CLÁUDIO FRANCISCO CHIPENETE

PREDIÇÃO DE DADOS DE ÁREA REFERENTES AO USO DE SEMENTES MELHORADAS DE MILHO EM MOÇAMBIQUE

LAVRAS – MG
2022

CLÁUDIO FRANCISCO CHIPENETE

PREDIÇÃO DE DADOS DE ÁREA REFERENTES AO USO DE SEMENTES MELHORADAS DE MILHO EM MOÇAMBIQUE

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

Prof. Dr. Renato Ribeiro de Lima
Orientador

Prof. Dr. Marcelo Silva de Oliveira
Coorientador

LAVRAS – MG
2022

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

Chipenete, Cláudio Francisco.
Predição de dados de área referentes ao uso de sementes melhoradas de milho em Moçambique / Cláudio Francisco Chipenete. - 2022.
57 p.

Orientador(a): Renato Ribeiro de Lima.
Coorientador(a): Marcelo Silva de Oliveira.
Tese (doutorado) - Universidade Federal de Lavras, 2022.
Bibliografia.

1. Modelos autorregressivos espaciais. 2. BLUP. 3. Matriz de ponderação espacial. I. de Lima, Renato Ribeiro. II. de Oliveira, Marcelo Silva. III. Título.

CLÁUDIO FRANCISCO CHIPENETE

PREDIÇÃO DE DADOS DE ÁREA REFERENTES AO USO DE SEMENTES MELHORADAS DE MILHO EM MOÇAMBIQUE

PREDICTION IN LATTICE DATA APPLIED TO THE UTILIZATION OF IMPROVED CORN SEEDS IN MOZAMBIQUE

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

APROVADA em 31 de março de 2022.

Dr. João Domingos Scalon	UFLA
Dr. Paulo Henrique Sales Guimarães	UFLA
Dr. Denismar Alves Nogueira	UNIFAL-MG
Dr. Elias Silva de Medeiros	UFGD

Prof. Dr. Renato Ribeiro de Lima
Orientador

Prof. Dr. Marcelo Silva de Oliveira
Coorientador

LAVRAS – MG
2022

À minha esposa Gisela;
aos meus queridos meninos:Elington, Ludwin e Lindsley.
DEDICO!

AGRADECIMENTOS

À minha família, Gisela, Elington, Ludwin e Lindsley por todo apoio, compreensão e abraços, naqueles momentos que tanto precisava, por terem vivenciado cada pedaço de todo tempo dispendido neste trabalho!

Aos meus pais Issaia e Benigna Chipenete. Aos meus irmãos Liliana, Hortêncio, Loudovino, Dorcelina, Rodane, pelo apoio! A mana Maria de Lourdes Chipenete (in memoriam).

À Universidade Federal de Lavras, em particular, ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, meu muito obrigado pela oportunidade concedida em tornar possível esta realização.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

À Universidade Eduardo Mondlane (UEM) pelo auxilio financeiro especialmente no período da pandemia (COVID-19). Ao Gabinete de Planificação, Qualidade e Estudos Institucionais (GaPQEI), por autorizarem a minha formação! Ao Ministério da Ciência, Tenologia e Ensino Superior de Moçambique (MCTES).

Ao meu orientador, Prof. Dr. Renato Ribeiro de Lima. Mais do que um orientador, foi um amigo para todas as ocasiões ao estilo de *Prov. 17:17* (Tradução do Novo Mundo da Bíblia Sagrada)! Agradeço pela confiança, compreensão, ensino, dedicação, disponibilidade, enfim... por ser humano e entendedor dos problemas, desafios, mesmo em períodos de extrema tensão! Obrigado do fundo do coração!

Ao meu co-orientador, Prof. Dr. Marcelo Silva de Oliveira, pelo acompanhamento e excelentes sugestões!

A todos os professores do Departamento de Estatística (DES), pelos conhecimentos transmitidos e constante disponibilidade no auxílio em todas as dificuldades. Aos demais funcionários do Departamento.

Aos meus colegas da turma: Carlos, Rafael, Cristian, Eleanderson, Ernandes, Luciano, Rodnei, Rodrigo, Vânia e Patrícia por toda amizade durante esses anos.

A todos os demais que, de alguma forma, foram parte da edificação desta tese.

Por fim, ao criador de todas as coisas, JEOVÁ (Salmos 83:18)!

"Aqueles que não se lembram do passado,
estão condenados a repeti-lo".
(G. Santayana)

RESUMO

Em análise de dados espaciais de área, tem-se entre outras finalidades, avaliar a presença ou não de agrupamentos em um determinado fenômeno, ajustar um modelo de regressão linear aos dados de modo a estimar seus parâmetros, e predição de valores não observados em algumas áreas. Contudo, um desafio imposto está relacionado com a especificação da matriz de ponderação espacial **W**, um componente sempre presente, independente de qual seja a finalidade do estudo. Essa matriz estabelece a relação de vizinhança entre um par de áreas ou conjunto delas. No entanto, uma vez que existe diferentes critérios na sua especificação, uma escolha não assertiva pode afetar os resultados finais. Por exemplo, ao predizer o valor de uma observação em certa área não amostrada, a partir dos preditores derivados do modelo autorregressivo de defasagem espacial (SAR), os parâmetros utilizados em tais preditores, podem ter sido superestimados ou subestimados, afetando a qualidade do preditor. Portanto, um dos objetivos nesta tese, é avaliar o efeito da matriz de ponderação espacial na qualidade de predição. Como forma de contornar o problema na especificação da matriz **W** devido aos vários critérios, uma alternativa é o uso de preditores de krigagem. Alguns estudos demonstraram que, ao se utilizar o preditor de krigagem, em que a matriz de autocorrelação espacial é ajustada baseando-se no modelo autorregressivo do erro espacial (SEM), foi possível obter melhores resultados em termos de predição. Portanto, o principal objetivo neste trabalho é avaliar a eficiência dos preditores, considerando o modelo SAR e a krigagem, na predição de valores ausentes referente ao uso de sementes melhoradas de milho pelos agricultores em Moçambique em 2012. Além disso, serão consideradas duas matrizes de ponderação espacial, especificadas por dois critérios distintos na avaliação do efeito delas sobre os preditores. Para o alcance de tal objetivo, são apresentados os resultados de predição pelos preditores do modelo SAR e de krigagem, a partir dos dados obtidos por simulação, considerando diferentes parâmetros de autocorrelação espacial (ρ). A eficiência de cada um deles, foi avaliada, comparando os preditores entre si através da raiz do erro quadrático médio (REQM) e eficiência relativa (ER). Os principais resultados são de que o preditor proposto de krigagem foi o que apresentou maior eficiência; a especificação da matriz **W** tem efeito sobre a eficiência dos preditores.

Palavras-chave: Modelos autorregressivos espaciais. BLUP. Matriz de ponderação espacial. Simulação.

ABSTRACT

In spatial statistical analysis for area or lattice data, some of the purposes are to evaluate the presence or not of clusters in the study of a given phenomenon, to adjust a linear regression model to the data and to predict unobserved values in some evaluated areas. However, one of the challenges is to define correctly the neighbourhood specifications, which is represented in an spatial weighting matrix, generally called **W** matrix. The **W** matrix is a component that always is present in this kind of analysis. However, since there are different criteria in its specification, a non-assertive choice may affect the final results. For example, a wrong choice of this matrix can lead to overestimate or underestimate the parameters of the spatial lag autoregressive model (SAR), which can compromise the prediction. An alternative to solve this prediction problem is to use the kriging predictors, which is used in Geostatistical analysis. Some studies have shown that by using the kriging predictor, it was obtained better results in the prediction. The objectives of this thesis were to evaluate the efficiency to predict missing values by considering the kriging and the SAR models with different spatial weighting matrix. In this study it was used simulated data and maize production data from farmers which used improved maize seeds in Mozambique in 2012. The efficiency of the different prediction methods was evaluated by using the root mean square error (RMSE) and relative efficiency (RE). The results showed us that the proposed kriging predictor presented highest efficiency and the different **W** matrices had an effect on the predictors efficiency.

Keywords: Spatial Autoregressive Models. BLUP. Spatial weighing matrix. Simulation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A estatística espacial tem sido utilizada na compreensão de fenômenos em que as observações são georreferenciadas no espaço. Tais fenômenos requerem atenção na sua análise, uma vez que, aplicando as técnicas conhecidas da estatística clássica, sem levar em conta a estrutura de variabilidade espacial, pode-se incorrer ao erro, por exemplo, de especificação do modelo de regressão linear, reduzindo sua eficiência, afetando a qualidade de estimação dos seus parâmetros ou de predição. Diante disso, métodos de estatística espacial tem sido desenvolvidos, conforme o tipo de dados a serem analisados, dividindo-se em padrões de pontos, de superfície contínua no contexto de geoestatística e dados de área, estes dois últimos, considerados neste trabalho.

Em relação a dados de área, um problema que tem sido objeto de estudos está relacionado com a especificação da matriz de ponderação espacial $\mathbf{W}$, um componente presente em todas as estruturas envolvidas, quer sejam nos modelos de regressão linear utilizados no processo de estimação dos parâmeros envolvidos ou na predição de observações ausentes ou não amostradas, ou ainda, nos indicadores que tornam possível analisar os agrupamentos ou *clusters* presentes, por possuir diferentes critérios para ser especificada. Citam-se estudos realizados por Getis e Aldstadt (2004), Chen (2012), Timmins et al. (2013) em que procuraram ajustar modelos de regressão espacial sob diferentes especificações da matriz de proximidade, comparando sua eficiência.

Dubin (1998), Getis e Aldstadt (2004) ajustaram a matriz de proximidade espacial diretamente dos dados, a partir do modelo do semivariograma, como o esférico, exponencial e Gaussiano e, em seguida, utilizaram no modelo autorregressivo de defasagem espacial (SAR). O resultado foi que, por essa via, o modelo apresentou melhor qualidade no ajuste aos dados quando comparado com outras matrizes especificadas por demais critérios.

Contudo, esses resultados não tiram a importância dos demais critérios presentes em dados de área. Como observa Getis (2010), a escolha de um determinado critério, depende das características da região e do fenômeno em estudo. Na maioria das vezes, a escolha do critério a ser utilizado, muito dependerá da experiência do pesquisador.

Por último, em dados de área, há interesse em se conhecer observações ausentes, perdidos ou não amostrados. Nesse caso, existe a possibilidade de se recorrer aos interpoladores

de modo a predizer tais observações. Um problema que surge, é que esses interpoladores são determinísticos, e não oferecem o grau de certeza ou erro da estimativa.

Diante da existência de diferentes critérios para especificação da matriz **W**, o uso de krigagem tem sido visto como uma ótima alternativa na predição, embora se saliente que os preditores do modelo SAR são tidos como melhores ou tão bons quanto os de krigagem, desde que o critério escolhido para especificação da matriz **W** seja assertiva. Por exemplo, Mojiri et al. (2018) compararam as predições utilizando os métodos da krigagem e de dados de área. Os resultados do estudo de simulação mostram que as previsões feitas pelos preditores do modelo SAR são bons concorrentes para o método de krigagem. Griffith (2017), Kelejian e Prucha (2007) compararam o preditor de krigagem e os de dados de área derivado dodo modelo autorregressivo do erro espacial (SEM), com diferentes graus de dependência espacial dado por índice de autocorrelação espacial (ρ). Esses autores, embora observassem a superioridade do preditor de krigagem em termos de eficiência em relação aos preditores do modelo SEM e SAR, a diferença numérica entre sí, foi mínima.

Neste trabalho foi definida a matriz **W** utilizando os critérios existentes em dados de área e, em seguida, utilizada no preditor de krigagem. Além disso, há interesse em se avaliar a eficiência dos preditores derivados do modelo SAR e de krigagem por simulação, considerando duas matrizes de ponderação espacial especificadas por critérios distintos. Para ilustrar a utilização dos diferentes métodos de predição, estes serão aplicados a dados reais referente ao uso de sementes melhoradas de milho em Moçambique no ano de 2012 pelos agricultores, com a finalidade de se predizer observações em algumas áreas. Essa predição é importante uma vez que conhecer o valor dessas observações ausentes, especialmente em culturas de sementes melhoradas de milho, é de extrema importância para avaliar e estimular o uso pelos agricultores, de modo a incrementar a produção e beneficiar as comunidades existentes nessas áreas.

Portanto, neste trabalho, tem se por objetivos:

a) avaliar por meio simulação, a eficiência dos preditores de krigagem e os derivados do modelo SAR, considerando diferentes valores do parâmetro de autocorrelação espacial (ρ) e duas matrizes de ponderação espacial;
b) propôr alternativas na predição em dados de área;
c) aplicar os resultados obtidos em dados reais.

2 REFERENCIAL TEÓRICO

A proposição e o desenvolvimento da estatística espacial foi uma grande contribuição na análise de dados, pois foi incorporada a localização espacial dos dados nessas análises, complementando o que era feito até então na Estatística tradicional ou clássica. Dentre várias características de interesse, destacam-se a distribuição das observações de um determinado fenômeno de interesse $Y(\mathbf{s})$ no espaço $\mathbf{s} = [s_1, \ldots, s_n]' \in D$, onde $\mathbf{s}$ representa o vector de coordenadas geográficas, distribuídas ao longo de um campo aleatório de domínio $D \subset \mathbb{R}^d$, sendo $\mathbb{R}^d$ o espaço indexador das coordenadas, com $d \geq 1$ representando a dimensão. No presente estudo, considera-se $d = 2$, que corresponde a uma superfície plana bi-dimensional $\mathbb{R}^2$, onde cada elemento do vector $\mathbf{s}$ é representado por respectivas coordenadas cartesianas $s_i = (x, y)$. Uma variável regionalizada $Y(\mathbf{s})$ é definida pela seguinte expressão matemática:

$$\left\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2\right\}. \tag{2.1}$$

As observações da variável $Y(\mathbf{s})$ são também denominados por dados espaciais, e se distinguem do que normalmente é considerado nas análises estatísticas clássicas, por apresentarem duas características principais: a primeira, é a representação de cada observação em termos de sua localização na região de estudo; a segunda é que essas observações estão correlacionadas entre si.

