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

DOI:10.34117/bjdv8n3-279

Recebimento dos originais: 14/02/2022
Aceitação para publicação: 22/03/2022

Cláudio Francisco Chipenete
Doutorando na Universidade Federal de Lavras
Endereço: Aquenta Sol, Lavras - MG, CEP: 37200-900
E-mail: claudiochipenete@hotmail.com

Gisela Hélnia Nunes Chipenete
Doutoranda na Universidade Federal de Lavras
Endereço: Aquenta Sol, Lavras - MG, CEP: 37200-900
E-mail: ghnchipenete@gmail.com

Renato Ribeiro de Lima
Doutor
Instituição: Universidade Federal de Lavras
Endereço: Aquenta Sol, Lavras - MG, CEP: 37200-900
E-mail: rrlima@ufla.br

RESUMO

Uma variável regionalizada representada por dados de área é aquela em que as observações possuem uma referência geográfica e provem de regiões como aldeias, localidades, municípios, distritos, províncias ou alguma área delimitada no espaço. Para cada uma dessas regiões, esses dados, em geral, se apresentam na forma de média, taxas, proporções, dentre outras. Geralmente, esses tipos de dados espaciais são denominados simplesmente por dados de área. Em estudos com esse tipo de dado, se o interesse é ajustar modelos de regressão ou outro tipo de modelo, deve-se levar em conta a existência de dependência espacial entre as observações. Nesse caso, modelos clássicos de regressão linear (OLS) podem não ser apropriados. Em tais casos, a opção tem sido o uso de modelos indicados para dados de área, como os autorregressivos de defasagem espacial (SAR) ou de erros espaciais correlacionados (SEM). Neste artigo, o objetivo foi avaliar de forma prática, a qualidade do ajuste desses três modelos: SAR, SEM e OLS. Além disso, foi avaliado o efeito da matriz de ponderação espacial W na qualidade de ajuste, um componente essencial nos dois primeiros modelos. Quanto aos dados, são provenientes de um inquérito agrícola, referente ao uso de sementes melhoradas de milho em Moçambique. Também foi avaliada a contribuição de algumas covariáveis de interesse para os agricultores que utilizam tais sementes. O principal resultado, é que o modelo SAR foi aquele que melhor se ajustou aos dados, seguido do SEM, e por último OLS. Além disso, foi observado que a especificação da matriz W pode influenciar na qualidade do ajuste do modelo.

Palavras-chave: modelos autorregressivos SAR e SEM, dependência espacial, modelo OLS.

ABSTRACT
A regionalized variable represented by area data is one in which the observations have a geographic reference and come from regions such as villages, localities, municipalities, districts, provinces or some delimited area in space. For each of these regions, these data are generally presented in the form of averages, rates, proportions, among others. These types of spatial data are often referred to simply as area data. In studies with this type of data, if the interest is to adjust regression models or another type of model, the existence of spatial dependence between the observations must be taken into account. In this case, classical linear regression (OLS) models may not be appropriate. In such cases, the option has been to use models indicated for area data, such as spatial lag autoregressive (SAR) or correlated spatial error (SEM) models. In this article, the objective was to evaluate, in a practical way, the quality of fit of these three models: SAR, SEM and OLS. In addition, the effect of the spatial weighting matrix W on the goodness of fit, an essential component in the first two models, was evaluated. As for the data, they come from an agricultural survey, referring to the use of improved maize seeds in Mozambique. The contribution of some covariates of interest to farmers using such seeds was also evaluated. The main result is that the SAR model was the one that best fitted the data, followed by SEM, and finally OLS. In addition, it was observed that the specification of the W matrix can influence the quality of the model's fit.

Keywords: autoregressive SAR and SEM models, spatial dependence, OLS model.

1 INTRODUÇÃO

Em situação de crise econômica, social e política, a maximização da produção de alimentos em regiões mais afetadas pode reduzir os efeitos adversos consequentes. Por exemplo, atualmente cerca de 1,9 milhão de pessoas vive em situação de insegurança alimentar aguda em Moçambique. Entre as razões estão questões político-sociais, chuvas escassas ou irregulares, o aumento de preços dos alimentos e as medidas restritivas impostas pela pandemia da COVID-19 (IPC, dezembro 2021).

Com isso, mais de 1,8 milhões de pessoas encontram-se deslocadas de suas regiões de origem, aumentando a pressão e a demanda por alimentos nos locais acolhedores (IPC, dezembro 2021). Para fazer face a eventos como estes ou para alimentar a população mundial tão crescente, há necessidade de aumentar a produção e produtividade dos alimentos.

O incremento na produção e produtividade pode ser alcançado melhorando as tecnologias de produção. Dentre essas, a semente é o mais importante insumo pois, ela não pode render além do seu potencial. Sementes com alto potencial fisiológico são mais efetivas na mobilização de reservas energéticas, permitindo uma germinação mais rápida

e uniforme, que é a chave para um estabelecimento adequado em campo (FINCH-SAVAGE; BASSEL, 2016; KHAN et al., 2012).

O melhoramento genético de plantas tem sido utilizado com sucesso para aumentar a produtividade de forma sustentável e ecologicamente equilibrada. Melhoristas incorporam tecnologias de manejo avançadas, integrando questões ambientais e sociais, para obter maior produtividade, resistência a doenças, insetos, qualidade nutricional, tolerância a condições adversas do clima e solo, com menor pressão sobre o ambiente (BORÉM; MIRANDA; FRITSCHE-NETO, 2021).

