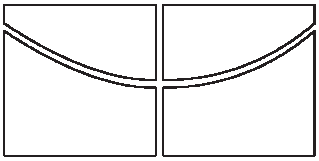

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Relação entre o Índice I de Moran e a Quantidade de Dados Observados

por

Tomaz Back Carrijo

Orientador: Prof. Dr. Alan Ricardo da Silva

Maio de 2015

Tomaz Back Carrijo

Relação entre o Índice I de Moran e a Quantidade de Dados Observados

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Maio de 2015

TERMO DE APROVAÇÃO

Tomaz Back Carrijo

Relação entre o Índice I de Moran e a Quantidade de Dados Observados

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Prof. Dr. Alan Ricardo da Silva
Departamento de Estatística - EST/UnB
Orientador

Prof. Dr. Raul Yukihiro Matsushita
Departamento de Estatística - EST/UnB

Prof. Dr. Pedro Henrique Melo Albuquerque
Departamento de Administração - ADM/UnB

Brasília, 12 de Maio de 2015

Ficha Catalográfica

CARRIJO, TOMAZ

Relação entre o Índice I de Moran e a Quantidade de Dados Observados, (UnB - IE, Mestre em Estatística, 2015)

Dissertação de Mestrado - Universidade de Brasília.
Departamento de Estatística - Instituto de Ciências Exatas.
Orientação: Alan Ricardo da Silva.

1. Autocorrelação espacial 2. Índice I de Moran
3. Quantidade de dados observados 4. Valores extremos

"Estudar ainda mais, me importar ainda menos. Rir mais, fazer mais graça, me manter sempre alegre. Ter cada vez mais disciplina, ter mais atitude, ser mais cético e menos cego. Manter minha honestidade, ser menos preocupado, menos vaidoso e mais consciente."

Agradecimentos

Acredito que estamos nessa vida para aprender. A busca ao conhecimento, em todos os níveis, é fundamental para a formação e evolução do ser humano. Sem ela, o indivíduo fica a deriva, perdido sem objetivos ou metas. Graças a Deus tive o privilégio de nascer em uma família que, desde pequeno, me mostrou esse caminho. Me incentivando a superar quaisquer obstáculos e sempre buscar meu aperfeiçoamento moral e intelectual.

Em decorrência disso, o principal agradecimento vai ao papai e a mamãe. Agradeço meu pai por ele ser meu maior ídolo e exemplo intelectual. Se eu estou aqui hoje e se eu tenho meus objetivos para o futuro, o principal responsável é você. É em você que eu me espelho. É te observando quando eu mais aprendo. Agradeço minha mãe por toda preocupação e pelo amor incondicional que ela nos dá. Por sempre abrir mão de tudo para cuidar de mim e de meus irmãos. Por ser meu maior exemplo de pessoa batalhadora. Não importou quantas vezes a vida "fechou" a porta para você. Você sempre seguiu em frente. Você merece mais do que ninguém o sucesso que você tem hoje.

Não seria um ser humano completo sem meus irmão. Definitivamente só sou o que sou por causa deles. É a melhor coisa que tenho na vida. Papai e mamãe fizeram um ótimo trabalho. Tobias, eu não poderia ter um irmão mais velho melhor. Hugo, você sempre será meu orgulho e maior exemplo de dedicação e de disciplina. Tuzza, só Deus sabe como você me aguenta. Agradeço todos os dias por isso. Eu preciso disso. Eu preciso de vocês. É bom demais tê-los como irmãos.

Queria agradecer a Paulinha, mulher da minha vida. A única pessoa que eu confio para ser mãe dos meus filhos. Você me incentiva a ser melhor, a ser mais disciplinado. Você me traz paz e me acalma da maneira mais simples e gostosa possível. Não teria

conseguido terminar em tão pouco tempo se não fosse você.

Aos meus grandes amigos: André, Bernardo e Zulba. Foi com vocês que eu aprendi a ser estatístico, a amar minha profissão, a querer me capacitar cada vez mais. São vocês que eu procuro quando tenho dúvidas. Além do mais, vocês são meus companheiros de folia. Quero manter essa amizade para o resto da vida.

Finalmente agradeço meu orientador Alan. Obrigado por você ter me oferecido esse tema. Por ter aceitado me orientar mesmo estando no pós-doutorado. Obrigado por me cobrar e me incentivar a terminar o mais rápido possível. Tenho por você uma grande amizade.

Just because someone stumbles and loses their path, doesn't mean they're lost forever.

Sumário

Lista de Tabelas

Lista de Figuras

Resumo

Devido a sua simplicidade, o índice I de Moran é a estatística mais famosa e largamente utilizada quando se deseja mensurar a autocorrelação em dados georreferenciados. A primeira parte do trabalho revisou o estado da arte desse índice, objetivando entender sua evolução e características intrínsecas. Logo em seguida, demonstrou-se que essa estatística apresenta limitações quando a quantidade de dados no sistema é pequena. Além disso, propôs uma modificação do índice I de Moran, introduzindo no coeficiente de correlação de Pearson o conceito de modelagem espacial autorregressiva de primeira ordem. Os resultados dessa proposta se mostraram bastante coerentes, principalmente quando o sistema possui poucos dados.

Palavras Chave: *Autocorrelação espacial, Índice* I *de Moran, Quantidade de dados observados , Valores extremos.*

Abstract

The Moran's I is the most famous and widely used spatial statistic when we want to measure spatial autocorrelation in geo-referenced data. The first part of this work reviewed the state of the art of Moran's I, in order to understand its evolution and intrinsic characteristics. Next, we showed that this statistic has limitations when the sample size of the system is small. In addition, we present a proposal of modification on Moran's I, introducing on the Pearson correlation coefficient the concept of the first order autoregressive spatial modelling. The results of this proposal were very consistent, especially when the system has few data.

Key Words: *Spatial autocorrelation, Moran's* I, *Sample Size, Extreme Values.*

Introdução

A estatística espacial é definida por Bailey e Gatrell (1995) como o conjunto de técnicas que considera o arranjo espacial dos dados na análise ou na interpretação dos resultados. Cliff e Ord (1981) definem "se a ocorrência de um determinado fenômeno, em alguma região, influenciar, para mais ou para menos, na ocorrência desse mesmo fenômeno em regiões vizinhas, pode-se afirmar que esse sistema apresenta autocorrelação espacial".

A análise de autocorrelação espacial é útil e aplicável nas mais diversas áreas do conhecimento. São exemplos a biologia, na análise de simulações de modelos evolutivos (Diniz-Filho e Santos, 2012); a geografia, na distribuição de mestres e doutores por microrregião brasileira (Viotti et al., 2012); a epidemiologia, no estudo do número de casos cólera nas proximidades de um poço de água (Snow, 1936); a engenharia de transportes, na atração por viagens de ônibus na cidade de Manaus/AM (Silva, 2006); a agronomia, na análise da concentração de metais pesados em solos agrícolas em Beijing, China (Huo e Zhang, 2011); e a economia, no agrupamento de regiões da Europa quanto a dinâmica de inovação e crescimento econômico (Verspagen, 2010).

O presente trabalho objetiva estudar a relação entre o índice I de Moran, definido por Moran (1950), a quantidade de dados observados e seus valores extremos. Esse índice é a estatística mais famosa e largamente utilizada quando se deseja mensurar a autocorrelação de uma determinada variável alocada em um espaço bidimensional.

Para tanto, apresentar-se-á primeiramente, o conceito e as características do coeficiente de correlação linear de Pearson. Depois, será exposta a forma de surgimento do conceito de autocorrelação em séries temporais. A seguir, definir-se-á o índice I de Moran e sua derivação a partir do coeficiente de Pearson e por fim, serão feitos estudos de caso para mostrar a problemática de sua aplicação em pequenas amostras.

Um aspecto muito importante, e inicial em todo tipo de análise, é a verificação do grau de relacionamento entre as variáveis analisadas. A análise de correlação fornece um indicador que estabelece a existência, ou não, de uma relação linear entre as variáveis analisadas, sem que para isso, seja necessário o ajuste de uma função matemática (Lira, 2004).

O coeficiente de correlação linear de Pearson é conhecido como o quociente da covariância de duas variáveis pelo produto dos seus respectivos desvios padrão (Neter e Kutner, 2005). Supondo duas variáveis aleatórias normalmente distribuídas, X e Y, o estimador de máxima verossimilhança da correlação da distribuição normal bivariada resultante dessas duas variáveis é:

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/n}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2/n}} \tag{1}$$

O coeficiente de correlação linear também pode ser interpretado como um indicador que mensura a proximidade dos dados a uma reta. Com um pouco de álgebra, é possível verificar que ele varia entre -1 e $+1$, ver Apêndice A (Lira, 2004). Na prática, $|\rho| = 1$ quando os dados obedecem perfeitamente à relação $y_i = \beta_0 + \beta_1 x_i$. Assim, pode-se dizer que a correlação entre as variáveis X e Y é $+1$ se $\beta_1 > 0$ e -1 se $\beta_1 < 0$.

Vale ressaltar que não existe relação entre o coeficiente de correlação de Pearson e a quantidade de dados observados. Mesmo em uma problemática com $n = 3$, se (x_1, y_1), (x_2, y_2) e (x_3, y_3) pertencerem a mesma equação de reta, o valor absoluto da correlação entre essas variáveis será unitário em módulo.

Existem algumas situações em que a aplicação desse coeficiente não é adequada. Exemplos dessas situações são: quando as variáveis são medidas em escala ordinal; quando os dados não apresentam associação linear; quando as variáveis não forem normalmente distribuídas e quando há presença de *outliers*. Além do mais, vale a pena ressaltar que é possível obter variáveis dependentes que, no entanto, possuem correlação nula.

Um exemplo simples dessa situação é supondo duas variáveis aleatórias X e Y, onde X é normalmente distribuída com média zero e Y é definida de modo que

$Y = X^2$. Claramente, tem-se que X e Y são dependentes. Porém, quando se calcula o coeficiente de correlação de Pearson, obtém-se $\rho = 0$.

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X)(X^2)\right] - (0)E(X^2)}{\sigma_X \sigma_{X^2}} = \frac{E(X^3)}{\sigma_X \sigma_{X^2}} = 0 \qquad (2)$$

Uma variante análoga da correlação convencional, definida por Pearson, é a correlação entre os valores de uma série temporal, ou autocorrelação. Ela é definida como a relação entre as observações de uma única variável de acordo com sua ordenação (Fuller, 1996).

Uma série temporal é qualquer conjunto de observações ordenadas no tempo. Qualquer área do conhecimento produz observações ao decorrer do tempo. Exemplos de séries temporais são: cotação do dólar em relação ao real ao final de cada dia; consumo mensal de um determinado produto; quantidade anual de CO_2 liberado na atmosfera. A importância desse tipo de análise decorre da sua alta aplicabilidade e interesse em se realizar previsões.

Uma série temporal X_t é definida completamente, no sentido probabilístico, se sua distribuição acumulada for conhecida (Fuller, 1996). Define-se que X_t é estritamente estacionária se:

$$F_{X_{t_1}, X_{t_2}, \cdots, X_{t_n}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n}) = F_{X_{t_1+h}, X_{t_2+h}, \cdots, X_{t_n+h}}(x_{t_1}, x_{t_2}, \cdots, x_{t_n}) \qquad (3)$$

No entanto, na maioria das aplicações, não se conhece a distribuição acumulada. Uma alternativa é trabalhar apenas com os dois primeiros momentos da série temporal (Fuller, 1996). Por essa abordagem, define-se que uma série de tempo é fracamente estacionária se: (i) a esperança de X_t for constante ao longo do tempo e (ii) se $\mathrm{Cov}(X_{t_1}, X_{t_2}, \cdots, X_{t_n}) = \mathrm{Cov}(X_{t_1+h}, X_{t_2+h}, \cdots, X_{t_n+h})$ para qualquer valor de h, desde que o índice $t_i + h$ pertença a série temporal.

Na prática, em sua maioria, as séries não são estacionárias. Porém, existem vários tipos de transformações que podem ser realizadas para transformar uma série não-estacionária em estacionária. Isso porque é mais conveniente analisar séries estacionárias. Dentre as mais diversas análises, existe uma função que mensura a capacidade de previsão de uma série no momento $t + h$ dado seus valores no momento t. Ela é

denominada função de autocorrelação

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\text{Cov}(X_{t+h}, X_t)}{\text{Cov}(X_t, X_t)} = \frac{E[(X_{t+h} - \mu)(X_t - \mu)]}{\text{Var}(X_t)} \tag{4}$$

sendo $\gamma(\cdot)$ denominada função de autocovariância. Assim como o coeficiente de correlação de Pearson, $\rho(h)$ varia entre -1 e $+1$, ver Apêndice B (Fuller, 1996).

Fuller (1996) apresenta dois estimadores para $\gamma(h)$ quando a média da série temporal é desconhecida. Ambos estimadores são viesados, porém o estimador $\widehat{\gamma}(h)$, em alguns tipos de séries, apresenta menor erro quadrático médio. Por isso, ele é o estimador mais utilizado.

$$\gamma^{\dagger}(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} \left(X_{t+h} - \bar{X}_n\right)\left(X_t - \bar{X}_n\right) \tag{5}$$

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} \left(X_{t+h} - \bar{X}_n\right)\left(X_t - \bar{X}_n\right) \tag{6}$$

Finalmente, Fuller (1996) relata que o melhor estimador para a função de autocorrelação de uma série temporal com média desconhecida é

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}\left(X_{t+h} - \bar{X}\right)\left(X_t - \bar{X}\right)}{\sum_{t=1}^{n}\left(X_t - \bar{X}\right)^2} \tag{7}$$

Uma dificuldade desse estimador é que seu numerador apresenta $n-h$ termos, ao passo que seu denominador apresenta n termos. Uma consequência direta é o fato de que, para $|h| > 0$, a equação (7) só atinge seus valores extremos quando n é grande.

Uma aplicação da teoria exposta pode ser realizada ao se observar a série temporal da Figura 1. É possível notar que ela satisfaz os dois critérios de estacionariedade fraca definidos por Fuller (1996). Note também que $n = 12$, apresentando uma sazonalidade perfeita para $h = 4$. Assim, tem-se que os valores $t + 4$ são exatamente iguais aos valores t para $t = 1, 2, \cdots, 8$.

É intuitivo inferir que $\rho(4) = 1$, uma vez que a série se repete a cada 4 espaços de tempo. Contudo, usando a equação (7), obtém-se apenas $\widehat{\rho}(4) = 0.66667$. Mesmo aumentando significativamente o tamanho amostral para 600 observações, a estimativa da autocorrelação não atinge seu valor máximo, sendo $\widehat{\rho}(4) = 0.99333$. Sobretudo,

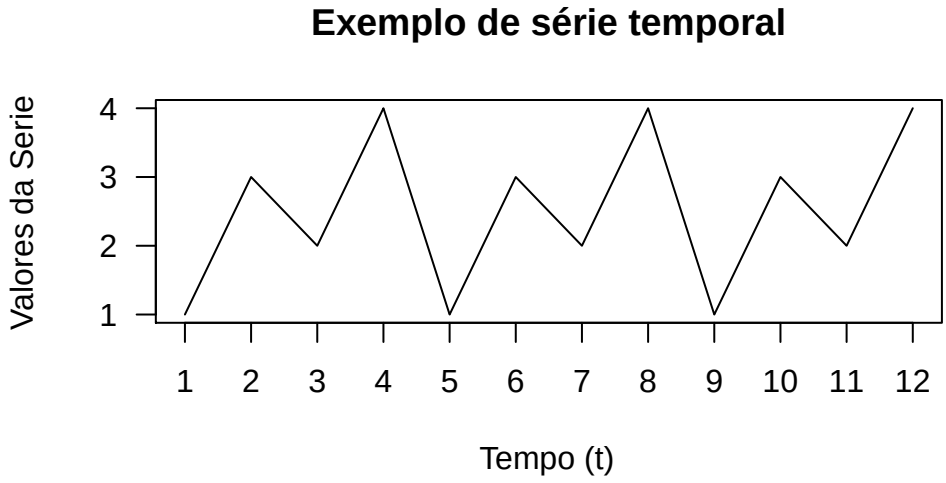

Figura 1: Exemplo de uma série temporal

nesse contexto, se fosse utilizado a equação (5) de modo a estimar as autocovariâncias, ao invés da equação (6), tanto para $n = 12$ quanto para $n = 600$, obtém-se $\rho^{\dagger}(4) = 1$.

Em estatística espacial, a autocorrelação pode ser definida como a correlação entre os valores de uma única variável dada sua posição em relação aos demais valores em um espaço bidimensional.

Cliff e Ord (1981) discutem a diferença entre determinar se dados geográficos são espacialmente autocorrelacionados com a problemática de medir a autocorrelação em séries temporais estacionárias. Isto ocorre fundamentalmente porque a variável de uma série temporal é influenciada apenas por seus valores anteriores, enquanto que um processo de dependência espacial se estende em todas as direções, como observado por Whittle (1954).

"Em uma série temporal, em qualquer instante, tem-se a distinção entre passado e futuro. O valor da variável depende apenas dos valores passados. Assim, a dependência em séries temporais se estende apenas em uma direção: para trás. Agora, supondo um problema espacial, por exemplo, um pasto de uma fazenda. Se um pouco de fertilizante for aplicado em qualquer ponto desse pasto, ele afetará a fertilidade do solo em todas as direções" (Whittle, 1954).

A proximidade entre as regiões é identificada usando uma matriz de proximidade $\mathbf{W} = \|c_{ij}\|$. Segundo Getis e Aldstadt (2004) existem pelo menos doze maneiras de $\mathbf{W}$. As mais conhecidas são: por meio de matriz binária ou com o uso de matriz de

distância. Na primeira situação, $c_{ij} = 1$ se as regiões i e j forem vizinhas, e $c_{ij} = 0$ caso contrário. Lembrando que $c_{ii} = 0 \; \forall \; i = 1, ..., n$ uma vez que uma região não pode se relacionar com ela mesma.

Ainda nesse contexto, existem três metodologias usuais para se definir o tipo de vizinhança: *rook*, *queen* e *bishop*. Essas metodologias recebem esses nomes por causa da movimentação das peças de xadrez, torre, rainha e bispo, respectivamente. A torre se movimenta para frente e para trás, para a direita e para a esquerda. A rainha se movimenta para frente ou para trás, para direita ou para a esquerda, ou nas diagonais. Já o bispo se movimenta apenas nas diagonais. As Figuras 2, 3 e 4 apresentam as regiões vizinhas a região A de acordo com cada uma dessas metodologias.

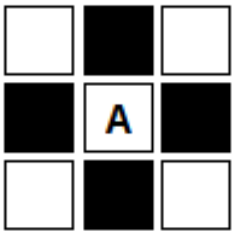

Figura 2: Metodologia *Rook*

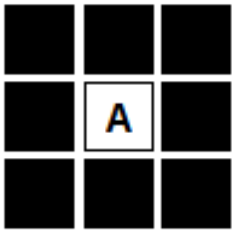

Figura 3: Metodologia *Queen*

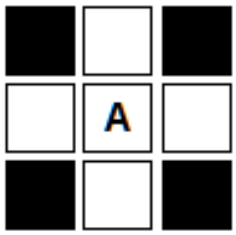

Figura 4: Metodologia *Bishop*

Adotar a metodologia *rook* significa procurar autocorrelação espacial sobre as linhas e colunas da estrutura. Já a metodologia *queen* busca autocorrelação ao longo das linhas, colunas e diagonais da estrutura. E a metodologia *bishop* procura autocorrelação apenas nas diagonais. É importante notar que a metodologia *queen* não acarreta nenhum viés direcional (Cliff e Ord, 1981).

