LOURENÇO MANUEL

SIMULAÇÃO DE DADOS DE ÁREA EM GEE E ANÁLISE DA ADOÇÃO DE VARIEDADES MELHORADAS DE MILHO EM MOÇAMBIQUE

LAVRAS – MG
2019

LOURENÇO MANUEL

SIMULAÇÃO DE DADOS DE ÁREA EM GEE E ANÁLISE DA ADOÇÃO DE VARIEDADES 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, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Doutor.

Prof. Dr. João Domingos Scalon
Orientador

LAVRAS –MG
2019

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

Manuel, Lourenço.

Simulação de dados de área em GEE e análise da adoção de variedades melhoradas de milho em Moçambique / Lourenço Manuel. - 2019.

99 p.

Orientador(a): João Domingos Scalon.

Tese (doutorado) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. Equações de estimação generalizadas. 2. Eficiência relativa. 3. Autocorrelação espacial. I. Scalon, João Domingos. II. Título.

LOURENÇO MANUEL

SIMULAÇÃO DE DADOS DE ÁREA EM GEE E ANÁLISE DA ADOÇÃO DE VARIEDADES MELHORADAS DE MILHO EM MOÇAMBIQUE

SIMULATION OF SPATIAL LATTICE DATA IN GEE AND ADOPTION ANALYSIS OF IMPROVED MAIZE VARIETIES IN MOZAMBIQUE

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

APROVADA em 12 de Março de 2019

Dr. Danilo Machado Pires UNIFAL - MG
Dr. Renato Ribeiro de Lima UFLA
Dra. Thelma Sáfadi UFLA
Dr. Marcelo Ângelo Cirilo UFLA

Prof. Dr. João Domingos Scalon
Orientador

LAVRAS –MG
2019

Aos meus filhos Kaid e Klaus, pelo imensurável amor que sinto por eles.

A minha falecida avó Tacinissa, ao meu falecido pai Manuel Uasse e minha falecida irmã Laura Manuel, que em vida sempre me impulsionaram na carreira estudantil.

Aos meus irmãos Luísa Manuel, Carlota Manuel, Fernando Manuel, Adelina Manuel e Álvaro Manuel que, apesar da distância, demonstraram sempre o seu amor, carinho e amizade.

Aos meus sobrinhos Bernardo Massango, Resalda Massango, Charla Dimbane, Doddy, Kevin, Cheminha, Názira e Bruninho, pelo amor e carinho.

Em especial a minha querida mãe Marta Xavier Manhisse

DEDICO

AGRADECIMENTOS

Em primeiro lugar à Deus pela vida e pela saúde. Em segundo a todos que de forma direta e indireta contribuíram para a efetivação deste trabalho.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES) pela oportunidade concedida para a realização do doutorado.

Ao Professor Doutor João Domingos Scalon pela orientação, amizade e pelas contribuições científicas inestimáveis dadas para o enriquecimento deste trabalho.

Aos membros da banca, os professores, Renato Ribeiro de Lima, Thelma Sáfadi, Danilo Machado, Marcelo Cirilo, Fernando Luíz, Deive Ciro, Marcelo Oliveira e Marcelo Costa pelas contribuições dadas para efetivação do trabalho.

Aos professosres do Departamento de Estatística, do Programa de Pós-graduação em Estatística e Experimentação Agropecuária (PPGEE) pelos conhecimentos transmitidos durante esta caminhada.

À Nádia, secretária do PPGEE pela prestatividade nos assuntos de natureza administrativa.

Aos colegas e amigos do DES, Joel Nuvunga, Neto Pascoal, Joaquim Mazunga, Alberto Jane, Elias, Carol, Edilson, Pablo, Victor, Kelly, dentre outros, pela amizade e convivência que tivemos.

Aos amigos Jonas Massuque, Matias, Mário Tuzine, António Taula, e outros aqui não citados, pela amizade e convivência.

À Jackelya Araújo, pela amizade, respeito, admiração e pelos momentos de descontração.

À Universidade Eduardo Mondlane (UEM) e ao Departamento de Economia e Desenvolvimento Agrário (DEDA) por terem homologado a licença de formação para o nível de doutoramento.

Ao Ministério da Ciência Tecnologia Ensino Superior e Técnico Profissional (MCTESTP) de Moçambique em parceria com o Banco Mundial pela concessão da bolsa de estudos, essencial para esta conquista.

MUITO OBRIGADO!

RESUMO

As equações de estimação generalizadas (EEG), amplamente usadas no estudo de dados longitudinais, constituem uma extensão dos modelos lineares generalizados (MLG) para modelagem de fenômenos com correlação temporal. Sua aplicação em dados espaciais de área é realizada por meio da modelagem da matriz de correlação espacial de trabalho. Essa matriz tem sido construída com base na estrutura do semivariograma, que é uma metodologia aplicada a dados de superfície contínua (geoestatística). Neste trabalho propõe-se a aplicação das EEG para dados espaciais em áreas, por meio da modelagem da matriz de correlação de trabalho, baseada no índice de Moran que é próprio para descrever a estrutura de correlação espacial nesse tipo de dados. Nesta tese é apresentado o desenvolvimento teórico das EEG, assim como os resultados baseados não somente em simulação de dados, mas também em dados reais referentes ao estudo de adoção de variedades melhoradas de milho em Moçambique. Para o caso de dados simulados, foram geradas 1000 amostras de uma variável aleatória no intervalo (0, 1) considerando diferentes valores do índice de Moran. Além disso, foram também geradas, aleatoriamente, duas covariáveis, uma binária e outra contínua. Em cada amostra foram ajustados dois modelos, um ignorando a dependência espacial (MLG) e o outro aplicando as EEG aqui propostas. Para cada amostra foi determinada a eficiência relativa do estimador baseado em EEG em relaçao ao estimador MLG. Além disso, foram usados os critérios de seleção de matriz de correlação de trabalho. Os resultados mostraram que a modelagem baseada na aplicação das EEG aumentou a eficiência dos estimadores, uma vez que uma boa especificação da matriz de correlação de trabalho traduz-se em ganho na diminuição da variância dos estimadores. Em relação aos dados reais, foi usada como variável resposta a proporção de produtores que usou sementes melhoradas de milho e um conjunto de 13 variáveis explicativas identificadas como fatores sócio-demográficos, econômicos, institucionais e tecnológicos. A dependência espacial da variável resposta foi avaliada com base nas estatísticas global e local de Moran. Na modelagem dos dados foram ajustados dois modelos, o modelo logístico e o uso das EEG aqui propostas. Os resultados obtidos usando os dados reais foram consistentes com os resultados obtidos para os dados simulados, isto é, a aplicação das EEG apresentou melhores resultados em relação ao modelo logístico. A modelagem baseada na aplicação das EEG permitiu aferir que a disponibilidade de mão de obra, a idade média do produtor, o uso de tração animal, o acesso a informação, a posse de celeiros melhorados, o acesso ao crédito e aos serviços de extensão, constituem fatores determinantes de adoção de sementes melhoradas de milho em Moçambique. Além disso, a dependência espacial presente na variável resposta também mostrou uma influência positiva significativa na decisão de usar variedades melhoradas de milho.

Palavras chave: Equações de estimação generalizadas. Eficiência relativa. Matriz de correlação espacial de trabalho. Estatística espacial. Autocorrelação espacial. Índice de Moran.

ABSTRACT

Generalized Estimating Equations (GEE) are extension of Generalized Linear Models (GLM), widely applied in longitudinal data analysis. GEE are also applied in spatial data analysis through modelling the working correlation matrix based on semivariogram structure widely applied in random fields (geostatistics). In this paper we propose application of GEE for spatial lattice data modelling the working correlation matrix using the Moran's index which is the most common index used to depict spatial autocorrelation between observations in this type of data. We present results for simulated and real data as well. For the former case, 1000 samples of a random variable defined in (0,1) interval were generated using different values of the Moran's index. In addition, a binary and a continuous variable were also randomly generated as covariates. In each sample two models were fitted, one ignoring the spatial dependency (GLM) and other using the proposed GEE approach. Two measures of model performance were used, the relative efficiency of GEE estimator to its counterpart GLM and the working correlation selection criterions. Results showed that our proposed GEE approach have improved the efficiency of estimators due the well specification of the working correlation matrix. For real data case, the proportion of small farmers who did use improved maize varieties was considered as the response variable and a set 13 variables were used as covariates. These predictors were classified as economic, demographic, institutional and technologic factors. The spatial dependency of the response variable was assessed by global and local Moran indexes. Two models were fitted, the logistic regression model and the GEE here proposed. Results for real data showed consistence with those obtained for simulation study, i.e, the application of GEE proposed has generated better results compared to its competitor. Using the GEE approach for spatial lattice data it was possible to claim that the household size, hired labour, household head age, animal traction, information access, ownership of improved grain storage system, credit and extension services access are the main factors affecting adoption of improved maize varieties in Mozambique. Furthermore, the spatial dependence between observations of the response variable has also showed a significant positive influence in adoption of improved maize seeds.

Key words: Generalized estimating equations. Relative efficiency. Spatial working correlation matrix. Spatial statistics. Spatial autocorrelation. Moran index.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A estatística espacial é um ramo da estatística que estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possam ser entendidos como processos estocásticos, definidos como conjunto de variáveis aleatórias $\{Y(s_i): s_i \in A \subset R^2\}$, em que $Y(s_i)$ é a variável aleatória na coordenada s_i e A corresponde a região de estudo. A grande peculiaridade da estatística espacial é que a informação espacial onde ocorre o fenômeno que está sendo analisado é incorporada nas análises.

Devido ao grande desenvolvimento computacional nos últimos anos, as técnicas de estatística espacial são aplicadas nas mais variadas áreas de conhecimento, tais como ciências agronômicas, epidemiologia, geologia, economia, mineração, entre outras. Cressie (1993) sugere que as técnicas de estatística espacial podem ser agrupadas em três grandes áreas, dependendo do tipo de dados: configurações pontuais, geoestatística e áreas (lattice).