Com apoio de entidades parceiras, programas de desenvolvimento e lançamento de variedades melhoradas tem sido estabelecido em Moçambique. Alguns desses programas incluem demonstrações de variedades de sementes melhoradas para aumentar o conhecimento, a diversidade e a procura dessas sementes de qualidade. Além disso, fornecem aos agricultores subsídios agrícolas para facilitar o acesso a tais variedades e insumos agrícolas, apoiando a expansão de redes de agro comerciantes, para disponibilizarem esses insumos nas áreas rurais (FAO, 2020).

Estudos de avaliação do impacto desses incentivos mostraram que famílias de agricultores beneficiárias de programas como estes obtiveram melhoria substancial da produção e da produtividade, impulsionando atividades de agronegócio e a expansão da economia local. Essas famílias aumentaram a área de cultivo do milho e o rendimento aumentou em até 22%, em relação a famílias não beneficiadas (FAO, 2020).

O milho ocupa o terceiro lugar nas culturas mais produzidas em Moçambique, com pouco mais de 1,63 milhão de toneladas produzidas no ano de 2020 (FAOSTAT, 2021). Embora seja um dos principais alimentos básicos nas comunidades moçambicanas, a produção e produtividade do milho em Moçambique é ainda muito baixa quando comparadas à média de produção dos países na região da África Subsaariana (MANGO et al., 2018).

Nos Estados Unidos da América, China e Brasil que ocupam o primeiro, o segundo e o terceiro lugar na produção mundial de milho, com uma produção de 296,6; 171,8 e 55,9 milhões de toneladas, respectivamente, o aumento da produção nos últimos 20 anos foi impactado em grande parte pela substituição de variedades tradicionais por variedades melhoradas (FAOSTAT, 2021). No Brasil, por exemplo a produção do milho aumentou 193% nos últimos 20 anos (ARTUZO et al., 2019).

No entanto, vários fatores levam à adoção ou não de variedades melhoradas pelos agricultores. Muzari, Gatsi e Muvhunzi (2012) mencionaram que fatores como renda,

vulnerabilidade, conscientização, trabalho e inovação podem constituir a base para a adoção ou não de novas tecnologias. O conhecimento das razões que levam à adoção ou não de tecnologias constituem uma ferramenta útil que auxilia na formulação de políticas e de intervenções mais eficazes para garantir o aumento da produção e produtividade.

Diante disso, técnicas da estatística, como o uso de modelos de regressão na análise do comportamento, auxiliam na tomada de decisões. Em alguns casos, pesquisadores utilizam o modelo de regressão linear clássico (OLS) para modelar e analisar o efeito de diferentes fatores (covariáveis) em estudos sobre a adoção de tecnologias. No entanto, quando os dados são regionalizados, isto é, referenciados espacialmente, deve-se levar em consideração a possibilidade da existência de dependência espacial entre as unidades amostrais. Nesse caso, pode não ser apropriado o uso do modelo OLS (ANSELIN & BERA, 1998).

Para dados referenciados espacialmente é importante que se utilize métodos apropriados de estatística espacial, que são classificados conforme o tipo de dados a serem analisados: (i) padrão de pontos ou processos pontuais; (ii) dados de área (ou látice); (iii) dados de superfície continua, no contexto da geoestatística. Neste artigo, o interesse está no segundo, ou seja, dados de área.

Dados espaciais de área, corresponde a um conjunto formado de áreas (A_i) dentro de um conjunto D, ou seja, $A_i \subset D \in R^2$. Em cada uma das A_i áreas têm-se observações em forma de média, taxas, proporções de um determinado fenômeno, e essas áreas representam municípios, localidades, distritos, dentre outras formas de delimitação espacial. Neste artigo, serão considerados 128 distritos de Moçambique e ajustados modelos de regressão aos dados.

No ajuste de modelos de regressão linear espaciais a dados de área, deve-se considerar estrutura da dependência espacial. Essa estrutura, geralmente é definida por meio de uma matriz de ponderação espacial, definida por matriz W, que estabelece a vizinhança entre áreas presentes no conjunto D, que será considerada na existência da dependência espacial. No entanto, existem diferentes critérios para sua especificação e estas podem levar a resultados diferentes.

Bhattacharjee e Jensen-Butler (2013) e Getis e Aldstadt (2004) ao utilizaram diferentes critérios na especificação de W. Esses autores, concluíram que esta pode afetar a qualidade do ajuste do modelo e, como consequência, reduzir sua eficiência, por subestimar ou superestimar os seus parâmetros. Neste artigo serão utilizados dois critérios para especificação da matriz W para efeitos de comparação.

2 MATERIAL E MÉTODOS

2.1 DADOS UTILIZADOS

Os dados utilizados provêm de uma base secundária dos 128 distritos de Moçambique, obtida do Trabalho do Inquérito Agrícola (TIA) de 2012, uma pesquisa do Ministério de Agricultura em parceria com o Instituto Nacional de Estatística de Moçambique (INE), cuja finalidade é fornecer informações relativas às zonas rurais, tanto a nível distrital e provincial, assim como a nível nacional. Cada um dos distritos representa uma área $A_i \subset D \in R^2, i = 1,\ldots,128$, e contém informações em forma proporção de cada distrito.