Geralmente, a estatística espacial é dividida em três áreas principais, conforme o tipo de dados a serem analisados: (i) padrão de pontos ou processos pontuais; (ii) dados de área (ou *lattice*); (iii) dados de superfície contínua, no contexto da geoestatística. Embora todos eles tenham em comum a regionalidade, diferem na forma em que as observações se apresentam no seu domínio e nos métodos de análise dos dados. Neste trabalho, aborda-se os tipos de dados mencionados em (ii) e (iii), com a finalidade de se ajustar modelos de predição. Portanto, serão apresentados alguns conceitos sobre estes dois tipos de dados, que servirão de sustentação para o alcance dos objetivos propostos.

2.1 Dados de área

Seja um conjunto $A = \{A_i : i = 1, ..., n\}$ em seu domínio D, sendo A_i a *i*-ésima área ou unidade espacial, de forma regular ou não, é fixa e pertence a um domínio $D = \bigcup_{i=1}^{n} A_i$.

Cada uma das *i*-ésimas áreas é representada por seu centroide $c_i \in A_i$ que, além de concentrar toda a informação da variável aleatória Y, representa sua posição geográfica. Seja $y(A_i)$ uma realização da variável Y na *i*-ésima área, esta pode ser em forma do valor médio, taxa, contagem, proporções entre outras. A expressão geral que corresponde a dados de área é, geralmente, definida como um processo estocástico, ou seja, $\{y(A_i) : A_i \in D \subset \mathbb{R}^2\}$.

2.2 Análise de dados de área

Em estudos envolvendo dados de área ou *lattice*, na maioria das vezes, têm-se como objetivos principais explorar a distribuição espacial da variável de interesse e ajustar modelos de regressão linear espacial para predizer valores não observadas em determinadas áreas.

Em relação a análise de distribuição espacial da variável de interesse, busca-se igualmente, identificar, caso existam, os agrupamentos formados, utilizando indicadores como o índice de Moran (I), de Geary (C), de Getis e Ord (G, G^*) entre outros. Não serão detalhados todos esses índicadores, apenas o índice de Moran. Para mais informações pode-se consultar Anselin e Bera (1998), Getis (1995), Cressie (1993).

A título ilustrativo, considera-se o índice de Moran global (I) um valor escalar que geralmente varia de -1 a 1, e expressa o grau de associação linear entre o vetor **y** e a média das observações das áreas vizinhas ponderadas pelo seu peso, com a seguinte classificação, conforme Pace e Lesage (2016): 0 a 0,25 insignificante; 0,25 a 0,7 fraca; 0,7 a 0,9 moderado; 0,9 a 1 forte. É dado por

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

em que

- n representa o número total de áreas no domínio D;
- y_i é o valor da variável aleatória na área *i*;
- y_j é o valor da variável aleatória na área *j*;
- $\bar{y}$ é o valor da média amostral da variável aleatória em toda região, ou seja, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$;
- w_{ij} são os elementos da matriz de proximidade espacial **W** que, em geral, tem sido normalizada em linha, de modo que, a soma dos elementos nela seja igual a 1.

Outra importância em dados de área é o ajuste de modelos de regressão espacial aos dados para posterior predição, um dos propósitos nesta tese, e por isso, considerado com maiores detalhes na seção 2.5. Em relação à predição, essencialmente tem se por objetivo conhecer o valor de uma ou mais observações não amostradas, com base nos valores da sua vizinhança.

Um exemplo disso é quando se pretende avaliar a produção média anual de uma determinada área, ao utilizar sementes melhoradas por produtores locais em Moçambique. É possível que não se tenha informação em certos distritos (áreas) em um determinado ano. Neste caso, há necessidade de se predizer esses valores, o que é objeto neste estudo.

Um problema que surge, é que os modelos espaciais de dados de área, incorporam uma matriz de ponderação ou vizinhança espacial **W**, que na maioria dos casos, assume-se a priori, uma estrutura de dependência que pode corresponder de perto a realidade ou não, de tal vizinhança. Além disso, deve-se levar em consideração o fato de que a especificação depende de vários critérios estabelecidos na literatura especializada, conforme observam Bhattacharjee e Jensen-Butler (2013), Getis (2010), Getis e Aldstadt (2004), Anselin e Bera (1998).

Citam-se exemplos de Getis e Aldstadt (2004), que avaliaram 12 critérios na especificação de **W**. Como resultado, os autores concluíram que o modelo de autocorrelação espacial era sensível à forma de se especificar a matriz **W**. Conforme observa Getis (2010), a especificação da matriz **W** é exógena ao processo e depende de vários critérios existentes na literatura especializada, o que de certo modo, pode influenciar na qualidade do modelo onde esteja inserida. A seguir serão apresentados detalhes em relação à matriz **W** e alguns critérios utilizados na sua definição.

2.3 Matriz de ponderação espacial

A matriz de proximidade ou vizinhança espacial **W**, de ordem $n \times n$, também representada por $\mathbf{W}^t$, em que t representa a ordem da vizinhança, estabelece a relação de proximidade entre um par de áreas nas *i*-ésimas posições, representada por

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix}^t , \tag{2.2}$$

sendo w_{ij} os elementos da matriz que correspondem a cada par de áreas nas *i,j*-ésimas posições; *t* representa a ordem da matriz, ou seja, se da primeira ordem ($t = 1$) quando uma área é vizinha imediata da outra, segunda ordem ($t = 2$) quando uma área é vizinha do vizinho imediato, e assim sucessivamente. Neste trabalho, considera-se a matriz **W** quando $t = 1$.

Conforme Anselin e Bera (1998), algumas características da matriz **W** são: (i) os seus elementos são sempre maiores ou iguais a zero, $w_{ij} \geqslant 0$; (ii) apresenta uma simetria em relação a diagonal, ou seja, $w_{ij} = w_{ji}$; (iii) os elementos na diagonal são iguais a zero, isto é, diag$(w_{ii}) = 0$. Ainda segundo esses autores, há necessidade de se normalizar a matriz de proximidade espacial, de modo que a soma dos elementos na linha seja 1 ($w_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$). Conforme afirmam, uma vantagem disso é que facilita os cálculos, possibilita comparações, assegura a compatibilidade entre os modelos e facilita a interpretação dos pesos, como uma média ponderada dos valores vizinhos.

2.4 Especificação da matriz W

Nesta seção, apresentam-se apenas dois critérios utilizados na especificação da matriz **W**, conforme Cressie (1993), Anselin e Bera (1998), Getis e Aldstadt (2004), Chen (2012): o de fronteira comum e o da distância entre os centroides.

Pelo critério de compartilhamento de fronteira, a vizinhança entre pares de áreas é definida pela existência de fronteira comum. Um par de áreas são consideradas vizinhas se estas apresentam fronteira comum e, assim, tem-se que w_{ij}=1; caso contrario $w_{ij} = 0$. O resultado é uma matriz binária composta de zeros $(0s)$ e uns $(1's)$. Assim, tem-se

$$w_{ij} = \begin{cases} 1, & \text{se compartilharem fronteira comum,} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.3}$$

Pelo critério baseado na distância entre par de centroides, leva-se em consideração a distância entre os centroides das áreas envolvidas. Essa distância tem sido, em geral, a euclidiana dada por $d_{ij} = ||c_i - c_j||$, sendo respectivamente c_i, c_j os centroides das áreas A_i e A_j. Ao critério é associado uma função $f(\cdot)$, isto é, $w_{ij} = f(d_{ij})$. Alguns dos critérios são descritos a seguir.

a) Critério de máxima distância: considera que duas áreas são vizinhas, se elas estiverem a uma distância menor que o raio de ação ou de corte δ, previamente estabelecido, ou seja,

$$w_{ij} = \begin{cases} 1, & \text{se } 0 < d_{ij} \leq \delta, \\ 0, & \text{se } d_{ij} > \delta. \end{cases} \tag{2.4}$$

b) Critério da distância inversa: reflete a 1ª lei da Geografia. Quanto mais próxima a distância entre duas observações, mais parecidas serão, e por consequência, o peso entre elas será maior. É apresentado na equação (2.5), em que α é um parâmetro de amortecimento da influência da distância sobre a força de interação.

$$w_{ij} = \frac{1}{d_{ij}^{\alpha}}. \tag{2.5}$$

c) Critério da distância exponencial negativa: é uma alternativa ao critério da distância inversa, definido em (2.5), por se comportar melhor em distâncias menores. Está definido na equação (2.6). As funções de distância inversa e exponencial negativa, são frequentemente combinadas com um critério de corte de distância δ, de modo que $w_{ij} = 0$ para $d_{ij} > \delta$. Na prática, os parâmetros raramente são estimados, mas definidos por um valor fixo, como $\alpha = 1$ em (2.5) ou $k = 2$ em (2.6).

$$w_{ij} = \frac{1}{e^{k\, d_{ij}}}, \text{ sendo o parâmetro } k > 0. \tag{2.6}$$

d) Critério de distância envolvendo k vizinhos mais próximos: dado um par de áreas, estas são vizinhas caso se encontrem dentro de uma distância de corte, dado em (2.7), sendo $d_i(k)$ a distância de corte para i, a fim de que se tenha k vizinhos.

$$w_{ij}(k) = \begin{cases} 1, \text{ se } & d_{ij} \leq d_i(k) \\ 0, \text{ se } & d_{ij} > d_i(k) \end{cases}, \tag{2.7}$$

e) Critério dupla potência: é uma função mais flexível, uma vez que incorpora um raio de influência com comprimento máximo $d \leq d_{ij}$ e por possuir funções em forma de sino.

Para cada número inteiro positivo do parâmetro k, define-se a classe de pesos por

$$w_{ij} = \begin{cases} [1 - (\frac{d_{ij}}{d})^k]^k & , \quad 0 \leq d_{ij} \leq d \\ 0 & , \quad d_{ij} \geq d \end{cases} \tag{2.8}$$

Como se observa, os pesos baseados na distância dependem não apenas do valor do parâmetro e da forma funcional, mas também da métrica usada para a distância. Como os pesos são inversamente relacionados à distância, valores grandes para o último resultarão em valores pequenos para o primeiro e vice-versa. Isso pode ser um problema na prática quando as distâncias são excessivamente grandes, ou seja, que os pesos de distância inversa correspondentes se tornam próximos de zero, resultando possivelmente numa matriz de pesos espaciais zero.

2.5 Ajustes de modelos em dados de área

Os modelos espaciais descritos nesta seção são os que incorporam na sua estrutura, um componente de dependência espacial. Serão considerados dois modelos autorregressivos: de defasagem e do erro espacial. No entanto, será aqui apresentado um terceiro modelo que resulta de ausência de autocorrelação espacial em cada um dos dois primeiros citados, ou seja, quando os parâmetros de autocorrelação espacial são iguais a zero.

Um modelo autorregressivo espacial, conhecido como modelo SARAR(1,1) (Spatial Autoregressive Model with Autocorrelation Error) (KELEJIAN; PRUCHA, 2007), é definido como

$$\mathbf{y} = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{y} + \boldsymbol{\nu}, \text{ sendo } \nu = \lambda\mathbf{W}\nu + \boldsymbol{\varepsilon}, \tag{2.9}$$

em que

- $\mathbf{y}$ é um vetor de ordem $nx1$ da variável dependente que contém n observações;
- $\mathbf{X}$ é uma matriz não estocástica de ordem $n \times (p+1)$ contendo n observações da p variáveis preditoras mais a primeira coluna composta por vetor de uns (**1**'s);
- $\mathbf{W}$ é uma matriz de ponderação (vizinhança) não estocástica de ordem $n \times n$;
- $\mathbf{Wy}$ é um vetor de ordem $n \times 1$ da defasagem espacial para a variável $\mathbf{y}$;
- $\boldsymbol{\beta}$ é um vetor de parâmetros de ordem $(p+1) \times 1$;
- $\boldsymbol{\nu}$ um vetor de ordem $n \times 1$ do erro espacialmente dependente;
- $\rho,\ \lambda$ são os coeficientes de autocorrelação espacial;

$\boldsymbol{\varepsilon}$ é o vetor de ordem $n \times 1$ da variável do erro aleatório normalmente distribuído, com média zero e variância constante $\varepsilon \sim N(\mathbf{0},\ \sigma^2\mathbf{I})$.

O modelo (2.9) pode igualmente ser escrito em sua forma reduzida por

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\nu},$$
$$\text{com } \boldsymbol{\nu} = (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \tag{2.10}$$

ou de forma equivalente,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \rho \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np}, \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} +$$

$$+ \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \lambda \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \right)^{-1} \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_n \end{bmatrix},$$

sendo,

$$\begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_n \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \lambda \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \right)^{-1} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Para que o modelo seja viável, algumas condições devem ser satisfeitas, nomeadamente: a) os elementos na diagonal da matriz $\mathbf{W}$ são iguais a zero; b) a matriz $(\mathbf{I} - \alpha\mathbf{W}),\ \alpha = \{\rho,\ \lambda\},\ |\alpha| < 1$, deve ser não singular; c) o vetor do erro aleatório é normalmente distribuído com média zero e variância constante $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Conforme Anselin e Bera (1998), a expressão que justifica a restrição nos coeficientes de autocorrelação espacial no intervalo entre -1 e 1 em b) é $(1/\omega_{\text{max}}) < \alpha < 1/\omega_{\text{min}}$, sendo $\omega_{\text{max}},\ \omega_{\text{min}}$ os autovalores máximos e mínimos respectivamente, da matriz $\mathbf{W}$. Com a normalização em linha da matriz $\mathbf{W}$, tem-se a seguinte restrição: $(\frac{1}{\omega_{\text{max}}}) < \alpha < 1$.