A segunda maneira de se definir $\mathbf{W}$ é utilizando uma matriz de distância. Nesse contexto, define-se $c_{ij} = g\left(d_{ij}\right)$ em que $g\left(\cdot\right)$ é uma função qualquer e d_{ij} é a distância entre os centros geográficos das regiões do sistema, por exemplo. Vale a pena ressaltar que a detecção de autocorrelação espacial depende criticamente da forma de $\mathbf{W}$, (Getis e Aldstadt, 2004). Esse assunto será aprofundado no Capítulo 2.

Finalmente, autocorrelação espacial pode ser expressa em termos do coeficiente de correlação de Pearson, equação (1), substituindo x_i pelos vizinhos de y_i (Griffith, 2003). Essa substituição pode ser expressa algebricamente inserindo o conceito da

matriz de proximidade.

$$\mathrm{I} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\,(y_j - \bar{y})\,(y_i - \bar{y}) \,/\, \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \,/ n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \,/ n}} \tag{8}$$

Enquanto o numerador de (1) é uma média aritmética, o numerador de (8) é uma média ponderada e, por isso, deve ser dividido pelo somatório de todos os pesos, ao invés de ser dividido por n. Essa expressão é conhecida como o índice I de Moran (Moran, 1950). Cliff e Ord (1969), utilizando a desigualdade de Cauchy-Schwarz, mostram que, se $\mathbf{W}$ estiver padronizada pela linha, $\sum_{j\in J} w_{ij} = 1$, essa estatística também varia entre -1 e $+1$. Esse assunto será aprofundado no Capítulo 1.

A fim de aumentar os esclarecimentos dos objetivos do presente trabalho, as Figuras 5 e 6 apresentam duas situações de dependência espacial direta de acordo com o tipo de vizinhança em um sistema regular com $n = 25$. Em ambas situações, o valor máximo se encontra na região de número 1 e vai decaindo constantemente a cada novo grupo de vizinhos. Por exemplo, na Figura 6, a região 1 tem valor 50; as regiões 2, 6 e 7 têm valor 40; as regiões 3, 8, 11, 12, 13 têm valor 30 e assim por diante.

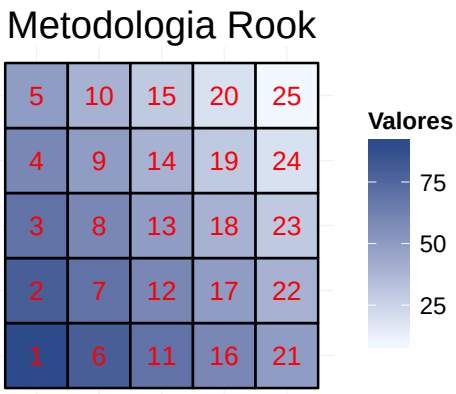

Figura 5: Sistema com 25 regiões, metodologia *rook*

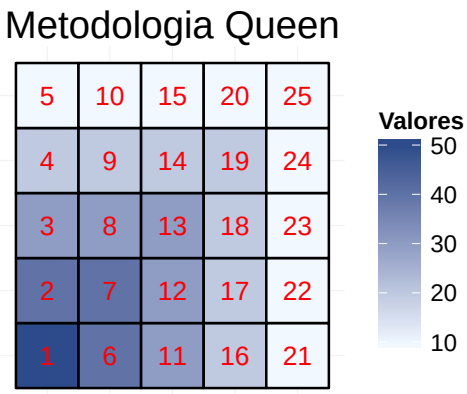

Figura 6: Sistema com 25 regiões, metodologia *queen*

O primeiro fato a ser notado é a diferença entre as dinâmicas de dependência espacial em ambos sistemas. Note que, enquanto a metodologia *rook* gerou 9 vizinhanças, a metodologia *queen* gerou apenas 5. Uma consequência direta de tal fato é que o valor do índice de Moran na Figura 5 é 0.84, ao passo que na Figura 6 é apenas 0.68.

O segundo fato digno de nota é que, mesmo definindo uma relação direta entre os vizinhos, o índice de Moran não atinge seu valor máximo. A Tabela 1 apresenta esse mesmo exercício em sistemas regulares com 9, 49, 100 e 196 regiões. Note que o índice I de Moran aparentemente tem relação com a quantidade de dados observados e que em nenhuma das situações ele atinge seu valor extremo. Utilizando a metodologia *queen*, com $n = 9$ o valor do índice é apenas 0.238, enquanto que para $n = 196$, o valor é 0.954.

Tabela 1: Comparação entre o índice de Moran e a quantidade de dados observados

Sistema	Observações	Vizinhança	I de Moran
3×3	9	*rook*	0.5556
		queen	0.2385
7×7	49	*rook*	0.9183
		queen	0.8303
10×10	100	*rook*	0.9600
		queen	0.9135
14×14	196	*rook*	0.9795
		queen	0.9547

Dessa forma, o presente trabalho possui dois objetivos: (i) revisar o estado da arte do índice I de Moran, estudando sua relação com a quantidade de dados observados e seus valores extremos; (ii) propor um índice de Moran modificado, de modo a tornar a interpretação da autocorrelação espacial mais simples e objetiva.

O primeiro objetivo corroborará para aumentar o conhecimento sobre as características intrínsecas do índice de I Moran. Em decorrência da dificuldade desse índice de atingir seus valores extremos, muitos trabalhos podem estar sub-representando a verdadeira relação espacial dos dados. O segundo objetivo, por sua vez, está relacionado ao índice I de Moran modificado. Sua proposta é justificada de modo a tornar a interpretação da autocorrelação espacial mais simples.

O presente trabalho está dividido da seguinte forma: o Capítulo 1 apresenta o estado da arte do índice I de Moran. O Capítulo 2 apresenta a matriz de proximidade $\mathbf{W}$ e discute sua influência sobre o índice em questão. O Capítulo 3 propõe uma modificação no índice I de Moran, apresenta uma discussão sobre os valores extremos desse índice e expõem uma demonstração da dependência do índice de Moran com a quantidade de dados observados. O Capítulo 4 contém aplicações das propostas do

presente trabalho. As considerações finais, limitações do trabalho e propostas de trabalhos futuros encontram-se no Capítulo 5. Os Apêndices contêm as demonstrações de que o coeficiente de correlação de Pearson, a função de autocorrelação em séries temporais e a generalização do índice de Moran (Cliff e Ord, 1969) variam entre -1 e $+1$.

Capítulo 1

I de Moran

O presente capítulo objetiva apresentar todas as formas do índice I de Moran propostas até o momento, bem como apresentar os artigos mais relevantes já publicados sobre o tema em questão. Para tanto, primeiramente, será introduzido o contexto histórico antes do trabalho de Moran (1950). Logo em seguida, será apresentado o estado da arte do índice I de Moran, além de discussão e crítica sobre alguns desses trabalhos.

1.1 Contextualização histórica

Determinadas pesquisas citam o trabalho de Moran (1950) como o marco do início da presente problemática (Verspagen, 2010; Huo e Zhang, 2011; Chen, 2013). Contudo Moran (1950) é um trabalho direcionado para fenômenos estocásticos em espaços de duas ou três dimensões e trata de dois tópicos principais: "relações entre processos contínuos e descontínuos" e "teste de hipótese da existência de fenômenos estocásticos em espaços de duas dimensões". A discussão da primeira parte se estende por aproximadamente 75% do trabalho. O índice de Moran é definido apenas na segunda parte, tornando evidente que é uma generalização da metodologia aplicada em Moran (1948).

Na realidade, a problemática se iniciou com Wishart e Hirschfeld (1936), que consideraram a questão de calcular a distribuição de probabilidade do número de ligações entre pontos "pretos" e "brancos" contidos em uma reta. Na problemática em

questão, o ponto recebe cor "preta", caracterizando a presença de uma determinado atributo, com probabilidade p, e recebe cor "branca", caracterizando a ausência desse atributo, com probabilidade $q = 1 - p$. Além de apresentar a solução exata para este problema, Wishart e Hirschfeld (1936) mostram que essa distribuição tende à normalidade para n grande.

Moran (1947) generalizou a problemática de Wishart e Hirschfeld (1936) para a segunda e a terceira dimensões, Contudo continuou abordando apenas variáveis binárias. Enunciou-a como: suponha um retângulo com mn pontos, cada um podendo ser da cor "preta" com probabilidade p e cor "branca" com probabilidade $q = 1 - p$.

Moran (1947), então, definiu o número de ligações entre células "pretas", BB , apenas para a metodologia *rook*. Dacey (1968) definiu BB para o caso *queen*.

$$\mathrm{BB}_{rook} = \sum_{i=1}^{m}\sum_{j=1}^{n-1} x_{i,j}x_{i,j+1} + \sum_{i=1}^{m-1}\sum_{j=1}^{n} x_{i,j}x_{i+1,j} \tag{1.1}$$

$$\mathrm{BB}_{queen} = \mathrm{BB}_{rook} + \sum_{i=2}^{m}\sum_{j=1}^{n-1} x_{i,j}x_{i-1,j+1} + \sum_{i=1}^{m-1}\sum_{j=1}^{n} x_{i,j}x_{i+1,j+1} \tag{1.2}$$

Os respectivos autores também definiram o número de ligações entre células "pretas" e "brancas", BW , para ambos os casos.

$$\mathrm{BW}_{rook} = \sum_{i=1}^{m}\sum_{j=1}^{n-1} (x_{i,j+1} - x_{i,j})^2 + \sum_{i=1}^{m-1}\sum_{j=1}^{n} (x_{i+1,j} - x_{i,j})^2 \tag{1.3}$$

$$\mathrm{BW}_{queen} = \mathrm{BW}_{rook} + \sum_{i=2}^{m}\sum_{j=1}^{n-1} (x_{i-1,j+1} - x_{i,j})^2 + \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} (x_{i+1,j+1} - x_{i,j})^2 \tag{1.4}$$

Moran (1947) calculou o primeiro e o segundo momentos das distribuições de probabilidade de BB e BW e demonstrou que ambas tendem à normalidade para n e m grandes. Ele estende todos esses resultados para a terceira dimensão supondo um cubo de lados $l \times m \times n$.

Moran (1948) é a continuação do estudo publicado por ele em 1947. Inicia seu trabalho afirmando que é fundamental considerar o arranjo espacial dos dados dependendo da análise. Argumenta que existem dois questionamentos usuais: (i) é possível afirmar estatisticamente que a ocorrência de um determinado fenômeno depende da sua localidade? (ii) é possível afirmar que a ocorrência de um fenômeno, em uma

região, influência para mais (ou para menos) sua ocorrência em regiões vizinhas?

Moran (1948) discute testes de hipótese para averiguar se a distribuição espacial desses fenômenos é aleatória e, igualmente aos trabalhos de Wishart e Hirschfeld (1936) e Moran (1947), sua metodologia se restringe a situações que apresentam variáveis binárias e sistemas regulares.

1.2 Índice I de Moran

Moran (1950) objetivou testar a existência de correlação espacial entre variáveis vizinhas. Diferentemente dos estudos anteriores, agora as variáveis podem ser numéricas ao invés de serem apenas binárias. Ele enuncia o problema supondo um espaço bidimensional regular, dividido por mn regiões, com cada região contendo uma informação x_{ij} $(i = 1, ..., m; j = 1, ..., n)$. Define o coeficiente de correlação espacial como

$$\begin{aligned} \mathrm{r}_{\texttt{Original}} &= \left(\frac{mn}{2mn - m - n}\right) \frac{\sum_{i=1}^{m} \sum_{j=1}^{n-1} z_{ij} z_{i,j+1} + \sum_{i=1}^{m-1} \sum_{j=1}^{n} z_{ij} z_{i+1,j}}{\sum_{i,j} z_{ij}^2} \\ &= \left(\frac{mn}{2mn - m - n}\right) \mathrm{I}_{\texttt{Original}} \end{aligned} \tag{1.5}$$

no qual $mn\bar{x} = \sum_{i,j} x_{ij}$ e $z_{ij} = x_{ij} - \bar{x}$.

O fator inicial de (1.5) foi convencionalmente introduzido porque há mn termos no denominador e $2mn - m - n$ termos no numerador de $\mathrm{I}_{\texttt{Original}}$. Assim, Moran (1950) indica que, para grandes amostras, $\mathrm{r}_{\texttt{Original}}$ pode ser considerado um estimador apropriado para mensurar a correlação entre vizinhos próximos. Caso o objetivo seja testar se as variáveis são espacialmente correlacionadas, é suficiente considerar apenas $\mathrm{I}_{\texttt{Original}}$ como estatística do teste em questão. Moran (1950) finaliza seu trabalho demonstrando que, quando m e n aumentam, a distribuição de probabilidade $\mathrm{I}_{\texttt{Original}}$ tende à normalidade.

Em um sistema irregular, como a malha das mesorregiões do estado de São Paulo, Figura 1.1, as regiões não são alinhadas geometricamente como ocorre nas Figuras 5 e 6. Consequente a isso, há uma dificuldade em se definir uma fórmula fechada para se calcular o número de ligações e, assim, o próprio índice de Moran.

Um artifício (Dacey, 1968; Cliff, 1969) que facilita essa contagem é a criação de

Figura 1.1: Mesorregiões do Estado de São Paulo

matriz de proximidade binária $\mathbf{\Delta} = \|\delta_{ij}\|$. Nela $\delta_{ii} = 0 \; \forall \; i = 1, ..., n$; $\delta_{ij} = 1$ se as regiões i e j forem vizinhas; e, $\delta_{ij} = 0$ caso contrário. Definindo-se L_i como o número de regiões que fazem fronteira com a região i, tem-se que a equação (1.6) representa o número total de ligações do sistema.

$$A = \frac{1}{2} \sum_{i=1}^{n} L_i \tag{1.6}$$

Assim, Dacey (1968)[1] generalizou a problemática definida por Moran (1950) para uma sistema irregular com n regiões onde x_i representa o valor da variável na região i e $z_i = x_i - \bar{x}$. Adicionando o conceito de matriz de proximidade $\mathbf{\Delta}$, Dacey (1968) reescreveu o índice de Moran como:

$$\begin{aligned} \mathrm{r}_{\mathtt{Moran}} &= \left(\frac{n}{A}\right) \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \delta_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \\ &= \left(\frac{n}{A}\right) \mathrm{I}_{\mathtt{Moran}} \end{aligned} \tag{1.7}$$

É simples notar que tanto (1.5) quanto (1.7) apresentam a forma clássica do coeficiente de autocorrelação: o numerador mensura a covariância entre as variáveis x_i, ao passo que o denominador mensura a variância.

[1] Vários artigos (Cliff, 1969; Cliff e Ord, 1969, 1970, 1973) fazem referência a esse trabalho citando Dacey (1965). Dacey (1965) e Dacey (1968) são o mesmo artigo. A diferença é que Dacey (1965) é uma Nota Técnica do Departamento de Geografia da Universidade de Northwestern e, consequentemente, sua publicação foi restrita.

Dacey (1968) ressaltou que a estatística $\mathrm{I}_{\texttt{Moran}}$ tem duas limitações importantes: invariância topológica e o fato de duas regiões só serem consideradas vizinhas se estiverem fisicamente ligadas. Invariância topológica é o fato da matriz de pesos espaciais não agregar nenhuma informação sobre o tamanho ou sobre o formato das regiões. Ela também não agrega nenhuma informação sobre como elas estão interligadas, por exemplo, se a estrada que liga dois municípios é asfaltada ou não.

Essa limitação pode ser ilustrada da seguinte maneira: primeiramente considere um sistema dividido em n regiões não sobrepostas denominado R_0. Esse sistema tem uma matriz de conexão, $\mathbf{\Delta} = \|\delta_{ij}\|$, e um conjunto valores, x_i, fazendo com que a estatística (1.7) atinja o valor I_0. Sem mudar $\|\delta_{ij}\|$ ou x_i, é possível transformar R_0 topologicamente a fim de se produzir um novo sistema, R_1, completamente diferente do sistema anterior, porém, $\mathrm{I}_1 = \mathrm{I}_0$. A Figura 1.2 exemplifica três sistemas que apresentam essa limitação.

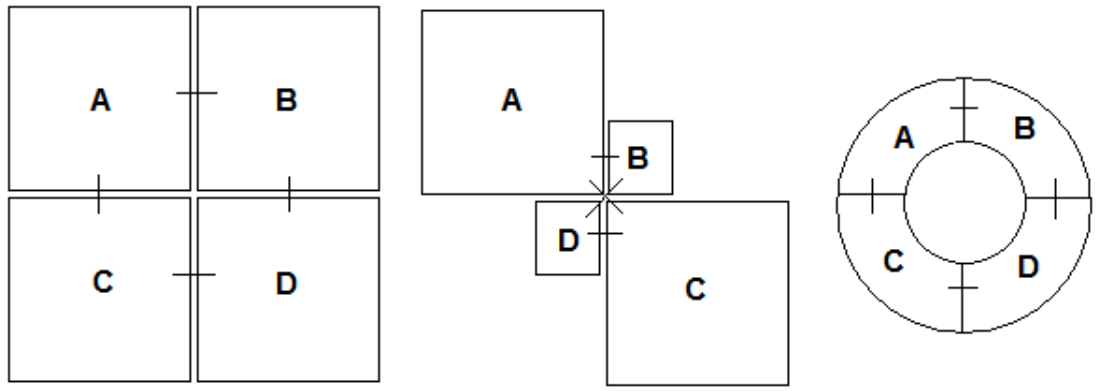

Figura 1.2: Três sistemas com a mesma estrutura de vizinhos

Para superar essa limitação, Dacey sugeriu a estatística $\mathrm{r}_{\texttt{Dacey}}$ para medir autocorrelação espacial.

$$\mathrm{r}_{\texttt{Dacey}} = \frac{n}{A}\left[\frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}\alpha_i\beta_{i(j)}z_iz_j}{\sum_{i=1}^{n}\alpha_i z_i^2}\right] = \frac{n}{A}\mathrm{R}_{\texttt{Dacey}} \tag{1.8}$$

$$\alpha_i = \frac{a_i}{\sum_{i=1}^{n} a_i} \qquad\qquad \beta_{i(j)} = \frac{b_{ij}}{\sum_{j=1}^{n} b_{ij}}$$

nela a_i é a área da região i e b_{ij} é número de ligações entre as regiões i e j. Logica-

mente, em decorrência de $\delta_{ii} = 0$, tem-se também $b_{ii} = 0 \; \forall \; i = 1, ..., n$.

A segunda limitação é representada pelo fato de que a estatística $\mathrm{I}_{\texttt{Moran}}$ mensura apenas a autocorrelação espacial entre os primeiros vizinhos mais próximos. Ela não determina como a função de autocorrelação decai ao longo do espaço. Para superar tal limitação, Dacey (1968) propõe a utilização de mais vizinhos, por exemplo, os segundos vizinhos mais próximos. Ela define que as regiões i e k são segundos vizinhos mais próximos se suas fronteiras não tiverem interseção em comum, contudo, existe um região j que é interligada com i e k.

Os momentos de (1.5) foram obtidos por Moran (1950) apenas para metodologia *rook*, em um sistema regular e sob a suposição de que a variável aleatória x_i é normalmente distribuída. Dacey (1968) generalizou os resultados obtidos por Moran para sistemas irregulares ao utilizar o artifício da matriz de proximidade binária a partir da equação (1.7). Contudo, ele não calculou os momentos de $\mathrm{r}_{\texttt{Dacey}}$ e $\mathrm{R}_{\texttt{Dacey}}$ pela dificuldade de fazer qualquer suposição sobre as distribuição de a_i e b_{ij}.

Cliff e Ord (1969), objetivando superar a segunda limitação ressaltada por Dacey (1968), propõem que, ao invés de uma matriz de proximidade binária, fosse utilizada um matriz de proximidade generalizada. Isso significa que se pode definir $\omega_{ij} = 1$ mesmo que i e j não sejam fisicamente interligados porém, por suposições iniciais, julga-se que i e j exercem dependências entre si. Tal generalização viabiliza a inclusão da segunda, terceira ou quarta classe de vizinhos mais próximos e permite ainda analisar dados sem fazer qualquer tipo de afirmação sobre a distribuição dos dados, bem como a análise de dados nominais e ordinais.