A variável resposta (y) é a proporção de agricultores pequenos e médios que fizeram o uso de sementes de milho melhoradas geneticamente, ou seja, y= 100x(sementes melhoradas/agricultores). Uma transformação na variável y foi necessária de modo a garantir a normalidade dos dados por y=log(y). As covariáveis avaliadas estão descritas na Tabela 1.

Tabela 1: Covariáveis avaliadas e que podem afetar a proporção de agricultores que utilizaram sementes de milho melhoradas geneticamente

Covariáveis	n	Média	Desvio padrão	Mínimo	Máximo
Idade da família	128	44,2	4,7	36	56
Tamanho da família	128	5,3	0,9	4	10
Trabalhadores efetivos	128	2,3	2,9	0	10
Trabalhadores eventuais	128	6,4	6,5	0	40
Tração animal	128	12,2	22,9	0	151
Sementes melhoradas	128	4,4	5,7	1	30
Acesso ao crédito	128	0,5	0,9	0	5

Fonte: Dos autores (2022) e adaptado de dados do Trabalho de Inquérito agrícola, 2012.

2.2 MODELOS AJUSTADOS

Foram ajustados três modelos, com base em um modelo geral, dado por:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{W}\boldsymbol{v} + \boldsymbol{\varepsilon}, \quad \textbf{(1)}$$

em que $\boldsymbol{y}$ é o vetor de valores observados da variável resposta, de ordem *n*x1, sendo *n* o número de observações; ρ, λ são os parâmetros de autocorrelação espacial entre observações da variável $\boldsymbol{y}$ e os resíduos respectivamente; **W é** uma matriz de proximidade espacial de ordem *n x n*, que representa a relação entre as áreas dentro do conjunto D ∈ R^2, ou seja, representa a vizinhança que será considerada na dependência espacial; **X** é

a matriz que contém as covariáveis, de ordem *n* x (p+1), onde p < *n* representa o número de parâmetros presentes; **β** é o vetor dos parâmetros de ordem (p+1)x1; $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios independentes e identicamente distribuídos, tal que $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$; $\boldsymbol{\upsilon}$ é o vetor dos resíduos normalmente distribuídos, não independentes, com média zero e matriz de variância covariância $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$, sendo $\mathbf{V} = \left((\mathbf{I} - \rho \boldsymbol{W}')(\mathbf{I} - \rho \boldsymbol{W})\right)^{-1}$ a matriz de autocorrelação espacial entre esses erros, de ordem n x n. No caso dos dados analisados neste estudo, *n = 128* e *p = 7*.

Considerando o modelo em (1) e assumindo que os parâmetros de autocorrelação espacial sejam zeros ou não, derivam-se três submodelos, ou seja, $\rho = 0$ (modelo SEM), $\lambda = 0$ (modelo SAR) e $\rho = \lambda = 0$ (modelo OLS), descritos a seguir, conforme Anselin e Bera (1998) e Bivand et. al (2021).

2.2.1 Modelo SAR

O modelo autorregressivo de defasagem espacial SAR surge quando o parâmetro $\lambda = 0$. Este modelo considera que o processo espacial está presente na variável dependente y em qualquer posição e é uma função entre observações de si mesma com outras ao seu redor ou bem próximas (vizinhança). A vizinhança é definida pela matriz **W**. O modelo é dado por $\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ou de forma equivalente por:

$$y = (\mathbf{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{X}\boldsymbol{\beta} + (\mathbf{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{\varepsilon}, \qquad (2)$$

em que $(\mathbf{I} - \rho \boldsymbol{W})^{-1}$ é a matriz inversa de $(\mathbf{I} - \rho \boldsymbol{W})$, não singular, sendo $\boldsymbol{I}$ uma matriz identidade; ρ o coeficiente de autocorrelação espacial; **W** uma matriz de proximidade espacial de ordem *n x n*; **X** é a matriz que contém as covariáveis, de ordem *n* x (p+1), onde p < *n* representa o número de parâmetros presentes; **β** é o vetor dos parâmetros de ordem (p+1)x1; $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios independentes e identicamente distribuídos, tal que $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$. O valor esperado da variável dependente é $E(\boldsymbol{y}) = (\mathbf{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{X}\boldsymbol{\beta}$; enquanto que a matriz de variância é expressa por $\mathrm{Var(y)} = \left((\mathbf{I} - \rho \boldsymbol{W}')(\mathbf{I} - \rho \boldsymbol{W})\right)^{-1} \sigma^2$. Os parâmetros são estimados pelo método de máxima verossimilhança e são obtidos por $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' A\mathbf{y}$, enquanto que $\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$, sendo que $\boldsymbol{A} = \boldsymbol{I} - \rho \boldsymbol{W}$.

2.2.2 Modelo SEM

O modelo autorregressivo de erros espaciais correlacionados surge quando o parâmetro $\rho = 0$. Esse modelo considera que a dependência espacial está presente entre os resíduos, ou seja, $\boldsymbol{\upsilon} = \lambda \boldsymbol{W\upsilon} + \boldsymbol{\varepsilon}$. O modelo é dado por

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\upsilon}, \qquad (3)$$

em que $\mathbf{X}$ é a matriz que contém as covariáveis, de ordem *n* x (p+1), onde p < *n* representa o número de parâmetros presentes; $\boldsymbol{\beta}$ é o vetor dos parâmetros de ordem (p+1)x1; $\boldsymbol{\upsilon}$ é o vetor dos resíduos normalmente distribuídos, não independentes, com média zero e matriz de variância covariância $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$, sendo $\mathbf{V} = \left((\mathbf{I} - \rho \boldsymbol{W}')(\mathbf{I} - \rho \boldsymbol{W})\right)^{-1}$ a matriz de autocorrelação espacial entre esses erros, de ordem n x n. Diante disso, tem-se $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$. As estimativas dos parâmetros do modelo são obtidas pelo método de máxima verossimilhança, por $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{y}$; enquanto que $\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$.