A análise espacial de dados de área é aplicada a fenômenos que se apresentam agregados em unidades geográficas, como municípios, bairros, etc, como é o caso da proporção de produtores que usam sementes melhoradas de milho em Moçambique, cuja informação é apresentada a nível distrital. Nesse tipo de dados, dentre outras hipóteses de interesse, procura-se identificar padrões espaciais do fenômeno em estudo com vista a avaliar a ocorrência da autocorrelação espacial presente nos dados.

No caso de dados de área onde a variável de interesse ocorre simultaneamente com algumas covariáveis, a construção de modelos torna-se crucial no estudo do inter relacionamento que possa existir entre essas covariáveis e a variável resposta, visando uma melhor compreensão do fenômeno em estudo. Com a ocorrência da dependência espacial, o processo de modelagem, além de incluir o efeito das covariáveis, deverá levar em consideração a autocorrelação espacial presente nos dados. Isto permite obter estimativas mais confiáveis, uma vez que a autocorrelação espacial, quando presente, altera o poder explicativo do modelo.

Quando a variável resposta é quantitativa (com distribuição normal), a modelagem em dados de área, na presença da dependência espacial, é feita recorrendo-se ao modelo espacial autoregressivo (SAR) ou o modelo de erro espacial (CAR). O primeiro considera que os efeitos espaciais são atribuídos à variável dependente por meio da inclusão de uma variável “lag” espacializada, enquanto que o segundo atribui os efeitos espaciais por meio de um ruído.

No entanto, quando a variável resposta constitui uma contagem ou proporção, os modelos SAR e CAR não podem ser utilizados. Porém, Manuel et al. (2018) mostram que o uso apropriado de uma transformação na variável resposta permite a aplicação dos modelos SAR e CAR em casos em que a variável resposta constitui uma contagem.

No caso em que a variável resposta corresponde a uma proporção, o processo de modelagem dos dados poderia ser feito recorrendo-se aos Modelos Lineares Generalizados (MLG) usando a distribuição binomial. Contudo, o MLG não pressupõe a incorporação da estrutura de correlação quando esta se encontra presente nos dados. Uma alternativa para contornar o problema é a utilização das Equações de Estimação Generalizadas (EEG). Esse método é uma extensão dos MLG que permite incluir a matriz de covariância que explique a autocorrelação temporal presente nos dados.

As equações de estimação generalizadas foram propostas pela primeira vez por Liang e Zeger (1986) para modelar dados longitudinais (com autocorrelação temporal). Porém, alguns estudos feitos por Lin e Clayton (2005) e Albert e McShaine (1995) mostram que as EEG podem ser aplicadas a dados binários com dependência espacial. Waller e Gotway (2004) também descreveram a aplicação das EEG para dados com estrutura espacial.

Os trabalhos de Lin e Clayton (2005), Albert e McShaine (1995) e Waller e Gotway (2004) descrevem a aplicação das EEG por meio da modelagem da matriz de correlação de trabalho usando uma função de autocorrelação espacial baseada na estrutura do semivariograma que é uma metodologia apropriada para descrever a correlação espacial em dados de superfície contínua (geoestatística).

Assim, esta tese defende a proposta de que a modelagem da matriz de correlação de trabalho pode ser realizada utilizando algum índice de autocorrelação espacial, próprio para dados de área, como é o caso do índice de Moran. Nesse sentido, no presente trabalho propõe-se uma metodologia da construção da referida matriz quando são considerados dados espaciais em áreas por forma a obter maior eficiência dos estimadores dos parâmetros. Além disso, o trabalho propõe o uso do modelo de erro espacial na estimação do parâmetro de associação das EEG.

Os modelos baseados em EEG propostos e desenvolvidos teoricamente neste trabalho foram utilizados para identificar os fatores determinantes do uso de sementes melhoradas da cultura de milho em Moçambique. Foi usada como variável resposta a proporção de produtores que usa sementes melhoradas de milho e um conjunto de 13 variáveis explicativas definidas em fatores sócio-demográficos, econômicos, institucionais e tecnológicos.

1.1 Objetivo geral

- Propor análise de dados de área utilizando Equações de Estimação Generalizadas.

1.2 Objetivos específicos

- Propor a modelagem da matriz de correlação de trabalho utilizando o índice de Moran;
- Propor métodos de simulação para dados de área baseados no índice de Moran;
- Propor a estimação do parâmetro de associação das Equações de Estimação Generalizadas por meio do uso do modelo de erro espacial;
- Utilizar os métodos propostos para identificar os fatores determinantes do uso de variedades melhoradas de milho em Moçambique;
- Utilizar os métodos propostos para descrever o padrão de distribuição espacial da proporção de produtores que usam variedades melhoradas de milho.

2 REFERENCIAL TEÓRICO

Esse capítulo encontra-se subdividido em três partes. Primeiro são abordados aspectos relacionados com a caracterização da cultura do milho em Moçambique bem como os principais trabalhos de adoção de tecnologias agrícolas. Na segunda seção é abordada a análise de dados de áreas e na terceira são apresentados os aspectos teóricos inerentes às Equações de Estimação Generalizadas.

2.1 A cultura do milho em Moçambique e a adoção de tecnologias agrárias

A agricultura constitui um dos mais importantes setores do desenvolvimento sócioeconômico em Moçambique, empregando mais de 80% da população ativa que vive nas zonas rurais e têm a agricultura como a principal fonte de subsistência e de geração de renda familiar. Segundo o INE (2017), o setor de agricultura contribui com cerca de 23% no Produto Interno Bruto (PIB) de Moçambique.

Dentre as principais culturas agrícolas praticadas em Moçambique, o milho é o cereal de maior importância alimentar e econômica no país, seguido do arroz, trigo, sorgo e mapira (DIAS, 2013). Dados mais recentes indicam que a produção nacional de milho foi de aproximadamente 1,2 milhões de toneladas em 2016 enquanto que o arroz contabilizou uma produção de cerca de 154 mil toneladas (INE, 2017).

De acordo com Maculuve (2011), o milho é a cultura alimentar básica para a maior parte da população moçambicana, contribuindo com mais de 40% do total da dieta calórica na nutrição. Sua contribuição nas despesas de consumo em regiões urbanas é de cerca de 13,4%, enquanto que para outros cereais como o arroz e trigo, a contribuição na despesa é estimada em 8,4% e 7,5%, respectivamente (DONAVAN; TOSTÃO, 2010).

O cultivo do milho é feito em todas as regiões do país em sistema de consórcio com outras culturas tais como o feijão e a mandioca, contabilizando cerca de 1/3 do total da área cultivada (MACULUVE, 2011). Dessa área, cerca de 95% é utilizada pelo setor familiar, cuja produção é, na sua maioria, orientada para o consumo (HOWARD et al., 2000). Apesar disso, existem alguns produtores de média e larga escala que produzem milho não só para comercialização em mercados locais mas também para exportação em países vizinhos tais como o Malawi e o Zimbabwe (MUDEMA et al., 2012).

Apesar da grande importância que a cultura de milho representa para o país, seus níveis de produtividade continuam muito abaixo do desejável (cerca de 0,97 ton/ha), quando

comparados com outros países da África sub-sahariana, como por exemplo, a África de Sul, cuja produtividade média do milho está em torno de 3,93 ton/ha (RENAPRI, 2017).

Para países em desenvolvimento, como é o caso de Moçambique, o uso de tecnologias agrícolas melhoradas pode conduzir ao aumento da produção e produtividade das culturas e, consequentemente, uma melhoria da renda familiar nas zonas rurais (UAIENE, 2006).

Uaiene (2009) afirma que em Moçambique encontram-se disponíveis grande parte das tecnologias agrícolas melhoradas, como é o caso das sementes melhoradas de milho. Todavia, apesar da disponibilidade dessas tecnologias, o acesso a elas por parte dos pequenos agricultores é bastante limitado.

A Tabela 1 descreve o percentual de adoção de algumas tecnologias agrárias entre 2002 a 2012.

Tabela 1 - Percentual de adoção de tecnologias agrárias entre 2002 a 2012.

	Anos						
Tecnologia	**2002**	**2003**	**2005**	**2006**	**2007**	**2008**	**2012**
Semente melhoradas de milho	-	-	5,6	9,3	10	9,9	8,7
Fertilizantes	3,8	2,6	3,9	4,7	4,1	4,1	2,8
Pesticidas	6,8	5,3	5,6	5,5	4,2	3,8	6,3
Tração animal	11,4	11,3	9,5	12,8	12	11,3	7,7
Irrigação	10,9	6,1	6	8,4	9,9	8,8	8,1

Fonte: Trabalho de Inquérito Agrícola (2002 a 2012) citado por Cavane et al. (2013).

Dados do último censo agropecuário (2009/2010) indicam que em Moçambique existem cerca de 3,83 milhões de explorações agropecuárias classificadas em pequenas, médias e grandes explorações (INE, 2011). Confrontando essa informação com os dados da da Tabela 1, verifica-se claramente que os percentuais de adoção de tecnologias agrárias no país ao longo dos 10 anos em análise é bastante baixo. Assim, com o propósito de compreender as razões do baixo percentual de adoção de tecnologias agrárias, vários autores procuraram em diferentes estudos identificar os fatores determinates da adoção de tecnologias.

Autores como Mwangi e Kariuki (2015) afirmam que a adoção de tecnolgias agrárias nos países em desenvolvimento está condicionada a um conjunto de fatores de ordem econômica, institucional e humana.

Uaiene (2009), num estudo de determinantes de adoção de tecnologias agrárias em Moçambique, usando o modelo probit, identificou o acesso ao crédito e a participação em alguma associação agrícola como os principais fatores determinantes de adoção para a maior parte das tecnologias agrárias avaliadas.

Zavale et al. (2005), estudando a adoção de variedades melhoradas de milho por parte dos pequenos agricultores em Moçambique, concluiram que uma série de fatores influencia na decisão do produtor adotar uma variedade melhorada. Os autores utilizaram os modelos probit e logit e identificaram os seguintes fatores: nível de educação do produtor, gênero, tamanho do agregado familiar, nível de escolaridade, renda fora da atividade agrária, acesso a eletricidade, acesso ao crédito e a celeiros melhorados.