Cliff e Ord (1969) enunciaram a problemática supondo um espaço de duas ou mais dimensões dividido em n regiões regulares ou irregulares não sobrepostas

$$\mathrm{r}_{\texttt{Cliff}} = \frac{n \sum_{(2)} \omega_{ij} z_i z_j}{W \sum_{i=1}^{n} z_i^2} \tag{1.9}$$

com $\sum_{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n}$ para $i \neq j$ e $W = \sum_{(2)} \omega_{ij}$.

Suponha que seja definido que a interação entre duas regiões dependa da distância, d_{ij}, entre os seus centros geográficos e a da porcentagem de união das suas fronteiras, $q_i(j)$. Assim, pode-se também definir $\omega_{ij} = g\left[d_{ij}, q_i(j)\right]$ em que $g\left(\cdot\right)$ é uma função qualquer. Raramente nessa situação $\omega_{ij} = \omega_{ji}$, contudo isso não representa problema.

É muito comum a utilização da matriz de proximidade na sua forma padronizada. Existe uma concesso na literatura de que a padronização deve ser feita em relação à linha (Cliff e Ord, 1981; Anselin, 1988), ou seja,

$$w_{ij}* = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{w_{i.}} \Rightarrow \sum_j w_{ij}* = 1 \qquad\qquad i = 1, \cdots, n \tag{1.10}$$

com $j \in J(i)$ o conjunto das regiões vizinhas à região i. Além do que, Ord (1975) discute que, para a autocorrelação espacial ter uma interpretação natural, ou seja, $\mathrm{r}_{\texttt{Cliff}} < 1$, é necessário que $\sum \omega_{ij} = 1$ seja essa soma sobre i ou sobre j.

Além de discutir sobre todas as vantagens de (1.9), Cliff e Ord (1969) apresentaram uma limitação importante dessa estatística. Ela só pertencerá ao intervalo de $[-1, +1]$ se, e somente se,

$$W^2 Var\left(z_i\right) \geq n^2 Var\left(\sum_{j \in J} w_{ij} z_j\right) \tag{1.11}$$

Caso contrário, o máximo absoluto de (1.9) é definido como

$$\max\left|\mathrm{r}_{\texttt{Cliff}}\right| = \frac{n}{W}\left[\frac{Var\left(\sum_j w_{ij} z_j\right)}{Var\left(z_i\right)}\right]^{\frac{1}{2}} \tag{1.12}$$

A demonstração desse resultado se encontra no Apêndice C. Nessas situações, Cliff e Ord (1981) apresentam que $\widetilde{\mathrm{r}}$ promove uma estimativa de melhor interpretação, já que está contido no intervalo $[-1, +1]$.

$$\widetilde{\mathrm{r}} = \frac{\mathrm{r}_{\texttt{Cliff}}}{\max\left|\mathrm{r}_{\texttt{Cliff}}\right|} \tag{1.13}$$

Cliff e Ord (1970) definem o problema de autocorrelação espacial e exemplifica situações nas quais ela pode ser empregada. Exemplos dessas situações são: determinar a existência de correlação espacial entre os dados originais ou entre os resíduos de uma regressão espacial; determinar se a relação entre as regiões muda ao longo do tempo; ou ainda usar as medidas de associação para definir *clusters*. Cliff e Ord (1970) apresentam a expressão mais geral de uma estatística que mensura a autocorrelação

espacial:

$$u \propto \frac{\sum_{(2)} \omega_{ij} f_{ij} g_{ij}}{\left(\sum_{(2)} a_{ij} f_{ij}^2 \sum_{(2)} b_{ij} g_{ij}^2 \right)^{\frac{1}{2}}} \tag{1.14}$$

com a_{ij} e b_{ij} são funções de ω_{ij}; f_{ij} e g_{ij} funções de z_i e z_j; e, $z_i = x_i - \bar{x}$.

Após calcular os respectivos a_{ij}, b_{ij}, f_{ij} e g_{ij}, demonstrando que a equação (1.14) se transforma em $\mathrm{r}_{\texttt{Moran}}$, $\mathrm{r}_{\texttt{Dacey}}$ e $\mathrm{r}_{\texttt{Cliff}}$, Cliff e Ord (1970) finalizam seu trabalho aplicando os coeficientes de autocorrelação nas situações enunciadas inicialmente.

Cliff e Ord (1971) estudam, via simulações de Monte Carlo, as diversas formas da distribuição de probabilidade de $\mathrm{r}_{\texttt{Cliff}}$. A motivação provém do fato de que a distribuição de $\mathrm{r}_{\texttt{Cliff}}$ só apresenta uma aproximação razoável para a normalidade quando $n > 50$ (Cliff e Ord, 1969). Argumentam que existem quatro fatores que afetam a forma da distribuição de probabilidade. São elas:

- $\{\mathrm{A}/n\}$: o número médio de ligações em cada região;
- $\{\mathbf{W}\}$: a matriz de proximidade;
- A distribuição da variável $\{x_i\}$;
- O tamanho do sistema $\{n\}$.

Nas simulações, Cliff e Ord (1971) mantém três dos quatro fatores supracitados fixos e varia o quarto de modo a determinar o efeito que ele exerce sobre a distribuição de probabilidade de $\mathrm{r}_{\texttt{Cliff}}$. Em cada uma das quatro situações, a simulação consiste em: (i) permutar aleatoriamente os valores de x_i nas regiões do sistema, (ii) calcular o valor de $\mathrm{r}_{\texttt{Cliff}}$ e (iii) repetir os passos $(i) - (ii)$ t vezes com o objetivo de construir uma distribuição de frequência de $\mathrm{r}_{\texttt{Cliff}}$.

Ao final, Cliff e Ord (1971) apresentam uma regra empírica para que a distribuição de $\mathrm{r}_{\texttt{Cliff}}$ se aproxime da normalidade. Sua intenção é que, ao aplicar o teste de hipótese, o pesquisador tenha certeza de que o nível de significância está sendo respeitado.

Cliff e Ord (1973) apresentam o índice de Moran em sua forma matricial. Seja

$\mathbf{x}' = (x_1, \cdots, x_n)$ e $\mathbf{z}' = (z_1, \cdots, z_n)$ definidos de maneira que

$$\mathbf{z} = \mathbf{M}\mathbf{x} \qquad\qquad \mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}'\mathbf{1} \tag{1.15}$$

com $\mathbf{I}_n$ sendo matriz identidade de dimensão n e $\mathbf{1}' = (1, \cdots, 1)$ sendo um vetor de dimensões $(1 \times n)$. Assim define-se o índice I de Moran como

$$\mathrm{I} = \frac{\mathbf{z}'\mathbf{W}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \tag{1.16}$$

no qual $\mathbf{W} = \|w_{ij}\|$ é a matriz de proximidade padronizada, ou seja, $\sum_{(2)} w_{ij} = n$.

Cliff e Ord (1973) generalizam os resultados de Moran (1947) e Dacey (1968) para variáveis binárias. Para tanto, Cliff e Ord (1973) reescreveram BB e BW, equações de (1.1) a (1.4), adicionando o conceito de matriz de proximidade.

$$\mathrm{BB} = \frac{1}{2}\mathbf{z}'\mathbf{W}\mathbf{z} \tag{1.17}$$

$$\mathrm{BW} = \frac{1}{2}\mathbf{z}'\mathbf{W}\left(\mathbf{1} - \mathbf{z}\right) \tag{1.18}$$

Como um dos objetivos da autocorrelação é mensurar se a ocorrência de um determinado fenômeno, em uma região, influencia para mais (ou para menos) sua ocorrência em regiões vizinhas, Cliff e Ord (1973) defini o índice de Moran, para variáveis binárias, em função das estatísticas BB e BW

$$\mathrm{I}_{\texttt{Binario}} = -\frac{2n\mathrm{BW}}{(n - \mathrm{r})\,\mathbf{1}'\mathbf{W}\mathbf{1}} + \frac{2n\,(n - 2\mathrm{r})\,\mathrm{BB}}{\mathrm{r}\,(n - \mathrm{r})\,\mathbf{1}'\mathbf{W}\mathbf{1}} + \frac{\mathrm{r}\mathbf{1}'\mathbf{W}\mathbf{1}}{n - \mathrm{r}} \tag{1.19}$$

com $r = \mathbf{z}'\mathbf{1}$.

Cliff e Ord (1975) realizaram simulações para estudar a distribuição amostral da estatística $(\bar{x}_1 - \bar{x}_2) \,/\, (\widehat{\sigma}_{\bar{x}_1 - \bar{x}_2})$ quando X_1 e X_2 forem autocorrelacionados pelas quantidades ρ_1 e ρ_2, respectivamente, através dos primeiros vizinhos mais próximos em quadrados regulares de vários tamanhos. O objetivo principal dos autores é testar se as médias dessas duas populações diferenciam entre si. Cliff e Ord (1975) consideraram apenas populações normalmente distribuídas com variâncias idênticas.

A diferença da abordagem supracitada para o convencional teste $t-$Student é a presença de autocorrelação espacial em X_1 e em X_2. Uma maneira de solucioná-la

é tentar remover a autocorrelação de ambas populações, com posterior aplicação do método convencional de análise. A dificuldade é que isso não é eficiente. Assim, para essas situações Cliff e Ord (1975) propõem uma alteração no teste t.

A relevância do trabalho de Cliff e Ord (1975) para o presente estudo ocorre porque a estatística do teste proposto depende das estimativas de ρ_1 e ρ_2. Desse modo, mesmo sendo um estimador viesado e não consistente, o autor usa a equação (1.13) para estimá-los.

De forma a aumentar os esclarecimentos do viés do estimador utilizado, Cliff e Ord (1975) realizaram outro experimento. Primeiramente eles criaram uma variável X obedecendo a equação (1.20), com ρ conhecido e calcularam $\widetilde{\mathtt{r}}_{\mathtt{Cliff}}$ através da equação (1.13). Cliff e Ord (1975) repetem esse experimento 600 vezes em 4 tamanhos de quadrados. A média dos resultados de cada uma das simulações pode ser vista na Tabela 1.1.

$$X_i = \rho \sum_j w_{ij} X_j + \epsilon_i \tag{1.20}$$

Fazendo isso, sutilmente Cliff e Ord (1975) apresentarm outra utilidade para o trabalho de (Cliff e Ord, 1969): o uso da equação (1.13) para corrigir o índice de Moran. Entretanto, como esse não era o objetivo do seu trabalho, tal correção acaba passando despercebida, não estando presente em diversos livros de estatística espacial (Anselin, 1988; Anselin e Florax, 2004; Gaetan e Guyon, 2010; Bivand e Pebesma, 2013).

Para finalizar, Cliff e Ord (1975) mostraram que o teste proposto pode ser aplicado caso o tamanho amostral de cada uma das duas populações for de pelo menos 25 observações. Essa última simulação teve apenas o intuito de fundamentar sua metodologia; nenhuma discussão sobre a aplicação do índice de Moran em pequenas amostras foi feita.

Kooijman (1976), por sua vez, usa a ideia de correlograma, definida por Kendall (1976) no estudo de séries temporais, para definir um coeficiente de autocorrelação espacial "maximizado". A partir de uma especificação inicial de $\mathbf{W}$, sua proposta se baseia em maximizar I em função dos coeficientes da matriz de proximidade.

Kooijman (1976) categoriza as distâncias d_{ij} do sistema em k classes, denominadas

Tabela 1.1: Comparação entre $\overline{\overline{\mathrm{r}}}$ e ρ em vários tamanhos de quadrados

Autocorrelação, ρ	Média[a] das estimativas de autocorrelação, $\overline{\overline{\mathrm{r}}}$ [b]			
	3×3	5×5	7×7	10×10
-0.9	-0.72	-0.81	-0.83	-0.86
-0.5	-0.52	-0.49	-0.47	-0.48
0.0	-0.26	-0.06	-0.03	0.00
0.5	0.00	0.34	0.42	0.47
0.9	0.05	0.61	0.76	0.83

[a] Cada média é baseada em 600 simulações
[b] Nos cálculos foi utilizado a estatística $\mathrm{r}_{\texttt{Cliff}}$

c_k, usando uma função $S(\cdot)$. Define $a_k = \sum_{d_{ij} \in c_k} z_i z_j$ onde o somatório é tomado ao longo dos n_k pares distintos. Finalmente, maximiza $\mathbf{z'Wz}$ em relação a $w_{ij} = u_k$, sujeito a $\sum_i \sum_j w_{ij} = 1$ e $\sum_i \sum_j w_{ij}^2 = 1$ e obtém

$$\mathrm{I}_{\texttt{Kooijman}} = \frac{-1}{n-1} + \left(\frac{n^2 - n - 1}{n^2 - n}\right)^{1/2} \left(\frac{-n}{n-1} + \frac{n^2}{\left(\sum_i z_i^2\right)^2} \sum_k \frac{a_k^2}{n_k}\right)^{1/2} \tag{1.21}$$

Para se obter os coeficientes de $\mathbf{W}$ que geram $\mathrm{I}_{\texttt{Kooijman}}$ usa-se a seguinte equação

$$u_k = \frac{a_k}{n_k} \left(\frac{n^2 - n - 1}{n + n\,(n-1)\,\mathrm{I}_{\texttt{Kooijman}}}\right) + \frac{n + \mathrm{I}_{\texttt{Kooijman}}}{n + n\,(n-1)\,\mathrm{I}_{\texttt{Kooijman}}} \tag{1.22}$$

de Jong et al. (1984) realizaram uma discussão sobre os valores extremos do índice de Moran. Eles caracterizam qualquer sistema como um grafo ou uma *network*, ou seja, representam as regiões por pontos e as relações de vizinhança por linhas. de Jong et al. (1984) lançam mão do índice em sua forma matricial, equação (1.16); contudo, utilizam uma matriz de proximidade binária $\mathbf{\Delta}$ não padronizada

$$\mathrm{I} = \frac{n}{\sum_i \sum_j \delta_{ij}} \frac{\mathbf{z'\Delta z}}{\mathbf{z'z}} \tag{1.23}$$

com $z_i = x_i - \overline{x}$. Define também o vetor $\mathbf{c'} = (1, \cdots, 1)$ de dimensões $(1 \times n)$ de modo que $\mathbf{z'c} = 0$.

Para achar os valores extremos da equação (1.23), de Jong et al. (1984) iniciam sua discussão a partir dos resultados do Teorema 1.2.1 (Bellman, 1970).

Teorema 1.2.1 (Courant-Fischer)**.** *Seja* $\mathbf{T}$ *uma matriz* $n \times n$ *simétrica com os au-*

tovalores $\lambda_1 > \lambda_2 > \cdots > \lambda_n$. *Então*

$$\lambda_1 = \lambda_{\max} = \max_{\|\mathbf{x}\|=1} \mathbf{x}'\mathbf{T}\mathbf{x} = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{T}\mathbf{x}}{\mathbf{x}'\mathbf{x}}$$
$$\lambda_n = \lambda_{\min} = \min_{\|\mathbf{x}\|=1} \mathbf{x}'\mathbf{T}\mathbf{x} = \min_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{T}\mathbf{x}}{\mathbf{x}'\mathbf{x}}$$

Se $\boldsymbol{\Delta}$ for simétrica, valem os resultados do Teorema 1.2.1. Contudo, se $\boldsymbol{\Delta}$ for assimétrica, deve-se definir a matriz $\mathbf{S} = (\boldsymbol{\Delta} + \boldsymbol{\Delta}')/2$. Os valores extremos passarão a ser o maior e o menor autovalores da matriz $\mathbf{S}$.

Em geral, os autovetores de $\boldsymbol{\Delta}$ não satisfazem a condição de serem ortogonais ao vetor $\mathbf{c}$. de Jong et al. (1984) com o objetivo de incorporar a condição $\mathbf{z}'\mathbf{c} = 0$, consideraram a matriz $\mathbf{C} = \mathbf{c}\mathbf{c}'/N$ e definiram a matriz $\mathbf{P} = \mathbf{I} - \mathbf{C}$, onde $\mathbf{I}$ é matriz identidade.

Assim eles passaram a considerar a matriz simétrica $\mathbf{P}\boldsymbol{\Delta}\mathbf{P}$ e, consequentemente, seus autovalores $m_1 > m_2 > \cdots > m_n$. É fácil notar, portanto, que tais valores extremos são, ao mesmo tempo, as soluções para o problema de maximizar $\mathbf{z}'\boldsymbol{\Delta}\mathbf{z}/\mathbf{z}'\mathbf{z}$ com $\mathbf{z}'\mathbf{c} = 0$.

$$\mathbf{z}'\mathbf{P}\boldsymbol{\Delta}\mathbf{P}\mathbf{z} = (\mathbf{P}\mathbf{z})'\boldsymbol{\Delta}(\mathbf{P}\mathbf{z}) = \mathbf{z}'\boldsymbol{\Delta}\mathbf{z} \tag{1.24}$$

Finalmente, apresentam-se os valores extremos da equação (1.23) utilizando os resultados do Teorema 1.2.1.

$$I_{min} = \frac{n}{\sum_i \sum_j \delta_{ij}} m_{\min} \qquad\qquad I_{max} = \frac{n}{\sum_i \sum_j \delta_{ij}} m_{\max} \tag{1.25}$$

de Jong et al. (1984) obtiveram seus resultados sem trazerem nenhum exemplo, aplicação prática ou discussões acerca da matriz de proximidade. Sua metodologia se baseia apenas na matriz $\mathbf{W}$; contudo não fica claro se ela pode ser aplicada em matrizes padronizadas ou em matrizes de distância.

Getis e Ord (1992) apresentaram a estatística G e a compararam com o índice I de Moran. A primeira diferença entre ambos índices é que, no primeiro, a variável X não é centralizada, ao passo que no segundo, ela é. A segunda diferença diz respeito à definição $\mathbf{W}$: cada elemento $\omega_{ij}(d)$ será diferente de zero se as regiões i e j estiverem

distantes entre si até no máximo uma distância d.

$$\mathrm{G}\,(d) = \frac{\sum_{(2)} \omega_{ij}(d) x_i x_j}{\sum_{(2)} x_i x_j} \tag{1.26}$$

Denominando $\mathrm{I}\,(d)$ como o índice de Moran definindo a partir de uma $\mathbf{W}$ onde cada elemento será 1 se as regiões i e j estiverem distantes entre si até no máximo uma distância d, as estatísticas em questão podem ser escritas como:

$$\begin{aligned} \mathrm{G}\,(d) &= \mathrm{K}_1 \sum \sum \omega_{ij} x_i x_j \\ \mathrm{I}\,(d) &= \mathrm{K}_2 \sum \sum \omega_{ij} \left(x_i - \bar{x}\right) \left(x_j - \bar{x}\right) \\ &= \frac{\mathrm{K}_1}{\mathrm{K}_2} \mathrm{G}\,(d) - \mathrm{K}_2 \bar{x} \sum \left(\omega_{i.} + \omega_{.j}\right) x_i + \mathrm{K}_2 \bar{x}^2 W \end{aligned} \tag{1.27}$$

Na prática, $\mathrm{G}\,(d)$ e $\mathrm{I}\,(d)$ serão diferentes quando $\sum \omega_{i.} x_i$ e $\sum \omega_{.i} x_i$ forem diferentes de $W\bar{x}$.

Getis e Ord (1992) discutem que $\mathrm{I}\,(d)$ não deve ser usada quando d for muito grande ou muito pequeno. Nessas situações, o teste de hipótese, sobre $\mathrm{G}\,(d)$, baseado em uma aproximação da distribuição normal, torna-se inapropriado. $\mathrm{G}\,(d)$ também é ineficiente ao se estudar resíduos de uma regressão, por exemplo. Isso ocorre pois a variável X deve ser positiva.