2.2.3 Modelo OLS

No modelo OLS, as observações e os erros são considerados não correlacionados espacialmente e, assim, os parâmetros de autocorrelação espacial são nulos, ou seja, $\rho = \lambda = 0$. Nesse caso, tem-se o modelo de Gauss-Markov ou de regressão clássico, sem a presença de dependência espacial. Neste, os erros aleatórios são normais, independentes e identicamente distribuídos com média zero e variância constante, ou seja $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. Portanto, a variável **y** tem distribuição normal $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sendo $\boldsymbol{\mu} = \boldsymbol{X\beta}$; e a matriz de variâncias e covariâncias dada por $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. O modelo é dado por

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \qquad (3)$$

em que X é a matriz que contém as covariáveis, de ordem *n* x (p+1), onde p < *n* representa o número de parâmetros presentes; $\boldsymbol{\beta}$ é o vetor dos parâmetros de ordem (p+1) x1; $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios independentes e identicamente distribuídos $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$. Os parâmetros incluídos no modelo são estimados pelo método dos mínimos quadrados ordinários, obtendo-se $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, enquanto $\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$.

Os modelos foram ajustados utilizando o pacote estatístico *spatialreg* (BIVAND; PIRAS 2021) do software R (R Core Team, 2021). Estes foram comparados utilizando o critério AIC, R2 ajustado e o teste dos multiplicadores de Lagrange (*Lagrange multipliers* – LM.). Conforme Anselin & Bera (1998), o teste LM segue uma distribuição qui-quadrada, com um grau de liberdade, sob a hipótese nula de que o parâmetro de autocorrelação espacial é igual a zero (H_0: $\rho = \lambda = 0$). O teste LM para o caso do modelo SEM é LMerr, enquanto que para o modelo SAR é o LMlag. No entanto, em caso de má especificação do modelo, a distribuição do teste será qui-quadrada não centrada, implicando assim na rejeição da hipótese nula mais frequentemente, do que especificado no nível do teste. Diante disso, testes mais robustos de LM são recomendados, como o RLMerr e RLMlag para os modelos SEM e SAR respectivamente (todos os testes LM estão inclusos no pacote *spatialreg* acima mencionado. A regra na escolha do melhor modelo pelo teste LM é a seguinte: 1) na ausência de significância estatística ($p < 0,05$) dos parâmetros ρ, λ, assume-se o modelo OLS; 2) caso o teste LMerr seja significativo e o LMlag não significativo, assume-se o modelo SEM; caso contrário, o modelo SAR; 3) se os testes LMerr e LMlag forem significativos, utiliza-se testes mais robustos RLMerr para o modelo SEM e RLMlag para o modelo SAR, e seleciona-se o modelo com maior significância estatística dos respectivos testes (YWATA & ALBURQUERQUE, 2011; ANSELIN & BERA, 1998).

2.3 COEFICIENTE DE AUTOCORRELAÇÃO ESPACIAL

Coeficientes de autocorrelação espacial são medidas utilizadas e úteis para se avaliar a existência e quantificar a dependência espacial. Neste trabalho, será utilizado o índice de Moran (I) para avaliar a presença de agrupamentos na variável resposta. Será ainda utilizado o índice de Moran local (l_i), denominado por "LISA". Este é mais específico, uma vez que indica a contribuição de uma área A_i no I de Moran global (ANSELIN, 1988). Esse indicador, classifica os agrupamentos ou *clusters* em quatro grupos: Alto-Alto e Baixo-Baixo que representam uma associação positiva entre regiões que apresentam agrupamentos formados por áreas com valores que altos ou baixos respetivamente; ou Alto-Baixo e Baixo-Alto, que correspondem uma associação negativa com agrupamentos formados de áreas com valores altos e baixos ou vice-versa. Esses resultados serão apresentados, de forma visual, através do Lisa Map na identificação de regiões com maior (Alto-Alto) ou menor (Baixo-Baixo) utilização dessas sementes de milho.

2.4 MATRIZ DE PROXIMIDADE OU VIZINHANÇA ESPACIAL

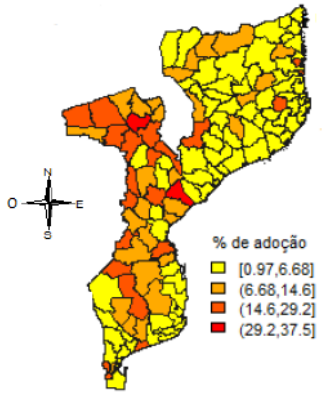
Como pode-se observar, na estrutura dos modelos definidos em (1), (2) e (3) está presente a matriz de proximidade espacial $\boldsymbol{W}$, de ordem *n*x*n*. Esta estabelece a relação de vizinhança ou proximidade entre um par de áreas, com base em critérios já estabelecidos na literatura. Alguns desses critérios são: 1) partilha de fronteira, pelo método de rainha (método mais comum), bispo e torre; 2) distância entre os centroides das áreas A_i e A_j. Entre alguns métodos tem-se a distância inversa; exponencial; *k*-distância sendo k a quantidade de áreas vizinhas próximas de A_i; máxima distância, em que se considera por vizinho toda área dentro de um raio ou distância crítica d_c. Em geral, ela tem sido normalizada na linha de modo que a soma dos elementos na linha seja igual a um (ANSELIN, 1988; CRESSIE, 1993; ALDSDAT e GETIS, 2004; CHEN, 2012).