A partir do modelo (2.9), derivam-se os seguintes três modelos parciais, apresentados conforme Fischer e Wang (2011), Kelejian e Prucha (2007), Pace e Lesage (2004) e Anselin e Bera (1998): autorregressivo de defasagem espacial SAR (neste trabalho mantém-se as siglas em inglês nos quais são conhecidos); do erro espacial SEM; de regressão linear (MRL), descritos a seguir.

a) Quando $\lambda = 0$ em (2.9), dá origem ao modelo autorregressivo de espacial SAR, dado nas suas duas formas por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \varepsilon,\ \ \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2.11}$$

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon. \tag{2.12}$$

Conforme Bivand (2002), tem-se a seguinte terminologia para os três termos a direita (2.11): tendência, sinal e ruído. Já, os termos a direita na expressão (2.12), tem-se a combinação de dois termos: o sinal e tendencia $(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$, e sinal e ruido $(\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon$.

O modelo SAR, considera uma relação autorregressiva entre as observações da variável Y, porém, em posições distintas, um processo denominado por defasagem espacial, nome pelo qual também é conhecido. Considerando que cada uma das *i*-ésimas áreas da matriz $\mathbf{W}$ são fixas e não estocásticas, bem como as observações na matriz $\mathbf{X}$, então, tem-se que

$$E\left[(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}\right] = (\mathbf{I} - \rho\mathbf{W})^{-1}E\left[\boldsymbol{\varepsilon}\right] = 0.$$

Diante disso, a variável aleatória Y é normalmente distribuída $\mathbf{Y} \sim N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, com a média condicional $\boldsymbol{\mu} = E(\mathbf{y}|\mathbf{X}) = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$, enquanto que, a estrutura da matriz de variância e covariância resultante $\boldsymbol{\Sigma}$ é dada por

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{V}, \tag{2.13}$$

sendo $\mathbf{V}$ a matriz de autocorrelação espacial definida por

$$\mathbf{V} = (\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})^{-1}. \tag{2.14}$$

Neste trabalho, é conveniente utilizar a matriz de precisão espacial como o inverso da matriz de variância covariância $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ (2.15), uma vez que será utilizada mais adiante, quando se tratar do preditor de krigagem.

$$\mathbf{Q} = \frac{1}{\sigma^2}\left((\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})\right) == \frac{1}{\sigma^2}\left[\mathbf{I} - \rho(\mathbf{W}' + \mathbf{W}) + \rho^2\mathbf{W}'\mathbf{W}\right]. \tag{2.15}$$

b) Modelo SEM: quando $\rho = 0$ em (2.9) dá origem ao modelo de erro espacial. Esse modelo, considera que os efeitos espaciais estão presentes na componente dos resíduos ν, que segue uma distribuição normal com média zero e a matriz de variância covariância $\mathbf{\Sigma} = \sigma^2\mathbf{V}$ (2.13), ou seja, $\boldsymbol{\nu} = \lambda\mathbf{W}\nu + \varepsilon = (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}$. É dado por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \tag{2.16}$$

ou de forma equivalente por

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np}, \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \lambda \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \right)^{-1} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

c) Modelo MRL: quando $\rho,\ \lambda = 0$ em (2.9), dá origem ao modelo de regressão linear de Gauss-Markov, definido em (2.17).

O componente do erro aleatório segue $\varepsilon \sim N(\mathbf{0}, \mathbf{\Sigma})$ sendo $\mathbf{\Sigma} = \boldsymbol{\sigma}^2\,\mathbf{V}$. Ao se supôr que existe independência entre os resíduos, então tem-se que $\mathbf{V} = \mathbf{I}$. O valor esperado da variável aleatória $E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\beta$, e a variância $Var(Y) = \sigma^2$. O modelo MRL é dado por,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.17}$$

ou, de forma equivalente por

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np}, \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Descritos os três modelos (SAR, SEM, MLR) a partir do modelo em (2.9), algumas observações de interesse sobre eles, são apresentadas a seguir. A primeira observação é que, conforme menciona Fischer e Wang (2011), o modelo SEM pode ser visto como uma combinação com o modelo de regressão linear clássico MRL, em que o valor esperado de ambos, são iguais. Para grandes amostras, as estimativas pontuais para os parâmetros $\boldsymbol{\beta}$ do modelo SEM e MRL tendem a ser as mesmas.

Ainda segundo esses autores, em pequenas amostras, há maior probabilidade de ganho de eficiência da modelagem espacial nos termos de erro de SEM em relação ao MRL. Entretanto, o mesmo não acontece para o modelo SAR que contém um componente de defasagem espacial **Wy**, fazendo com que o valor esperado seja diferente do modelo de regressão clássico. Diante disso, é essencial que se tenha um bom modelo SAR para pequenas amostras.

A segunda observação é que, a escolha entre qual modelo utilizar para ajustar aos dados, se SAR ou SEM, em geral, é feita mediante testes específicos. Um deles é o teste do multiplicador de Lagrange, em que se compara a qualidade do ajuste aos dados. Não se fará menção desses testes aqui. Mais informações podem ser obtidas consultando, por exemplo Anselin (1988), Baltagi e Yang (2013).

2.6 Estimação dos parâmetros

Os parâmetros no modelo SAR são estimados pelo método de máxima verossimilhança. Este método, consiste em maximizar a função densidade de probabilidade $f = (y|\theta)$, sendo $\theta = (\boldsymbol{\beta}, \sigma^2, \rho)$ os parâmetros a serem estimados, obtendo-se, a partir da amostra, o estimador "mais verossímil" dos parâmetros. A função log da verossimilhança é dada por:

$$l(\boldsymbol{\theta}|y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

A obtenção dos parâmetros consiste em maximizar a função $l\left(\hat{\boldsymbol{\theta}}|y\right) = \text{máx.}_{\mathbb{R}^k} l\left(\boldsymbol{\theta}|y\right))$, derivando em relação aos seus parâmetros, sendo k o número de parâmetros, igualando a zero, e resolvendo o sistema de equações resultantes. Os estimadores de $\boldsymbol{\beta}, \sigma^2$, são respectivamente dados por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{2.18}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.19}$$

Para a estimação do parâmetro de autocorrelação espacial ρ, deve-se explorar a decomposição da matriz $\ln|\mathbf{I} - \rho\mathbf{W}| = \ln\left[\Pi_{i=1}^{n}\left(1 - \rho\omega_i\right)\right]$, sendo ω_i os autovalores da matriz $\mathbf{W}$, ρ o parâmetro de autocorrelação espacial estimado a partir de métodos iterativos.

Já, para o modelo MLR, uma vez que o erro aleatório é normal, independente e identicamente distribuído, ou seja, $\boldsymbol{\varepsilon} \sim \ N(\mathbf{0},\ \mathbf{I}\sigma^2)$, então, $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{I}\boldsymbol{\sigma}^2)$. O estimador de $\boldsymbol{\mu}$ é $\hat{\mu} = \mathbf{X}\hat{\boldsymbol{\beta}}$, sendo $\boldsymbol{\beta}$ estimado por $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ pelo método de mínimos quadrados ordinários.

2.7 Matriz de proximidade espacial particionada

Em modelos utilizados em dados de área como o SAR e SEM, a matriz de ponderação espacial pode ser expressa de forma particionada ($\mathbf{W}^*$), conforme apresentada em Pace e Lesage (2016), Kato (2008), LeSage e Pace (2004). Essa forma de representar a matriz $\mathbf{W}$, torna mais fácil a representação de áreas com e sem observações dentro do domínio $D \subset R^2$. Supondo que o número total de áreas é expressa por $n = n_d + n_f$, sendo que n_d representa o número de áreas que contém observações e n_f é o número de áreas que não contém observações. A matriz particionada é então dada por

$$\mathbf{W}^* = \begin{bmatrix} \mathbf{W}_{dd} & \mathbf{W}_{df} \\ \mathbf{W}_{fd} & \mathbf{W}_{ff} \end{bmatrix}, \tag{2.20}$$

em que,

$\mathbf{W}_{dd}$ é uma sub-matriz de ordem $n_d \times n_d$ onde n_d representa o número das unidades espaciais que contém observações;

$\mathbf{W}_{ff}$ é uma sub-matriz de ordem $n_f \times n_f$ onde n_f representa o número de unidades espaciais que não contém amostra;

$\mathbf{W}_{df}$ é uma sub-matriz de ordem $n_d \times n_f$ e representa o número de unidades espaciais que contém observações com aquelas que não contém;

$\mathbf{W}_{fd}$ é uma sub-matriz de ordem $n_f \times n_d$ e representa o número de unidades espaciais que não contém observações com aquelas que é.

No modelo visto em (2.12) a estrutura da matriz $\mathbf{I}_n - \rho\mathbf{W}$, considerando $\mathbf{W}^*$, é apresentada da seguinte forma:

$$\mathbf{W}^* = \begin{bmatrix} \mathbf{I}_d - \rho\mathbf{W}_{dd} & -\rho\mathbf{W}_{df} \\ -\rho\mathbf{W}_{fd} & \mathbf{I}_f - \rho\mathbf{W} \end{bmatrix}.$$

Como se pode observar, além de poder identificar áreas sem informações, possibilita estimar o valor médio em cada uma das A_i áreas por $\hat{\boldsymbol{\mu}} = (\boldsymbol{I}_d - \hat{\rho}\mathbf{W}_{dd})\mathbf{X}\hat{\boldsymbol{\beta}}$, ou ainda predizer um valor ausente $\hat{\boldsymbol{\mu}} = (\boldsymbol{I}_d - \hat{\rho}\mathbf{W}_{df})\mathbf{X}\hat{\boldsymbol{\beta}}$, uma vez que $\mathbf{W}_{df}$ corresponde a sub-matriz com algumas áreas que contêm observações da variável Y e outras não, e precisam ser preditas. Uma vez que predição é o objetivo neste trabalho, esse assunto será desenvolvido na seção 2.8.

2.8 Predição em dados de áreas

Para se estimar o valor não observado em uma unidade espacial em dados de área, tem sido utilizado o método de interpolação e os preditores lineares, descritos a seguir.

2.8.1 Interpoladores espaciais

Os preditores espaciais, buscam predizer o valor não amostrado $\hat{y}(s_0)$ no ponto s_0, ou seja,

$$\hat{y}(s_0) = \frac{\sum_{j=1}^{n} w_j y(s_j)}{\sum_j w_j}, \tag{2.21}$$

sendo $\hat{y}(s_0)$ o valor a ser predito no ponto s_0; $w_j y(s_0)$ o valor ponderado das observações vizinhas $y(s_j)$ à volta do ponto a ser predito; w_j o peso ponderado entre $y(s_0)$ e $y(s_j)$.

Os pesos ou contribuição w_j são obtidos por meio de interpoladores, dentre eles: triangulação, inverso das distâncias, polinomial local e global, funções de base radial, estimação via *kernel* entre outros. Serão destacados dois deles, o inverso de distância e a função kernel gaussi-

ana. Para mais detalhes sobre interpoladores, sugere-se a consulta de Cressie (1993), Goovaerts (2006).

A distância inversa é um interpolador local, bastante utilizado e considera que, os pesos entre um par de áreas próximas ou vizinhas entre si, se assemelham, ou seja, quanto mais próximas uma da outra, mais se assemelham. É expresso por $w_j = d_{ij}^k$, com $k = 1, 2$. Uma vez obtido o ponderador, faz-se a substituição direta em (2.21).

A função *kernel* gaussiana, é igualmente bastante utilizada por ser mais abrangente e incluir também a dimensão ou tamanho da área. É dado em (2.22), onde τ corresponde ao tamanho do raio de influência. Obtido o ponderador, substitui-se w_j por $k(c_i, \tau)$ na equação (2.21).

$$k(c_i, \tau) = \frac{1}{2\pi\tau} exp\left(-\frac{d_{ij}^2}{2\tau^2}\right), \; d_{ij} \leq \tau. \tag{2.22}$$

Entretanto, esses dois interpoladores são determinísticos e não oferecerem qualquer grau de incerteza na modelagem. Como opção podem ser utilizados métodos que envolvam modelos probabilísticos, tornando possível que tais incertezas sejam modeladas.

2.8.2 Preditores - Modelo SAR

Seja o modelo geral (2.9) reescrito em sua forma expandida, considerando a matriz particionada $\mathbf{W}^*$, por

$$\begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_f \end{bmatrix} = \begin{bmatrix} \mathbf{X}_d \\ \mathbf{X}_f \end{bmatrix} \boldsymbol{\beta} + \rho \begin{bmatrix} \mathbf{W}_{dd} & \mathbf{W}_{df} \\ \mathbf{W}_{fd} & \mathbf{W}_{ff} \end{bmatrix} \begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_f \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_d \\ \boldsymbol{\varepsilon}_f \end{bmatrix}, \tag{2.23}$$

ou ainda, em sua forma parametrizada por

$$\begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_f \end{bmatrix} = \left(\mathbf{I} - \rho \begin{bmatrix} \mathbf{W}_{dd} & \mathbf{W}_{df} \\ \mathbf{W}_{fd} & \mathbf{W}_{ff} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{X}_d \\ \mathbf{X}_f \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_d \\ \boldsymbol{\varepsilon}_f \end{bmatrix}. \tag{2.24}$$

Kelejian e Prucha (2007) consideram alguns preditores construídos a partir dos modelos (2.23) e (2.24), condicionados as informações disponíveis ($\mathbf{X}$, $\mathbf{W}$, $\mathbf{y}$) presentes na amostra $n = n_d + n_f$. Portanto, serão considerados dois cenários.