Oden (1995) propôs uma modificação no índice de Moran para estudos nos quais se deseja estudar se a taxa de ocorrência de um determinado evento, por exemplo o número de casos de uma epidemia por mil habitantes, é correlacionado com sua posição. Sua motivação decorre do fato de que indivíduos presentes em regiões com populações maiores estão mais expostos ao risco do que indivíduos presentes em regiões com populações menores.

$$\mathrm{I}^*_{\mathtt{Pop}} = \frac{n^2 \sum_{ij} M^*_{ij} \left(e_i - d_i\right) \left(e_j - d_j\right) - n\left(1 - 2\bar{b}\right) \sum_i M^*_{ii} e_i - n\bar{b} \sum_i M^*_{ii} d_i}{\bar{b}\left(1 - \bar{b}\right) \left(x^2 \sum_{ij} M^*_{ij} d_i d_j - x \sum_i M^*_{ii} d_i\right)} \tag{1.28}$$

nesse caso, n_i representa o número de casos do evento em questão, $n = \sum n_i$, x_i corresponde à população que está exposta ao evento, sendo $x = \sum x_i$, $\bar{b} = x/n$, $e_i = n_i/n$, $d_i = x_i/x$ e $M^*_{ij} = M_{ij}/\sqrt{d_i d_j}$. O termo M_{ij} pode ser interpretado como a generalização da matriz $\mathbf{W}$ aceitando-se $M_{ii} \neq 0$.

Por meio de simulações, Oden (1995) apresenta que o teste de hipótese derivado de $I^*_{\texttt{Pop}}$ é mais poderoso que o teste de hipótese sobre o índice de Moran quando os tamanhos das populações expostas ao evento forem muito diferentes.

Assunção e Reis (1999) objetivaram comparar o índice I de Moran, o índice $I^*_{\texttt{Pop}}$ e o índice proposto por eles, $I_{\texttt{EBI}}$, em situações em que se deseja estudar taxas baseadas em regiões com diferentes tamanhos de populações. Eles iniciam a discussão afirmando que, na problemática em questão, existem três possibilidades de configuração espacial: (A) taxas são constantes no espaço, (B) taxas são heterogêneas, porém não há correlação espacial e (C) taxas são heterogêneas e correlacionadas espacialmente. Assunção e Reis (1999) argumentam que a principal desvantagem do teste derivado de $I^*_{\texttt{Pop}}$ é que as hipóteses desse teste são H_0 : A contra H_1 : B $\cup$ C. Já as hipóteses do teste de Moran de são H_0 : A $\cup$ B contra H_1 : C.

Assunção e Reis (1999), usando uma abordagem Bayesiana, propõem o índice $I_{\texttt{EBI}}$. O trabalho mostra que o teste de hipótese derivado desse índice é mais poderoso que o teste $I_{\texttt{Cliff}}$ de Moran e não apresenta as mesmas deficiências do teste $I^*_{\texttt{Pop}}$.

$$I_{\texttt{EBI}} = \frac{m}{\sum \omega_{ij}} \frac{\sum \omega_{ij} z_i z_j}{\sum (z_i - \bar{z})^2} \tag{1.29}$$

com $z_i = (p_i - b)/\sqrt{v_i}$, $p_i = n_i/x_i$, $b = n/x$, $v_i = a + b/x_i$, $a = s^2 - b/(x/m)$ e $s^2 = \sum x_i (p_i - b)^2/x$.

Li e Calder (2007) discutem que, devido à simplicidade do índice de Moran, ele é frequentemente utilizado na análise exploratória de dados georreferenciados. Porém, o índice referido apresenta uma série de limitações que muitos desconhecem. Afim de ilustrar algumas delas, Li e Calder (2007) comparam o resultado do índice em dados simulados a partir do modelo da equação (1.20) e concluem que $I_{\texttt{Cliff}}$ só é um bom estimador de ρ quando esse está próximo de zero.

Li e Calder (2007) propõem o índice $I_{\texttt{APLE}}$ como uma alternativa para o índice I de Moran. Eles nomearam o índice de *APLE* por se tratar de uma aproximação do estimador de máxima verossimilhança de ρ que, como apresentado por Ord (1975),

não apresenta uma forma fechada.

$$\mathrm{I}_{\texttt{APLE}} = \frac{\mathbf{Z}' \left[\left(\mathbf{W} + \mathbf{W}' \right) /2 \right] \mathbf{Z}}{\mathbf{Z}' \left(\mathbf{W}'\mathbf{W} + \lambda'\lambda \mathbf{I}/n \right) \mathbf{Z}} \tag{1.30}$$

sendo $\mathbf{W}$ padronizada pela linha, e λ um vetor com todos os autovalores dessa matriz.

Li e Calder (2007) finalizam o trabalho mostrando que $\mathrm{I}_{\texttt{APLE}}$ é um bom estimador para ρ, atingindo valores próximos aos valores do estimador de máxima verossimilhança.

Jackson e Huang (2010) propõem uma modificação no índice de Moran incluindo informação sobre a matriz $\mathbf{W}$ no denominador do seu índice. Via simulações, eles concluíram que o poder do teste de hipótese do seu índice é maior que os poderes do teste dos índices $\mathrm{I}_{\texttt{Cliff}}$ e $\mathrm{I}^{*}_{\texttt{Pop}}$. A desvantagem de $\mathrm{I}_{\texttt{Jackson}}$ é que ele pode atingir valores maiores que 1.

$$\mathrm{I}_{\texttt{Jackson}} = \frac{\sum_{(2)} \omega_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i \leq i < j < N} \omega_{ij}(y_i - y_j)^2} \tag{1.31}$$

Objetivando esclarecer o intervalo do $\mathrm{I}_{\texttt{Jackson}}$, realizaram-se as mesmas simulações apresentadas na Introdução. Os resultados estão presentes na Tabela 1.2. O que se pode concluir é que, em praticamente todos os casos, o índice não pertence ao intervalo -1 e $+1$. Consequentemente, sua interpretação não é trivial.

Tabela 1.2: Valores do índice de Jackson

Sistema	Observações	Vizinhança	I de Jackson
3×3	9	*rook*	1.4815
		queen	0.3619
7×7	49	*rook*	14.6939
		queen	6.3786
10×10	100	*rook*	31.68
		queen	14.082
14×14	196	*rook*	63.6735
		queen	28.5216

Chen (2013) inicia seu trabalho afirmando que a fórmula para o índice de Moran é complicada e propõe uma reconstrução para ela. Primeiramente, padroniza a variável X ao invés de apenas centralizá-la, ou seja, $z = (x - \mu)/\sigma$. Depois, ele define três

propriedades para $\mathbf{W}$: (i) ela deve ser simétrica; (ii) sua diagonal deve ser zero e (iii) a normalização deve obedecer a $\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij} = 1$. Contudo, para que o resultado seja igual ao resultado da equação (1.16), deve-se padronizar $\mathbf{W}$ pela linha e depois dividi-la por n. Assim, reescreve o índice de Moran como

$$\mathrm{I}_{\texttt{Chen}} = \mathbf{z}'\mathbf{W}\mathbf{z} \tag{1.32}$$

Além do formato da equação (1.32), Chen (2013) apresenta outras três maneiras de se calcular o índice de Moran, todas elas, em função de $\mathbf{z}\mathbf{z}'\mathbf{W}$.

1.3 Tópicos Conclusivos

O coeficiente de autocorrelação espacial fornece um ponto de partida para qualquer tipo de análise envolvendo dados georreferenciados. Esse indicador é útil em diversas situações como: determinar a existência de correlação espacial em um sistema bidimensional; determinar se essa correlação muda ao longo do tempo; ou ainda, ser usado para agrupar os dados em *clusters*.

Devido às características gerais e intrínsecas do índice I de Moran, entender a dinâmica dos seus valores extremos se torna uma tarefa fundamental. Por exemplo, é extremamente simples interpretar o coeficiente de Pearson, vide equação (1), ao se estudar a correlação linear entre duas variáveis. Entretanto, o mesmo não ocorre quando utiliza-se o índice de Moran para mensurar autocorrelação espacial, como pode ser visto nas Figuras 5 e 6.

Inexiste na literatura um trabalho que elenca todas as apresentações do índice de Moran, ou ainda um que relaciona o índice de Moran, a quantidade de dados observados e seus valores extremos. Publicações recentes, como os trabalhos de Chen (2013) e Jackson e Huang (2010), revelam que não existe um consenso sobre o tema e que ainda existem questões em aberto.

Capítulo 2

Matriz de Proximidade

Esse capítulo objetiva apresentar uma discussão sobre como se deve definir a matriz de proximidade e como ela interfere diretamente nos valores extremos do índice I de Moran. Para tanto, primeiramente serão apresentados os artigos mais relevantes que tratam sobre esse tema. Depois, será discutido o tipo de padronização de $\mathbf{W}$, finalizando com a apresentação de um exemplo utilizando $\mathbf{W}$ em diferentes formatos. O objetivo é analisar sua influência sobre os valores do índice de Moran.

2.1 Revisão Bibliográfica

Como foi apresentado na Introdução, o índice de Moran em sua forma original não apresentava uma matriz de proximidade. Informação sobre a matriz $\mathbf{W}$ só foi introduzida a partir dos trabalhos de Dacey (1968) e Cliff (1969). A motivação desses trabalhos foi generalizar o $\mathrm{I}_{\texttt{Original}}$ para sistemas irregulares.

Arora e Brown (1977) apresentaram que o conceito da matriz $\mathbf{W}$ já estava presente no estudo de séries temporais. Por isso, para se entender a problemática de autocorrelação espacial, primeiramente é preciso se familiarizar com o conceito de autocorrelação em séries temporais. Assim, supondo a problemática em que os erros de uma regressão linear são autocorrelacionados com o tempo, tem-se o modelo autorregressivo temporal

$$\epsilon = \rho \mathbf{W} \epsilon + \mathbf{u} \tag{2.1}$$

em que ϵ e $\mathbf{u}$ são vetores de erros de dimensão $T \times 1$, $\mathbf{W}$ é a matriz de pesos $T \times T$ e $|\rho| < 1$.

Em séries temporais, normalmente adota-se que erros de uma regressão seguem processos de Markov de primeira ou de segunda ordem (Fuller, 1996). Para processos de primeira ordem, $\mathbf{W}$ é especificada como

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{2.2}$$

Contudo, como observado por Whittle (1954), ao contrário do que ocorre com a matriz apresentada em (2.2), a dependência de dados espaciais é multi-dimensional. Isso torna a análise espacial muito mais complexa.

Arora e Brown (1977) apresentaram quatro métodos para estimar $\mathbf{W}$ a partir da variável x_i. Contudo, sua abordagem requer que se conheça x_i durante vários períodos de tempo e que, em cada um desses períodos, as observações sejam normalmente distribuídas.

Anselin (1988) discute que a definição da matriz $\mathbf{W}$ é parte fundamental de qualquer modelagem espacial. É a matriz de proximidade que expressa a dependência espacial entre as regiões. Logo, a escolha dos pesos de $\mathbf{W}$ deve ser feita com cautela para se evitar correlação espúria (Arora e Brown, 1977; Anselin, 1988; Griffith, 2003). O exemplo da Figura 2.3 mostra que, simplesmente alterando $\mathbf{W}$, consegue-se obter desde uma correlação positiva forte até uma correlação negativa fraca.

Griffith (1996) propôs algumas dicas para se especificar $\mathbf{W}$:

- "É melhor usar uma matriz com pesos razoáveis do que assumir a independência entre as regiões". Griffith (1996) argumenta que os melhores resultados são obtidos quando a distância é levada em consideração. Isso porque a distância ilustra o decaimento das relações entre as regiões. O mesmo não ocorre quando se adota um matriz de proximidade binária;

- "Em regiões sem fronteiras específicas, por exemplo o pasto de uma fazenda,

é melhor particioná-lo usando subregiões regulares quadráticas ou hexagonais". Griffith (1996) argumenta que subregiões com menos de quatro ou com mais de seis vizinhos não contextualiza muito bem a problemática;

- "É razoável ter um número relativamente grande de regiões dentro do sistema". Griffith (1996) discute que, em decorrência da lei dos grandes números, é interessante ter um $n > 60$.

- "Em geral, é preferível usar um número menor de vizinhos." Griffith (1996) argumenta que a superespecificação das vizinhanças, por exemplo vizinhos de terceiro ou quarto grau, reduz o poder do teste e não caracteriza muito bem a questão proposta.

Getis e Aldstadt (2004) listam doze maneiras diferentes de se especificar a matriz de proximidade. Já Getis (2007) apresenta que, dessas doze, utiliza-se basicamente quatro metodologias: (i) matriz binária ($\omega_{ij} = 1$ se as regiões exercerem influência em si e $\omega_{ij} = 0$ caso contrário); (ii) matriz do decaimento da distância ($1/d_{ij}^{\alpha}$); (iii) matriz representando as fronteiras comuns entre as regiões (s/p sendo s é o número de fronteiras comuns e p o perímetro dessas intersecções); (iv) matriz com as estatísticas locais de autocorrelação espacial (G_i ou I_i). Essa última metodologia foi apresentada por Getis e Aldstadt (2004) e sua utilização é recomendada em sistemas que apresentam *clusters*.

Logicamente, a especificação de $\mathbf{W}$ dependerá do estudo em questão. Dado seu conhecimento *a priori*, o pesquisador pode escolher qualquer função $g(\cdot)$ para definir os pesos de $\mathbf{W}$. Por exemplo, Silva e Yamashita (2007) os definiu utilizando o tempo médio gasto para ir do município i ao município j através das rodovias que interligam essas regiões.

Getis (2009) discute que ideologicamente existem três maneiras de se definir $\mathbf{W}$: (i) a teórica; (ii) a topológica e (iii) a empírica. A teórica implica que os pesos de $\mathbf{W}$ são exógenos a qualquer sistema e baseiam-se em uma estrutura pré-concebida, geralmente baseada na distância entre os centroides. A topológica surgiu da necessidade de descrever, da maneira mais realista possível, o formato do sistema. Nessa perspectiva, $\mathbf{W}$ apresenta informações como o número de vizinhos mais próximos, a proporção de perímetro em comum e a área de cada região. A forma empírica é a

ideologia mais versátil. Nela está presente a abordagem de Arora e Brown (1977) e situações nas quais o pesquisador destaca as características de cada vizinhança da maneira que ele julgar mais relevante.

2.2 Tipos de Padronização

Há um consenso na literatura acerca de $\mathbf{W}$ ser padronizada pela linha (Cliff e Ord, 1981; Anselin, 1988). A única exceção ocorre em Tiefelsdorf e Griffith (1999), o qual defende que a padronização deve ser global. Argumenta que a padronização pela linha fornece pesos maiores a regiões com poucos vizinhos. Já a padronização global dá enfase justamente às regiões com mais vizinhos.

Pela definição, suponha que $\mathbf{\Delta} = \|\delta_{ij}\|$ seja a matriz não-padronizada e $\mathbf{W} = \|\omega_{ij}\|$ seja a matriz padronizada. A relação entre ω_{ij} e δ_{ij} padronizando $\mathbf{\Delta}$ pela linha e globalmente é, respectivamente:

$$\omega_{i\cdot} = \frac{\delta_{ij}}{\sum_j \delta_{ij}} = \frac{\delta_{ij}}{\delta_{i\cdot}} \qquad\qquad \omega_{ij} = \frac{\delta_{ij}}{\sum_i \sum_j \delta_{ij}} \tag{2.3}$$

Inexiste na literatura o fato que, se a matriz $\mathbf{\Delta}$ for simétrica, ou seja, $\delta_{ij} = \delta_{ji}$, o índice de Moran em sua forma clássica, equação (8), apresenta os mesmos resultados quando $\mathbf{\Delta}$ for padronizada pela linha ou pela coluna.

Para demonstrar esse fato, basta mostrar que os numeradores em ambas as padronizações são iguais, uma vez que o denominador do índice de Moran, equação (8), não depende de $\|\delta_{ij}\|$. Assim, o primeiro passo é mostrar que $\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij} = n$ para ambos os casos.

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\omega_{ij}\right) = \sum_{i=1}^{n}\omega_{i\cdot} = \sum_{i=1}^{n}1 = n \tag{2.4}$$

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\omega_{ij}\right) = \sum_{j=1}^{n}\omega_{\cdot j} = \sum_{j=1}^{n}1 = n \tag{2.5}$$

Agora, para a que a demonstração esteja concluída, basta mostrar que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} z_i \left(\frac{\delta_{ij}}{\delta_{i\cdot}}\right) z_j = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i \left(\frac{\delta_{ij}}{\delta_{\cdot j}}\right) z_j \tag{2.6}$$

Sem perda de generalidades, pela definição do índice I de Moran, a variável z_i apresenta média zero. Desenvolvendo o numerador da equação (8), levando em conta $\boldsymbol{\Delta}$ despadronizada, obtém-se

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} z_i\delta_{ij}z_j = {} & z_1\delta_{12}z_2 + \cdots + z_1\delta_{1,n-1}z_{n-1} + z_1\delta_{1n}z_n + \\ & + z_2\delta_{21}z_1 + \cdots + z_2\delta_{2,n-1}z_{n-1} + z_2\delta_{2n}z_n + \cdots + \\ & + z_{n-1}\delta_{n-1,1}z_1 + z_{n-1}\delta_{n-1,2}z_2 + \cdots + z_{n-1}\delta_{n-1,n}z_n + \\ & + z_n\delta_{n1}z_1 + z_n\delta_{n2}z_2 + \cdots + z_n\delta_{n,n-1}z_{n-1}\end{aligned} \tag{2.7}$$

É fácil notar que cada par g e h aparece duas vezes, a primeira como $z_g\delta_{gh}z_h$ e a segunda como $z_h\delta_{hg}z_g$. Note que $g \neq h$ e cada um assume valores entre $1, \cdots, n$. Finalmente, basta mostrar que a soma desses dois termos gera o mesmo resultado em ambas padronizações.

$$\begin{aligned}z_g z_h\left(\frac{\delta_{gh}}{\delta_{g\cdot}} + \frac{\delta_{hg}}{\delta_{h\cdot}}\right) &= z_g z_h\left(\frac{\delta_{gh}}{\delta_{g1}+\cdots+\delta_{gn}} + \frac{\delta_{hg}}{\delta_{h1}+\cdots+\delta_{hn}}\right) \\ = z_g z_h\left(\frac{\delta_{hg}}{\delta_{1g}+\cdots+\delta_{ng}} + \frac{\delta_{gh}}{\delta_{1h}+\cdots+\delta_{nh}}\right) &= z_g z_h\left(\frac{\delta_{hg}}{\delta_{\cdot h}} + \frac{\delta_{gh}}{\delta_{\cdot g}}\right)\end{aligned} \tag{2.8}$$

Note que a segunda igualdade só é possível porque $\boldsymbol{\Delta}$ é simétrica.

2.3 Exemplo

Essa seção apresenta como a matriz $\mathbf{W}$ pode influenciar o índice I de Moran. Os Capítulos 03 e 04 já incluem muitas simulações utilizando sistemas regulares com as metodologias de vizinhança *queen* e *rook*. Por esse motivo, será abordado um exemplo envolvendo um sistema irregular.

Geralmente, em sistemas irregulares, como as mesorregiões do Estado de São Paulo, não há diferença entre as metodologias *queen* e *rook*, como pode ser visto na Figura 1.1. Contudo, podem existir alguns exemplos que apresentam diferença, como é o caso da região de Columbus, capital de Ohio, EUA (Figuras 2.1 e 2.2). Entretanto, são diferenças sutis que não influenciam muito o índice de Moran: $\mathrm{I}_{\texttt{Queen}} = 0.501$ e $\mathrm{I}_{\texttt{Rook}} = 0.523$. Para mais detalhes veja Tabela 4.4 do Capítulo 04.