Neste trabalho foram consideradas duas matrizes, especificadas com base em dois critérios: W1 foi utilizado o critério que considera a quantidade de vizinhos que uma área deve possuir. Nesse caso, definiu-se k = 6 vizinhos; uma segunda W2 foi definida com base em pesos proporcionais ao inverso da distância, ou seja $1/d_{ij}$, em que d_{ij} é a distância euclidiana entre os centroides das áreas A_i e A_j, $d_{ij} = \left\|A_i - A_j\right\|^2$, sendo esta considerada até um limite igual a 100 km. Assim, se d_{ij} for menor ou igual a 100 km, $w_{ij} = 1$; caso contrário, $w_{ij} = 0$. O valor máximo de 100 km foi definido com base no conhecimento prévio da região de estudo D, ou seja, de Moçambique. Ambas as matrizes foram normalizadas nas linhas e para a sua especificação utilizou-se o pacote *spdep* do R (R Core Team, 2021). Todas as demais análises estatísticas foram realizadas utilizando o software R (R Core Team, 2021) e Geoda (2018).

3 RESULTADOS E DISCUSSÃO

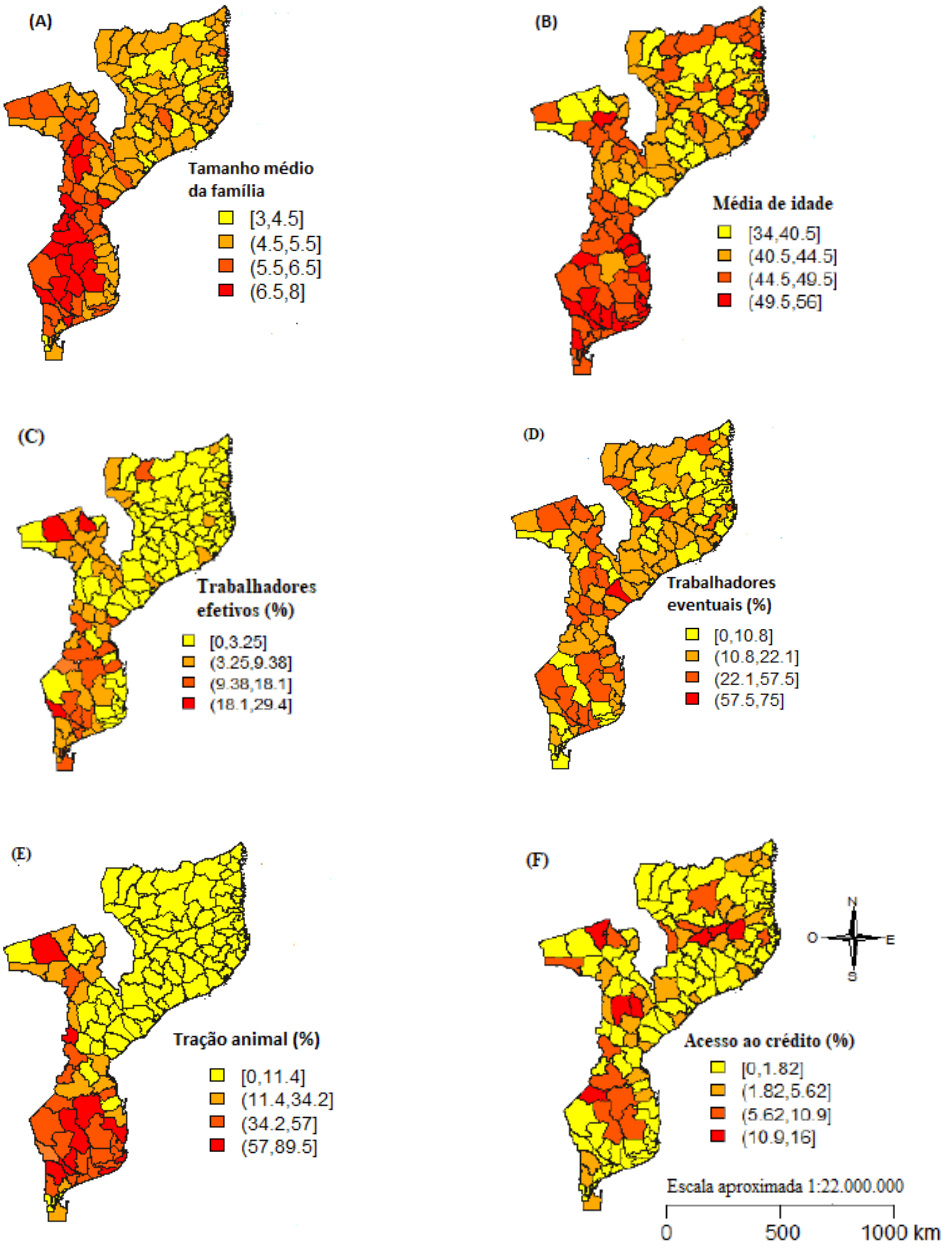
Na Figura 1, tem-se a distribuição da proporção dos agricultores de Moçambique que utilizaram sementes melhoradas de milho no ano de 2012. Uma maior concentração dos agricultores que fizeram uso dessa semente situa-se na região central e centro-oeste, representando entre 29 e 38% de agricultores. Na região do extremo sul e norte, se observa pouco uso de tais sementes abaixo de 7% de agricultores.