O primeiro cenário corresponde a uma situação em que se tem nas n_d áreas, informações das variáveis resposta $\mathbf{y}$, das preditoras $\mathbf{X}$ e pretende-se predizer o valor de $\mathbf{y}$ em tais áreas após

ajustar-se o modelo aos dados. Na prática, esse processo equivale a estimar o valor da variável **y** na área A_i. Na realidade, o objetivo é avaliar o poder preditivo do preditor ou acurácia, uma vez que se conhece o verdadeiro valor. Neste caso, a sub matriz que acomoda esse cenário é $\mathbf{W}_{dd}$. Diante disso, são apresentados dois preditores:

$$\hat{\mathbf{y}}_d = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}_{dd}\mathbf{X}_d\hat{\beta}; \tag{2.25}$$

$$\hat{\mathbf{y}}_d = \mathbf{X}_d\hat{\beta} + \hat{\rho}\mathbf{W}_{dd}\mathbf{y}_d. \tag{2.26}$$

O segundo cenário corresponde a uma situação em que para um total de n áreas, observa-se nas n_d áreas as unidades dentro da amostra, a variável dependente y_d e preditora $\mathbf{X}_d$; enquanto que nas n_f áreas, apenas se observa a variável preditora $\mathbf{X}_f$. Portanto, o objetivo é predizer observações ausentes da variável resposta $\mathbf{y}_f$ a partir do conhecimento de $\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f$. Neste caso, a sub matriz que acomoda esse cenário é $\mathbf{W}_{df}$. Diante disso, são apresentados dois preditores:

$$\hat{\mathbf{y}}_f = (\mathbf{I} - \hat{\rho}\mathbf{W}_{df})^{-1}\mathbf{X}_f\hat{\beta}; \tag{2.27}$$

$$\hat{\mathbf{y}}_f = \mathbf{X}_f\hat{\beta} + \hat{\rho}\mathbf{W}_{fd}\mathbf{y}_d. \tag{2.28}$$

2.9 Geoestatística

Conforme Cressie (1993) em Geoestatística, a variável aleatória $Y(\mathbf{s})$ rege-se por um processo estocástico. Suas realizações $y(s_1), \ldots, y(s_n)$ distribuem-se continuamente em todo o domínio $D \subset \mathbb{R}^d$, obtidas por amostragem em número finito de locais, nas coordenadas $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]'$. Este processo é descrito em (2.1), ou seja,

$$\left\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\right\}.$$

2.9.1 Hipóteses de estacionariedade

Ao se analisar dados espacias em geoestatística, há necessidade de que as hipóteses de estacionariedade sejam satisfeitas. Um processo é dito ser estacionário se este ocorrer de forma homogênea, oscilando aleatória e continuamente a volta de um valor médio. Para que o processo seja estacionário, duas hipóteses são consideradas a seguir, conforme Journel e Huijbregts (1978) e Cressie (1993) e Bailey e Gatrell (1995).

Um processo é dito ser estacionário de segunda ordem, se a esperança da variável aleatória existir $E[Y(\mathbf{s})] = \boldsymbol{\mu}$ e não depender da posição **s**; além disso, a função de covariância entre duas observações existe $Cov[y(\mathbf{s}_i), y(\mathbf{s}_j)] = C(s_i - s_j)$, e depende apenas da distância que separa as observações nos pontos s_i e s_j, ou seja,

$$Cov[y(\mathbf{s}_i), y(\mathbf{s}_j)] = E[y(\mathbf{s}_i)y(\mathbf{s}_j)] - \mu^2 = C(d_{ij}).$$

O estimador da função de covariância é dado por

$$\hat{C}(d_{ij}) = \frac{1}{n(d_{ij})} \sum_{i=1}^{n(d_{ij})} \left[(y(\mathbf{s})_i - \mu) (y(\mathbf{s}_j) - \mu \right], \tag{2.29}$$

em que

$\hat{C}(d_{ij})$ é o estimador de $C(d_{ij})$;

$n(d_{ij})$ corresponde o número de pares de valores observados separados por uma distância d_{ij};

$Y(\mathbf{s}_i),\ Y(\mathbf{s}_j)$ são respectivamente as realizações da variável aleatória Y, nas coordenadas s_i e s_j;

μ é a média aritmética inerente aos dados $\mu = \bar{Y} = \frac{\sum_{i=1}^{n} y(s_i)}{n}$.

Entretanto, a hipótese de estacionariedade de 2ª ordem considera existência da variância finita, por vezes não satisfeita, como em casos em que ela é dispersa. Diante disso, adota-se a estacionariedade intrínseca, que relaxa essa restrição.

Um processo é dito ser estacionário intrínseco, quando a esperança da diferença da variável $Y(\mathbf{s})$ nas posições $\mathbf{s}_i$ e $\mathbf{s}_j$; é nula,

$$E[Y(\mathbf{s}_i) - Y(\mathbf{s}_j)] = 0;$$

e a variância da diferença $[Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]$ existe e depende somente da distância entre si, ou seja,

$$\gamma(d_{ij}) = \frac{1}{2}\mathrm{E}\left[(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2\right], \tag{2.30}$$

sendo $\gamma(d_{ij})$ uma medida de dependência espacial denominada semivariância. O estimador de $\gamma(d_{ij})$ é dado por

$$\hat{\gamma}(d_{ij}) = \frac{1}{2n(d_{ij})} \sum_{n(d_{ij})} \left[y(\mathbf{s}_i) - y(\mathbf{s}_j)\right]^2, \tag{2.31}$$

em que

$\hat{\gamma}(d_{ij})$ é o estimador de $\gamma(d_{ij})$;

$n(d_{ij})$ corresponde o número de pares de valores observados separados por uma distância d_{ij};

$y(s_i),\ y(s_j)$ são respectivamente as realizações da variável aleatória Y, nas coordenadas s_i e s_j.

O semivariograma é a ferramenta básica da Geoestatística para caracterizar a dependência espacial dos dados, considerando todas as combinações de pares de pontos medidos no espaço, separados por uma distância d_{ij}. Por sua vez, conhecendo essa característica, é possível obter predições da variável aleatória na região D, por meio de krigagem, um interpolador linear.

Em um processo estacionário, diga-se, de 2ª ordem, há uma relação estabelecida entre a covariança $C(d_{ij})$ e semivariância $\gamma(d_{ij})$. Essa relação é expressa por

$$\gamma(d_{ij}) = C(0) - C(d_{ij}) = \sigma^2 - C(d_{ij}). \tag{2.32}$$

Uma outra medida é a autocorrelação espacial $\rho(d_{ij})$ que descreve a associação entre as variáveis, é dada por

$$\rho(d_{ij}) = \frac{C(d_{ij})}{C(0)} = \frac{C(0) - \gamma(d_{ij})}{C(0)} = 1 - \frac{\gamma(d_{ij})}{\sigma^2}, \tag{2.33}$$

em que $C(0)$ é a covariância quando a distância é zero $d_{ij} = 0$, e corresponde a variância σ^2 em um processo estacionário de segunda ordem; $\rho(d_{ij})$ é a autocorrelação espacial, que é a razão entre covariância $C(d_{ij})$ e a variância σ^2.

2.9.2 Krigagem

Segundo Cressie (1993), a krigagem é uma técnica que permite conhecer o valor de uma observação em um ponto não amostrado, dentro do domínio D. Portanto, krigagem é um

interpolador linear conhecido por ótimo ou BLUP (melhor preditor linear não viesado, traduzido do inglês).

Para que o interpolador de krigagem seja ótimo, num processo estacionário, deve satisfazer as seguintes condições (JOURNEL; HUIJBREGTS, 1978):

a) média não viesada,

$$E\left[Y(s_0) - \hat{Y}(s_0)\right] = 0; \tag{2.34}$$

b) variância mínima (VM),

$$\text{Var}[\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)] = E\left\{[\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)]^2\right\} = \text{VM}. \tag{2.35}$$

São listados a seguir, alguns interpoladores de krigagem:

a) krigagem simples: assume que a função média do processo estocástico é conhecida, portanto, não é estimada a partir das realizações;
b) krigagem ordinária: envolve apenas o variograma, tornando-a a mais utilizada e cuja função média não é conhecida, portanto, estimada a partir dos dados;
c) krigagem universal: quando se desconhece a média e está é uma função de tendência, o que viola a hipótese de estacionariedade. Neste caso, a remoção da tendência é feita no ajuste de polinômios de baixo grau.

Neste trabalho, serão destacados apenas os dois últimos, krigagem ordinária e universal, com mais ênfase, no último. Além disso, a ideia que se pretende trazer desse preditor é descrito por Bailey e Gatrell (1995), ser um preditor espacial generalizado. Daqui em diante, será esse o termo aqui utilizado.

2.9.3 Krigagem ordinária

Krigagem ordinária destina-se a calcular os pesos ou ponderadores ótimos, que minimizem a variância do erro de estimação. Esse preditor é expresso por

$$\hat{y}(s_0) = \sum_{i=1}^{n} \lambda(s_i) y(s_i), \tag{2.36}$$

sendo $\hat{y}(s_0)$ o valor a ser predito no ponto s_0; enquanto que $\lambda(s_i)$ são os pesos associados a cada valor; $y(s_i)$ são realizações vizinhas nos *i*-ésimos pontos.

Os pesos ótimos $\lambda_i = 1, \ldots, n$ são obtidos conforme Journel e Huijbregts (1978) por:

$$\boldsymbol{\lambda} = \mathbf{C}^{-1}\mathbf{c}, \tag{2.37}$$

ou, de forma equivalente, por

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \alpha \end{bmatrix} = \begin{bmatrix} C(s_1,s_1) & C(s_1,s_2) & \ldots & C(s_1,s_n) & 1 \\ C(s_2,s_1) & C(s_2,s_2) & \ldots & C(s_2,s_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(s_n,s_1) & C(s_n,s_2) & \ldots & C(s_n,s_n) & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} c(s_1,s_0) \\ c(s_2,s_0) \\ \vdots \\ c(s_n,s_0) \\ 1 \end{bmatrix}$$

sendo $\mathbf{C}$ uma matriz de ordem $n \times n$ das covariância dos valores amostrados envolvidos na estimativa de $\hat{y}_0$; α o multiplicador de Lagrange; $\mathbf{c}$ um vetor de ordem $n \times 1$ que contêm as semivariâncias entre o valor amostrado no ponto s_j e aquele a ser estimado no ponto s_0; $\boldsymbol{\lambda}$ é o vetor de ordem $n \times 1$ associados a cada valor $y(\mathbf{s}_i)$. Para satisfazer (2.34) e (2.35), o somatório dos pesos deve ser igual a 1, ou seja $\sum_{i=1} \lambda_i = 1$.

2.9.4 Preditor espacial generalizado

Segundo Bailey e Gatrell (1995), o preditor do modelo espacial generalizado tem sido utilizado quando a média dos valores amostrados é desconhecida e apresenta características não estacionárias do processo. Sendo uma extensão de krigagem ordinária, as estimativas são resultado de uma combinação linear ponderada dos dados amostrados. A tendência apresentada pelos dados é calculada simultaneamente e estimada implicitamente. Uma vez removida a tendência, é possível utilizar a krigagem ordinária na estimação de um valor não observado.

Considere-se o seguinte modelo linear generalizado, dado por

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}). \tag{2.38}$$

sendo $\mu = \mathbf{x}'\boldsymbol{\beta}$ a tendência do processo com o vetor dos atributos $\mathbf{x} = [x_{i0},\ x_1,\ x_2,\ \ldots,\ x_n]^T$, com x_{i0}=1 que representa o intercepto β_0 no modelo; $\boldsymbol{\varepsilon}$ é o erro aleatório espacialmente dependente com média zero e matriz de variância covariância dada por $\Sigma = \sigma^2\mathbf{C}$.

A matriz $\boldsymbol{\Sigma}$ é estimada, considerando a covariância entre as observações dos resíduos $\sigma_{ij} = C\left[\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j)\right] = Cov\left[y(\mathbf{s}_i), y(\mathbf{s}_j)\right] = C(d_{ij})$ num processo estacionário de segunda ordem por

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{1n} & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{C}.$$

Além disso, considerando a relação $\rho_{ij} = \sigma_{ij}/\sigma^2$ vista em (2.32), então, a matriz de covariância é reescrita como

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{1n} & \dots & 1 \end{bmatrix}. \tag{2.39}$$

Diante do exposto e considerando o modelo (2.38) o preditor de krigagem é dado por (BAILEY; GATRELL, 1995)

$$\hat{\mathbf{y}}_0 = \mathbf{x}'\hat{\beta} + \mathbf{c}'\mathbf{C}^{-1} \times (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.40}$$

sendo $\mathbf{x}'$ um vetor de ordem $n \times 1$ de atributos no local a ser predito; $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$ o estimador dos resíduos $\boldsymbol{\varepsilon}$; $\hat{\boldsymbol{\beta}}$ é o estimador de $\boldsymbol{\beta}$ obtido pelo método dos quadrados mínimos generalizados por $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}$; $\mathbf{X}$ é uma matriz de ordem $n \times (p+1)$ contendo p covariáveis preditoras e uma coluna de uns (1's).

Griffith (2017) apresenta uma relação algébrica entre dados de área e técnicas geoestatística na predição de um valor não observado, levando em conta a matriz de precisão espacial $\mathbf{Q}$ ($\mathbf{Q} = \Sigma^{-1} = \mathbf{V}^{-1}\sigma^{-2}$) conforme as equações (2.13) e (2.15) respectivamente. Essa relação se baseia na substituição da matriz $\mathbf{c}'\mathbf{C}^{-1}$ na equação (2.40) por $-\mathbf{Q}_{ff}^{-1}\mathbf{Q}_{fd}$.