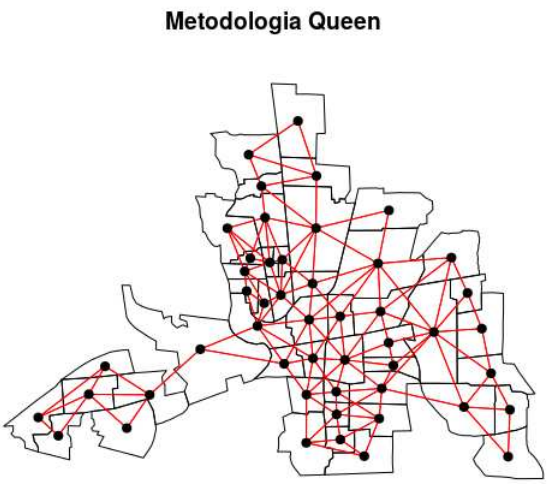

Figura 2.1: Vizinhanças *queen* na região de Columbus, EUA

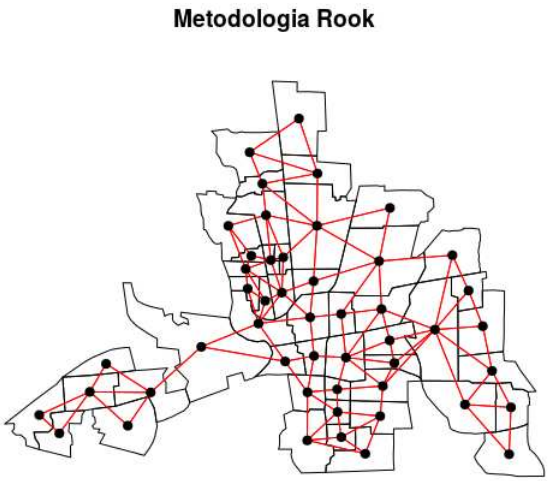

Figura 2.2: Vizinhanças *rook* na região de Columbus, EUA

O exercício dessa seção se baseou em calcular o índice I de Moran para os dados simulados da Figura 2.3, utilizando seis tipos de matrizes $\mathbf{W}$: (i) *rook*, (ii) *queen*, (iii) inverso da distância entre os centróides das regiões (d_{ij}^{-1}), (iv) d_{ij}^{-2}, (v) d_{ij}^{-5} e (vi) variograma gaussiano. Optou-se por essas escolhas a fim de reproduzir a discussão de Getis e Aldstadt (2004). Os resultados encontram-se na Tabela 2.1.

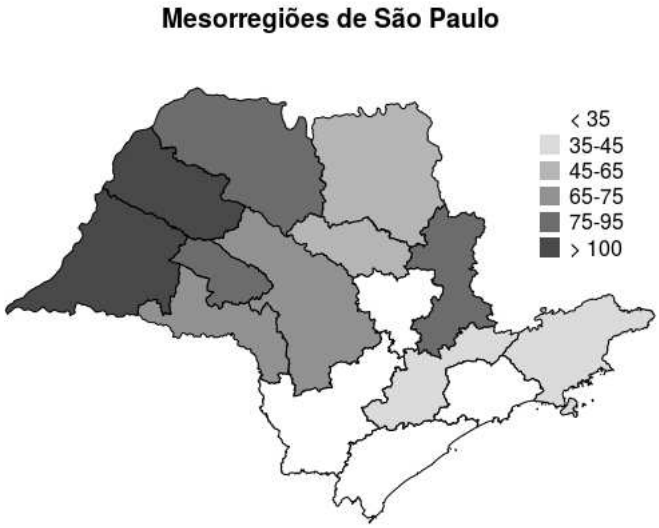

Figura 2.3: Mesorregiões do Estado de São Paulo

Como era de se esperar, o índice de Moran obteve os mesmos valores quando utiliza-se a metodologia *rook* e *queen*, $\mathrm{I}_{\texttt{Rook}} = \mathrm{I}_{\texttt{Queen}} = 0.62$. Quando utiliza-se d_{ij}^{-5}, quanto maior for a distância entre as regiões, mais próximo d_{ij}^{-5} estará de 0. Assim, nesse exemplo, $\mathbf{W}$ se assemelha a estrutura *queen*, $\mathrm{I}_{d_{ij}^{-5}} = 0.63$.

A título de ilustração, realizou-se um exercício diminuindo ainda mais o valor do expoente de d_{ij}^{α}. O interessante é que os resultados convergem. A partir de certo ponto, independente do quanto se diminua α, os valores do índice não mudam consideravelmente: $\mathrm{I}_{d_{ij}^{-10}} = 0.701$, $\mathrm{I}_{d_{ij}^{-15}} = 0.7125$, $\mathrm{I}_{d_{ij}^{-30}} = 0.7126$ e $\mathrm{I}_{d_{ij}^{-100}} = 0.711$.

Na prática, isso ocorre por causa da padronização de $\mathbf{W}$. Para cada região, a diminuição do expoente de d_{ij}^{α} faz com que o menor d_{ij} produza um valor proporcionalmente muito maior que todos os outros d_{ij}. Assim, todas as matrizes $\mathbf{W}$ do parágrafo acima são basicamente as mesmas. Elas dizem que cada região interage apenas com a região mais próxima a ela.

Tabela 2.1: Valores do Índice de Moran em relação ao formato da matriz de proximidade

Rook	*Queen*	d_{ij}^{-1}	d_{ij}^{-2}	d_{ij}^{-5}	d_{ij}^{-15}	$1-\exp\left(-d_{ij}^{2}\right)$
0.6271	0.6271	0.1813	0.3875	0.6305	0.7125	-0.1114

d_{ij}: distância euclidiana entre os centroides das regiões

Todavia, os resultados mais impressionantes ocorrem quando utiliza-se d_{ij}^{-1} e d_{ij}^{-2}. Ambas metodologias são largamente utilizadas por estarem presentes em várias equações de modelos físicos gravitacionais. Contudo, o índice atinge valores muito inferiores aos expostos anteriormente, $\mathrm{I}_{d_{ij}^{-1}} = 0.18$ e $\mathrm{I}_{d_{ij}^{-2}} = 0.38$.

Para mostrar quão influente é a matriz de proximidade, utilizou-se a função $g\left(d_{ij}\right) = 1 - \exp\left(-d_{ij}^{2}\right)$ para se definir os pesos de $\mathbf{W}$. Para $d_{ij} > 0$, a função $g\left(\cdot\right)$ assumirá valores entre 0 e 1. Quanto maior for d_{ij}, mais próxima essa função será de uma unidade. Adotar uma $\mathbf{W}$ dessa maneira, significa dizer que uma região interage mais com regiões mais afastadas. Assim, nessa situação, o índice atinge seu valor mais destoante: $\mathrm{I}_{g(d_{ij})} = -0.11$.

Simplesmente mudando $\mathbf{W}$, consegui-se obter valores bastantes discrepantes. O valor mais alto ocorreu quando utilizou-se uma matriz com o decaimento da distância $\mathrm{I}_{d_{ij}^{-30}} = 0.71$. Já o valor mais baixo ocorreu quando utilizou-se um variograma gaussiano $\mathrm{I}_{g(d_{ij})} = -0.11$. Note que a ínterpretação da autocorrelação espacial fornecida pelo índice de Moran mudou de positiva forte para negativa fraca.

2.4 Tópicos Conclusivos

A especificação de $\mathbf{W}$ dependerá do estudo em questão. O pesquisador, dado seu conhecimento *a priori*, pode escolher qualquer função $g(\cdot)$ de modo a se definir os pesos de $\mathbf{W}$. Todavia, tal escolha deve ser feita com cautela para que se evite correlação espúria.

Se a matriz de proximidade $\mathbf{\Delta}$ for simétrica, o índice de Moran apresenta os mesmos resultados, independentemente se ela for padronizada pela linha ou pela coluna.

Capítulo 3

Índice de Moran Modificado

Esse capítulo está dividido em três seções e apresenta os resultados obtidos na presente pesquisa. A primeira seção propoe uma modificação no índice I de Moran introduzindo no coeficiente de correlação de Pearson o conceito de modelagem espacial autoregressiva de primeira ordem. A segunda trás um discussão sobre os valores extremos dos índices $I_{\texttt{Moran}}$ e $I_{\texttt{Proposto}}$, bem como uma demonstração da dependência do índice de Moran com a quantidade de dados observados. A terceira apresenta um primeiro exercício a fim de se estudar a distribuição de probabilidade do índice proposto.

3.1 Proposta

Como apresentado na Introdução do presente trabalho, Griffith (2003) expõe uma simples explicação de como deduzir o índice de Moran a partir do coeficiente de correlação de Pearson. Basicamente, Griffith (2003) aponta que deve-se substituir, no numerador do coeficiente de Pearson, os y_i pelos vizinhos de x_i. Contudo, não traz nenhuma discussão sobre o que deve ser feito no denominador da equação (1). Apenas define que o denominador passa a ser igual à variância da variável X.

Para aumentar os esclarecimentos, ainda na introdução deste trabalho, foi apresentado que o coeficiente de correlação linear de Pearson será, em módulo, igual a 1 toda vez que as variáveis X e Y obedecerem à expressão $Y = \alpha + \beta X$, com $\beta \neq 0$. Intencionando demonstrar o aspecto supracitado, basta substituir a relação linear na

equação (1).

$$\begin{aligned}\rho &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i-\alpha-\beta\bar{x})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{n}(\alpha+\beta x_i-\alpha-\beta\bar{x})^2\right]^{1/2}} \\ &= \frac{\beta\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}\left[\beta^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1\end{aligned} \tag{3.1}$$

Ord (1975) foi um dos primeiros a estudar métodos de estimação para modelos que tentam descrever a interação entre as variáveis e suas posições no espaço. Na época, sua motivação se baseou na dificuldade de se maximizar a função de máxima verossimilhança devido às dimensões da matriz **W**. LeSage e Pace (2009) apresentam uma solução para essa problemática utilizando a função log-verossimilhança e os autovalores de **W**. Em ambos estudos, um dos modelos estudados foi o modelo espacial autoregressivo de primeira ordem

$$X_i = \alpha + \beta\sum_j \omega_{ij}X_j + \epsilon_i \qquad\qquad i = 1,\cdots,n \tag{3.2}$$

sendo $j \in J(i)$ o conjunto das regiões vizinhas à região i; α e β os parâmetros a serem estimados; ω_{ij} os pesos da matriz de proximidade e ϵ_i são os erros i.i.d do modelo.

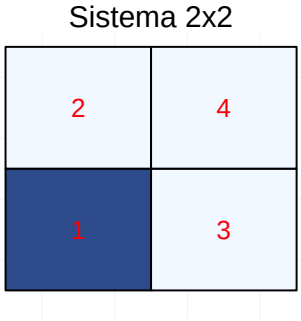

Figura 3.1: Exemplo de um sistema 2×2

O objetivo do referido modelo é estimar o valor da variável X_i por meio de uma combinação linear das variáveis vizinhas à região i. Logo, supondo a metodologia

queen, a estimativa do valor da variável na região 1, na Figura 3.1, é

$$\widehat{x}_1 = 3\widehat{\alpha} + \widehat{\beta}\left(\omega_{12}x_2 + \omega_{13}x_3 + \omega_{14}x_4\right) \tag{3.3}$$

Analogamente, se as variáveis obedecerem exatamente à expressão (3.2), ou seja, se $\epsilon_i = 0 \; \forall \; i = 1, \cdots, n$, e se for adotada tal relação espacial ao invés da relação linear na equação (3.1), ter-se-á

$$\rho = \frac{\sum_{i=1}^{n}\left[x_i - \bar{x}\right]\left[\alpha + \beta\sum_j \omega_{ij}x_j - \frac{1}{n}\sum_{i=1}^{n}\left(\alpha + \beta\sum_j \omega_{ij}x_j\right)\right]}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^{1/2}\left[\sum_{i=1}^{n}\left(\alpha + \beta\sum_j \omega_{ij}x_j - \frac{1}{n}\sum_{i=1}^{n}\left(\alpha + \beta\sum_j \omega_{ij}x_j\right)\right)^2\right]^{1/2}} \tag{3.4}$$

De maneira a facilitar a apresentação dos cálculos, a média amostral das estimativas do modelo espacial autorregressivo será desenvolvido em separado:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\alpha + \beta\sum_j \omega_{ij}x_j\right) &= \frac{n\alpha}{n} + \beta\left[\frac{\left(\sum_j \omega_{1j}x_j\right) + \left(\sum_j \omega_{2j}x_j\right) + \cdots + \left(\sum_j \omega_{nj}x_j\right)}{n}\right] \\ &= \alpha + \beta\left[\frac{x_1\left(\sum_j \omega_{j1}\right) + x_2\left(\sum_j \omega_{j2}\right) + \cdots + x_n\left(\sum_j \omega_{jn}\right)}{n}\right]\end{aligned} \tag{3.5}$$

Assim, para que o resultado da equação (3.5) seja igual a $\alpha + \beta\bar{x}$, é necessário adotar $\sum_j \omega_{j\xi} = 1 \; \forall \; \xi = 1, \cdots, n$. Ou seja, a matriz de proximidade deve ser padronizada em relação à coluna.

Para aumentar os esclarecimentos, sobre o aspecto discutido, será calculada a média dos valores esperados do modelo espacial autorregressivo aplicado no sistema

da Figura 3.1, adotando uma matriz de proximidade padronizada na coluna.

$$
\begin{aligned}
n\bar{x} &= [\alpha + \beta(\omega_{12}x_2 + \omega_{13}x_3 + \omega_{14}x_4)] + [\alpha + \beta(\omega_{21}x_1 + \omega_{23}x_3 + \omega_{24}x_4)] + \\
&+ [\alpha + \beta(\omega_{31}x_1 + \omega_{32}x_2 + \omega_{34}x_4)] + [\alpha + \beta(\omega_{41}x_1 + \omega_{42}x_2 + \omega_{43}x_3)] \\
n\bar{x} &= n\alpha + \beta x_1(\omega_{21} + \omega_{31} + \omega_{41}) + \beta x_2(\omega_{12} + \omega_{32} + \omega_{42}) + \\
&+ \beta x_3(\omega_{13} + \omega_{23} + \omega_{43}) + \beta x_4(\omega_{14} + \omega_{24} + \omega_{34}) \\
\bar{x} &= \frac{n\alpha}{n} + \frac{\beta(x_1 + x_2 + x_3 + x_4)}{n} = \alpha + \beta\bar{x}
\end{aligned} \tag{3.6}
$$

Finalmente, substituindo o resultado de (3.5) na equação (3.4) tem-se

$$
\begin{aligned}
\rho &= \frac{\sum_{i=1}^{n}[x_i - \bar{x}]\left[\alpha + \beta\sum_j \omega_{ij}x_j - \alpha - \beta\bar{x}\right]}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{n}\left(\alpha + \beta\sum_j \omega_{ij}x_j - \alpha - \beta\bar{x}\right)^2\right]^{1/2}} \\
&= \frac{\beta\sum_{i=1}^{n}[x_i - \bar{x}]\left[\sum_j \omega_{ij}x_j - \bar{x}\right]}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}\left[\beta^2\sum_{i=1}^{n}\left(\sum_j \omega_{ij}x_j - \bar{x}\right)^2\right]^{1/2}} \\
&= \frac{\sum_{i=1}^{n}[x_i - \bar{x}]\left[\sum_j \omega_{ij}x_j - \bar{x}\right]}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{n}\left(\sum_j \omega_{ij}x_j - \bar{x}\right)^2\right]^{1/2}}
\end{aligned} \tag{3.7}
$$

Retornando à discussão de Griffith (2003), o resultado da equação (3.7) mostra que, para se deduzir um coeficiente de autocorrelação espacial a partir do coeficiente de correlação de Pearson, deve-se substituir, tanto no numerador quanto no denominador, os y_i pelos vizinhos de x_i.

A partir de tudo o que foi apresentado, chega-se à conclusão de seis tópicos primordiais:

- o índice $\mathrm{I}_{\texttt{Proposto}}$, definido na equação (3.7), pode-se ser utilizado para se mensurar a autocorrelação espacial em um sistema bidimensional, uma vez que ele mensura a correlação entre os vetores $\mathbf{x}$ e $\mathbf{Wx}$;

- a matriz de proximidade deve ser padronizada em relação à coluna de modo que os vetores $\mathbf{x}$ e $\mathbf{Wx}$ apresentem a mesma média amostral;

- $\mathrm{I}_{\texttt{Proposto}}$ mensura a relação entre o valor da variável na região i com a capacidade

de esse valor ser previsto pelo do modelo autorregressivo espacial. Ou seja, se $x_i = \sum_j \omega_{ij} x_j \ \forall \ \xi = 1, \cdots, n$ então $\mathrm{I}_{\texttt{Proposto}} = 1$;

- $\mathrm{I}_{\texttt{Proposto}}$ é invariante com alterações de escala ($Y_i = bX_i$ produz os mesmos resultados que X_i). Entretanto, não é invariante com alterações de localização ($Y_i = a + X_i$ produz os resultados diferentes de X_i);

- Como $\mathrm{I}_{\texttt{Proposto}}$ é derivado do coeficiente de correlação de Pearson, ele também varia entre -1 e $+1$. A demonstração é idêntica à apresentada no Apêndice A.

- $\widetilde{\mathrm{r}}_{\texttt{Cliff}}$ produzirá o mesmo resultado que o $\mathrm{I}_{\texttt{Proposto}}$ quando: (i) a variável $\mathbf{x}$ apresentar média zero, por exemplo quando se estiver analisando resíduos de uma regressão; (ii) no calculo de $\widetilde{\mathrm{r}}_{\texttt{Cliff}}$, a matriz $\mathbf{W}$ também deve ser padronizada em relação a coluna.

$$\begin{aligned}\widetilde{\mathrm{r}}_{\texttt{Cliff}} = \frac{\mathrm{r}_{\texttt{Cliff}}}{\max|\mathrm{r}_{\texttt{Cliff}}|} &= \frac{Cov\left(\sum_j \omega_{ij} z_j, z_i\right)}{\left[Var\left(z_i\right) Var\left(z_i\right)\right]^{\frac{1}{2}}} \times \left[\frac{Var\left(z_i\right)}{Var\left(\sum_j \omega_{ij} z_j\right)}\right]^{\frac{1}{2}} \\ &= \frac{Cov\left(\sum_j \omega_{ij} z_j, z_i\right)}{\left[Var\left(\sum_j \omega_{ij} z_j\right) Var\left(z_i\right)\right]^{\frac{1}{2}}} = \mathrm{I}_{\texttt{Proposto}} \end{aligned} \tag{3.8}$$

3.2 Valores Extremos

O terceiro resultado apresentado na seção anterior afirma que se $x_i = \sum_j \omega_{ij} x_j \ \forall \ \xi = 1, \cdots, n$, então $\mathrm{I}_{\texttt{Proposto}} = 1$. Nessas situações, o modelo espacial autorregressivo está explicando perfeitamente o arranjo espacial dos dados.

A título de ilustração, suponha uma situação em que a variável $\mathbf{x}$ apresenta média nula. Logicamente, tem-se que a soma dos elementos do vetor $\mathbf{x}$ será igual a zero

$$x_1 + x_2 + x_3 + \cdots + x_n = 0 \tag{3.9}$$

Da equação (3.9) infere-se que a soma de todos os elementos, exceto do elemento $x_\xi \ \forall \ \xi = 1, \cdots, n$, será igual a $-x_\xi$. Por exemplo, para $\xi = 1$

$$(0)\, x_1 + (1)\, x_2 + (1)\, x_3 + \cdots + (1)\, x_n = (-1)\, x_1 \tag{3.10}$$

Replicando isso para todos as n observações, obtém-se uma matriz $\mathbf{W}$ do seguinte formato:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 0 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 0 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \tag{3.11}$$

Essa matriz vai contra a Primeira Lei da Geografia, a qual revela que: "tudo está relacionado com tudo, mas as coisas mais próximas estão mais relacionadas entre si que as coisas mais distantes" (Tobler, 1979).