Figura 1. Proporção de agricultores que utilizaram sementes melhoradas nos 128 distritos de Moçambique



A visualização da distribuição espacial das covariáveis incluídas no modelo, são apresentadas na Figura 2. Em relação ao tamanho médio de família (FIGURA 2 (A)), o maior número se concentra na região centro-sul, variando entre 6 e 8 membros por família; já para a região norte esse número é menor, entre 3 e 4,5 membros por família. Quanto a idade (FIGURA 2 (B)), os agricultores entre 44 e 56 anos foram os que mais adotaram essas sementes. Esse fato pode ser devido a experiência acumulada ao longo dos anos e na necessidade de experimentar novas tecnologias. Já, entre a idade mais jovem (34 e 44 anos), a aderência a nova tecnologia foi menor. Em relação a mão de obra, os trabalhadores eventuais se distribuem um pouco por todo o país, com maior concentração entre as regiões centro-sul variando entre 60 a 80% (FIGURA 2 (D)). A cultura é sazonal, portanto, em épocas de plantio e colheita, há necessidade de mais trabalhadores. Por outro lado, observa-se na Figura 2 (C) que, entre os trabalhadores efetivos, o maior número situa-se entre 18 a 29%, havendo regiões com poucos ou nenhum trabalhador efetivo, com uma variação entre 0 e 3%. Na Figura 2 (E) observa-se que a região sul é que detêm ou fazem maior uso de animais de tração, variando entre 55 e 90%; já, entre a região central e norte, esse número é bem menor, entre 0 e 4%. Por último, quanto ao acesso ao crédito (FIGURA 2 (F)) Apenas 2,25% de produtores têm em média acesso ao crédito. Por outro lado, um número bem pequeno de distritos tem acesso ao crédito entre 10%, e 16%. Esse fato pode ser devido a fraca disponibilidade por parte de instituições financeiras no investimento no sector agrário, e uma razão pode ser o elevado risco de retorno do capital cedido por essas instituições.

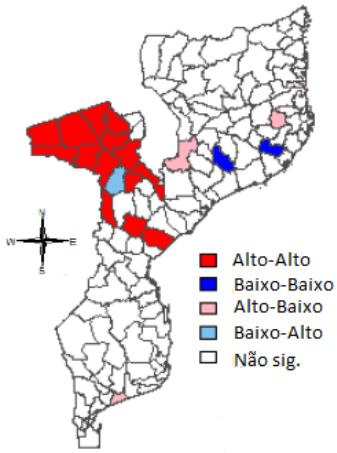
Figura 2. Distribuição espacial das covariáveis: (A) tamanho médio da família; (B) trabalhadores eventuais; (C) tração animal



A estimativa do I de Moran para a proporção dos agricultores que utilizaram sementes melhoradas de milho é de 0,37 ($p < 0,05$), o que corresponde a uma associação positiva e se caracteriza por presença de agrupamentos formados do tipo Alto-Alto ou Baixo-Baixo. No Moran map (FIGURA 3) se observa na região centro-oeste, agrupamentos com a cor vermelha, do tipo Alto-Alto, que corresponde ao maior uso de sementes melhorada de milho. Na região periférica do agrupamento, um distrito se destaca por fazer pouco uso de tais semente entre os distritos que fizeram maior uso (Baixo-Alto). No entanto, também se observa que tal distrito faz fronteira com um outro

que não fez uso de tais sementes (p>0,05). Portanto, é possível que este, seja afetado por tal vizinhança e não se analisou neste artigo as razões para tal situação. Já na região centro-norte, tem-se agrupamento do tipo Baixo-Baixo, que representam áreas que fizeram pouco uso dessas sementes pelos agricultores.

Figura 3. LISA map da proporção de agricultores que utilizaram sementes melhoradas.



Na Tabela 2 têm-se as estimativas dos parâmetros dos três modelos, considerando duas matrizes de vizinhança W1 e W2, além dos critérios de avaliação e comparação dos ajustes dos modelos. Verifica-se que o modelo SAR foi bem ajustado aos dados pelo critério AIC, seguido do modelo SEM e por último o OLS. Quando se compara a qualidade do ajuste em função da matriz de ponderação espacial W, tem-se os seguintes resultados: o desempenho do modelo SAR foi melhor quando se utilizou a matriz W2 (AIC=141,4). Já, para o modelo SEM a matriz W1 mostrou-se aquela que permitiu o melhor ajuste (AIC= 147,1).

Esses resultados (TABELA 2) mostram que a qualidade do ajuste dos modelos autorregressivos espaciais é afetada pela forma como a matriz W é especificada. Getis e Aldstadt (2004) observaram que, ao se ajustar o modelo SAR, especificando a matriz de ponderação espacial com diferentes critérios, a qualidade de ajuste do modelo foi impactada, no sentido de subestimar ou superestimar os seus parâmetros. Embora eles tenham estudado apenas para o modelo SAR, esse resultado também se aplica ao modelo SEM, como foi observado neste trabalho. Conforme Getis (2010) observa, há que se levar em conta o tipo de dados, o fenômeno a ser analisado, entre outros critérios, na escolha de uma matriz W que melhor descreva a realidade em termos de vizinhança e dependência espacial das áreas.