Seja considerado a matriz de precisão espacial $\mathbf{Q}$ na sua forma particionada dada por

$$\mathbf{Q} = \frac{1}{\sigma^2}(\mathbf{I} - \rho(\mathbf{W}' + \mathbf{W}) + \rho^2\mathbf{W}'\mathbf{W}) = \begin{bmatrix} \mathbf{Q}_{dd} & \mathbf{Q}_{df} \\ \mathbf{Q}_{fd} & \mathbf{Q}_{ff} \end{bmatrix}; \tag{2.41}$$

e os sub-vetores da variável resposta $\mathbf{y}$ e de covariáveis $\mathbf{X}$ dados por

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_f \end{bmatrix} \; ; \; \mathbf{X} = \begin{bmatrix} \mathbf{X}_d \\ \mathbf{X}_f \end{bmatrix} .$$

Assim, o preditor linear que incorpora ambas as técnicas de dados espaciais de área e krigagem é dado considerando dois cenários. O primeiro é que se tem todas informações disponíveis na amostra, ou seja, $\mathbf{y}, \mathbf{X}_d, \mathbf{W}_{dd}$. Então, o preditor de krigagem é dado por (GRIFFITH, 2017)

$$\hat{\mathbf{y}}_d = \mathbf{X}_d\boldsymbol{\beta} - \text{diag}(\hat{\mathbf{Q}}_{dd})^{-1}\tilde{\hat{\mathbf{Q}}}_{dd} \times \left(\mathbf{y}_d - \hat{\mathbf{X}}_d\hat{\boldsymbol{\beta}}\right), \tag{2.42}$$

sendo o estimador da matriz $\mathbf{Q}_{dd}$ dado por

$$\hat{\mathbf{Q}}_{dd} = \frac{1}{\hat{\sigma}^2}\left[(\mathbf{I} - \hat{\rho}\mathbf{W}_{dd})'(\mathbf{I} - \hat{\rho}\mathbf{W}_{dd})\right] \;\; \text{com} \;\; \tilde{\hat{\mathbf{Q}}}_{dd} = \hat{\mathbf{Q}}_{ss} - \text{diag}(\hat{\mathbf{Q}}_{dd}).$$

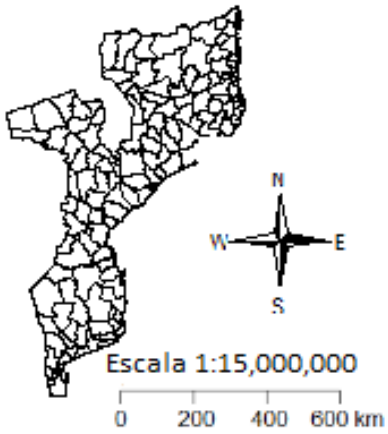
O segundo cenário representa uma situação em que não se conhece o valor de observações da variável resposta $\mathbf{y}$ em algumas áreas. Nesse caso, o conjunto de informações disponíveis na amostra é composto por $\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f, \mathbf{W}_{df}$. O preditor de krigagem, ainda conforme Griffith (2017), é dado por

$$\hat{\mathbf{y}}_f = \mathbf{X}_f\hat{\boldsymbol{\beta}} - \hat{\mathbf{Q}}_{ff}^{-1}\hat{\mathbf{Q}}_{df} \times (\mathbf{y}_d - \mathbf{X}_d\hat{\boldsymbol{\beta}}). \tag{2.43}$$

3 METODOLOGIA

Esta seção destina-se à descrição da metodologia que será utilizada para a execução desse trabalho. Serão utilizados dados reais de 128 distritos de Moçambique (FIGURA 3.1). Serão igualmente simulados dados para cada um desses distritos. Essencialmente, a utilização da simulação tem como finalidade comparar os preditores do modelo SAR existentes na literatura com o proposto nesta tese, que será apresentado mais adiante, de modo a avaliar sua eficiência. A simulação dos dados será executada considerando a disposição das áreas iguais à dos dados reais, ou seja, considerando os 128 distritos de Moçambique.

Figura 3.1 – Mapa de Moçambique dividido em 128 distritos



Fonte: Do autor (2022)

3.1 Processo de simulação

Foram obtidos dados para a variável Y por simulação, à partir do software R (R Core Team, 2021), considerando o seguinte modelo espacial autorregressivo:

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1} (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) . \tag{3.1}$$

que, de forma equivalente, representando por

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \rho \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \right)^{-1} \times$$

$$\times \left(\begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np}, \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \right),$$

em que

- **y** é o vetor da variável dependente de ordem $nx1$ com *n* observações;
- ρ é o parâmetro escalar de autocorrelação espacial;
- **W** é a matriz de proximidade espacial de ordem $n \times n$ e define a relação entre as áreas dentro do conjunto $D \in R^2$;
- **X** é a matriz de variáveis preditoras de ordem $n \times (p+1)$, onde $p < n$ representa o número de parâmetros presentes, sendo a primeira coluna com uns 1's;
- $\boldsymbol{\beta}$ é o vetor dos parâmetros de ordem $(p+1) \times 1$;
- $\boldsymbol{\varepsilon}$ é o vetor dos erros aleatórios, normalmente distribuído com média zero e variância constante $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$.

Os valores da variável aleatória espacial Y foram gerados por simulação, considerando três covariáveis: uma com distribuição normal; outra com a binomial; a última com distribuição uniforme. Atribuiu-se valores iniciais aos parâmetros $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3)$ e considerou-se $\sigma^2 = 1$, sem perda de generalidade.

Na combinação entre o parâmetro de autocorrelação espacial e matriz de ponderação espacial $\rho\mathbf{W}$, atribuiu-se ao ρ nove valores de modo a espelhar diferentes graus de dependência espacial ($\rho = -0,4; -0,2; 0,05; 0,2; 0,30; 0,5; 0,7; 0,8; 0,9$); especificou-se duas matrizes de ponderação espacial com base em dois critérios: a matriz **W**1 pelo critério que considera a quantidade de vizinhos que uma área deve possuir, sendo definido $k = 10$ vizinhos (2.7); a matriz **W**2 pela combinação do critério da distância inversa e máxima distância, fixada em 70 km. A Tabela 3.1 apresenta algumas informações referente às duas matrizes.

Tabela 3.1 – Dados gerais das matrizes de ponderação espacial (W1,W2)

	W1	**W**2
Nº de conexões	1280	16256
% de conexões	7,8	99,2
Média de conexões	10	127

Fonte: Do autor (2022)

De modo a avaliar a eficiência dos preditores, além dos 18 cenários já definidos, que correspondem às combinações de nove valores de ρ e dois tipos de matrizes **W**, ainda foram consideradas duas situações no processo de simulação descritas a seguir.

A primeira situação é que se tem informações da variável resposta **y** bem como das covariáveis envolvidas em todas as A_i áreas, ou seja, $\mathbf{X}_d,\ \mathbf{y}_d$. Esta corresponde à situação em que a matriz W seria composta apenas pela submatriz $\mathbf{W}_{dd}$, conforme partição apresentada em (2.20). Portanto, se está diante de um processo que equivale ajustar um modelo de regressão ao conjunto dos dados, com a finalidade de se estimar o valor médio $\boldsymbol{\mu}$. Na prática, uma vez que se conhece os valores observados em cada uma das n_d áreas, o objetivo é avaliar a acurácia do preditor em predizer uma observação em que já se conhece o valor. Com essa finalidade, foram geradas por simulação 1000 amostras de tamanho $n = 128$ para cada uma das 18 combinações $(\rho, \mathbf{W}, \sigma^2)$, a partir da função *sar* definida no pacote *hsar* (DONG; HARRIS; MIMIS, 2016) disponível no *software* R (R Core Team, 2021).

A segunda situação é aquela na qual não se conhece o valor de **y** em certas áreas. Nesse caso, se está diante de um processo em que se conhece observações da variável resposta **y** em algumas áreas. Já, em relação às observações das covariáveis **X**, estas estão presentes em todas as áreas. Neste caso, a matriz **W** pode ser representada em partições, conforme definido em (2.20) e os valores faltantes (y_f) serão preditos utilizando as equações apresentadas em (2.26) e (2.27), nos quais são consideradas as matrizes de ponderação W_{fd} e W_{df}.

Para a predição de observações faltantes da variável resposta **y** foi realizada uma simulação de 1000 repetições. Adotou-se o seguinte procedimento:

a) selecionou-se aleatoriamente algumas áreas sem observação com a finalidade de se estimar o seu valor. Considerou-se seis (6) tamanhos diferentes de amostras que representam áreas com observações faltantes $n_f = 1, 3, 6, 9, 12, 14$. Considerando as áreas com observações de **y** é ajustado o modelo SAR aos dados com a finalidade de se obter os valores estimados dos parâmetros $\hat{\boldsymbol{\beta}},\ \hat{\sigma},\ \hat{\rho}$ pelo método de máxima verossimilhança;

b) os parâmetros estimados em i) são utilizados nos preditores dados nas equações (2.27) e (2.28). Para cada 1000 amostras, utilizou-se por uma questão de economia apenas três (3) valores do parâmetro de autocorrelação espacial ρ ($\rho = -0,4; 0,02; 0,70$) de um total de nove, já apresentados anteriormente.

3.2 Preditores do modelo SAR e de krigagem

Neste trabalho serão utilizados os preditores do modelo SAR propostos por Kelejian e Prucha (2007). Os preditores parcial (SAR-P) e geral (SAR-G) para avaliação de acurácia são dados por

$$\text{SAR-P} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}_{dd}\mathbf{X}_d\hat{\beta}, \tag{3.2}$$

$$\text{SAR-G} = \mathbf{X}_d\hat{\beta} + \hat{\rho}\mathbf{W}_{dd}\mathbf{y}_d. \tag{3.3}$$

Esses preditores (SAR-P, SAR-G) são utilizados em situações que se têm valores observados em todas as áreas avaliadas, ou seja, consideram informações disponíveis $\mathbf{y}_d, \mathbf{X}_d, \mathbf{W}_{dd}$. Além desses dois estimadores, também foram utilizados dois preditores de observações faltantes, os parciais (PSAR-P) e geral (PSAR-G), os quais são utilizados em situações onde não se têm observações em todas as áreas avaliadas, baseando-se no conjunto de informações disponíveis $\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f, \mathbf{W}_{dd}, \mathbf{W}_{df}$. Estes são dados por

$$\text{PSAR-P} = (\mathbf{I} - \hat{\rho}\mathbf{W}_{df})^{-1}\mathbf{X}_f\hat{\beta}, \tag{3.4}$$

$$\text{PSAR-G} = \mathbf{X}_f\hat{\beta} + \hat{\rho}\mathbf{W}_{fd}\mathbf{y}_d. \tag{3.5}$$

Além dos preditores autorregressivos propostos por Kelejian e Prucha (2007), neste trabalho estão sendo propostos dois estimadores de krigagem, dados em (3.6) e (3.7), obtidos a partir da definição do preditor espacial generalizado, definidos em (2.40). A ideia consiste em modificar o termo de tendência $\mathbf{X}_d\hat{\boldsymbol{\beta}},\ \mathbf{X}_f\hat{\boldsymbol{\beta}}$ pelos preditores do modelo SAR. Assim, tem-se os seguintes preditores de krigagem:

$$\text{PK1} = \mathbf{X}_d\hat{\boldsymbol{\beta}} - \text{diag}(\hat{\mathbf{Q}}_{dd})^{-1}\tilde{\hat{\mathbf{Q}}}_{dd} \times \left(\mathbf{y} - \hat{\mathbf{X}}_d\hat{\boldsymbol{\beta}}\right), \tag{3.6}$$

$$\text{PK2} = \mathbf{X}_f\hat{\boldsymbol{\beta}} - \hat{\mathbf{Q}}^{-1}_{ff}\hat{\mathbf{Q}}_{df} \times (\mathbf{y}_d - \mathbf{X}_d\hat{\boldsymbol{\beta}}), \tag{3.7}$$

sendo que $\mathbf{X}_d\hat{\boldsymbol{\beta}} =$ SAR-P e $\mathbf{X}_f\hat{\boldsymbol{\beta}} =$ PSAR-P.

No caso do preditor PK1 as informações disponíveis na amostra são $\mathbf{y}, \mathbf{X}_d, \mathbf{W}_{dd}$, que serão utilizadas na avaliação da acurácia do mesmo. Já, para o preditor PK2 cuja finalidade é predizer valores ausentes em certas áreas, o conjunto de informações disponíveis são, respectivamente $\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f, \mathbf{W}_{df}$.

3.3 Avaliação da eficiência dos preditores

Na avaliação da eficiência dos preditores será utilizada a medida da raiz do erro quadrático médio (REQM) por

$$\text{REQM} = \sqrt{EQM}, \tag{3.8}$$

sendo que EQM $= \frac{1}{n_f}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, em que $y_i,\ \hat{y}_i$ representam os valores observados e preditos respectivamente; n_f é o número de observações ausentes na amostra. Será utilizado igualmente a eficiência relativa (ER). Neste caso, considera-se a razão entre preditor de krigagem de referência e cada um dos preditores do modelo SAR. Na avaliação da qualidade do ajuste do modelo, será utilizado o critério de Informação de Akaike (AIC). Todo o processo de simulação foi realizado utilizando o software estatístico R (R Core Team, 2021).