Cavane e Donavan (2011) aplicaram o modelo logístico num estudo de determinantes de adoção de fertilizantes químicos na província de Manica e identificaram a fonte de informação como o principal fator para decisão do uso de fertlizantes.

Outros autores como Mesfin (2005), Omonona et al. (2005) e Mignouna et al. (2011), reportam o efeito do gênero como um dos fatores determinantes na adoção de tecnologias agrárias pelo fato dos homens possuirem maior acesso aos fatores de produção em relação às mulheres devido a fatores sócio-culturais.

Os trabalhos de adoção de tecnologias agrárias outrora mencionados, não tomam em consideração o fato de produtores localizados em áreas próximas interagirem entre si e influenciarem-se uns aos outros sobre a decisão de adotar uma determinada tecnologia. Nesse sentido, espera-se que numa determinada região, os produtores que se encontram em áreas vizinhas, apresentem um padrão de adoção de tecnologias similar entre si comparativamente aos produtores localizados em áreas mais distantes. Assim, o estudo da adoção de tecnologias melhoradas deve ser feito tendo em conta esse padrão espacial usando métodos de estatística espacial, mais especificamente, métodos específicos para dados de área.

2.2 Análise de dados de áreas

Segundo Assunção (2001), os dados de área referem-se a um mapa particionado em áreas contíguas e disjuntas e, em cada uma delas medem-se uma ou mais variáveis aleatórias $Y(s_i)$ e possivelmente covariáveis de interesse $X(s_i)$, que supostamente afetam a distribuição de probabilidade de $Y(s_i)$, em que s_i é o centróide da área. Em outras palavras, a análise espacial de dados de áreas é adequada em situações nas quais a localização do evento a ser analisado está associado à áreas delimitadas por polígonos. Druck et al. (2004) afirmam que esse caso ocorre com muita frequência quando lidamos com eventos agregados por municípios, bairros ou setores censitários, onde não se dispõe da localização exata dos eventos, mas de um valor agregado por área.

Embora o valor da variável de interesse esteja associado com toda a área e não a um ponto particular, associa-se este valor a um ponto específico dentro da área. Esse ponto pode ser o centróide da área, que corresponde ao centro de massa do polígono que delimita a área (ASSUNÇÃO, 2001).

De acordo com Druck et al. (2004), uma forma usual de apresentação de dados agregados por áreas é por meio do uso de mapas coloridos que representam o padrão espacial do fenômeno analisado. Isso permite, por um lado, uma visualização clara sobre como o fenômeno se "comporta" no espaço, dando a possibilidade de identificar áreas de maior ou menor magnitude de ocorrência desse evento. Por outro lado, permite identificar áreas com características semelhantes ou diferentes desse fenômeno por meio da análise da dependência espacial.

2.2.1 Estrutura da dependência espacial

Uma das suposições básicas feitas na estatística clássica é que as observações de uma variável aleatória são independentes. Em situações em que a variável aleatória de interesse encontra-se espacializada, a suposição de independência entre as observações pode não ser verdadeira, isto é, os valores de uma variável aleatória em distâncias menores, tendem a ser mais parecidos do que observações em distâncias maiores (CRESSIE, 1993). Assim, surge uma necessidade de incorporar nas análises esse grau de similaridade ou dissimilaridade existente entre as observações.

Segundo Waller e Gotway (2004), na análise de dados de área, esse grau de similidaridade ou essa dependência espacial é avaliada utilizando a autocorrelação espacial que pode ser medida por meio de diversos índices tais como Moran, Geary, Getis, entre outros. Além disso, Druck et al. (2004) afirmam que a suavização espacial por meio da média móvel espacial também possui um papel preponderante na determinação do padrão espacial do fenômeno em estudo. Contudo, a aplicação dessa metodologia, assim como dos índices de autocorrelação espacial, dependem da definição de uma matriz de vizinhança ou matriz de proximidade espacial.

2.2.2 Matriz de proximidade espacial

Segundo Werneck (2008), a matriz de proximidade espacial ou de vizinhança (**W**) representa a estrutura da dependência espacial de uma variável aleatória em dados de áreas, ou seja, ela indica o grau de proximidade ou não entre observações.

Vários são os critérios usados para definir a matriz de vizinhança. Dado um conjunto de *n* áreas $\{A_1,..,A_n\}$, constrói-se a matriz **W** (*n x n*), em que cada um dos elementos w_{ij} representa uma medida de proximidade entre A_i e A_j.

Assunção (2001) aborda diferentes critérios utilizados na obtenção de **W**, tais como:

a) $w_{ij} = 1$, se a área A_i compartilha da mesma fronteira com a área A_j ($i \neq j$), $w_{ij} = 0$ caso contrário;

b) $w_{ij} = 1$, se o centróide de A_j dista menos de "*k*" quilômetros do centróide de A_i e $w_{ij} = 0$, caso contrário;

c) $w_{ij} = L_{ij}/L_i$, em que L_{ij} é o comprimento da fronteira entre A_i e A_j e L_i é o perímetro de A_i.

Na Figura 1 ilustra-se a construção da matriz **W** usando como critério a fronteira (critério a) entre as áreas no mapa de Moçambique.

Figura 1 - Matriz de proximidade espacial usando como critério a fronteira entre as áreas (critério a) no mapa de Moçambique, onde: M – Maputo, G – Gaza, I – Inhambane, S- Sofala, Ma - Manica, Z – Zambézia, T – Tete, NP – Nampula, NS – Niassa e CD – Cabo Delgado.

	M	G	I	S	Ma	Z	T	NP	NS	CD
M	0	1	0	0	0	0	0	0	0	0
G	1	0	1	0	1	0	0	0	0	0
I	0	1	0	1	1	0	0	0	0	0
S	0	0	1	0	1	1	1	0	0	0
Ma	0	1	1	1	0	0	1	0	0	0
Z	0	0	0	1	0	0	1	1	1	0
T	0	0	0	1	1	1	0	0	0	0
NP	0	0	0	0	0	1	0	0	1	1
NS	0	0	0	0	0	1		1	0	1
CD	0	0	0	0	0	0	0	1	1	0

Fonte: Do autor (2011).

Autores como Waller e Gotway (2004) e Druck et al. (2004) recomendam a normalização das linhas da matriz **W,** dividindo cada elemento da matriz pelo total da linha, de tal forma que $w_i = \sum_{j=1}^{n} w_{ij} = 1$, ou seja, os pesos w_{ij}, associados com a área A_i, somam um. Isso permite que cada área tenha a mesma contribuição no cálculo dos indicadores de associação espacial tais como o índice de Moran, Geary e a média móvel espacial. Além disso, a normalização permite uma melhor interpretação desses indicadores. A título de

exemplo, a média móvel espacial estimada para uma determinada área será interpretada como a média dos vizinhos dessa área. A Figura 2 ilustra a normalização da matriz **W** nas linhas.

Figura 2 - Matriz de proximidade espacial normalizada nas linhas, onde: M – Maputo, G-Gaza, I – Inhambane, S – Sofala, Ma - Manica, Z – Zambézia, T – Tete, NP – Nampula, NS – Niassa e CD – Cabo Delgado.

	M	G	I	S	Ma	Z	T	NP	NS	CD
M	0	1	0	0	0	0	0	0	0	0
G	1/3	0	1/3	0	1/3	0	0	0	0	0
I	0	1/3	0	1/3	1/3	0	0	0	0	0
S	0	0	1/4	0	1/4	1/4	1/4	0	0	0
Ma	0	1/4	1/4	1/4	0	0	1/4	0	0	0
Z	0	0	0	1/4	0	0	1/4	1/4	1/4	0
T	0	0	0	1/3	1/3	1/3	0	0	0	0
NP	0	0	0	0	0	1/3	0	0	1/3	1/3
NS	0	0	0	0	0	1/3	0	1/3	0	1/3
CD	0	0	0	0	0	0	0	1/2	1/2	0

Fonte: Do autor (2011).

2.2.3 Suavização espacial (média móvel espacial)

A suavização espacial é um método que permite avaliar a tendência espacial que possa existir no estudo de um determinado fenômeno em uma análise de dados de área. Uma forma simples de avaliar essa tendência é calculando a média dos vizinhos (média móvel espacial). Isso permite produzir uma superfície com menor flutuação em relação aos dados originais e por outro lado facilita a identificação de padrões do fenômeno em análise dentro da área de estudo (DRUCK et al., 2004).

Segundo Bailey e Gatrell (1995) a média móvel espacial é definida por:

$$\hat{\pi}_i = \sum_{j=1}^{n} w_{ij} y_j \, ,$$

em que:

$\hat{\pi}_i$ - é o estimador da média móvel na área *i*;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas.

2.2.4 Autocorrelação espacial

A autocorrelação espacial é uma medida que indica o grau de similaridade (dissimilaridade) de uma mesma variável no espaço. A ideia básica da autocorrelação espacial é determinar um indicador que mede a relação de uma determinada variável com ela mesmo no espaço, isto é, como essa variável se comporta numa determinada região geográfica.

Segundo Rogerson e Yamada (2009) e Waller e Gotway (2004), a autocorrelação espacial pode ser medida utilizando vários índices. Neste trabalho são apresentados o índice global de Moran (I), Geary (c) e o índice de Getis e Ord (G).

2.2.4.1 Índice de Moran

O índice de Moran é uma medida de autocorrelação espacial amplamente utilizada em análises espaciais de áreas (ROGERSON; YAMADA, 2009). Autores como Almeida et al. (2008) afirmam que essa medida incorpora a similiaridade entre valores de uma determinada variável avaliada em áreas localizadas a uma distância pré-definida.

Segundo Zhang et al. (2016), o índice de Moran é calculado por:

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

em que:

n - é o número de áreas ou de observações;

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 ($\bar{y} = \sum_{i=1}^{n} y_i/n$);

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas.