Contudo, de salientar que, a diferença observada entre os modelos SAR e SEM, não necessariamente se deve à matriz W ou aos critérios para sua especificação, mas na própria estrutura de cada modelo. No modelo SEM considera-se que os efeitos espaciais estão presentes na componente dos resíduos e no modelo SAR, considera-se que exista uma relação autorregressiva entre as observações da variável Y, a qual descreve a dependência espacial entre as áreas.

Em relação estimativas dos parâmetros de autocorrelação espacial ρ e λ dos modelos SAR e SEM, respectivamente, sob hipótese nula de ausência de dependência espacial ($H_0: \rho = \lambda = 0$) confirmam a presença de dependência espacial nos dados sobre a proporção dos agricultores que fizeram uso de sementes melhoradas. Para o modelo SAR a estimativa do parâmetro ρ foi de $0{,}28(0{,}08)$ e $0{,}72(0{,}16)$ ao se utilizar as matrizes W1 e W2 respectivamente. Já, para o modelo SEM, as estimativas de λ foram de $0{,}27(0{,}11)$ e $0{,}70(0{,}19)$ respectivamente para as matrizes W1 e W2. Ambos os coeficientes ρ, λ foram significativos a 5% ($p < 0{,}05$).

Tabela 2. Estimativas dos parâmetros dos modelos ajustados e erro padrão (EP) dessas estimativas (entre parênteses) e valores calculados dos critérios de avaliação da qualidade desses modelos

Covariáveis	**SAR**		**SEM**		**OLS**
	W1	**W2**	**W1**	**W2**	
Intercepto	2,449*	1,495*	2,886*	2,817*	2,753*
	(0,463)	(0,53)	(0,451)	(0,463)	(0,471)
Idade da família	-0,017*	-0,015*	-0,017*	-0,014	-0,011
	(0,009)	(0,009)	(0,009)	(0,009)	(0,01)
Tamanho da família	0,133*	0,127*	0,121*	0,127*	0,155*
	(0,042)	(0,042)	(0,041)	(0,042)	(0,045)
Trabalhadores efetivos	0,026	0,025	0,055*	0,041*	0,059*
	(0,021)	(0,019)	(0,021)	(0,02)	(0,02)
Trabalhadores eventuais	0,040*	0,040*	0,039*	0,039*	0,044*
	(0,008)	(0,008)	(0,008)	(0,008)	(0,008)
Tração animal	0,007*	0,008*	0,007*	0,007*	0,009*
	(0,002)	(0,002)	(0,002)	(0,002)	(0,003)
Acesso ao crédito	-0,027	-0,019	-0,015	-0,013	-0,001
	(0,047)	(0,046)	(0,047)	(0,047)	(0,05)
ρ	0,28	0,72			
	(0,08)	(0,16)			
λ			0,27	0,701	
			(0,11)	(0,19)	
R^2 ajustado	0,58	0,59	0.58	0,57	0,54
AIC	143,5	141,4	147,1	147,4	149,4
Teste F					25,423*

*Significativo (valor-p <0,05).

Em relação ao modelo OLS, o teste F foi significativo (valor-p < 0,05), o que indica efeito significativo de pelo menos uma das covariáveis. O coeficiente de determinação ajustado foi $R^2_{ajust.} = 0{,}54$, menor, comparando com os demais modelos. Além disso, avaliando o modelo pelo critério AIC, foi o que apresentou o maior valor dos demais (AIC=149,36). Observa-se uma superestimação dos parâmetros em relação aos modelos SAR e SEM que consideram a dependência espacial. Esses resultados, confirmam que, ao se utilizar modelos clássicos de regressão linear para dados espacialmente dependentes, incorre-se no erro de especificação do modelo (ANSELIN, BERA, 1998).

De modo a identificar qual dos três modelos SAR, SEM e OLS deve ser utilizado, comparou-se o teste dos multiplicadores de Lagrange (LM). (TABELA 3). Ambos modelos apresentam significância estatística no teste LMerr e LMlag (p < 0,05). Portanto, considerou-se os testes mais robustos, no qual, foi significativo apenas para o modelo SAR (p<0,05) com os seguintes resultados RLMlag=5,97 (W1) e RLMlag=14,64 (W2) respectivamente. Portanto, a opção na escolha do modelo a ser utilizado recai para o modelo SAR. Quanto a qual das duas matrizes escolher, pelo critério AIC observou-se que favorecia a matriz W2 (AIC= 141,4) em comparação com a matriz W1 (AIC=143,5) (TABELA 2). Portanto, o modelo SAR tem o melhor desempenho, em termos de qualidade de ajuste, quando se utiliza a matriz W2.

Tabela 3. Teste multiplicador de Lagrange (ML) em que se considera as matrizes W1 e W2

Teste ML	**Estatística**	
	W1	W2
LMerr	3,7482	4,5432*
LMlag	8,6859*	17,7057*
RLMerr	1,0324	1,4742
RLMlag	5,9702*	14,6368*
SARMA	9,7183*	19,1799*

* p-valor < 0.05. Fonte: Dos autores (2022).