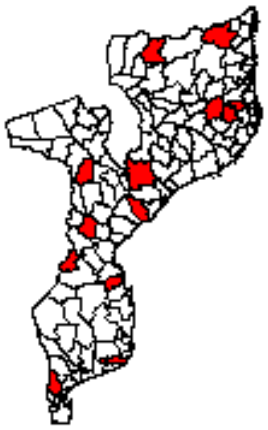
3.4 Dados reais

Os dados reais utilizados provêm de uma base secundária sobre o inquérito agrícola (TIA) de 2012, produzida pelo Ministério de Agricultura e Instituto Nacional de Estatística de Moçambique (INE), cuja finalidade é obter informações agropecuárias, características demográficas do setor agrário e o acesso à infraestrutura tanto a nível comunitário como ao nível do agregado familiar por todo país. Situado na África sub-subsahariana, Moçambique possui três macrorregiões: Sul, Centro e Norte. Divide-se em quatro níveis: províncias, distritais, postos administrativos, localidades, sendo o último, o nível mais baixo. Neste trabalho, considera-se o nível dois, com um total de 128 distritos. Considerou-se apenas os dados das pequenas e médias explorações pelo fato destas constituírem as potenciais produtoras de milho. Dentre várias covariáveis do inquérito, foram consideradas neste estudo apenas três: tamanho da família (TF) (numérica), trabalhadores efetivos (TE) (categórica), tração animal (TA) (categórica). A variável resposta **y** (numérica) é a proporção que corresponde ao número de produtores que usaram

sementes melhoradas de milho em cada distrito (N_{sm}) por número total de agricultores N, ou seja, $y = 100 \times \frac{N_{sm}}{N}$.

Aos dados, ajustou-se o modelo SAR, definido em (3.1) com a finalidade de se estimar os seus parâmetros, posteriormente utilizados nos preditores (3.2), (3.3), (3.6) na avaliação de acurácia; e utilizados igualmente nos preditores (3.4), (3.5), (3.7) que serão utilizados para a predição dos valores nos 13 distritos em que se supõem não existir informações da variável **y** (FIGURA 3.2).

Figura 3.2 – Distribuição espacial dos 13 distritos (à vermelho) sem informação da variável resposta.



Fonte: Do autor (2022)

Resumidamente, foram executados os seguintes passos:

a) ajuste do modelo SAR, definido em (3.1) considerando o conjunto de informações disponíveis, ou seja, $\mathbf{y}_d, \mathbf{X}_d, \mathbf{W}_{dd}$ de modo a se estimar os seus parâmetros, e em seguida, utilizar nos preditores para avaliar a acurácia dos mesmos. Neste caso, foram utilizados os preditores (3.2), (3.3), (3.6);

b) ajuste do modelo (3.1) considerando os dados disponíveis nas áreas que contêm observações de **y**, de modo a se estimar seus parâmetros. Estas estimativas serão consideradas nos preditores do modelo SAR, definidos em (3.4) e (3.5) e no preditor de krigagem proposto em (3.7), com o objetivo de predizer os valores faltantes da variável resposta em alguns distritos, considerando o conjunto de informações disponíveis, ou seja, $\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f, \mathbf{W}_{df}$.

4 RESULTADOS E DISCUSSÃO

Este capítulo está dividido em duas partes: a teórica, em que se apresenta os resultados de predição, obtidos por simulação, cujos preditores foram ajustados a partir do modelo SAR e de krigagem. Considera-se dois processos: o primeiro em que se prediz as observações em todas as áreas; na segunda parte, busca-se predizer valores da variável resposta em algumas áreas não observadas. Nas duas partes, tem-se por objetivo, comparar o desempenho dos preditores SAR e de krigagem. Consideram-se igualmente a especificação de duas matrizes $\mathbf{W}1, \mathbf{W}2$ por critérios distintos, para efeitos de comparação.

4.1 Avaliação da acurácia dos preditores

Neste tópico serão apresentados os valores da raiz do erro quadrático médio (REQM), que é uma medida da acurácia das predições. Nessa avaliação foram considerados os preditores obtidos com o ajuste dos modelos SAR: o preditor espacial autorregressivo geral (SAR-G), definido em (3.3), o preditor espacial autorregressivo parcial (SAR-P), definido em (3.2); e o preditor de krigagem (PK1), definido em (3.6). As predições foram realizadas utilizando os dados simulados, considerando diferentes valores da autocorrelação espacial (ρ) e as diferentes matrizes de vizinhança ($\mathbf{W}1$ e $\mathbf{W}2$), conforme definido anteriormente. Os resultados de REQM, considerando os diferentes cenários simulados e os diferentes modelos de ajuste estão apresentados na Tabela 4.1.

Tabela 4.1 – Médias dos valores da raiz do erro quadrático médio dos preditores SAR-P, SAR-G, PK1, considerando as matrizes $\mathbf{W}1$ e $\mathbf{W}2$ com diferentes valores de autocorrelação espacial (ρ) em dados simulados.

ρ	W1			W2		
	PK1	SAR-G	SAR-P	PK1	SAR-G	SAR-P
-0,4	0,961	0,975	1,000	0,968	0,981	0,993
-0,2	0,968	0,974	0,985	0,964	0,975	0,984
0,05	0,975	0,977	0,980	0,968	0,974	0,980
0,2	0,980	0,982	0,988	0,975	0,979	0,984
0,3	0,976	0,981	1,002	0,973	0,976	0,980
0,5	0,974	0,984	1,034	0,973	0,976	0,981
0,7	0,969	0,987	1,101	0,978	0,982	0,988
0,8	0,956	0,984	1,284	0,979	0,983	0,993
0,9	0,952	0,987	1,340	0,978	0,983	0,997

Fonte: Do autor (2022)

Na Tabela 4.1, ao se considerar os diferentes valores do parâmetro de autocorrelação espacial ρ, observa-se a superioridade do preditor PK1 em termos de acurácia quando comparado com os demais, com valores inferiores da REQM, para qualquer uma das duas matrizes. Já, o preditor SAR-P apresenta menor acurácia na predição. Além disso, observa-se para o SAR-P que, à medida que o valor do parâmetro ρ aumenta ($\rho > 0,3$), os valores da REQM também aumentam muito rapidamente na matriz **W1**. No entanto, esse fato não é observado quando se utiliza a matriz $\mathbf{W}2$, que apresenta uma estabilidade na matriz **W2**.

Ao se considerar os diferentes valores de autocorrelação espacial, tomando em consideração a matriz $\mathbf{W}1$, à medida que ρ aumenta ($\rho > 0,3$), observa-se no preditor PK1, uma diminuição gradual da REQM com valor bem menor, quando comparado aos demais. Essa diferença é expressiva quando $\rho = 0,9$, em que se supõe uma forte dependência espacial entre as observações. Ainda nesse valor, ocorre uma diminuição da REQM em 3,5% e 29,0% quando comparada PK1 com SAR-G e SAR-P respectivamente. Já, na matriz $\mathbf{W}2$ observou-se uma estabilidade entre todos os preditores.

Em geral, tem-se a seguinte classificação dos preditores em relação aos valores da REQM: PK1 < SAR-G < SAR-P, sendo este último o menos eficiente. Comparando o desempenho das matrizes, a $\mathbf{W}2$ apresenta valores inferiores da REQM (TABELA 4.1). Conforme Kelejian e Prucha (2007), duas razões são apontadas no desempenho das matrizes: a primeira é a quantidade de informação disponível no preditor que é menor em comparação com os demais $\{\mathbf{X}, \mathbf{W}\}$; a segunda é que a matriz $\mathbf{W}1$ é dispersa. Possivelmente não consiga captar melhor as informações ao redor da sua vizinhança, sendo responsável em cerca de 8% de conexões, que corresponde a 1280 do total das 128 áreas (TABELA 3.1).

Na Tabela 4.2 estão apresentados os valores da eficiência relativa (ER) do preditor de krigagem em relação aos preditores obtidos com o ajuste dos modelos SAR. Considerando os valores negativos do parâmetro de autocorrelação espacial $\rho < 0$, observa-se bom desempenho dos preditores SAR em termos de eficiência. Já, no intervalo entre $0,05 < \rho < 0,2$ em que se supõe por hipótese, ausência de dependência espacial ($H_0 : \rho = 0$), conforme a classificação em Pace e Lesage (2016), os valores da EP se aproximam entre si, independente da matriz, ou seja, PK1≈SAR-G≈SAR-P. Portanto, conforme afirmam Bivand, Millo e Piras (2021), na ausência de dependência espacial, é preferível utilizar preditores dos modelos de regressão Gauss Markov, que são mais simples e tão eficiente quanto os modelos considerando a informação espacial.

Tabela 4.2 – Eficiência relativa (ER) do estimador PK1 com base nos estimadores do modelo SAR, considerando as matrizes W1, W2 e diferentes valores do índice de autocorrelação espacial (ρ).

ρ	W1		W2	
	PK1/SAR-G	PK1/SAR-P	PK1/SAR-G	PK1/SAR-P
-0,4	0,970	0,922	0,973	0,950
-0,2	0,988	0,966	0,979	0,961
0,05	0,997	0,990	0,988	0,976
0,2	0,997	0,984	0,991	0,982
0,3	0,990	0,949	0,994	0,986
0,5	0,979	0,888	0,994	0,985
0,7	0,964	0,774	0,994	0,981
0,8	0,944	0,555	0,991	0,971
0,9	0,929	0,324	0,989	0,962

Fonte: Do autor (2022)

Getis e Aldstadt (2004) observaram que o ajuste de modelos SAR com diferentes especificações da matriz **W**, tem influencia na qualidade do ajuste. Uma das conclusões desses autores foi que com o uso da matriz **W**, definida pelo método de compartilhamento de fronteira, obteve-se um melhor ajuste do modelo. Contudo, os melhores resultados foram obtidos quando a matriz **W** foi obtida diretamente dos dados, considerando o ajuste de três modelos do semivariograma, exponencial, esférico e Gaussiano e, posteriormente, calculando os valores dos elementos da matriz **W**, definido em (2.33) por $\rho(d_{ij}) = 1 - \frac{\gamma(d_{ij})}{\sigma^2}$.

Bhattacharjee e Jensen-Butler (2013) e Kelejian e Prucha (2007) mencionaram que, na especificação da matriz **W** deve se levar em conta o tipo de dados a serem analisados, a distribuição espacial das áreas na região de estudo e particularidades como a quantidade de vizinhos, presença ou não de agrupamentos formados. Uma matriz **W** dispersa, resulta em menor eficiência dos preditores.

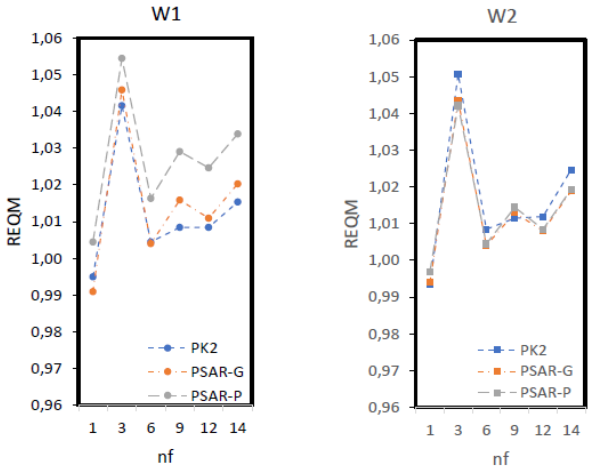
4.2 Eficiência dos preditores

Neste tópico serão apresentados resultados referentes às análises dos dados simulados, considerando o seguinte conjunto de informações na amostra: $\{\mathbf{y}_d, \mathbf{X}_d, \mathbf{X}_f, \mathbf{W}_{df}\}$.

No ajuste do modelo SAR, foram considerados três valores distintos do índice de autocorrelação espacial $\rho = -0.4, 0.02, 0.7$. Em todas as situações, manteve-se o valor fixo de $\sigma^2 = 1$, sem perda de generalidade. Enfatiza-se o fato de que a sub-matriz $(\mathbf{W}_{df})$ é uma das partições da matriz $\mathbf{W}^*$ conforme definido em (2.20). Para a predição de observações faltantes em algumas áreas (n_f) foram considerados os preditores obtidos com o ajuste dos modelos SAR: o preditor espacial autorregressivo parcial (SAR-P), definido em (3.4), o preditor espacial autorregressivo geral (SAR-G), definido em (3.5) e o preditor de krigagem (PK2), definido em (3.7).

O primeiro resultado, quando $\rho = -0,4$ é apresentado na Figura 4.1. Em relação a matriz $\mathbf{W}1$, o preditor PK2 apresenta maior acurácia, ou seja, tem valores menores da REQM, seguido do preditor PSAR-G. Já, em relação a matriz $\mathbf{W}2$, os valores da REQM são muito próximas entre os preditores envolvidos. Em relação ao tamanho da amostra (n_f), observa-se valores relativamente baixos da REQM quando $n_f = 1$. Esse resultado, também se aplica a outros valores de autocorrelação espacial $\rho = 0,02; 0,7$. Comparando o desempenho dos preditores em termos numéricos, todos se mostram mais eficientes ao se utilizar a matriz $\mathbf{W}2$.

Figura 4.1 – Raiz do erro quadrático médio (REQM) entre os preditores do modelo SAR e de krigagem, considerando as matrizes W1 e W2, com diferentes tamanhos de amostras ausentes (n_f) em dados simulados, com $\rho = -0.4$.