Note que uma matriz de proximidade nesse formato não traz nenhuma informação sobre a distribuição espacial dos dados. Note também que essa matriz implica que, independentemente dos dados observados, $-\mathbf{x} = \mathbf{W}\mathbf{x}$. Logicamente, espera-se obter $\mathrm{I}_{\texttt{Proposto}} = -1$. A Tabela 3.1 apresenta a média dos resultados das simulações do $\mathrm{I}_{\texttt{Proposto}}$ e do $\mathrm{I}_{\texttt{Moran}}$ aplicando $\mathbf{W}$ no formato (3.11). Em cada iteração, preenche-se o vetor $\mathbf{x}$ com valores aleatórios normalmente distribuidos.

Tabela 3.1: Média[a] do resultados do índice de Moran e do índice Proposto quando se utiliza uma matriz de proximidade completa

Tamanho do Sistema	Número de Observações	Índice de Moran	Índice Proposto
3×3	9	-0.1250	-1
5×5	25	-0.0416	-1
7×7	49	-0.0208	-1
10×10	100	-0.0101	-1

[a] Média é baseada em 100 simulações

Como era de se esperar, os resultados em cada uma das 100 simulações foram os mesmos. Utilizando uma matriz de proximidade completa, enquanto o valor do $\mathrm{I}_{\texttt{Proposto}}$ é sempre -1, o valor do índice de Moran é determinado pela padronização de $\mathbf{W}$. Cada linha da matriz tem $n-1$ observações unitárias. Note que os resultados do $\mathrm{I}_{\texttt{Moran}}$ são iguais aos valores de cada célula de $\mathbf{W}$ após ela ser padronizada em relação a sua linha, $1/8 = 0.1250$, $1/24 = 0.0416$, $1/48 = 0.0208$ e $1/99 = 0.0101$. Note também que para ter $\mathrm{I}_{\texttt{Proposto}} = 1$ basta usar a matriz (3.11) negativa.

Para demonstrar a dependência do $\mathrm{I}_{\texttt{Moran}}$ com a quantidade de dados observados, utilizar-se-á: (i) o índice em sua forma matricial, equação (1.16); (ii) uma matriz de proximidade completa, equação (3.11), padronizada em relação às suas linhas. Lembrando que o vetor de dados $\mathbf{z}$, pela definição, tem média zero.

Expandindo o denominador de (1.16), tem-se que

$$\mathbf{z}'\mathbf{z} = z_1^2 + z_2^2 + \cdots + z_n^2 \tag{3.12}$$

Expandindo o numerador de (1.16), tem-se

$$\begin{aligned}\mathbf{z}'\mathbf{W}\mathbf{z} = z_1\left(\frac{z_2 + \cdots + z_n}{n-1}\right) + z_2\left(\frac{z_1 + z_3 + \cdots + z_n}{n-1}\right) + \cdots + \\ + z_{n-1}\left(\frac{z_1 + \cdots + z_{n-2} + z_n}{n-1}\right) + z_n\left(\frac{z_1 + \cdots + z_{n-1}}{n-1}\right)\end{aligned} \tag{3.13}$$

Note que, em decorrência de $\sum_j z_j = 0$ para $j = 1, \cdots, n$, em cada numerador de (3.13) vale a propriedade (3.10). Assim,

$$\mathbf{z}'\mathbf{W}\mathbf{z} = \frac{1}{n-1}\left(-z_1^2 - z_2^2 - \cdots - z_n^2\right) \tag{3.14}$$

Finalmente, juntando os resultados de (3.12) e (3.14)

$$\mathrm{I}_{\texttt{Moran}} = \frac{\mathbf{z}'\mathbf{W}\mathbf{z}}{\mathbf{z}'\mathbf{z}} = \frac{-1}{n-1} \tag{3.15}$$

3.2.1 Exemplos de Valor Extremo $n = 9$

Na seção anterior foi apresentada uma situação hipotética e não-aplicável de quando o $\mathrm{I}_{\texttt{Proposto}}$ atinge seu valor extremo. Nessa situação, o mesmo não acontece com o $\mathrm{I}_{\texttt{Moran}}$, confirmando apenas sua dependência com a quantidade de dados do sistema. A presente seção irá apresentar dois exemplos, utilizando a metodologia de vizinhanças *queen*, em um sistema regular com poucas observações $n = 9$. Neles o $\mathrm{I}_{\texttt{Proposto}}$ atinge um valor bem próximo ao seu extremo.

O primeiro conjunto de dados objetivou simular uma análise descritiva na qual a variável em questão apresenta média igual a zero, como resíduos de uma regressão. Esses dados estão presentes na Figura 3.2 e os resultados dos índices estão na primeira

parte da Tabela 3.2. Como era de se esperar, o índice de Moran obteve o mesmo valor independentemente da padronização da matriz $\mathbf{W}$ e $\widetilde{\mathrm{r}}_{\texttt{Cliff}} = \mathrm{I}_{\texttt{Proposto}}$ apenas quando $\mathbf{W}$ é padronizada pela coluna.

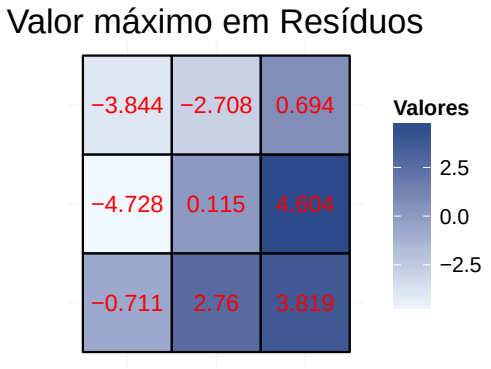

Figura 3.2: Exemplo de valor extremo analisando uma variável com média zero

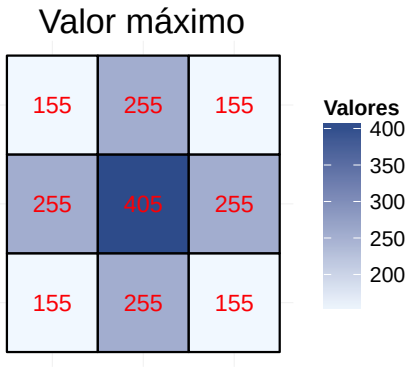

Figura 3.3: Exemplo de valor extremo em um sistema regular com $n = 9$

A partir do decaimento das cores da Figura 3.2, infere-se que esse sistema apresenta uma autocorrelação espacial bastante alta. O índice proposto consegue captar isso, atingindo o valor de 0.994. Já o índice de Moran, sem o fator de correção proposto do Cliff e Ord (1969), não consegue. Esse indica apenas a existência de uma autocorrelação baixa, $\mathrm{I}_{\texttt{Moran}} = 0.356$.

Tabela 3.2: Comparação dos valores extremos do índice Proposto e o do índice de Moran

	Figura	Índice Proposto	Índice de Moran	
			Linha[a]	Coluna[a]
I	(3.2)	0.99435	0.35654	0.35654
$\max \lvert \mathrm{r}_{\texttt{Cliff}} \rvert$		-	0.41668	0.35857
$\widetilde{\mathrm{r}}_{\texttt{Cliff}}$		-	0.85567	0.99435
I	(3.3)	0.99999	-0.44	-0.44
$\max \lvert \mathrm{r}_{\texttt{Cliff}} \rvert$		-	0.45607	0.44109
$\widetilde{\mathrm{r}}_{\texttt{Cliff}}$		-	-0.96476	-0.99751

[a] Tipo de padronização da matriz $\mathbf{W}$

O segundo conjunto de dados, Figura 3.3, objetivou simular a utilização e a interpretação do $\mathrm{I}_{\texttt{Proposto}}$ como um índice que mensura a autocorrelação espacial e compará-lo com o $\mathrm{I}_{\texttt{Moran}}$. Os resultados dos índices estão na segunda parte da Tabela 3.2.

Note que esse sistema é um exemplo de decaimento perfeito. O valor máximo da variável está situado no centro do sistema e, a cada nova vizinhança, esse valor muda igualmente. Intuitivamente espera-se uma autocorrelação espacial muito próxima de $+1$. Contudo, além do índice de Moran não atingir um valor próximo a esse, ele fornece um valor negativo, $\mathrm{I}_{\texttt{Moran}} = -0.44$.

Já o índice proposto, além de atingir um valor muito próximo do intuitivo $\mathrm{I}_{\texttt{Proposto}} = 0.9999$, ele fornece uma explicação bastante simples do porque ele não é $+1$. Como foi apresentado anteriormente, o índice proposto mensura a correlação entre o vetor $\mathbf{x}$ e a capacidade de um modelo espacial autorregressivo explicá-lo. Assim, como pode ser visto na equação (3.16), nessa problemática o modelo $\mathbf{Wx}$ explica quase que perfeitamente $\mathbf{x}$. Contudo, como $\mathbf{x} \neq \mathbf{Wx}$, o índice proposto não atinge seu extremo.

$$\begin{aligned} \mathbf{x}' &= (155,\ 255,\ 155,\ 255,\ 405,\ 255,\ 155,\ 255,\ 155) \\ \mathbf{Wx}' &= (152.6,\ 256,\ 152.6,\ 256,\ 410.7,\ 256,\ 152.6,\ 256,\ 152.6) \end{aligned} \tag{3.16}$$

3.3 Distribuição de Probabilidade Empírica

Uma vez que os resultados do índice proposto foram bastante coerentes, realizou-se, nessa seção, uma simulação que objetivou estudar sua distribuição de probabilidade e compará-la com a distribuição do índice de Moran. O objetivo é propor um novo teste de hipótese capaz de inferir a existência de autocorrelação espacial.

Em um sistema 7×7, utilizando a metodologia *queen*, (i) gerou-se aleatoriamente observações normalmente distribuidas com média 0 e desvio-padrão 1, (ii) distribuiu-se aleatoriamente essas observações pelo sistema e (iii) calculou-se os valores dos índices $\mathrm{I}_{\texttt{Moran}}$ e $\mathrm{I}_{\texttt{Proposto}}$. Repetiu-se os passos $(i) - (iii)$ 10.000.000 de vezes. A Figura 3.4 apresenta o histograma dos resultados do índice proposto enquanto que a Figura 3.5 apresenta os resultados do índice de Moran. Para ambos casos apresentou-se também a curva da distribuição de normal ajustada aos dados.

Em ambos os casos, aplicou-se o teste não-paramétrico de Shapiro-Wilk. O p-valor do índice proposto foi 0.2501 enquanto que, para o índice de Moran, foi inferior a 0.01. Logo, há indícios que a distribuição de probabilidade do índice proposto é

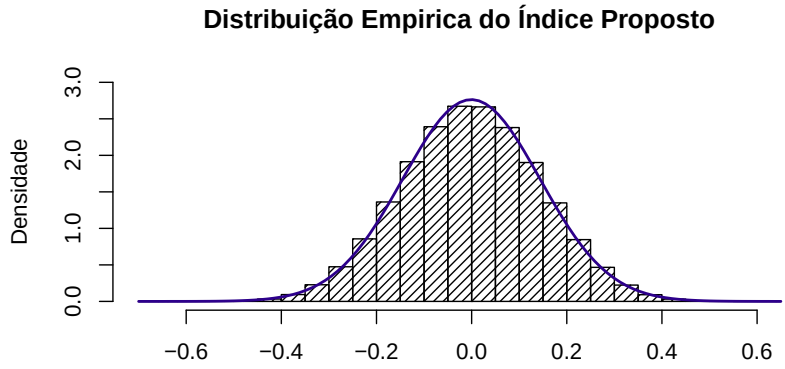

Figura 3.4: Distribuição Empírica do Índice Proposto $n = 49$

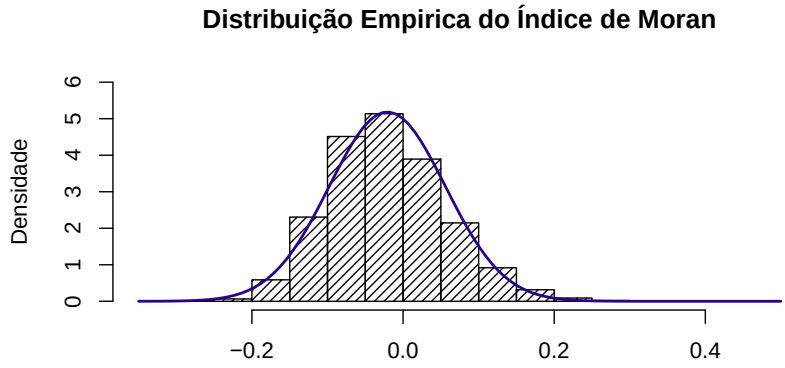

Figura 3.5: Distribuição Empírica do Índice de Moran $n = 49$

normal.

Uma explicação para o resultado do teste de hipotese do índice de Moran ter sido significativo é o fato de se ter utilizando um sistema com apenas 49 observações. A literatura apresenta que o índice de Moran obedece uma distribuição normal apenas quando o número de observações é muito grande (Tiefelsdorf e Boots, 1995).

3.4 Tópicos Conclusivos

O índice proposto mensura a relação entre o valor da variável na região i com a capacidade de esse valor ser previsto através de uma combinação linear dos valores vizinhos à região i. Logicamente, quanto menor for o erro de estimação, mais o índice proposto se aproximará de 1.

A matriz de proximidade $\mathbf{W}$ deve ser padronizada em relação à coluna de forma que os vetores $\mathbf{x}$ e $\mathbf{Wx}$ apresentem a mesma média, viabilizando sua comparação.

Há indícios que a distribuição de probabilidade do índice proposto seja normal e que, provavelmente, essa convergência ocorra mais rápido que a convergência da distribuição do índice de Moran.

Capítulo 4

Aplicações

Esse capítulo apresenta uma comparação da metodologia existente com a metodologia que está sendo proposta. O objetivo é aplicar o que foi desenvolvido em sistemas com poucas observações, de modo a aumentar os esclarecimentos. A primeira seção apresenta uma discussão utilizando dados simulados, enquanto a segunda apresenta duas aplicações em dados reais. A primeira aplicação foi responsável por despertar a motivação para o presente trabalho. São dados referentes ao número total de roubos residenciais e de veículos por mil domicílios, no ano de 1980, em cada bairro de Columbus, capital de Ohio, Estados Unidos (Anselin, 1988). A segunda abrange dados referentes ao número de títulos de mestrado concedidos no Brasil por Unidade da Federação no ano de 2012 (GEOCAPES/MEC).

4.1 Dados Simulados

As simulações apresentadas a seguir pretendem continuar o exercício apresentado na Introdução, aumentar os esclarecimentos acerca das características intrínsecas do índice I de Moran e testar o índice proposto. Objetivou-se entender a relação entre os valores dos índices, o tipo de vizinhança e o número de vizinhos (curvas de nível) geradas a partir da localização do pico. Por exemplo, a Figura 4.1 apresenta 4 vizinhanças. Calculou-se o valor do índice de Moran, o valor do índice proposto e a quantidade de vizinhanças existentes em sistemas quadráticos regulares utilizando as metodologias de vizinhança *queen* e *rook*. O exercício abrangeu sistemas de tamanho

Figura 4.1: Menor número de vizinhanças em um sistema 7×7

3×3 ($n = 9$) até sistemas de tamanho 14×14 ($n = 196$). Os valores das regiões x_i foram gerados após a definição da localidade do pico, replicando a metodologia da Figura 6. Todas as etapas, desde a criação dos sistemas quadráticos até os resultados das simulações, foram feitas usando o software **R 3.1.2**.

Tabela 4.1: Comparação entre o índice de Moran, o índice proposto e a quantidade de vizinhanças em um sistema 7×7

Pico	*Queen*			*Rook*		
	Curvas de Nível	Índice de Moran	Índice Proposto	Curvas de Nível	Índice de Moran	Índice Proposto
1	7	0.8303	0.8718	13	0.9184	0.9249
2	7	0.8406	0.8787	12	0.9139	0.9493
3	7	0.8452	0.8727	11	0.8926	0.9413
4	7	0.8464	0.8715	10	0.8813	0.9491
9	6	0.8331	0.9076	11	0.9058	0.9787
10	6	0.8056	0.8973	10	0.8754	0.9723
11	6	0.7919	0.8873	9	0.8582	0.9797
16	6	0.8056	0.8973	10	0.8754	0.9723
17	5	0.7265	0.9185	9	0.8211	0.9610
18	5	0.6510	0.9110	8	0.7855	0.9687
23	6	0.7919	0.8873	9	0.8582	0.9797
24	5	0.6510	0.9110	8	0.7855	0.9687
25	4	0.5193	0.9497	7	0.7335	0.9770

As Tabelas 4.1, 4.2, 4.3 e D.1 apresentam os resultados das referidas simulações para sistemas regulares com 49, 25, 9 e 196 regiões respectivamente. Os resultados

do sistema 14×14 estão presentes no Apêndice D. Optou-se pela referida alternativa já que a Tabela D.1 apresenta um elevado número de linhas.

Em suma, para um número de observações pequeno ou médio, como os sistema 3×3 e 5×5, o índice proposto atinge valores mais elevados e, consequentemente, mais coerentes que os valores de Moran. Na metodologia *queen*, o índice proposto obtém valores maiores em situações com menor número de vizinhanças, como é o caso da Figura 4.1. Com o índice de Moran, nessa mesma metodologia ocorre o contrário: obtêm-se valores maiores quando há maior número de vizinhanças. Quando utiliza-se vizinhanças *rook*, esse padrão não se mantém.

Tabela 4.2: Comparação entre o índice de Moran, o índice proposto e a quantidade de vizinhanças em um sistema 5×5

Pico	*Queen*			*Rook*		
	Curvas de Nível	Índice de Moran	Índice Proposto	Curvas de Nível	Índice de Moran	Índice Proposto
1	5	0.6850	0.7838	9	0.8400	0.8640
2	5	0.7019	0.8112	8	0.8110	0.9130
3	5	0.7046	0.8016	7	0.7719	0.9026
7	4	0.6359	0.8612	7	0.7474	0.9778
8	4	0.5215	0.8507	6	0.6642	0.9589
12	4	0.5215	0.8507	6	0.6642	0.9589
13	3	0.2283	0.8647	5	0.5095	0.9297

Quando são comparados os resultados das metodologias *queen* e *rook*, observa-se um número maior de vizinhanças em sistemas *rook*. Outro fato interessante é que nesses sistemas, os valores de ambos índices são significativamente maiores do que seus respectivos valores em sistemas *queen*. Por exemplo, em um sistema 5×5, quando o pico situa-se na posição 7, obtém-se $\mathrm{I}_{\texttt{Moran}} = 0.63$ e $\mathrm{I}_{\texttt{Proposto}} = 0.86$ na metodologia *queen* versus $\mathrm{I}_{\texttt{Moran}} = 0.74$ e $\mathrm{I}_{\texttt{Proposto}} = 0.97$ na metodologia *rook*.