Um valor positivo da estatística de Moran, indica que os valores da variável em áreas vizinhas tendem a ser similares entre si (autocorrelação espacial positiva) evidenciando ocorrência de agrupamentos na região de estudo. Já valores negativos indicam dissimilaridade entre os valores dessa variável em áreas vizinhas, ou seja, ocorrência de um padrão regular entre áreas vizinhas. Quando a estatístiva de Moran é proxima de zero, os valores da variável de interesse seguem um padrão aleatório na região de estudo (WALLER; GOTWAY, 2004).

O valor esperado da estatística de Moran na ausência de autocorrelação espacial, conforme Cliff e Ord (1981) é dado por:

$$E\left[\hat{I}\right] = \frac{-1}{n-1}, \tag{2}$$

que se aproxima de zero quando *n* aumenta.

Diferente de outros coeficientes de correlação, como o de Pearson e de Spearman, cujo valor do coeficiente encontra-se no intervalo [-1; 1], Rogerson e Yamada (2009) afirmam que o índice de Moran pode assumir qualquer valor no conjunto dos reais. Porém, na maior parte dos casos ele encontra-se no intervalo [-1; 1].

Uma vez calculado o índice de Moran, é importante fazer inferência sobre ele. De uma forma geral, o índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial, nesse caso, seu valor seria zero (GRIFFITH, 2010).

Segundo Cliff e Ord (1981), o índice de Moran segue assintoticamente sob hipótese de independência espacial uma distribuição Normal com média e variância dadas pelas equações (2) e (3), respectivamente.

$$Var\left[\hat{I}\right] = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2} - \left(\frac{1}{n-1}\right)^2, \tag{3}$$

com $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = {}^1\!/_2 \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2$,

$S_2 = \sum_{i=1}^{n}(w_{i+} + w_{+j})^2$, com $w_{i+} = \sum_{j=1}^{n} w_{ij}$ e $w_{+j} = \sum_{i=1}^{n} w_{ij}$.

Assim, a significância da estatística de Moran pode ser avaliada com base no teste de Wald cuja estatística é dada por:

$$z = \frac{\hat{I} - E[\hat{I}]}{\sqrt{Var[\hat{I}]}}, \tag{4}$$

em que $\hat{I}$ é dado pela equação (1), $E[\hat{I}]$ é dado por (2) e $Var[\hat{I}]$ é dado pela equação (3).

O valor de z obtido na equação (4), corresponde a um determinado quantil da distribuição normal padronizada, que está associado a um determinado v*alor p* estimado. O índice de Moran será considerado significativamente diferente de zero se o *valor p* for inferior ao nível nominal de significância previamente estabelecido. Alternativamente, pode-se fazer inferência sobre o índice de Moran com base no teste de permutação aleatória (WALLER; GOTWAY, 2004). Para tal, constrói-se a distribuição do índice, sob a hipótese de completa aleatoriedade espacial, incluindo o valor estimado da estatística de Moran e determina-se o *valor p*.

Segundo Waller e Gotway (2004), a estimação do *valor p* é feita da seguinte maneira: seja $\hat{I}_{obs}$ a estimativa do índice obtida com base nos valores observados e $\hat{I}_{(1)}, \dots, \hat{I}_{(P)}$ as estimativas do índice obtidas em cada processo de permutação, em que *P* indica o número de permutações. Um histograma construído com base nos valores ordenados de $\hat{I}_{(1)}, \hat{I}_{(2)}, \dots, \hat{I}_{(P)}$ corresponde a estimativa da distribuição do índice sob a hipótese de completa aleatoriedade espacial. Se, $\hat{I}_{(1)} >, \dots, > \hat{I}_{(k)} > \hat{I}_{(obs)} > \hat{I}_{(k+1)}$, isto é, apenas *k* estimativas são superiores ao valor observado do índice, então a estimativa do *valor p* é dada por:

$$valor\ p = \frac{k}{P+1}.$$

Mondini e Chiaravalloti (2008) afirmam que a estatística global de Moran fornece uma ideia global sobre a presença da autocorrelação espacial presente na variável, não identificando as áreas que são mais similares (dissimilares) entre si, ou seja, o índice global de Moran não indica o conjunto de áreas que podem formar agrupamentos. Assim, observa-se a necessidade de evidenciar também um indicador local de autocorrelação espacial denominado "LISA".

2.2.4.2 Índice local de autocorrelação espacial ("LISA")

No índice local de autocorrelação espacial, cada unidade de espaço (área) é caracterizada por um único valor do índice. Segundo Pouliou e Elliott (2009), esse índice indica a contribuição desse local no índice global da autocorrelação espacial (Moran), medida em todas as *n* áreas.

De acordo com Zhang et al. (2016), o "LISA" é calculado por:

$$\hat{I}_i = \frac{(y_i - \bar{y}) \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})}{\sum_{j=1}^{n}(y_j - \bar{y})^2}, \qquad (5)$$

em que:

$\hat{I}_i$ - é o índice de autocorrelação espacial na área *i*;

y_i - é o valor da variável aleatória na área *i*;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas.

Werneck (2008) afirma que esse índice é uma medida de autocorrelação espacial local entre o valor de uma variável numa determinada área e os valores de seus vizinhos, permitindo evidenciar padrões significativos de associação espacial local.

Em outras palavras, pode-se dizer que o “LISA” indica em que medida os valores de alguma observação são similares ou diferentes das observações vizinhas. Isso permite ao índice $\hat{I}_i$ estar associado a cada unidade espacial (área) e requer que haja especificação da matriz de vizinhança ou de proximidade espacial (SANTOS; SOUSA, 2007).

Segundo Almeida et al. (2008), igualmente ao índice global de Moran, a estatística “LISA” não está estritamente limitada ao intervalo [-1; 1], mas seu valor se afasta de zero à medida que aumenta o grau de correlação positiva ou negativa. Na ausência de dependência espacial, esse valor é próximo de zero.

Depois de obtido o índice local de autocorrelação espacial é necessário fazer inferência sobre ele, isto é, avaliar a sua significância por meio do teste de permutação. Flahaut et al. (2002) afirmam que, uma vez avaliada a significância estatística “LISA”, é útil gerar um mapa (“LISA cluster map”) indicando o grupo de áreas que podem formar agrupamentos na região.

2.2.4.3 Índice de Geary (c)

De acordo com Rogerson e Yamada (2009), o índice de Geary é também uma medida de autocorrelação espacial amplamente usada em dados de área. Igualmente ao caso do Moran, a aplicação do índice de Geary também está associada a definição de uma matriz de vizinhança. Segundo Feng et al. (2017), a estatística *c* de Geary é definida por:

$$\hat{c} = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (6)$$

em que:

$\hat{c}$ - é o índice de Geary;

n - é o número de áreas ou de observações;

y_i - é o valor da variável aleatória na área *i*;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas;

$\bar{y}$- é o valor da média amostral da variável aleatória em toda região ($\bar{y} = \sum_{i=1}^{n} y_i/n$).

Diferente do índice de Moran, a estatística “*c*” de Geary encontra-se definida no intervalo [0; 2]. Quando o valor da estatística é próximo de zero há indícios de ocorrência de associação espacial positiva, isto é, valores vizinhos de uma variável tendem a ser similares entre si na área de estudo. Quando a estatística de Geary é proxima de 2, significa ocorrência

de associação espacial negativa, apresentando dissimilaridade entre valores da variável em áreas vizinhas. Para um valor de *"c"* de Geary igual a 1, ocorre a completa aleatoriedade espacial (independência) da variável em estudo (ROGERSON; YAMADA, 2009).

Igualmente ao caso do Moran, o índice de Geary é apenas um indicador global, não identificando os locais que realmente apresentam associação espacial positiva ou negativa, havendo necessidade de determinar um indicar local de associação espacial que é definido de forma análoga ao LISA de Moran.

De acordo com Anselin (1995) a estatística local de Geary é definida por:

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

em que:

$\hat{c}_i$ - é o índice local de Geary na área *i*;

n - é o número de áreas ou de observações;

y_i - é o valor da variável aleatória na área *i*;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial normalizada nas linhas;

$\bar{y}$- é o valor da média amostral da variável aleatória em toda região ($\bar{y} = \sum_{i=1}^{n} y_i / n$).

A inferência sobre a estatística local de Geary é feita construíndo a distribuição da mesma sob a hipótese nula por meio do processo de permutação. Em cada processo de permutação é estimada a estatística "*c*" e são ordenados os diferentes valores obtidos incluíndo o "*c*" exato, construíndo-se uma distribuição sob a hipótese nula. O *valor p* é determinado de forma análoga ao caso do índice de Moran anteriormente descrito.

2.2.4.4 Estatística *G* de Getis e Ord

A estatística *G* de Getis e Ord é outra medida de associação espacial também usada em dados de área. Essa estatística indica em que medida valores de uma variável encontram-se concentrados numa determinada área para uma distância pré definida (GETIS; ORD, 1992).

Diferente dos índices global de Moran e de Geary, as quais indicam apenas a ocorrência de agrupamentos ou não na região, a estatístia *G* permite identificar o tipo de agrupamento existente na região, isto é, permite identificar a ocorrência de "*hot spots*" (locais onde valores altos formam agrupamentos entre si) e "*cold spots*" (locais onde valores baixos formam agrupamentos entre si).

Segundo Getis e Ord (1992) a estatística *G* é definida por:

$$\hat{G} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\, y_i y_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j}, \tag{7}$$

em que:

$\hat{G}$ - é o índice de Getis e Ord;

n - é o número de áreas ou de observações;

y_i - é o valor da variável aleatória na área *i*;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial compostos por 0 e 1.