Sendo o modelo SAR em que se utilizou a matriz W2 o que apresentou melhor ajuste (pelo critério AIC e o teste LM), será então, considerado as estimativas por ele produzidas. Os resultados apresentados na Tabela 2 são reproduzidos na Tabela 4 que, além das estimativas ($\hat{\beta}$) dos parâmetros, estas são apresentadas por $exp(\hat{\beta})$. Observa-se

que o tamanho da família, trabalhadores efetivos e eventuais, bem como tração animal, apresentaram coeficientes positivos ($\hat{\beta}$'s > 0) e foram 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, ou seja, aumentando-se valores destas, espera-se um aumento na proporção de uso de sementes melhoradas de milho. Portanto, um aumento de uma unidade no tamanho da família, espera-se um aumento em média de 13% no uso de sementes melhoradas.

Tabela 4. Estimativas dos parâmetros do modelo SAR utilizando a matriz W2

Parâmetros	Estimativa ($\hat{\beta}$)	Exp ($\hat{\beta}$)	p-valor
Intercepto	1,50	4,46	0,0050
Idade da família	-0,02	0,99	0,0970
Tamanho da família	0,13	1,14	0,0020
Trabalhadores efetivos	0,03	1,03	0,1970
Trabalhadores eventuais	0,04	1,04	0,0000
Tração animal	0,01	1,01	0,0010
Acesso ao crédito	-0,02	0,98	0,6740

Esse aumento é também verificado no caso da covariável trabalhadores eventuais (4%) e tração animal (1%). Famílias que podem contratar mais trabalhadores, o que é justificado pelo incremento na produção, provavelmente devido ao uso dessas sementes. Conforme afirmam Mignouna et al. (2011) e Bonana-Wabbi (2002) famílias com acesso a animais de tração, são mais propensas em utilizar as sementes em comparação com aquelas sem tração animal. Em relação aos trabalhadores efetivos e acesso ao crédito, não se observou diferenças no efeito sobre uso de sementes melhoradas (TABELA 4).

4 CONCLUSÃO

Este artigo mostrou que o uso do modelo OLS para variáveis regionalizadas, deve ser cauteloso, uma vez que, na possibilidade de existir a dependência espacial, incorre-se no erro de especificação, subestimando ou superestimando a estimativa dos seus parâmetros. Modelos autorregressivos espaciais como SAR e SEM, são mais adequados na presença de dependência espacial entre as áreas. Além disso, deve-se estar atento ao critério utilizado na especificação da matriz de vizinhança espacial W, uma vez que esta matriz pode influenciar na qualidade do ajuste. Sobre qual dos modelos escolher, se SAR e SEM, foi demonstrado utilizando o teste dos multiplicadores de Lagrange.

AGRADECIMENTOS

Os autores agradecem ao Ministério da Ciência e Tecnologia do ensino superior de Moçambique (MCTES), Instituto de Investigação Agronómica de Moçambique (IIAM), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Universidade Eduardo Mondlane (UEM). Um especial agradecimento para o Lourenço Manuel, por disponibilizar os dados parciais.

REFERÊNCIAS

ARTUZO, F. D. et al. O POTENCIAL PRODUTIVO BRASILEIRO: UMA ANÁLISE HISTÓRICA DA PRODUÇÃO DE MILHO. **Revista em Agronegócio e Meio Ambiente**, v. 12, n. 2, p. 151–540, 2019.

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.

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.; 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.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. [s.l.] Oficina de Textos, 2021.

FAO. 2019 FAO MOZAMBIQUE ANNUALREPORT. **Food and Agriculture Organization of the United Nations**, p. 25, 2020.

FAOSTAT. **UN Food and Agriculture Organization statistics [Online]**Available online at http://www.fao.org/faostat, , 2021.

FINCH-SAVAGE, W. E.; BASSEL, G. W. Seed vigour and crop establishment: extending performance beyond adaptation. **Journal of Experimental Botany**, v. 67, n. 3, p. 567–591, 2016.

KHAN, N. et al. Exploring the Natural Variation for Seedling Traits and Their Link with Seed Dimensions in Tomato. **PLOS ONE**, v. 7, n. 8, p. e43991, 2012.

MANGO, N. et al. Maize value chain analysis: A case of smallholder maize production and marketing in selected areas of Malawi and Mozambique. **Cogent Business & Management**, v. 5, p. 1–15, 2018.

MUZARI, W.; GATSI, W.; MUVHUNZI, S. The Impacts of Technology Adoption on Smallholder Agricultural Productivity in Sub-Saharan Africa: A Review. **Journal of Sustainable Development**, v. 5, n. 8, p. 69–77, 2012.

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.

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

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.

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. R Core Team. R: **A Language and Environment for Statistical Computing**. Vienna, Austria, 2021.

GEODA **Center for geospatial analysis and computation.** Version 1.12.1.129 Disponível em: < https://geodacenter.github.io/download_windows.html >. Acesso em: Março, 2022.

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

MIGNOUNA, B. et al. Determinants of adopting imazapyr-resistant maize technology and its impact on household Income in western Kenya: **AgBioforum**, v. 14, n. 3, p. 158-163, 2011.

BONABANA-WABBI, J. **Assessing factors affecting adoption of agricultural technologies: The case of integrated pest management (IPM) in Kumi District, Eastern Uganda**. 2002. 135 p. Msc. Thesis, Virginia Polytechnic Institute and State University, Virginia, 2002.