Para um número grande de observações, por exemplo um sistema 14×14, os valores do índice de Moran e do índice proposto são bastante semelhantes quando há um número elevado de vizinhanças. Por exemplo, quando o pico localiza-se na região de número 1, existem 14 vizinhanças *queen* e 27 vizinhanças *rook*. Nessa situação, encontraram-se valores muitos semelhantes para ambos os casos: $\mathrm{I}_{\texttt{Moran}} = 0.954$ e $\mathrm{I}_{\texttt{Proposto}} = 0.948$ na metodologia *queen* versus $\mathrm{I}_{\texttt{Moran}} = 0.979$ e $\mathrm{I}_{\texttt{Proposto}} = 0.974$ na metodologia *rook*. Todavia, quando a posição do pico não gera muitas vizinhanças,

como ocorre na região de número 91: os valores dos dados formam exatamente uma "pirâmide" e geram 8 vizinhanças *queen* e 15 vizinhanças *rook*. Nessa situação, os valores dos índices diferem entre si: $I_{\texttt{Moran}} = 0.855$ e $I_{\texttt{Proposto}} = 0.989$ na metodologia *queen* versus $I_{\texttt{Moran}} = 0.932$ e $I_{\texttt{Proposto}} = 0.993$ na metodologia *rook*.

Tabela 4.3: Comparação entre o índice de Moran, o índice proposto e a quantidade de vizinhanças em um sistema 3×3

Pico	*Queen*			*Rook*		
	Curvas de Nível	Índice de Moran	Índice Proposto	Curvas de Nível	Índice de Moran	Índice Proposto
1	3	0.2386	0.3896	5	0.5556	0.5524
2	3	0.2344	0.5408	4	0.4028	0.7510
5	2	-0.2667	0.7546	3	-0.1111	0.9615

Contudo, o grande destaque do índice definido no presente trabalho é percebido quando o número de observações no sistema não é grande. No caso do sistema 3×3, os resultados do índice de Moran são muito baixos, com exceção do caso em que o pico localiza-se na região de número 1 e utiliza-se a metodologia *rook*. Nessa situação, os valores dos dois índices são praticamente os mesmos: $I_{\texttt{Moran}} = 0.555$ e $I_{\texttt{Proposto}} = 0.552$.

Um aspecto a ser ressaltado é que quando o pico se localiza no centro do sistema, mesmo existindo uma relação direta entre tais dados, o índice de Moran é negativo em ambas metodologias. Situação semelhante à análise da Figura 3.3. Contudo, nessas situações é justamente quando o índice proposto atinge valores mais elevados, caracterizando melhor o arranjo espacial dos dados.

Por fim, realizou-se uma simulação a fim de descobrir o tamanho do sistema para o qual o valor máximo do índice de Moran seja aproximadamente igual ao máximo do índice proposto.

Como visto anteriormente, ambos índices atingem valores mais altos quando utiliza-se a metodologia *rook*. Em virtude disso, a convergência nessa situação é mais rápida que a convergência da metodologia *queen*.

Na metodologia *rook*, em um sistema de lado 20×20 ($n = 400$), ambos os índices já ultrapassaram a casa de 0.99, $I_{\texttt{Moran}} = 0.9901$ e $I_{\texttt{Proposto}} = 0.9968$. Contudo, mesmo utilizando um sistema 30×30 ($n = 900$), o índice de Moran não conseguiu alcançar o valor do índice proposto, $I_{\texttt{Moran}} = 0.9965$ e $I_{\texttt{Proposto}} = 0.9978$.

Pela metodologia *queen*, enquanto que índice proposto ultrapassa a casa do 0.99

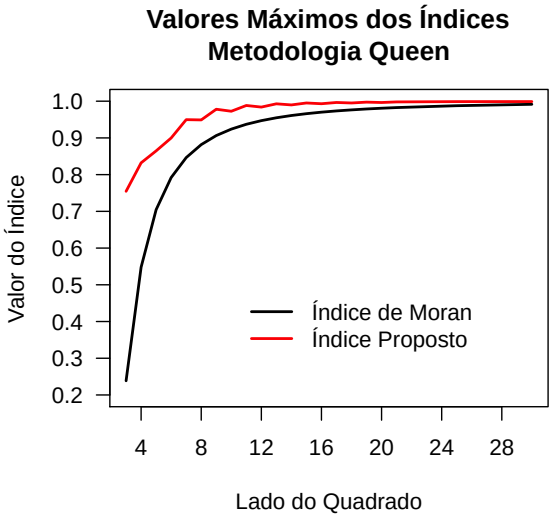

Figura 4.2: Convergência do valor máximo do índice de Moran em relação ao tamanho do sistema utilizando a metodologia *queen*

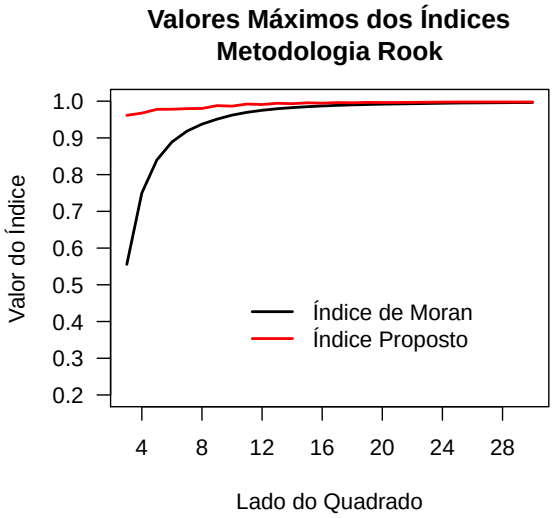

Figura 4.3: Convergência do valor máximo do índice de Moran em relação ao tamanho do sistema utilizando a metodologia *rook*

em um sistema 13×13 (na metodologia *rook* isso ocorreu em um 12×12) o índice de Moran só consegue atingir esse patamar em um sistema 30×30 ($n = 900$). Nessa situação, os valores do índices são: $\mathrm{I}_{\texttt{Moran}} = 0.9915$ e $\mathrm{I}_{\texttt{Proposto}} = 0.9989$.

4.2 Dados Reais

4.2.1 Roubos por mil domicílios em Columbus em 1980

Anselin (1988) apresenta dados sobre o total de roubos residenciais e de veículos por mil domicílios em cada um dos 49 bairros de Columbus, capital de Ohio, EUA.

Essa aplicação despertou a motivação para este trabalho, uma vez que é simples notar que nos bairros mais ao centro de Columbus há muito mais roubos do que nos bairros mais afastados, Figura 4.4. Essa dependência é tão direta que lembra simulações feitas anteriormente, como a expressa na Figura 4.1. Contudo, Anselin (1988) obteve apenas $\mathrm{I}_{\texttt{Moran}} = 0.50$, afirmando que o sistema apresenta uma dependência espacial moderada, sendo intuitivamente mais baixa que o esperado.

Quando calcula-se o índice proposto e a correção indicada por Cliff e Ord (1969) a tais dados, obtêm-se valores mais coerentes do que o indicado por Anselin (1988),

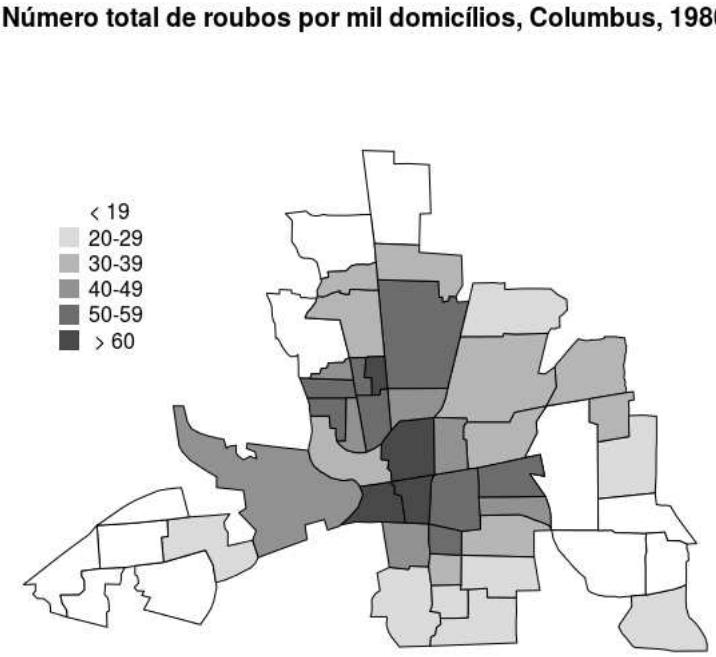

Figura 4.4: Número total de roubos residenciais e de veículos por mil domicílios, Columbus, 1980

0.69 e 0.68, respectivamente. Tal fator indica uma importante e mais alta correlação espacial.

Tabela 4.4: Valores do Índice de Moran em relação a metodologia da matriz de proximidade para o número total de roubos residenciais e de veículos por mil domicílios em cada bairro de Columbus, Ohio, 1980

I de Moran	*Queen*	*Rook*	$1 \,/\, d_{ij}^2$ *
Convencional	0.5001	0.5236	0.3688
Cliff (1969)	0.6837	0.7011	0.7327
Proposto	0.6910	0.6898	0.7579

* Distância Euclidiana entre os centroides das regiões

É importante ressaltar, entretanto, que resultados mais expressivos ocorrem quando se utiliza uma **W** com o inverso do quadrado da distância entre os centroides das regiões, Tabela 4.4. Como argumentado no Capítulo 2, em análises semelhantes a essa, é mais recomendado a utilização de uma matriz de proximidade que incorpore informação da distância entre as regiões pois ela caracteriza melhor o decaimento dos dados.

Com essa **W**, o índice de Moran atinge apenas 0.36, não sendo um valor plausível para os dados expostos. Já o índice proposto atinge 0.76, o que confirma que o número

de roubos em Columbus depende consideravelmente da localidade do bairro. Assim, quanto mais próximo ele for do centro, mais alto será o número de delitos.

4.2.2 Títulos de Mestrado no Brasil em 2012

A existência de uma população educada, com adequados níveis de qualificação profissional, capaz de se ajustar aos permanentes avanços tecnológicos do processo de trabalho e dos bens e serviços em geral é condição necessária para o desenvolvimento do país, para sua competitividade e para a própria qualidade de vida de seus cidadãos (Viotti et al., 2012).

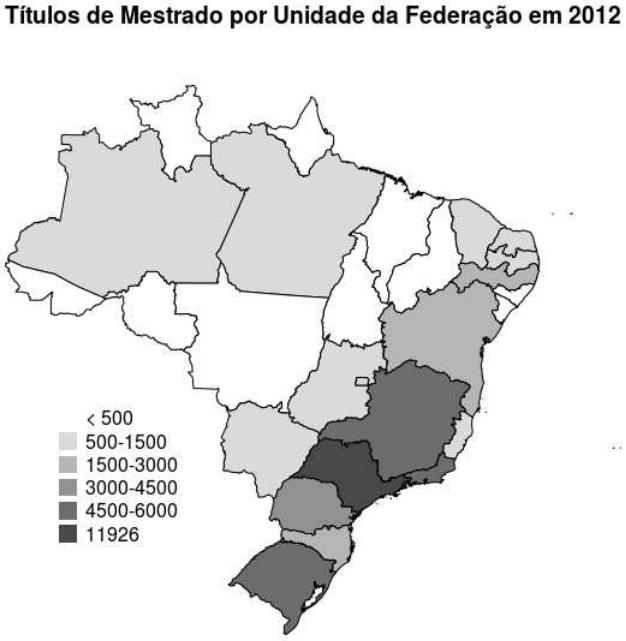

Figura 4.5: Número de títulos de Mestrado concedidos no Brasil por Unidade da Federação, 2012

Com o objetivo de identificar problemas ou necessidades e de orientar políticas públicas relativas à formação, ao treinamento, à absorção e ao emprego de recursos humanos, o Centro de Gestão e Estudos Estratégicos (CGEE) elabora desde 2010 estudos sobre a formação e o emprego de mestres e doutores titulados no Brasil. Esses estudos nasceram a partir da cooperação entre a Coordenação de Capacitação de Pessoal de Ensino Superior (Capes), o Ministério da Educação (MEC), o Ministério do Trabalho e Emprego (MTE), o Ministério da Ciência, Tecnologia e Inovação e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Viotti et al.,

2012).

Essa seção objetiva estudar a relação espacial do número de títulos de mestrado concedidos em 2012 por Unidade da Federação da Instituição de Ensino Superior. Como era esperado, a UF com maior número de títulos é São Paulo 11.926 (25.3%), seguida por Rio de Janeiro 5.921 (12, 6%) e Minas Gerais 4.925 (10, 4%). As UF's com menor número de títulos são Amapá 61 (0, 12%), Acre 48 (0, 10%) e Roraima 44 (0.09%). Todos os dados são apresentados na Figura 4.5 (CAPES/MEC, 2015). A média e o desvio-padrão deles são 1.745, 7 e 2.572, 5, respectivamente.

Tabela 4.5: Valores do Índice de Moran em relação à metodologia da matriz de proximidade para o número de títulos de mestrado por Unidade da Federação do Brasil, 2012

I de Moran	*Queen*	*Rook*	$1\ /\ d_{ij}^2$ *
Convencional	0.3299	0.3299	0.1751
Cliff (1969)	0.5485	0.5485	0.4272
Proposto	0.5375	0.5375	0.3826

* Distância Euclidiana entre os centroides das regiões

Diferentemente do que aconteceu na aplicação anterior, os dados de titulação não apresentam um decaimento homogêneo. Pelo contrário, são dados altamente agregados. A região Sudeste sozinha é responsável por 48.3% das titulações. Quando a metodologia *queen* é utilizada, obtém-se $\mathrm{I}_{\texttt{Moran}} = 0.33$ e $\mathrm{I}_{\texttt{Proposto}} = 0.54$. Novamente, o índice de Moran apresenta um valor inferir ao índice proposto. Esse, por sua vez, indica uma correlação espacial moderada entre os dados.

Um aspecto interessante é que, como as fronteiras entres as Unidades das Federações brasileiras são mais assimétricas que as fronteiras de Columbus, não existem diferenças entre se utilizar vizinhanças *queen* ou *rook*.

Como o Brasil é um país continental, mais uma vez é valido utilizar uma matriz de proximidade que incorpore a distância entre as regiões. Nesse caso, os valores são bem inferiores aos da metodologia *queen*, $\mathrm{I}_{\texttt{Moran}} = 0.16$ e $\mathrm{I}_{\texttt{Proposto}} = 0.38$. Esse aspecto termina por indicar correlação espacial fraca. Que é um resultado mais plausível, uma vez que os dados são altamente agregados e as regiões são muito distantes umas das outras.

Capítulo 5

Considerações Finais

O coeficiente de autocorrelação espacial fornece um ponto de partida para qualquer análise que envolve dados georreferenciados. Devido a sua simplicidade, o índice I de Moran é a estatística mais famosa e largamente utilizada para esse fim. Contudo, ela apresenta uma série de limitações que muitos desconhecem. Uma delas é sua relação com a quantidade de dados observados (demonstrada nesse trabalho) fazendo com que o $I_{\texttt{Moran}}$ não atinja valores extremos quando o número de observações do sistema é pequeno.

Inexistia na literatura um trabalho que elenca-se todas as apresentações do índice de Moran. Muito menos um trabalho que apresenta-se como o índice evoluiu da sua primeira apresentação, equação (1.5), para sua forma atual, equação (8).

Também inexistia o fato que, se a matriz de proximidade $\boldsymbol{\Delta}$ for simétrica, o índice de Moran apresenta os mesmos resultados, independente se $\boldsymbol{\Delta}$ for padronizada pela linha ou pela coluna.

A proposta de modificação do índice I de Moran, introduzindo no coeficiente de correlação de Pearson o conceito de modelagem espacial autorregressiva de primeira ordem, se mostrou bastante coerente. Os resultados mais expressivos são percebidos quando o sistema apresenta um número de observações pequeno ou médio. Isso é uma confirmação do fato de que muitos trabalhos podem estar sub-representando a verdadeira relação espacial entre os dados. Para sistemas com um número grande de observações, o índice de Moran e o índice proposto fornecem valores bastante semelhantes.

Uma vez que o índice proposto apresentou-se bastante coerente, realizou-se um primeiro exercício a fim de se estudar sua distribuição de probabilidade. Esse exercício indicou a possibilidade dessa distribuição ser normal. A continuação desse trabalho objetivará essa dedução algébrica. A intenção é propor um novo teste de hipótese capaz de inferir a existência de autocorrelação espacial.

Outra proposta de trabalho futuro é generalizar os resultados aqui apresentados para o índice C de Geary e o índice de Moran local.

Uma limitação do presente trabalho foi o fato da grande maioria das simulações incluirem apenas sistemas quadráticos regulares. Não foi feita uma busca exaustiva em sistemas com outro formato, muito menos em sistemas irregulares, onde há maior número ligações entre as regiões. No entanto, isso não invalida os resultados aqui apresentados.

Referências Bibliográficas

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

Anselin, L. & Florax, R. (2004). *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Springer-Verlag.

Arora, S. S. & Brown, M. (1977). Alternative approaches to spatial autocorrelation: An improvement over current practice. *International Regional Science Review*, 2:67–78.

Assunção, R. M. & Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18:2147–2162.

Bailey, T. C. & Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. John Wiley and Sons.

Bellman, R. (1970). *Introduction to Matrix Analysis*. McGraw-Hill.

Bivand, R. S. & Pebesma, E. (2013). *Applied spatial data analysis with R*, (2nd ed.). Springer.

CAPES/MEC (2015). Geocapes dados estatísticos. http://geocapes.capes.gov.br.

Chen, Y. (2013). New Approaches for Calculating Moran's Index of Spatial Autocorrelation. *PLoS ONE*, 8(7).

Cliff, A. D. (1969). *Some Measures of Spatial Association in Areal Data*. University of Bristol. Tese de Doutorado.

Cliff, A. D. & Ord, J. K. (1969). The problem of spatial autocorrelation. In: *Studies in Regional Science*, A. Scott, ed., pages 25–55. Pion Press.

Cliff, A. D. & Ord, J. K. (1970). Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography*, (46):269–292.

Cliff, A. D. & Ord, J. K. (1971). Evaluating the percentage points of a spatial autocorrelation coefficient. *Geographical Analysis*, (3):51–62.

Cliff, A. D. & Ord, J. K. (1973). *Spatial autocorrelation*. Pion Press.

Cliff, A. D. & Ord, J. K. (1975). The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning A*, 7(6):725–734.

Cliff, A. D. & Ord, J. K. (1981). *Spatial processes: models and applications*. Pion.

Dacey, M. F. (1965). A review of measures of contiguity for two and k-color maps. *Technical Report no 2, Spatial Diffusion Study*. Department of Geography, Northwestern University.

Dacey, M. F. (1968). A review of measures of contiguity for two and k-color maps. In: *Spatial Analysis: a Reader in Statistical Geography*, B. J. L. Berry & D. F. Marble, ed., pages 479–495. Prentice-Hall.

de Jong, P., Sprenger, C., & van Veen, F. (1984). On extreme values of moran's i and geary's c. *Geographical Analysis*, (16):17–24.

Diniz-Filho, J. A. F. & Santos, T. (2012). A comparison of metrics for estimating phylogenetic signal under alternative evolutionary models. *Genetics and Molecular Biology*, 35:673–679.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, (2nd ed.). John Wiley.

Gaetan, C. & Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer Series in Statistics. Springer.

Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, 37(4):491–496.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis*, 41(4):404–410.

Getis, A. & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2):90–104.

Getis, A. & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial models. In: *Pratical Handbook of Spatial Statistics*, S. L. Arlinghaus, ed. CRC.

Griffith, D. A. (2003). *Spatial autocorrelation and spatial filtering.* Springer.

Huo, X. N. & Zhang, W. W. (2011). Spatial pattern analysis of heavy metals in beijing agricultural soils based on spatial autocorrelation statistics. *International Journal of Environmental Research and Public Health*, 8(6):2074–2089.

Jackson, M. C. & Huang, L. (2010). A modified version of moran's i. *International Journal of Health Geographics*, 9(1).