A interpretação da estatística "*G"* está diretamente ligada ao seu valor esperado. Quando o valor estimado de "*G"* é largamente superior ao seu valor esperado, espera-se que haja ocorrência de "*hot spots*" na região. Já, quando "*G*" é extremamento inferior que o valor esperado, há indicação para a ocorrência de "*cold spots*". Caso o valor da estatística "G" seja próxima do seu valor esperado, então a distribuição da variável aleatória na região de estudo ocorre de forma aleatória. Segundo Getis e Ord (1992) o valor esperado de *G* é dado por:

$$E\left[\hat{G}\right] = \frac{W}{n(n-1)}, \tag{8}$$

em que $W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, para $i \neq j$ e w_{ij} são os elementos da matriz de proximidade espacial.

Igualmente ao caso das estatísticas global de Moran e de Geary, a estatística *G* por si só não é capaz de indicar que conjunto de áreas formam "*hot spots*" ou "*cold spots*", sendo necessário usar indicadores locais. A ideia é similar aos índices locais de Moran e Geary onde para cada área é atribuído um índice. Segundo Getis e Ord (1992) a estatística local G_i é definida por:

$$\hat{G}_i = \frac{\sum_{j=1}^{n} w_{ij}\, y_j}{\sum_{j=1}^{n} y_j}, \quad i \neq j, \tag{9}$$

em que:

$\hat{G}_i$ - é o índice local de Getis e Ord;

y_j - é o valor da variável aleatória na área *j*;

w_{ij} – são os elementos da matriz de proximidade espacial compostos por 0 e 1.

Igualmente ao caso da estatística de Geary, a inferência sobre a estatística local e global de Getis e Ord é realizada com base na construção de uma distribuição, sob a hipótese nula, através do processo de permutação.

2.3 Equações de Estimação Generalizadas (EEG)

As equações de estimação generalizadas (EEG) são uma extensão dos modelos lineares generalizados (MLG) utilizados no estudo de fenômenos em que as observações de uma variável resposta apresentam-se correlacionadas. Liang e Zeger (1986) apresentaram a proposta do uso de equações de estimação generalizadas na modelagem de dados longitudinais.

Seja $\boldsymbol{Y}_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^T$ um vetor n_i x *1* de variáveis resposta e $\mathbf{X}_\mathrm{i} = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i})^T$ uma matriz n_i x p de valores das covariáveis para o i-ésimo indivíduo (i = 1, ... ,K). Assumindo que se conhece apenas a distribuição marginal de Y_{it}, com t = 1, 2,...,n_i, e esta pertencente à família exponencial, isto é, dada por:

$$f(y_{it}; \theta_{it}, \phi) = exp[\phi\{y_{it}\theta_{it} - b(\theta_{it})\} + c(y_{it}, \phi)], \tag{10}$$

em que $E[Y_{it}] = \mu_{it} = b'(\theta_{it})$, $\mathrm{Var}[Y_{it}] = \phi^{-1}\mathrm{V_{it}}$, $\mathrm{V_{it}} = b''(\theta_{it})$, é a função de variância e $\phi^{-1} > 0$ em que ϕ é o parâmetro de dispersão (geralmente desconhecido). Segundo McCullagh e Nelder (1989), define-se um modelo linear generalizado para cada instante *t*, acrescentando a parte sistemática em (10) dada por:

$$g(\mu_{it}) = \eta_{it}\ , \tag{11}$$

em que $g(.)$ é a função de ligação, $\eta_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ é um vetor de parâmetros desconhecidos a serem estimados e $\boldsymbol{x}_{it} = (x_{it1}, \dots, x_{itp})^T$ representa os valores das variáveis explicativas observadas no i-ésimo indivíduo no tempo *t*.

Hardin e Hilbe (2013) destacam que a função escore e a matriz de informação para $\boldsymbol{\beta}$, na ausência de correlação dentro do indivíduo em tempos distintos, são definidos respectivamente, por:

$$\boldsymbol{\Psi_\beta} = \phi \sum_{i=1}^{n} \mathbf{D_i^T V_i^{-1}} (\boldsymbol{y_i} - \boldsymbol{\mu_i}) \tag{12}$$

e

$$\boldsymbol{\Psi_{\beta\beta}} = \phi \sum_{i=1}^{n} \mathbf{D_i^T V_i^{-1} D_i}\ , \tag{13}$$

em que $\mathbf{D_i} = \partial \boldsymbol{\mu_i}/\partial \boldsymbol{\beta_j}$, com $j = 0,...,p$ denota a matriz de derivadas parciais de $\boldsymbol{\mu_i}$ em relação ao vetor de parâmetros $\boldsymbol{\beta}$, $\mathbf{V_i} = diag\{\mathrm{V_{i1}}, \ldots, \mathrm{V_{in_i}}\}$, representa a matriz da função de variância, $\boldsymbol{y_i} = \left(y_{i1}, \ldots, y_{in_i}\right)^T$ é o vetor das observações da variável resposta; $\boldsymbol{\mu_i} = \left(\mu_{i1}, \ldots, \mu_{in_i}\right)^T$ é o vetor de médias e $\boldsymbol{\phi}$ é o parâmetro de dispersão.

Segundo McCullagh e Nelder (1989), as distribuições Normal, Poisson, Binomial e Gama, constituem alguns exemplos das principais distribuições pertencentes a família exponencial e suas características encontram-se descritas na Tabela 2.

Tabela 2 - Exemplos de distribuições pertencentes a família exponencial.

Distribuição	Parâmetro canônico (θ)	Parâmetro de dispersão (ϕ)	Função de variância (V_{it})	Ligação canônica
Normal	μ	σ^2	1	$\mu = \eta$
Poisson	$Log(\mu)$	1	μ	$Log(\mu) = \eta$
Binomial	$Log[\mu(1-\mu)^{-1}]$	1	$\mu(1-\mu)$	$Log[\mu/(1-\mu)] = \eta$
Gama	$-\mu^{-1}$	ν^{-1}	μ^2	$\mu^{-1} = \eta$

Fonte: McCullagah e Nelder (1989).

Quando utiliza-se a função de ligação canônica, as funções escore e matriz de informação ficam definidas por:

$$\boldsymbol{\Psi_\beta} = \phi \sum_{i=1}^{n} \mathbf{X_i^T}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) \tag{14}$$

e

$$\boldsymbol{\Psi_{\beta\beta}} = \phi \sum_{i=1}^{n} \mathbf{X_i^T V_i X_i}\,, \tag{15}$$

respectivamente, em que $\boldsymbol{\phi}$, $\mathbf{V}_\mathrm{i}$, $\boldsymbol{y}_\mathrm{i}$ e $\boldsymbol{\mu}_\mathrm{i}$ são definidos tal qual em (12) e (13) e $\mathbf{X}_\mathrm{i}$ é uma matriz n_i x p contendo as observações das covariáveis (MCCULLAGH; NELDER, 1989). A estimação dos parâmetros é feita igualando a zero a função escore ($\boldsymbol{\Psi_\beta} = \mathbf{0}$) e resolvendo o sistema de equações daí resultante. Esse estimador é consistente e assintoticamente normal (HARDIN; HILBE, 2013).

Supondo que a distribuição de Y_{it} não é conhecida e que V_{it} é uma função da média, porém não caracteriza a distribuição de Y_{it}, ou seja, não se conhece a distribuição de Y_{it}, porém assume-se que a média $\boldsymbol{\mu_i}$ é uma função de um vetor de parâmetros desconhecidos $\boldsymbol{\beta}$ e que a variância é proporcional a média, então tem-se um modelo de quasi-verosimilhança (WEDDERBURN, 1974).

De acordo com Wedderburn (1974) a função log quasi-verosimilhança é descrita como:

$$Q(\mu, y) = \int_y^{\mu} \frac{y - t}{\phi V(t)} dt \, .$$

Para o caso das funções da média apresentadas na Tabela 2, Wedderburn (1974) mostrou que a função de quase verossimilhança tem as mesmas características com a função de verossimilhança, isto é, com base na função de quase verossimilhança é possível recuperar a função de verossimilhança para a maior parte das distribuições conhecidas pertencentes a família exponencial, ou seja $Q(\boldsymbol{\mu}; \boldsymbol{y}) \propto L(\boldsymbol{\mu}; \boldsymbol{y})$. A título de exemplo, considerando uma amostra de tamanho n = 1 tem-se:

i. Distribuição Normal

Se Y é uma variável aleatória com distribuição $N(\mu, \phi)$, então o logarítmo da função de verossimilhança é dado por:

$$l(\mu, y) = -\tfrac{1}{2\phi}(y - \mu)^2 - \ln(\sqrt{\phi 2\pi}).$$

A função de variância é descrita por V(μ) = 1. Assim, a função log quasi-verossimilhança é definida por:

$$Q(\mu, y) = \int_y^{\mu} \frac{y - t}{\phi V(t)} dt = \frac{1}{\phi}\left(yt - \frac{t^2}{2}\right)|_y^{\mu} = \frac{1}{\phi}\left(y\mu - \frac{\mu^2}{2} - y^2 + \frac{y^2}{2}\right) = -\frac{1}{2\phi}(y - \mu)^2,$$

que é proporcional ao logarítmo da função de verossimilhança de uma variável aleatória Y com distribuição $N(\mu, \phi)$.

ii. Distribuição Binomial

A função log verossimilhança de uma variável aleatória $Y \sim Bin(n, \mu)$ é dada por:

$$l(\mu, y) = y * \ln\left(\frac{\mu}{1 - \mu}\right) + n * \ln(1 - \mu) + \ln\binom{n}{y}.$$

Para essa distribuição, a função de variância é definida por $V(\mu) = \mu(1 - \mu)$. Deste modo, o logarítmo da função quase verossimilhança é dado por:

$$Q(\mu, y) = \int_y^{\mu} \frac{y - t}{\phi V(t)} dt = \frac{1}{\phi}\int_y^{\mu} \frac{y - t}{t(1 - t)} dt = \frac{1}{\phi}\int_y^{\mu} \frac{y}{t(1 - t)} + \frac{1}{\phi}\ln(1 - t)\,|_y^{\mu}$$

$$= \frac{1}{\phi} y \ln\left(\frac{t}{1-t}\right)|_y^\mu + \frac{1}{\phi}\ln(1-t)|_y^\mu$$

$$= \frac{1}{\phi}\left\{y\left[\ln\left(\frac{\mu}{1-\mu}\right) - \ln\left(\frac{y}{1-y}\right)\right] + \ln(1-\mu) - \ln(1-y)\right\}$$

$$= \frac{1}{\phi}\{y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu) - y\ln\left(\frac{y}{1-y}\right) - \ln(1-y)\}.$$