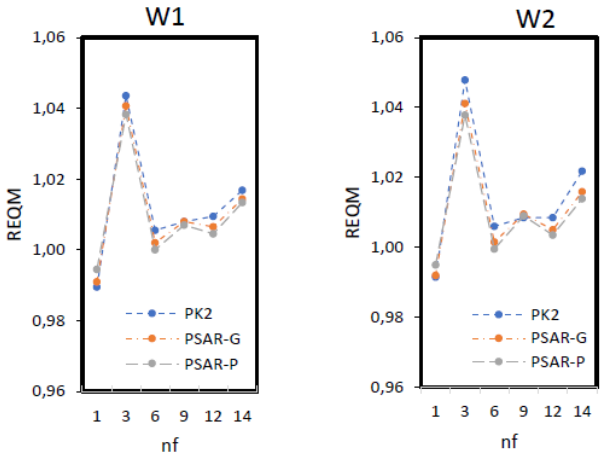


Fonte: Do autor (2022)

O segundo resultado, quando $\rho = 0,02$ é apresentado na Figura 4.2. Neste caso, tem-se um processo em que se supõe, por hipótese, ausência de dependência espacial $(H_0 : \rho = 0)$ e, portanto, as observações da variável **y** distribuem-se aleatoriamente no domínio D. Observa-se um comportamento similar entre os preditores, com valores da REQM bem próximos entre sí nas duas matrizes, **W**1 e **W**2.

Esse resultado era previsível, uma vez que se está diante de um modelo de regressão linear de Gauss-Markov (2.17), sendo que o valor esperado na *i*-ésima área da variável aleatória **y**, considerando informações presentes na amostra, é $\hat{\boldsymbol{\mu}}|\mathbf{X} = \mathbf{X}\hat{\boldsymbol{\beta}}$, onde $\hat{\beta}$ é estimado pelo método de mínimos quadrados ordinário, sendo este o melhor estimador linear não viesado (BLUE), conforme observado em Chipenete, Chipenete e Lima (2022).

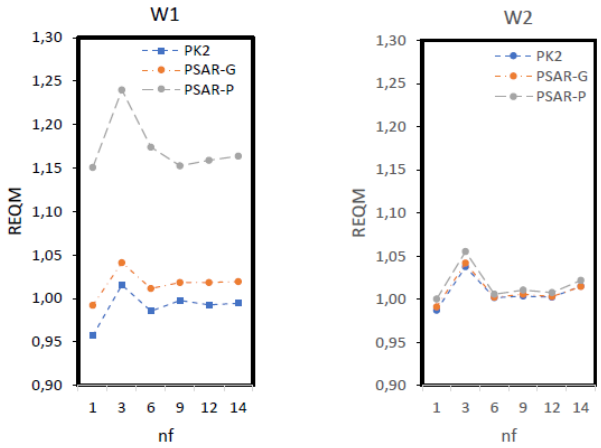
Figura 4.2 – Raiz do erro quadrático médio (REQM) entre os preditores do modelo SAR e de krigagem, considerando as matrizes W1 e W2, com diferentes tamanhos de amostras ausentes (n_f) em dados simulados, com $\rho = 0,02$.



Fonte: Do autor (2022)

O terceiro resultado, quando $\rho = 0,7$, é apresentado na Figura 4.3. Os valores da REQM dos preditores PK2, PSAR-G e PSAR-P são bem próximos entre si, ao se utilizar a matriz **W**2. Além disso, observa-se que a matriz **W**2 com melhor desempenho, com menor amplitude dos valores da REQM, situando-se aproximadamente entre 0,95 e 1,25. Contudo, ao se utilizar a matriz **W**1, essa variação dos valores de REQM é maior, com amplitude aproximada entre 1,0 e 1,5.

Figura 4.3 – Raiz do erro quadrático médio (REQM) entre os preditores do modelo SAR e de krigagem, considerando as matrizes W1 e W2, com diferentes tamanhos de amostras ausentes (n_f) em dados simulados, com $\rho = 0,70$.



Fonte: Do autor (2022)

Os resultados das eficiências relativas (ER) dos preditores PSAR em relação ao PK2, considerando os valores fixos de autocorrelação espacial ($\rho = -0,04; 0,02; 0,7$) e usando uma das duas matrizes de ponderação espacial $\mathbf{W}1$, $\mathbf{W}2$ são apresentados na Tabela 4.3. Observa-se uma menor variação em termos percentuais dos valores da ER na matriz $\mathbf{W}2$, ou seja, de aproximadamente 3%. Já, para a matriz $\mathbf{W}1$ essa variação situa-se por volta de 30% para o preditor PSAR-P, no valor mais alto da correlação espacial ($\rho = 0,7$). Além disso, o PSAR-P foi o que apresentou pior ER. Ainda na Tabela 4.3, quando considerado a autocorrelação espacial negativa, a eficiência dos preditores se assemelham em termos numéricos, independente da matriz $\mathbf{W}$ utilizada.

Os resultados do processo de simulação são consistentes com os apresentados na literatura. Kato (2008), LeSage e Pace (2004) embora estivessem apenas a lidar com o modelo SEM, definido em (2.16), chegaram a conclusões similares quanto ao desempenho do preditor de krigagem PK2.

Getis e Aldstadt (2004), ao comparar o desempenho das matrizes de ponderação ou vizinhança espacial, observaram que, quando ajustadas utilizando o modelo do semivariograma tiveram melhor ou igual desempenho em comparação com os demais critérios. Kelejian e Prucha (2007) comparou diferentes preditores do modelo SAR e krigagem, considerando o modelo SEM, entre outros e chegaram à conclusão de que é sempre mais fácil e preferível utilizar o preditor de krigagem.

Tabela 4.3 – Valores das eficiências relativas (ER), PK2/PSAR-G e PK2/PSAR-P, nos três valores fixos de autocorrelação espacial (ρ), considerando as matrizes de vizinhança **W**1 e **W**2.

W	ρ	ER	Tamanho de amostras faltantes (nf)					
			1	3	6	9	12	14
W1	-0,04	$\frac{\text{PK2}}{\text{PSAR-G}}$	1,000	0,985	0,991	0,991	0,992	0,998
		$\frac{\text{PK2}}{\text{PSAR-P}}$	1,008	0,992	1,001	0,985	0,995	0,990
	0,02	$\frac{\text{PK2}}{\text{PSAR-G}}$	0,997	1,006	1,007	1,000	1,006	1,005
		$\frac{\text{PK2}}{\text{PSAR-P}}$	0,990	1,010	1,011	1,002	1,010	1,007
	0,7	$\frac{\text{PK2}}{\text{PSAR-G}}$	0,931	0,951	0,949	0,959	0,950	0,952
		$\frac{\text{PK2}}{\text{PSAR-P}}$	0,692	0,671	0,705	0,749	0,733	0,730
W2	-0,04	$\frac{PK2}{PSAR-G}$	1,000	0,991	0,997	0,993	1,000	1,000
		$\frac{\text{PK2}}{\text{PSAR-P}}$	0,999	1,014	1,009	0,997	1,008	1,012
	0,02	$\frac{\text{PK2}}{\text{PSAR-G}}$	0,999	1,013	1,009	0,998	1,007	1,012
		$\frac{\text{PK2}}{\text{PSAR-P}}$	0,993	1,019	1,013	0,999	1,010	1,016
	0,7	$\frac{\text{PK2}}{\text{PSAR-G}}$	0,992	0,992	0,999	0,996	0,998	1,001
		$\frac{\text{PK2}}{\text{PSAR-P}}$	0,974	0,967	0,992	0,986	0,989	0,987

Fonte: Do autor (2022)

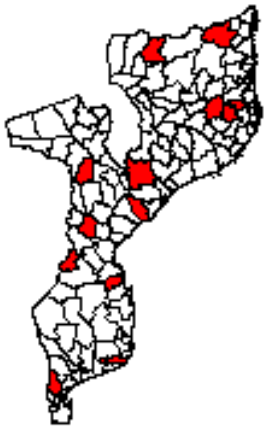
Por fim, neste trabalho, utilizou-se o modelo SAR pra essa comparação e propõe-se o preditor PK2 com modificações no termo de tendência. Um outro resultado observado é que os preditores PK2, PSAR-G e PSAR-G são mais eficientes quando se utiliza a matriz **W**2. Uma vantagem em se utilizar a matriz de ponderação ou de vizinhança espacial no preditor PK2 é que, minimiza o esforço na busca de um critério que possa espelhar de perto a realidade, conforme afirma Bhattacharjee e Jensen-Butler (2013).

4.3 Análise de dados reais

Os resultados apresentados a seguir são baseados em dados reais, referentes ao uso de sementes melhoradas de milho em Moçambique. Inicialmente, será ajustado o modelo SAR aos dados referente apenas a 115 distritos, sem levar em consideração os restantes 13, com valores

faltantes. Neste caso, os preditores utilizados são os definidos em 3.3 e 3.2. Neste caso, a matriz de ponderação espacial utilizada é $\mathbf{W}_{dd}$. Em seguida, com a finalidade de predizer valores nos treze distritos em que se supõe não conhecer o verdadeiro valor da variável resposta, cujo mapa está apresentado na Figura 4.4, serão utilizados os preditores definidos em 3.4, 3.5 e 3.7. Neste caso, a matriz de ponderação espacial que inclui os treze distritos é definido em (2.20), ou seja, $\mathbf{W}_{df}$, um elemento da matriz $\mathbf{W}^*$.

Figura 4.4 – Distribuição espacial dos 13 distritos (à vermelho) sem informação da variável resposta.



Fonte: Do autor (2022)

O modelo SAR ajustado aos dados reais é dado, matricialmente, por $\mathbf{y} = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{y} + \boldsymbol{\epsilon}$, em que $\mathbf{y}$ é o vetor de dados observados, $\mathbf{X}$ é a matriz de incidência dos efeitos fixos ou covariáveis, $\boldsymbol{\beta}$ é o vetor de efeitos fixos ou das covariáveis, ρ é o coeficiente de autocorrelação espacial, $\mathbf{W}$ é a matriz de vizinhança e $\boldsymbol{\epsilon}$ é o vetor de erros, tal que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. O vetor $\boldsymbol{\beta}$ é composto pela constante ou intercepto ($\boldsymbol{\mu}$) e efeitos das três covariáveis ($\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$), que correspondem a: tamanho da família, número de trabalhadores ativos e uso de tração animal.

Na Tabela 4.4 são apresentadas as estimativas dos parâmetros das covariáveis obtidas ajustando-se o modelo SAR, apresentado em (3.1), considerando a matriz $\mathbf{W}1$. Já, na Tabela 4.5 essas estimativas foram obtidas utilizando a matriz $\mathbf{W}2$. Pelo critério AIC, o modelo ajustou-se melhor aos dados ao se utilizar a matriz $\mathbf{W}2$ (AIC=659,2), quando comparado com a matriz $\mathbf{W}1$ (AIC=661,81).

Nas Tabelas 4.4 e 4.5 se observam que, o tamanho da família, trabalhadores efetivos e tração animal, apresentam coeficientes positivos ($\hat{\beta}'s > 0$), significativos a um nível de 5% de significância ($p < 0,05$). Esses resultados mostram que essas covariáveis exercem um efeito positivo no uso de sementes melhoradas de milho pelos agricultores.

Conforme observaram Manuel e Scalon (2020), famílias com maior número de membros, que podem representar maior disponibilidade de mão de obra no processo de produção, tem um impacto direto sobre a decisão de adoção de variedades melhoradas do milho. Esse resultado é percebido nos dois modelos ajustados, tal que, o incremento de uma unidade no número de membros da família, produz um aumento médio em 1,147 quilogramas por hectare (kg/ha) (W1) e 1,079 kg/ha (W2), respectivamente, no uso das sementes de milho, mantendo constante outras variáveis.

Tabela 4.4 – Estimativas dos parâmetros do modelo SAR considerando a matriz **W**1.

Covariáveis	**Estimativas**	**Erro padrão**	**Z**
Intercepto	10,256	1,745	5,878**
Tamanho da família	1,147	0,32	3,578**
Trabalhadores efetivos	0,442	0,152	2,914**
Tração animal	0,108	0,016	6,559**
ρ	0,156	0,106	1,47**
AIC	661,81		

**p < 0,05

Fonte: Do autor (2022)

Tabela 4.5 – Estimativas dos parâmetros do modelo SAR considerando a matriz **W**2.

Covariáveis	**Estimativas**	**Erro padrão**	**Z**
Intercepto	6,719	2,312	2,906**
Tamanho da família	1,079	0,317	3,405**
Trabalhadores efetivos	0,387	0,135	2,855**
Tração animal	0,11	0,016	6,775**
ρ	0,579	0,216	2,679**
AIC	659,2		

**p < 0,05

Fonte: Do autor (2022)

Verificam-se também que, o aumento de uma unidade de trabalhadores efetivos produz um aumento no uso de tais sementes em 0,442 kg/ha (W1) e 0,387 kg/ha (W2), respectivamente. A presença desses trabalhadores é uma garantia de continuidade do trabalho, principalmente em épocas de semeadura e colheita, ainda que nessa fase de produção, a maioria dos agricultores optem pela contratação de trabalhadores temporários, uma vez que nessas épocas existe uma maior demanda por mão de obra.

Para o caso da covariável tração animal, o aumento em uma unidade de animal de tração produz um incremento no uso de sementes melhoradas de milho em 0,108 kg/ha (W1) e 0,11

kg/ha (W2), respectivamente. A opção no uso de animais de tração, principalmente nas regiões sul e centro, deve-se à existência de maior efetivo desses animais e ser economicamente mais viável na maioria dos agricultores de baixa e média renda.

Ao se supor existência de áreas sem valores observados da variável resposta (FIGURA 3.2), o objetivo é predizer tais valores. Os resultados das estimativas dos parâmetros nas Tabelas 4.4 e 4.5, serão utilizados nos preditores PK2, PSAR-G, PSAR-P, definidos em 3.4, 3.5 e 3.7, respectivamente. Neste caso, a matriz de ponderação espacial que inclui os treze distritos é definido em (2.20), ou seja, $\mathbf{W}_{df}$, um elemento da matriz $\mathbf{W}^{*}$.