Kendall, M. (1976). *Time Series*, (2nd ed.). Charles Griffin and Co.

Kooijman, S. A. L. M. (1976). Some remarks on the statistical analysis of grids especially with respect to ecology. In: *Annals of Systems Research*, B. Van Rootselaar, ed., volume 5, pages 113–132. Springer.

LeSage, J. & Pace, R. K. (2009). *Introduction to Spatial Econometrics.* Chapman and Hall.

Li, H. & Calder, C. A. (2007). Beyond moran's i: Testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, 39:357–375.

Lira, S. A. (2004). *Análise de correlação: abordagem teorica e de construção dos coeficientes com aplicações.* Universidade Federal do Paraná. Dissertação de Mestrado.

Moran, P. A. P. (1947). Random associations on a lattice. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43:321–328.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, 37(2):243–251.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, (37):17–23.

Neter, J. & Kutner, M. H. (2005). *Applied Linear Statistical Models*, (5th ed.). McGraw Hill.

Oden, N. (1995). Adjusting moran's i for population density. *Statistics in Medicine*, 14(1):17–26.

Ord, J. K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, (70):120–126.

Silva, A. R. (2006). *Avaliação de Modelos de Regressão Espacial para Análise de Cenários do Transporte Rodoviário de Carga*. Universidade de Brasília. Dissertação de Mestrado.

Silva, A. R. & Yamashita, Y. (2007). *Análise da Matriz de Proximidades Espacail para Problemas de Transportes*. Associação Nacional de Pesquisa e Ensino em Transportes. Rio de Janeiro: ANPET.

Snow, J. (1936). *Snow on cholera: being a reprint of two papers*. Humphrey Milford, Oxford University Press.

Tiefelsdorf, M. & Boots, B. (1995). The exact distribution of moran's i. *Environment and Planning A*, 27(6):985–999.

Tiefelsdorf, M. & Griffith, D. A. (1999). A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A*, 31(1):165–180.

Tobler, W. R. (1979). Cellular geography. In: *Philosophy in Geography*, S. Gale, ed., pages 379–386. Dordrecht: Reidel.

Verspagen, B. (2010). The spatial hierarchy of technological change and economic development in europe. *The Annals of Regional Science*, 45(1):109–132.

Viotti, E. B., Daher, S., Queiroz, A. S., & Carrijo, T. B. (2012). *Mestres 2012: Estudos da demografia da base técnico-científica brasileira*. Centro de Gestão e Estudos Estratégicos, CGEE.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, (41):434–449.

Wishart, J. & Hirschfeld, H. O. (1936). A theorem concerning the distribution of joins between line segments. *Journal of the London Mathematical Society*, 11(3):227–235.

Apêndice A

Valores Extremos do Coeficiente de Correlação de Pearson

Como foi apresentado na introdução do presente trabalho, na equação (1), a correlação entre as variáveis X e Y é definida por

$$\rho = \frac{\mathrm{cov}\,(X,Y)}{\sigma_X \sigma_Y} \tag{A.1}$$

σ_X é o desvio padrão de X, σ_Y é o desvio padrão de Y e $\mathrm{cov}\,(X,Y)$ é a covariância entre X e Y. Sabe-se, por definição, que a variância de qualquer variável é sempre positiva. Assim, pode-se escrever

$$\mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0 \tag{A.2}$$

Usando a propriedade da variância *da soma de duas variáveis aleatórias*, pode-se reescrever (A.2) como

$$\begin{aligned}
\mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\mathrm{cov}\left(\frac{\mathrm{X}}{\sigma_\mathrm{X}}, \frac{\mathrm{Y}}{\sigma_\mathrm{Y}}\right) &\geq 0 \qquad \text{(A.3)}\\
\frac{1}{\sigma_X^2}\mathrm{Var}\,(X) + \frac{1}{\sigma_Y^2}\mathrm{Var}\,(Y) + \frac{2}{\sigma_X \sigma_Y}\mathrm{cov}\,(\mathrm{X},\mathrm{Y}) &\geq 0\\
1 + 1 + \frac{2}{\sigma_X \sigma_Y}\mathrm{cov}\,(\mathrm{X},\mathrm{Y}) &\geq 0\\
1 + \frac{\mathrm{cov}\,(\mathrm{X},\mathrm{Y})}{\sigma_X \sigma_Y} &\geq 0\\
\rho = \frac{\mathrm{cov}\,(X,Y)}{\sigma_X \sigma_Y} &\geq -1
\end{aligned}$$

De maneira análoga também existe a inequação

$$\mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) \geq 0 \tag{A.4}$$

Aplicando a mesma metodologia que (A.3), obtém-se

$$\begin{aligned} \mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) - 2\mathrm{cov}\left(\frac{\mathrm{X}}{\sigma_\mathrm{X}}, \frac{\mathrm{Y}}{\sigma_\mathrm{Y}}\right) &\geq 0 \\ 1 - \frac{\mathrm{cov}\,(\mathrm{X}, \mathrm{Y})}{\sigma_X \sigma_Y} &\geq 0 \\ \rho = \frac{\mathrm{cov}\,(X, Y)}{\sigma_X \sigma_Y} &\leq 1 \end{aligned} \tag{A.5}$$

Finalmente, a partir de (A.3) e (A.5)

$$-1 \leq \rho \leq 1 \tag{A.6}$$

Apêndice B

Valores Extremos da Autocorrelação em Séries Temporais

Uma função $f(x)$, definida em $x \in \mathfrak{X}$, é dita ser semidefinida positiva se

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k f(t_k - t_j) \geq 0 \tag{B.1}$$

aqui $a_i \in \mathbb{R}$ para $i = 1, \cdots, n$ e $t_i - t_j \in \mathfrak{X}$ para todo par ordenado (i, j).

Seja X_t uma série temporal com $t \in \mathfrak{T}$. Sem perda de generalidades, será adotado que X_t tem média nula, ou seja, $E\{X_t\} = 0$. Assim, defini-se $(t_1, t_2, \cdots, t_n) \in \mathfrak{T}$ e, pela equação (4) da Introdução, sabe-se que $\gamma(t_k - t_j)$ é a covariância entre X_{t_j} e X_{t_k}. Como a variância de uma variável aleatória é sempre não negativa, tem-se

$$\begin{aligned} 0 \leq \mathrm{Var}\left\{\sum_{j=1}^{n} a_j X_{t_j}\right\} &= \mathrm{E}\left\{\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k X_{t_j} X_{t_k}\right\} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \gamma(t_k - t_j) \end{aligned} \tag{B.2}$$

Definindo $n = 2$ em (B.2), obtém-se

$$0 \leq a_1^2\gamma(0) + a_2^2\gamma(0) + 2a_1 a_2 \gamma(t_1 - t_2) \tag{B.3}$$

Com alguma álgebra obtém-se

$$\frac{-a_1 a_2 \gamma(t_1 - t_2)}{\gamma(0)} \leq \frac{(a_1^2 + a_2^2)}{2} \tag{B.4}$$

Primeiramente defini-se $t_1 - t_2 = h$. Depois, definindo $-a_1 = a_2 = 1$ obtém-se

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \leq 1 \tag{B.5}$$

Agora definindo $a_1 = a_2 = 1$ obtém-se

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \geq -1 \tag{B.6}$$

Finalmente, a partir de (B.5) e (B.6)

$$-1 \leq \rho(h) \leq 1 \tag{B.7}$$

Apêndice C

Valores Extremos da Índice de Cliff e Ord (1969)

Como foi apresentado na introdução do presente trabalho, Cliff e Ord (1969) propôs uma generalização da estatística proposta por Moran (1950) para mensurar a autocorrelação espacial em um espaço bidimensional

$$\mathrm{r}_{\texttt{Cliff}} = \frac{n \sum_{(2)} \omega_{ij} z_i z_j}{W \sum_{i=1}^{n} z_i^2} \tag{C.1}$$

com $\sum_{(1)} = \sum_{i=1}^{n}$; $\sum_{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n}$ para $i \neq j$ e $W = \sum_{(2)} \omega_{ij}$. Além da generalização, Cliff e Ord (1969) calculou os valores extremos que (C.1) atinge. Para tanto, primeiro definiu o numerador de (C.1) como

$$T = \sum_{(2)} w_{ij} z_i z_j \tag{C.2}$$

Definiu $k = n(i-1) + j$, com $k = 1, ..., n^2$, assim

$$\varpi_k = \omega_{ij} \qquad \gamma_k \equiv \gamma_{(i-1)n+j} = z_j \qquad \nu_k \equiv \nu_{j-n+1-ni} = z_i \tag{C.3}$$

Logo, (C.2) pode ser escrita como $\sum_k \varpi_k \gamma_k \nu_k$. Usando a desigualdade de Cauchy-Schwarz tem-se que

$$|T| = \sum_k \varpi_k^{\frac{1}{2}} \gamma_k \varpi_k^{\frac{1}{2}} \nu_k \leq \left\{ \left(\sum_k \varpi_k \gamma_k^2 \right) \left(\sum_k \varpi_k \nu_k^2 \right) \right\}^{1/2} = T^* \tag{C.4}$$

Isso implica que o valor absoluto de $\mathrm{r}_{\texttt{Cliff}}$ obedece a inequação

$$|\mathrm{r}_{\texttt{Cliff}}| \leq \frac{T^*}{\sum_{i=1}^{n} z_i^2} \tag{C.5}$$

Para simplificar a demonstração, de agora em diante, sem perda de generalidade, será adotado que $W = n$. Para que $\mathrm{r}_{\texttt{Cliff}}$ pertença ao intervalo $[-1; +1]$, o numerador de (C.5) deve ser menor que denominador

$$\left(\sum_{(1)} \omega_{i\cdot} z_i^2\right)\left(\sum_{(1)} \omega_{\cdot i} z_i^2\right) \leq \left(\sum_{(1)} z_i^2\right) \tag{C.6}$$

A inequação (C.6) se transforma em uma igualdade quando $\omega_{i\cdot} = \omega_{\cdot i} = 1$. A estatística (C.1) pode ser reescrita como

$$\begin{aligned} \mathrm{r}_{\texttt{Cliff}} &= \frac{\sum_i \left[\left(\sum_j \omega_{ij} z_j\right)(z_i)\right]}{\sqrt{\sum_i z_i^2}\sqrt{\sum_i z_i^2}} \\ &= \frac{Cov\left(\sum_j \omega_{ij} z_j, z_i\right)}{\left[Var\left(z_i\right) Var\left(z_i\right)\right]^{\frac{1}{2}}} \times \left[\frac{Var\left(\sum_j \omega_{ij} z_j\right)}{Var\left(\sum_j \omega_{ij} z_j\right)}\right]^{\frac{1}{2}} \\ &= \frac{Cov\left(\sum_j \omega_{ij} z_j, z_i\right)}{\left[Var\left(\sum_j \omega_{ij} z_j\right) Var\left(z_i\right)\right]^{\frac{1}{2}}} \times \left[\frac{Var\left(\sum_j \omega_{ij} z_j\right)}{Var\left(z_i\right)}\right]^{\frac{1}{2}} \end{aligned} \tag{C.7}$$

Com $J \equiv J(i)$ represento o conjunto das regiões vizinhas à região i. Assim, $\max|\mathrm{r}_{\texttt{Cliff}}| \leq 1$ sempre que

$$Var\left(\sum_j \omega_{ij} z_j\right) \leq Var\left(z_i\right) \tag{C.8}$$

Suponha agora que, cada observação seja composta por duas componentes, por exemplo $z_i = a_i + b$ com $i = 1, ..., n$. A componente b é comum para todas regiões e $Cov\left(a_i, b\right) = 0$ para $\forall i$. A componente a_i é não correlacionada com as outras componentes a_j para $i \neq j$. Definindo-se $Var\left(a_i\right) = \sigma^2$ e $Var\left(b\right) = \alpha^2$ tem-se

$$\begin{aligned} Var\left(z_i\right) &= \sigma^2 + \alpha^2 \\ Var\left(\sum_j \omega_{ij} z_j\right) &= n^{-1}\sigma^2 \sum_{(2)} \omega_{ij}^2 + \alpha^2 \end{aligned} \tag{C.9}$$

Substituindo (C.9) em (C.8) obtém-se que $\max|\mathrm{r}_{\texttt{Cliff}}| \leq 1$ sempre que

$$\sum_{(2)} \omega_{ij}^2 \leq n \tag{C.10}$$

A equação (C.10) será verdade sempre que a matriz de pesos estiver padronizada em

relação a suas linhas, ou seja, $\omega_{i\cdot} = 1$. Assim, usando uma matriz padronizada, a equação (C.8) implica que

$$\max |\mathrm{r}_{\texttt{Cliff}}| = \left[\frac{Var\left(\sum_j w_{ij} z_j\right)}{Var\left(z_i\right)} \right]^{\frac{1}{2}} \tag{C.11}$$

Finalmente, sem perda de generalidade, se a matriz de proximidade não estiver padronizada, o máximo absoluto de (C.1) é definido como

$$\max |\mathrm{r}_{\texttt{Cliff}}| = \frac{n}{W} \left[\frac{Var\left(\sum_j w_{ij} z_j\right)}{Var\left(z_i\right)} \right]^{\frac{1}{2}} \tag{C.12}$$

Apêndice D

Simulação em um Sistema 14×14

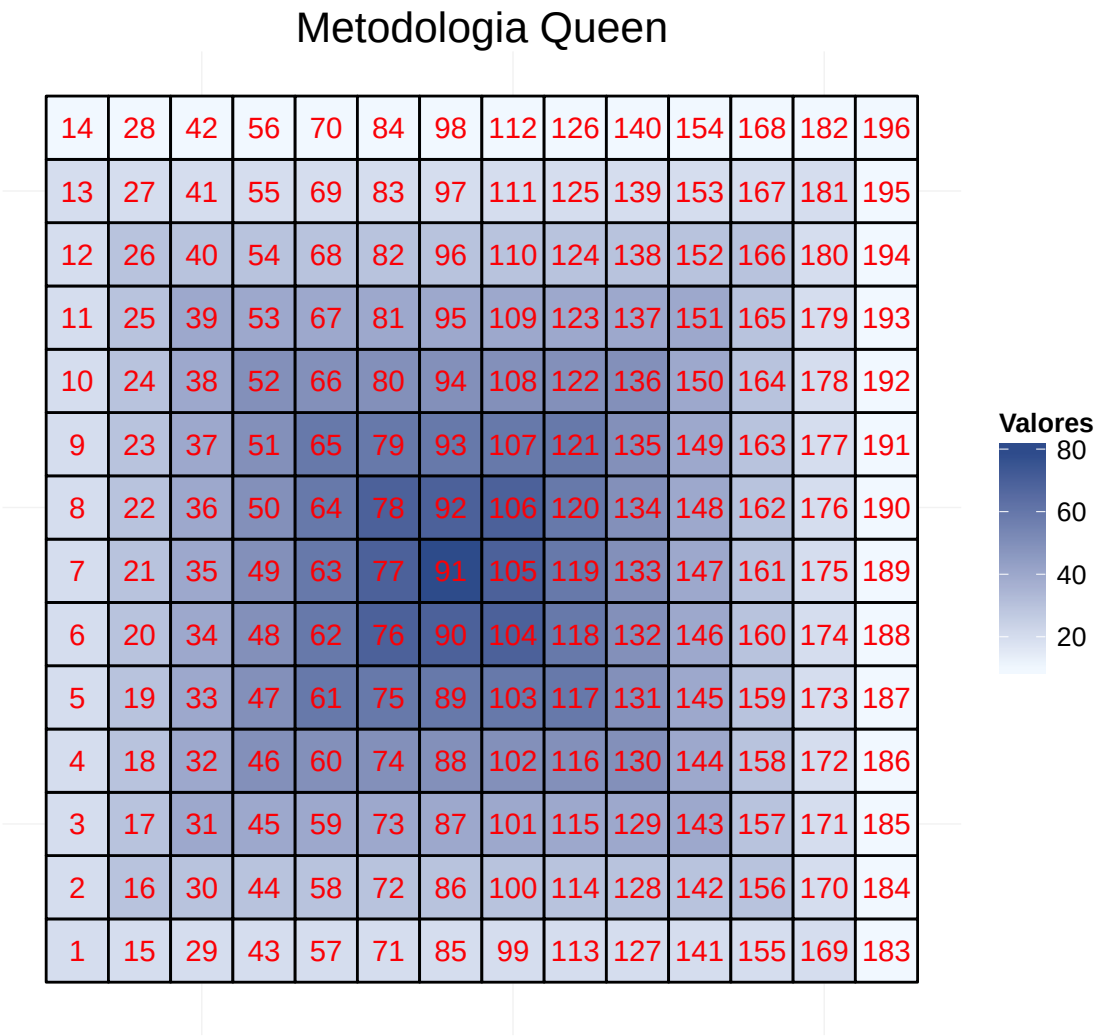

Figura D.1: Sistema 14×14 com pico na região 91

Tabela D.1: Comparação entre o índice de Moran, o índice proposto e a quantidade de vizinhanças em um sistema 14×14

Pico	*Queen*			*Rook*		
	Curvas de Nível	Índice de Moran	Índice Proposto	Curvas de Nível	Índice de Moran	Índice Proposto
1	14	0.9547	0.9485	27	0.9796	0.9741
2	14	0.9566	0.9475	26	0.9810	0.9799
3	14	0.9580	0.9454	25	0.9791	0.9784
4	14	0.9591	0.9441	24	0.9766	0.9788
5	14	0.9600	0.9433	23	0.9738	0.9794
6	14	0.9607	0.9428	22	0.9714	0.9803
7	14	0.9610	0.9426	21	0.9700	0.9820
16	13	0.9576	0.9545	25	0.9825	0.9859
17	13	0.9574	0.9526	24	0.9805	0.9847
18	13	0.9569	0.9502	23	0.9780	0.9856
19	13	0.9564	0.9487	22	0.9752	0.9866
20	13	0.9561	0.9476	21	0.9727	0.9879
21	13	0.9559	0.9470	20	0.9712	0.9895
30	13	0.9574	0.9526	24	0.9805	0.9847
31	12	0.9560	0.9594	23	0.9781	0.9833
32	12	0.9533	0.9573	22	0.9750	0.9840
33	12	0.9506	0.9546	21	0.9715	0.9849
34	12	0.9484	0.9528	20	0.9682	0.9861
35	12	0.9472	0.9518	19	0.9662	0.9878
44	13	0.9569	0.9502	23	0.9780	0.9856
45	12	0.9533	0.9573	22	0.9750	0.9840
46	11	0.9484	0.9643	21	0.9710	0.9848
47	11	0.9420	0.9619	20	0.9663	0.9857
48	11	0.9363	0.9590	19	0.9618	0.9869
49	11	0.9330	0.9575	18	0.9589	0.9887
58	13	0.9564	0.9487	22	0.9752	0.9866
59	12	0.9506	0.9546	21	0.9715	0.9849
60	11	0.9420	0.9619	20	0.9663	0.9857
61	10	0.9312	0.9698	19	0.9600	0.9866
62	10	0.9189	0.9676	18	0.9537	0.9879
63	10	0.9109	0.9654	17	0.9495	0.9898
72	13	0.9561	0.9476	21	0.9727	0.9879
73	12	0.9484	0.9528	20	0.9682	0.9861
74	11	0.9363	0.9590	19	0.9618	0.9869
75	10	0.9189	0.9676	18	0.9537	0.9879
76	9	0.8984	0.9777	17	0.9451	0.9893
86	13	0.9559	0.9470	20	0.9712	0.9895
87	12	0.9472	0.9518	19	0.9662	0.9878
88	11	0.9330	0.9575	18	0.9589	0.9887
89	10	0.9109	0.9654	17	0.9495	0.9898
90	9	0.8809	0.9770	16	0.9392	0.9913
91	8	0.8552	0.9898	15	0.9320	0.9932