Assumindo que $\phi = 1$, tem-se:

$$Q(\mu, y) = y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu) - y\ln\left(\frac{y}{1-y}\right) - \ln(1-y),$$

que é proporcional ao logarítmo da função de verossimilhança de uma variável aleatória Y com distribuição binomial.

iii. Distribuição de Poisson

Seja a variável aleatória $Y \sim P(\mu)$. A função log verossimilhança é definida por:

$$l(\mu, y) = y\ln(\mu) - \mu - \ln(y!).$$

A função de variância para a distribuição de Poisson é dada por $V(\mu) = \mu$. Assim, a funcão log quase verossimilhança será dada por:

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\phi V(t)}dt = \frac{1}{\phi}\int_y^\mu \frac{y-t}{t}dt = \frac{1}{\phi}\{y\ln\left(\frac{\mu}{y}\right) + y - \mu\}.$$

Assumindo que $\phi = 1$, tem-se:

$$Q(\mu, y) = y\ln(\mu) - \mu + y[1 - \ln(y)],$$

que é proporcional ao logarítmo da função log verossimilhança de uma variável aleatória $Y \sim P(\mu)$.

iv. Distribuição Gama

Seja a variável aleatória $Y \sim G(\mu, \phi)$. A função de densidade de probabilidade de Y é definida por:

$$f(y;\mu,\phi) = \frac{1}{\Gamma(\phi)}\mu^{-\phi}y^{\phi-1}\exp\left(-\frac{y}{\mu}\right).$$

O logarítimo da função de verossimilhança é dado por:

$$l(\mu;\phi,y) = -\frac{y}{\mu} - \phi\ln(\mu) + (\phi-1)\ln(y) - \ln\{\Gamma(\phi)\}.$$

A função de variância para a distribuição Gama é definida por $V(\mu) = \mu^2$. Assim, a função log quase verossimilhança é dada por:

$$Q(\mu,y) = \int_y^\mu \frac{y-t}{\phi V(t)}dt \int_y^\mu \frac{y-t}{\phi t^2}dt = \frac{1}{\phi}\frac{-y}{t}|_y^\mu - \frac{1}{\phi}\ln(t)\,|_y^\mu = \frac{1}{\phi}\{1 - \frac{y}{\mu} - \ln(\mu) + \ln(y)\}.$$

Assumindo que $\phi = 1$, tem-se:

$$Q(\mu,y) = -\tfrac{y}{\mu} - \ln(\mu) + \ln(y) + 1,$$

que é proporcional ao logarítmo da função log verossimilhança de uma variável aleatória $Y \sim G(\mu, 1)$.

Portanto, as equações quasi-escore e matriz quasi-informação são definidas segundo as equações (12) e (13) quando ignorada a correlação entre tempos distintos. Quando os dados apresentam uma estrutura correlacionada, o que deve ser levado em consideração na função quasi-escore, resultando num novo sistema de equações para estimar $\boldsymbol{\beta}$, conhecidas como equações de estimação generalizadas (LIANG; ZEGER, 1986).

Seja a matriz de variâncias e covariâncias para $\boldsymbol{Y_i}$ definida por:

$$\mathrm{Var}(\boldsymbol{Y_i}) = \phi^{-1}\mathbf{V_i^{1/2}R_iV_i^{1/2}}, \tag{16}$$

em que $\mathbf{R_i}$ é a matriz de correlação para o i-ésimo grupo e $\mathbf{V_i} = diag\{\mathrm{V_{i1}}, \ldots, \mathrm{V_{in_i}}\}$. No caso de dados não correlacionados a matriz de correlação ($\mathbf{R_i}$) será igual a matriz identidade ($\mathbf{R_i} = \mathbf{I}$), portanto $\mathrm{Var}(\boldsymbol{Y_i}) = \phi^{-1}\mathbf{V_i}$. Desse modo, a equação (16) representa uma generalização do MLG (LIANG; ZIGER, 1986).

Liang e Zeger (1986) propuseram uma matriz de correlação dada por $\mathbf{R_i(\alpha)}$, em que $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$ é um vetor de parâmetros de correlação que não dependem de $\boldsymbol{\beta}$.

Assim, substituindo $\mathbf{R_i}$ por $\mathbf{R_i(\alpha)}$ na equação (16), tem-se que a matriz de variâncias e covariâncias de $\boldsymbol{Y_i}$ será dada por:

$$\boldsymbol{\Omega_i} = \phi^{-1}\mathbf{V_i^{1/2}}\mathbf{R_i}(\boldsymbol{\alpha})\mathbf{V_i^{1/2}}, \tag{17}$$

em que $\mathbf{R_i(\alpha)}$ representa a verdadeira matriz de correlação entre os elementos de $\boldsymbol{Y_i}$, parametrizada pelo vetor $\boldsymbol{\alpha}$. Contudo, Liang e Zeger (1986) afirmam que, em situações práticas, a verdadeira matriz de correlação não é conhecida e desse modo $\mathbf{R_i(\alpha)}$ é definida como uma matriz de correlação de trabalho que depende de um número finito de parâmetros $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$. A estimação desses parâmetros é feita pelo método dos momentos cujos estimadores dependem do tipo de matriz de correlação de trabalho usada, como será ilustrado na seção 2.3.1.

Para estimar $\boldsymbol{\beta}$ resolve-se o seguinte sistema de equações:

$$\boldsymbol{\Psi_\beta}(\widehat{\boldsymbol{\beta}}_G) = \mathbf{0}, \tag{18}$$

denominado equações de estimação generalizadas (EEG), em que $\boldsymbol{\Psi_\beta}(\widehat{\boldsymbol{\beta}}_G) = \sum_{i=1}^{n} \mathbf{D_i^T}\boldsymbol{\Omega_i^{-1}}(\boldsymbol{y_i} - \boldsymbol{\mu_i})$.

2.3.1 Estimação dos parâmetros

Nessa abordagem metodológica existem 3 tipos de parâmetros a serem estimados: o vetor dos parâmetros de posição ($\boldsymbol{\beta}$), o parâmetro de dispersão (ϕ) e o vetor de parâmetros de correlação ($\boldsymbol{\alpha}$).

a. Estimação dos parâmetros de posição ($\boldsymbol{\beta}$)

O sistema de equações definido em (18) é não linear, necessitando de métodos iterativos para a sua resolução (LIANG; ZEGER, 1986). Assim, dadas as estimativas de $\widehat{\boldsymbol{\alpha}}$ e $\hat{\phi}$, obtidas por meio do método dos momentos, para a estimação dos parâmetros de posição em equações de estimação generalizadas é aplicado o método de escore Fisher que é dado por:

$$\boldsymbol{\beta}_G^{(m+1)} = \boldsymbol{\beta}_G^{(m)} + [\sum_{i=1}^{n} \mathbf{D_i}^{\mathrm{T(m)}}\boldsymbol{\Omega_i}^{-(\mathrm{m})}\mathbf{D_i}^{(\mathrm{m})}]^{-1}.[\sum_{i=1}^{n} \mathbf{D_i}^{\mathrm{T(m)}}\boldsymbol{\Omega_i}^{-(\mathrm{m})}(\boldsymbol{y_i} - \boldsymbol{\mu_i}^{(m)})], \tag{19}$$

com *m = 0, 1,...* definindo o número de iterações necessárias até a convergência.

Liang e Zeger (1986) afirmam que o processo de estimação definido em (19) é equivalente a aplicação de um processo iterativo de mínimos quadrados reponderados de uma regressão entre uma variável dependente modificada *Z* sobre **D**.

Seja $\mathbf{D} = (\mathbf{D_1}, \dots, \mathbf{D_k})^{\mathbf{T}}$ uma matriz de derivadas parciais de $\boldsymbol{\mu_i}$ em relação ao vetor de parâmetros $\boldsymbol{\beta}$, $\boldsymbol{s} = (\boldsymbol{s_1^T}, \dots, \boldsymbol{s_k^T})$ em que $\boldsymbol{s_i} = (y_{i1} - \mu_{i1}, \dots, y_{in_i} - \mu_{in_i})$ com *i = 1,2,...,k* um vetor de resíduos e $\boldsymbol{\Sigma}$ uma matriz bloco diagonal *nk x nk* cujos elementos da diagonal são compostos por $\boldsymbol{\Omega_i}$. Define-se uma variável dependente modificada *Z* dada por:

$$Z = \boldsymbol{D\beta}^{(m)} + \boldsymbol{S}. \tag{20}$$

A solução do processo iterativo descrito em (19) é equivalente a solução iterativa de mínimos quadrados reponderados usando $\boldsymbol{\Sigma^{-1}}$ como a matriz de pesos, isto é:

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{D^T \Sigma^{-1} D})^{-1} \mathbf{D^T \Sigma^{-1} Z}.$$

Liang e Zeger (1986) mostraram que o estimador de parâmetros $\widehat{\boldsymbol{\beta}}$ é consistente e assintoticamente normal. Além disso, quando a matriz de correlação de trabalho é especificada corretamente (conhecida), o estimador da matriz de variâncias e covariâncias do estimador dos parâmetros de posição será ótimo e definido por:

$$\widehat{\mathbf{V}}_{\mathbf{opt}} = \mathbf{H_1^{-1}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}), \tag{21}$$

em que $\mathbf{H_1}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\widehat{\mathbf{D}}_{\mathbf{i}}^{\mathbf{T}} \widehat{\boldsymbol{\Omega}}_{\mathbf{i}}^{\mathbf{-1}} \widehat{\mathbf{D}}_{\mathbf{i}})$.