Na Tabela 4.6 têm-se os valores preditos nos distritos identificados por "ID", considerando a matriz $\mathbf{W}1$. Já, na Tabela 4.7 são apresentados os valores preditos considerando a matriz $\mathbf{W}2$. No geral, ao se utilizar a matriz $\mathbf{W}2$ a variabilidade foi menor ao se considerar o desvio médio e desvio padrão comparando com os resultados obtidos ao se utilizar a matriz $\mathbf{W}1$. Esse resultado mostra que a predição é sensível ao critério utilizado na especificação da matriz e ponderação $\mathbf{W}$, conforme observaram Getis e Aldstadt (2004), Getis (2010), Bhattacharjee e Jensen-Butler (2013) e outros. Além disso, é possível que a quantidade de conexões presente na matriz $\mathbf{W}2$ (TABELA 3.1) seja responsável pelo seu bom desempenho ao comparar com as da matriz $\mathbf{W}1$.

Tabela 4.6 – Valores preditos da variável resposta nos treze distritos identificados por ID, utilizando os preditores PK2, PSAR-G, PSAR-P, considerando a matriz **W**1.

ID	TF	TE	TA	PK2	PSAR-G	PSAR-P
22571	5,1	0	4	5,556	5,616	5,573
22625	5,8	5	17	8,850	8,678	8,843
22618	5	4	18	6,419	6,515	6,530
22615	6,4	3	23	11,066	11,147	11,172
22557	4,6	0	0	5,079	5,100	5,096
22632	5,3	5	48	4,892	4,626	4,778
22592	5,2	2	0	7,307	7,440	7,388
22624	4,7	5	0	11,470	11,380	11,481
22566	4,9	0	0	8,737	8,715	8,751
22552	4,1	0	0	3,991	4,133	4,045
22563	4,2	0	0	4,439	4,258	4,428
22526	4,1	0	0	4,122	4,158	4,160
22527	4,4	0	0	4,683	4,720	4,733
Média				6,662	6,653	6,691
Desvio padrão				2,604	2,601	2,618
Desvio médio				2,172	2,169	2,182

TF=Tamanho da família; TE=Trabalhadores efetivos; TA=Tração animal.
Fonte: Do autor (2022)

Tabela 4.7 – Valores preditos da variável resposta nos treze distritos identificados por ID, utilizando os preditores PK2, PSAR-G, PSAR-P, considerando a matriz **W**2.

ID	TF	TE	TA	PK2	PSAR-G	PSAR-P
22571	5,1	0	4	5,464	5,524	5,481
22625	5,8	5	17	8,758	8,602	8,751
22618	5	4	18	6,327	6,410	6,438
22615	6,4	3	23	10,974	11,057	11,080
22557	4,6	0	0	4,987	5,013	5,004
22632	5,3	5	48	4,800	4,552	4,686
22592	5,2	2	0	7,215	7,353	7,296
22624	4,7	5	0	11,378	11,299	11,389
22566	4,9	0	0	8,645	8,613	8,659
22552	4,1	0	0	3,899	4,045	3,953
22563	4,2	0	0	4,347	4,170	4,336
22526	4,1	0	0	4,030	4,069	4,068
22527	4,4	0	0	4,591	4,629	4,641
Média				6,570	6,564	6,599
Desvio padrão				2,578	2,575	2,592
Desvio médio				2,146	2,144	2,156

TF=Tamanho da família; TE=Trabalhadores efetivos; TA=Tração animal.
Fonte: Do autor (2022)

Nas Tabelas 4.8 e 4.9 são apresentados os resultados da raiz do erro quadrático médio (REQM) e eficiência relativa (ER), respectivamente. Esses resultados, mostram diferença nu-

mérica entre o preditor PK2 em comparação com o PSAR-G e PSAR-P, com valores inferiores, sendo o mais eficiente, seguido de PSAR-G, ao se utilizar a matriz **W**2.

Tabela 4.8 – Raiz do erro quadrático médio (REQM) dos preditores PK2, PSAR-G, PSAR-P, utilizando as matrizes de vizinhança **W**1 e **W**2.

Preditores	**W**1	**W**2
PK2	1,903	1,864
PSAR-G	1,910	1,892
PSAR-P	1,966	1,939

Fonte: Do autor (2022)

Tabela 4.9 – Valores das eficiências relativas (ER), PK2/PSAR-G e PK2/PSAR-P, considerando as matrizes de vizinhança **W**1 e **W**2.

W	$\frac{\text{PK2}}{\text{PSAR-G}}$	$\frac{\text{PK2}}{\text{PSAR-P}}$
W1	0,996	0,968
W2	0,985	0,961

Fonte: Do autor (2022)

4.4 Considerações finais

Ao predizer uma observação ou conjunto delas em dados espaciais de área, a literatura especializada, oferece várias opções. Como se observou nesta tese, foram propostos ao longo dos anos diferentes formas para predição, desde interpoladores determinísticos até preditores probabilísticos e mais eficientes. Contudo, uma vez que a matriz de ponderação espacial está sempre presente como componente essencial, a especificação depende de vários critérios. Assim, uma escolha não assertiva pode comprometer a qualidade de predição, independentemente de quão bom seja o preditor.

Observou-se que preditores que utilizaram a matriz com maior número de conexões (**W**2) foram mais eficientes do que quando utilizaram a matriz com menor número de conexões (**W**1). Além disso, também se observou que a quantidade de informação disponível na amostra, ou seja, presença de observações das covariáveis e da variável resposta, afetou a eficiência dos preditores SAR, ou seja, quanto maior a quantidade de informações existente mais eficiente se torna o preditor. É o caso do preditor geral SAR-G, que na sua estrutura, está presente a matriz de vizinhança espacial composta por áreas com e sem observações $\mathbf{W}_{df}$, um vetor da variável resposta ($\mathbf{y}_d$) em algumas áreas, às covariáveis ($\mathbf{X}_d, \mathbf{X}_f$). Já, no preditor parcial PSAR-P não

está presente, por exemplo, a informação com respeito ao vetor de observações da variável resposta **y**.

Em relação ao preditor de krigagem, foi observado que oferece uma solução bem mais prática, uma vez que, é minimizado o problema da especificação e seleção da matriz de ponderação espacial **W**. Observou-se ainda que, mesmo considerando as duas matrizes, conseguiu superar ou se igualar aos preditores do modelo SAR. Contudo, também se observou limitações com a alteração nos valores do parâmetro de autocorrelação espacial (ρ), ou seja, no nível de relacionamento ou dependência espacial entre as áreas. Para alguns valores de ρ, os valores da REQM oscilava, algumas vezes com valores altos ao comparar com os preditores SAR. Portanto, como proposta nos próximos estudos, pretende-se avaliar com maior exatidão, uma zona ou região segura do intervalo do parâmetro (ρ) em que seja mais eficiente.

Por fim, esse estudo mostrou que é possível aplicar os resultados a uma situação real, como no caso, na predição de observações ausentes na proporção de agricultores que fizeram uso de sementes melhoradas de milho, com maior grau de certeza, utilizando o preditor de krigagem. Espera-se que problemas derivados de ausência de valores da variável resposta, que provêm de inquéritos agrícolas ou em alguma outra área, e que envolvem dados espacias de área, podem ser solucionados utilizando o preditor de krigagem.

5 CONCLUSÃO

A presente tese teve por objetivo propôr um preditor que pudesse incorporar técnicas de dados espaciais de área na krigagem através do termo de tendência e utilizando a matriz de ponderação espacial (**W**) especificada previamente, cuja finalidade é a predição de valores ausentes da variável resposta. Foram utilizadas duas matrizes **W**, uma especificada pelo critério que leva em conta a quantidade de vizinhos que uma determinada área deve possuir (**W**1), que teve menor número de conexões; a outra pela combinação dos critérios da distância inversa e de máxima distância (**W**2) que teve maior número de conexões entre áreas, que por sinal, foi a que os preditores foram mais eficientes. O uso dessas duas matrizes, contribuiu para aferir que diferentes critérios na especificação da matriz **W** afetam a qualidade do ajuste do modelo aos dados. Assim, em um processo de simulação, comparou-se o preditor de krigagem e aqueles derivados do modelo autorregressivo de defasagem espacial (SAR) considerando essas duas matrizes. Os resultados mostraram que o preditor de krigagem, com a modificação no termo de tendência (μ), proposto nesta tese, mostrou-se, na maioria dos casos estudados, foi mais eficiente ou se igualou aos preditores derivados do modelo SAR. Entre os preditores do modelo SAR, o SAR-G e PSAR-G, definidos como gerais, foram os mais eficientes. Esse resultado se deve à maior quantidade de informação disponível que apresentavam na sua estrutura em relação aos SAR-P e PSAR-P definidos como preditores parciais, ou seja, que possuíam apenas informações da covariável **X** e da matriz **W** na sua estrutura. Por fim, os resultados obtidos, especialmente no que diz respeito ao preditor de krigagem, permitiu aplicar em casos reais, cujo objetivo é predizer predizer observações faltantes da variável resposta.

REFERÊNCIAS

ANSELIN, L. A test for spatial autocorrelation in seemingly unrelated regressions. **Economics Letters**, Elsevier, v. 28, n. 4, p. 335–341, 1988.

ANSELIN, L.; BERA, A. Spatial dependence in linear regression models with an application to spatial econometrics. **Handbook of Applied Economics Statistics, Springer-Verlag, Berlin**, v. 21, p. 74, 1998.

BAILEY, T. C.; GATRELL, A. C. **Interactive Spatial Data Analysis**. [S.l.]: Routledge, 1995. ISBN 978-0582244931.

BALTAGI, B.; YANG, Z. Standardized lm tests for spatial error dependence in linear or panel regressions. **Econometrics Journal**, v. 16, p. 103–134, 11 2013.

BHATTACHARJEE, A.; JENSEN-BUTLER, C. Estimation of the spatial weights matrix under structural constraints. **Regional Science and Urban Economics**, Elsevier BV, v. 43, n. 4, p. 617–634, jul 2013.

BIVAND, R. Spatial econometrics functions in r: Classes and methods. **Journal of geographical systems**, Springer, v. 4, n. 4, p. 405–421, 2002.

BIVAND, R.; MILLO, G.; PIRAS, G. A review of software for spatial econometrics in r. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 9, n. 11, p. 1276, 2021.

CHEN, Y. On the four types of weight functions for spatial contiguity matrix. **Letters in Spatial and Resource Sciences**, Springer Nature, v. 5, n. 2, p. 65–72, jan 2012.

CHIPENETE, C. F.; CHIPENETE, G. H. N.; LIMA, R. R. de. Modelos de regressão ajustados a dados espaciais de áreas com sementes melhoradas de milho em moçambique regression models fitted to spatial area data which used improved maize seeds in mozambique. **Brazilian Journal of Development**, v. 8, n. 3, p. 20017–20034, 2022.

CRESSIE, N. A. Statistics for spatial data/noel ac cressie. **Wiley series in probability and mathematical statistics. Applied probability and statistics section.**, Wiley. New York. US, 1993.

HSAR: An R Package for Integrated Spatial Econometric and Multilevel Modelling. [S.l.]: GISRUK, 2016.

DUBIN, R. Predicting house prices using multiple listings data. **The Journal of Real Estate Finance and Economics**, v. 17, p. 35–59, 02 1998.

FISCHER, M.; WANG, J. **Spatial Data Analysis: Models, Methods and Techniques**. [S.l.: s.n.], 2011. ISBN 978-3-642-21719-7.

GETIS, A. Spatial autocorrelation. london: **Pion**. v. 19, p. 245–249, 1995. ISSN 0309-1325.

GETIS, A. **Spatial Autocorrelation**. 2010. 255-278 p.

GETIS, A.; ALDSTADT, J. Constructing the spatial weights matrix using a local statistic. **Geographical analysis**, Wiley Online Library, v. 36, n. 2, p. 90–104, 2004.

GOOVAERTS, P. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. **International journal of health geographics**, BioMed Central, v. 5, n. 1, p. 52, 2006.

GRIFFITH, D. A. **Spatial Statistics and Geostatistics: Basic Concepts.** 2017.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. [S.l.]: Academic press London, 1978. v. 600.

KATO, T. A further exploration into the robustness of spatial autocorrelation specifications. **Journal of Regional Science - J REG SCI**, v. 48, p. 615–639, 08 2008.

KELEJIAN, H. H.; PRUCHA, I. R. The relative efficiencies of various predictors in spatial econometric models containing spatial lags. **Regional Science and Urban Economics**, v. 37, n. 3, p. 363 – 374, 2007. ISSN 0166-0462.

LESAGE, J. P.; PACE, R. K. Models for Spatially Dependent Missing Data. **The Journal of Real Estate Finance and Economics**, v. 29, n. 2, p. 233–254, September 2004.

MANUEL, L.; SCALON, J. D. Generalized estimating equations approach for spatial lattice data: A case study in adoption of improved maize varieties in mozambique. **Biometrical Journal**, v. 62, n. 8, p. 1879–1895, 2020.

MOJIRI, A. et al. Comparison of predictions by kriging and spatial autoregressive models. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 47, n. 6, p. 1785–1795, 2018.

PACE, K.; LESAGE, J. Models for spatially dependent missing data. **The Journal of Real Estate Finance and Economics**, v. 29, p. 233–254, 09 2004.

PACE, R.; LESAGE, J. **Spatial Econometric Models, Prediction**. [S.l.: s.n.], 2016. 1-7 p.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

TIMMINS, T. L. et al. Developing spatial weight matrices for incorporation into multiple linear regression models: An example using grizzly bear body size and environmental predictor variables. **Geographical Analysis**, Wiley, v. 45, n. 4, p. 359–379, sep 2013.