, R^2, não é apropriado como um indicador de qualidade de ajuste (ALMEIDA, 2012).

Ao invés disso, usaremos índices que utilizam o valor da função de verossimilhança (LIK), que quanto maior, mais ajustado estará o modelo. Intuitivamente, a verossimilhança é a probabilidade do quanto uma certa distribuição de probabilidade se ajusta aos dados.

Temos assim, os cri´teiors de informação Akaike (AIC) e de Schwarts (SC) dados pelas fórmulas:

$$AIC = -2LIK + 2k$$
$$SC = -2LIK + k \cdot ln(n).$$

Como ambos os critérios usam o valor negativo de LIK, quanto menores os valores, melhores são os modelos.

E no fim, o modelo SEM acabou se ajustando melhor aos dados analisados, apresentando valores menores de AIC e SC.

5.2 Eleições

Eduardo Almeida também realiza um artigo em que o principal interesse é em quantificar a importância do fator "Agora é lula", *slogan* do candidato Lula nas eleições de 2002. Esse fator transcende aspectos socioeconômicos, políticos e ideológicos, trazendo fatores da personalidade do candidato e cenário eleitoral.

Para identificar os aspectos intangíveis e não observaveis, primeiramente seria quantificado os fatores determinantes e observáveis, como fator socioeconômico, fator demográfico, fator de vulnerabilidade à violência e fator político.

Mapa de clusters

5.2.1 Especificação, estimação e avaliação

Após a especificação e estimação, novamente temos um caso em que o modelo SEM se ajusta melhor aos dados. Curiosamente, o fator socioeconômico assume coeficiente positivo e altamente significativo, revelando que os votos de Lula foram maiores em munícipios com melhores condições socioeconômicas.

Mas ainda assim, esses fatores só correspondem à 47% da variação dos dados. Logo, 53% da variação dos dados vem do tal fator "Agora é Lula". Assim, a maior parte dos votos de Lula não é explicada pelos fatores socioeconômico, demográfico, de vulnerabilidade ou político-ideológico, mas sim pelo sentimento difuso, amplo e generalizado de que havia chegado a vez do candidato vencer as eleições.

Referências Bibliográficas

[1] ALMEIDA, Eduardo. *Econometria Espacial Aplicada*. Editora Alínea, Campinas - SP, 2012.

[2] ANSELIN, L. *Spatial econometrics: methods and models.* Boston: Kluwer Academic, 1988.

[3] DARMOFAL, D. *Spatial econometrics and political science.* Mimeo, Department of Political Science, University of South Carolina, Columbia, 2006.

[4] GOODCHILD, M. *The validity and usefulness of laws in geographic information science and geography.* Annals of the Association of American Geographers, v. 94, n. 2, p. 300-303, 2004.

[5] HAINING, R. *Spatial data analysis in the social and environment sciences.* Cambridge University Press, Cambridge, 1990.

[6] HYNDMAN, R.J., ATHANASOPOULOS, G. *Forecasting: principles and practice*, 3ª edição, OTexts: Melbourne, Australia. Disponível em https://otexts.com/fpp3/. Acesso em: 20 de dezembro de 2022.

[7] ODLAND, J. *Spatial autocorrelation.* Sage publications, Londres, 1988.

[8] PLAIGIN, C. *Exploratory study on the presence of cultural and institutional growth spillovers.* III World Conference of Spatial Econometrics, Barcelona, 2009.