Se a matriz de correlação de trabalho não é definida corretamente, o estimador dado em (21) pode ser inconsistente. Assim, Liang e Zeger (1986) propuseram o uso de um estimador robusto definido por:

$$\widehat{\mathbf{V}}_{\mathbf{G}} = \mathbf{H_1^{-1}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}) \mathbf{H_2}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}) \mathbf{H_1^{-1}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}), \tag{22}$$

em que $\mathbf{H_2}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \{ \widehat{\mathbf{D}}_{\mathbf{i}}^{\mathbf{T}} \widehat{\boldsymbol{\Omega}}_{\mathbf{i}}^{\mathbf{-1}} (\boldsymbol{y_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}) (\boldsymbol{y_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})^{\boldsymbol{T}} \widehat{\boldsymbol{\Omega}}_{\mathbf{i}}^{\mathbf{-1}} \widehat{\mathbf{D}}_{\mathbf{i}} \}$ e $\mathbf{H_1}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}})$ é definido em (21).

Autores como Brajendra e Kalyan (1999), assim como Wang e Carey (2003), afirmam que uma correta especificação da matriz de correlação de trabalho irá se traduzir em maior ganho na eficiência do estimador no processo de estimação. Portanto, quando a matriz de correlação de trabalho coincide com a verdadeira matriz de correlação, o estimador será ótimo. Contudo, em situações práticas a verdadeira matriz de correlação não é conhecida,

sendo necessário estimá-la. Assim, o uso de diferentes matrizes de correlação de trabalho irá conduzir a diferentes eficiências.

Wang e Lin (2005) definem o conceito de eficiência relativa do estimador em EEG como a razão $\hat{V}_{opt}/\hat{V}_G$,definidos em (21) e (22), para cada parâmetro do modelo. Por outro lado, Seber (2008) define a eficiência relativa do estimador $\hat{\gamma}_1$ em relação ao estimador $\hat{\gamma}_2$ como:

$$ER = \text{tr}[\text{Var}(\hat{\boldsymbol{\gamma}}_{\mathbf{1}}) - \text{Var}(\hat{\boldsymbol{\gamma}}_{\mathbf{2}})], \tag{23}$$

em que tr[.] é o traço da matriz e Var(.) é a matriz de covariância dos estimadores $\hat{\gamma}_2$ e $\hat{\gamma}_1$.

b. Estimação do parâmetro de dispersão (ϕ)

Segundo Liang e Zeger (1986) um estimador consistente para o parâmetro de dispersão é dado por:

$$\hat{\phi} = \frac{1}{(N-p)} \sum_{i=1}^{k} \sum_{j=1}^{ni} \frac{(y_{ij}-\mu_{ij})^2}{\hat{V}_{ij}}, \tag{24}$$

em que N é o número de observações e *p* é o número de parâmetros do modelo.

c. Estimação do vetor de parâmetros de correlação (***α***)

Liang e Zeger (1986) apresentaram quatro estruturas de matrizes de correlação de trabalho, nomeadamente a matriz permutável, a autoregressiva AR(1), M-dependent ("Toeplitz") e a não estruturada. Assim, o estimador de ***α*** irá depender da escolha e estrutura de matriz de correlação de trabalho a ser utilizada.

i. Matriz de correlação de trabalho permutável

Nesse tipo de matriz, considera-se que as ordens das unidades dentro do indivíduo não importa, pois assume-se que a correlação dentro dos indivíduos é a mesma. Neste caso, assume-se que $\mathbf{R_i} = \mathbf{R_i}(\boldsymbol{\alpha})$ em que *(j,j´)* – ésimo elemento da matriz $\mathbf{R_i}$ fica determinado por $r_{ijj'} = 1$, para $j = j'$ e $r_{ijj'} = \alpha$, para $j \neq j'$. Segundo Liang e Zeger (1986), um estimador consistente para *α*, é dado por:

$$\hat{\alpha} = \left\{ \sum_{i=1}^{K} \sum_{j \neq j'} \frac{(y_{ij}-\hat{\mu}_{ij})}{\sqrt{\hat{V}_{ij}}} \frac{(y_{ij}-\hat{\mu}_{ij'})}{\sqrt{\hat{V}_{ij'}}} \right\} \Big/ \left\{ \sum_{i=1}^{K} \frac{n_i(n_i-1)}{2} \right\}, \tag{25}$$

em que n_i é o número de unidades dentro do indivíduo. A matriz de correlação de trabalho para a estrutura permutável é apresentada da seguinte forma:

$$R_i = \begin{bmatrix} 1 & \alpha & ... & \alpha \\ \alpha & 1 & ... & ... \\ ... & ... & ... & \alpha \\ \alpha & ... & \alpha & 1 \end{bmatrix}.$$

ii. Matriz de correlação de trabalho autoregressiva, AR(1)

Na matriz de correlação de trabalho para a estrutura de correlação AR(1) a ordem das unidades dentro do indivíduo tem relevância. As observações em tempos mais próximos tendem a apresentar uma correlação maior em relação às medidas em tempos distantes, isto é, a correlação tende a diminuir com o tempo. Assim, tem-se que $\mathbf{R_i} = \mathbf{R_i}(\alpha)$ em que o (j,j') – ésimo elemento da matriz $\mathbf{R_i}$ é dado por $r_{ijj'} = 1$, para $j = j'$ e $r_{ijj'} = \alpha^{|j-j'|}$, para $j \neq j'$. Um estimador consistente para α conforme descrito por Liang e Zeger (1986), é dado por:

$$\hat{\alpha} = \frac{1}{K}\sum_{i=1}^{K} \frac{1}{(n_i-1)} \sum_{j=1}^{ni-1} \frac{(y_{ij}-\hat{\mu}_{ij})}{\sqrt{\hat{v}_{ij}}} \frac{(y_{i(j+1)}-\hat{\mu}_{i(j+1)})}{\sqrt{\hat{v}_{i(j+1)}}}. \quad (26)$$

A matriz de correlação de trabalho para a estrutura AR(1) é apresentada da seguinte forma:

$$R_i = \begin{bmatrix} 1 & \alpha & ... & \alpha^{|1-n|} \\ \alpha & 1 & ... & ... \\ ... & ... & ... & ... \\ \alpha^{|1-n|} & ... & ... & 1 \end{bmatrix}.$$

iii. Matriz de correlação de trabalho m-dependente (“toeplitz”)

A matriz de correlação de trabalho do tipo “toeplitz” apresenta uma estrutura mais flexível. Nessa estrutura, assume-se que as observações igualmente espaçadas no tempo apresentam a mesma correlação. Desse modo, as estruturas de correlação de trabalho permutável e AR(1), são casos particultares da estrutura “Toeplitz” (JANG, 2011). Para esse tipo de matriz de correlação, tem-se que $\mathbf{R_i} = \mathbf{R_i}(\alpha)$ em que em que (j,j') – ésimo elemento da matriz $\mathbf{R_i}$ fica dado por $r_{ijj'} = 1$, para $j = j'$ e $r_{ijj'} = \alpha_j$, com $j = 1, 2,...,n-1$, para $j \neq j'$. Um estimador consistente para α_j, é dado por:

$$\hat{\alpha_j} = \frac{\phi}{K-p}\sum_{i=1}^{K} \frac{(y_{ij}-\hat{\mu}_{ij})}{\sqrt{\hat{v}_{ij}}} \frac{(y_{i(j+1)}-\hat{\mu}_{i(j+1)})}{\sqrt{\hat{v}_{i(j+1)}}}, \quad (27)$$

em que K é o número de indivíduos (LIANG; ZEGER, 1986).

A matriz de correlação de trabalho “Toeplitz” é apresentada da seguinte forma:

$$R_i = \begin{bmatrix} 1 & \alpha_1 & \dots & \alpha_{n-1} \\ \alpha_1 & 1 & \dots & \dots \\ \dots & \dots & \dots & \alpha_1 \\ \alpha_{n-1} & \dots & \alpha_1 & 1 \end{bmatrix}.$$

iv. Matriz de correlação de trabalho não estruturada

Quando a matriz de correlação de trabalho é não estruturada tem-se $n_i(n_i - 1)/2$ parâmetros a serem estimados. Denotando $\mathbf{R_i} = \{\boldsymbol{\alpha_{jj'}}\}$, o *(j,j´)*-ésimo elemento de $\mathbf{R_i}$ pode ser estimado por :

$$\hat{\alpha}_{jj'} = \frac{1}{K}\sum_{i=1}^{K} \frac{(y_{ij} - \hat{\mu}_{ij})}{\sqrt{\hat{V}_{ij}}} \frac{(y_{ij'} - \hat{\mu}_{ij'})}{\sqrt{\hat{V}_{ij'}}}. \qquad (28)$$

A matriz de correlação de trabalho não estruturada é apresentada da seguinte forma:

$$R_i = \begin{bmatrix} 1 & \alpha_{1,2} & \dots & \alpha_{1,n} \\ \alpha_{2,1} & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \alpha_{n,1} & \dots & \dots & 1 \end{bmatrix}.$$

Hardin e Hilbe (2013), para além das estruturas de correlação descritas por Liang e Zeger (1986), abordam também o uso da matriz de correlação de trabalho fixa. Para esse caso, os autores reportam que essa especificação é feita quando o pesquisador possui alguma fonte de conhecimento da estrutura de correlação do fenômeno em estudo. Isso conferi uma flexibilidade aos pesquisadores em usar outras estruturas de correlação diferentes daquelas que foram anteriormente citadas.

Waller e Gotway (2004), Lin e Clayton (2005) e Albert e McShaine (1995) descrevem o uso da estrutura de correlação fixa para dados espaciais baseada na estrutura do semivariograma, amplamente usada na metodologia geoestatística.

Nesta tese propõe-se usar a estrutura de correlação espacial fixa baseada no índice de Moran ou índice de Geary amplamente usados para descrever a estrutura de correlação espacial em dados de áreas.

A existência de várias estruturas de matriz de correlação de trabalho introduziu, de certa forma, a subjetividade na escolha da melhor matriz de correlação de trabalho. Por essa razão, houve necessidade de desenvolvimento de critérios que auxiliam na escolha da melhor estrutura de correlação a ser usada. Com o propósito de minimizar essa subjetividade,

diversos autores como Pan (2001), Hin e Wang (2009) e Rotnitzky e Jewell (1990) desenvolveram diferentes critérios de seleção de matriz de correlação de trabalho.

2.3.2 Critérios de seleção de matriz de correlação de trabalho

Dentre os critérios de seleção de matriz de correlação de trabalho existentes na literatura, os critérios QIC, CIC e RJC são os que se encontram amplamente difundidos nas Equações de Estimação Generalizadas (INAN et al., 2018).

i. Critério QIC

O critério de informação de Akaike (AIC) é uma das medidas amplamente usadas no processo de seleção de modelos baseados na função de verossimilhança. Pan (2001), usando a mesma abordagem da definição do AIC, desenvolveu o critério de informação de quasi-verossimilhança (QIC) que é dado por:

$$\mathrm{QIC(R)} = -2Q(\hat{\mu}) + 2\mathrm{tr}(\mathbf{V_I^{-1}V_{G,R}}),$$

em que $Q(\hat{\mu})$ é a estimativa de log quasi-verossimilhança usando as estimativas do modelo; tr[.] é o traço da matriz; $\mathbf{V_I}$ é a matriz de covariâncias dos estimadores dos parâmetros usando a estrutura independente (MLG); $\mathbf{V_{G,R}}$ é a matriz de covariâncias usando o estimador robusto definido pela equação (22) considerando a estrutura de correlação em análise.

Igualmente ao caso do AIC, a escolha da melhor estrutura de matriz de correlação de trabalho é definida pela estrutura que apresenta o menor valor do QIC(R).

ii. Critério CIC

Hin e Wang (2009) definiram o critério de informação da matriz de correlação (CIC) para identificar a melhor estrutura de correlação nas EEG. Os autores propuseram o CIC baseados no QIC o qual é definido como a metade do segundo termo do critério QIC, isto é:

$$CIC(R) = \mathrm{tr}(\mathbf{V_I^{-1}V_{G,R}}),$$

em que $\mathbf{V_I}$ e $\mathbf{V_{G,R}}$ definem a matriz de covariâncias para a estrutura independente e para estrutura de correlação em análise, respectivamente.

O primeiro termo do critério QIC, que define a estimativa da log quasi-verossimilhança, não possui nenhuma informação relativa a verdadeira matriz de correlação

nem da matriz de correlação de trabalho. Não sendo informativo, em nada contribui para a escolha da matriz de correlação de trabalho (HIN; WANG, 2009).

Igualmente ao caso do QIC, será considerada a melhor estrutura de matriz de correlação de trabalho aquela que apresentar o menor valor de CIC(R).

iii. Critério RJC

Rotnitzky e Jewell (1990) apresentaram a proposta do critério RJC baseada nas estimativas da variância dos estimadores dos parâmetros. Se o 1º e 2º momentos da distribuição de $\boldsymbol{Y_i}$ são corretamente especificados, espera-se que as matrizes de covariâncias dos estimadores dos parâmetros $\widehat{\mathbf{V}}_{\mathbf{opt}}$ e $\widehat{\mathbf{V}}_{\mathbf{G}}$, definidos pelas equações (21) e (22), respectivamente, sejam idênticos para tamanhos de amostras grandes. Assim, Rotnitzky e Jewell (1990) definiram três medidas usadas para seleção da matriz de correlação de trabalho nas EEG, dadas por:

$$RJ1 = \text{tr}(\mathbf{V_G V_{opt}^{-1}})/p,$$

$$RJ2 = \text{tr}(\mathbf{V_G V_{opt}^{-1}})^2/p$$

e

$$RJ3 = \sqrt{(RJ1-1)^2 + (RJ2-1)^2}.$$

Todas as três medidas são usadas como critério de seleção da matriz de correlação de trabalho. Contudo, nesta tese será usada apenas a medida RJ1. Igualmente aos outros critérios anteriormente descritos, a melhor estrutura de correlação é aquela que fornece o menor valor de RJ.

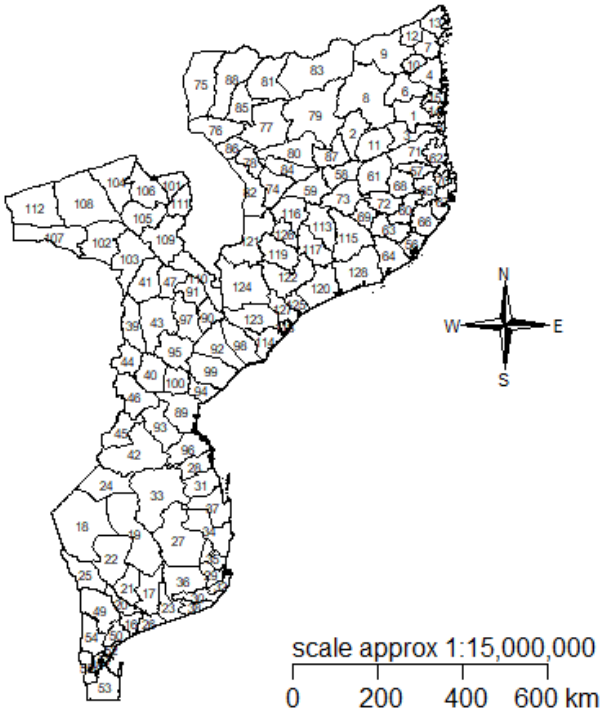
3 MATERIAL E MÉTODOS

Este capítulo encontra-se subdividido da seguinte forma: primeiro é feita a descrição do local do estudo e a fonte dos dados incluíndo os métodos de análise exploratória para a descrição das variáveis utilizadas no estudo. Em seguida é apresentada a proposta metodológica da aplicação das EEG para dados espaciais em áreas. Por fim é apresentado o procedimento utilizado na simulação dos dados.

3.1 Local de estudo

A pesquisa foi realizada com recurso ao banco de dados de Moçambique, que é um país da África sub-sahariana, cuja parte leste é delimitada pelo oceano Índico. Moçambique é constituído por 3 macroregiões, nomeadamente as regiões sul, centro e norte. A zona sul do país é composta pelas províncias de Maputo, Gaza e Inhambane. A zona centro é constituída pelas províncias de Sofala, Tete, Manica e Zambézia, enquanto que a região norte é composta pelas províncias de Nampula, Cabo Delgado e Niassa. Cada uma das províncias, encontra-se subdividida em diferentes distritos perfazendo um total de 128 distritos para todo o país. Na Figura 3 tem-se a região de estudo exibindo os 128 distritos que correspondem a unidade de área georeferenciada com base no centróide da área (distrito).

Figura 3 - Mapa de Moçambique dividido em 128 distritos.



Fonte: CENACARTA (1997).

3.2 Coleta de dados

Para o presente estudo recorreu-se ao banco de dados do “Trabalho do Inquérito Agrícola” de 2012 (TIA). O TIA é uma pesquisa planejada e implementada pelo Ministério de Agricultura em coordenação com o Instituto Nacional de Estatística de Moçambique (INE), que visa fornecer informações relativas às zonas rurais, tanto a nível distrital e provincial, assim como a nível nacional. Essa pesquisa é conduzida a três tipos de explorações (pequenas, médias e grandes) e possibilita a coleta de informação relativa a produção de culturas, atividade pecuária, características demográficas do setor agrário e o acesso à infra-estrutura tanto a nível comunitário como ao nível do agregado familiar. Neste trabalho foram considerados apenas os dados das pequenas e médias explorações pelo fato de estas constituirem as potenciais produtoras de milho.

3.3 Variáveis do estudo

Foi usada como variável resposta a proporção de produtores que adotou sementes melhoradas de milho em cada um dos distritos. Além disso, foi considerado um conjunto de 13 covariáveis, divididas em fatores sócio-demográficos, econômicos, institucionais e tecnológicos. Os fatores sócio-demográficos considerados no estudo foram os seguintes: (i) tamanho médio do agregado familiar; (ii) idade média do produtor; (iii) nível de educação média do produtor; (iv) uso de trabalhadores sazonais; (v) uso de trabalhadores efetivos e (vi) proporção de famílias chefiadas por homens. Quanto aos fatores econômicos, foram considerados: (i) posse de meio de transporte; (ii) posse de celeiros melhorados e (iii) acesso ao crédito. Os fatores institucionais correspondem a: (i) acesso aos serviços de extensão; (ii) pertence à uma associação agrícola e (iii) acesso a informação. Para o caso dos fatores tecnológicos foi usada apenas uma variável correspondente ao uso de tração animal. Todas as variáveis foram medidas ao nível distrital que corresponde à unidade de área georeferenciada com base no centróide da área (distrito).

3.4 Análise exploratória dos dados

Para o caso das variáveis independentes, a análise exploratória dos dados foi feita usando estatísticas descritivas bem como o seu mapeamento usando o método de Fisher-Jenks para definição das classes. Esse método consiste em minimizar a soma dos quadrados dos desvios em relação a média dentro de cada classe (FISHER, 1958). Já, para a variável resposta, além do uso das estatísticas descritivas para efetuar a análise exploratória dos dados, foram usadas medidas de autocorrelação espacial definidas na seção 2.2, com o propósito de

analisar a dependência espacial entre as observações da proporção dos produtores que usaram variedades melhoradas de milho em Moçambique.

A inferência sobre as estatísticas de Moran, Geary e Getis foi feita com base no teste de permutação aleatória para cada uma das estatísticas, usando 999 permutações. Para cada caso, foi construída a distribuição das estatísticas sob a hipótese de completa aleatoriedade espacial, incluindo o valor estimado da estatística em questão (Moran, Geary ou Getis) e determinou-se o *valor p* utilizando o procedimento anteriormente descrito.

Adicionalmente, foi construído o "LISA cluster map" com vista a evidenciar os locais que apresentam padrões de autocorrelação espacial na região de estudo.

3.5 Modelagem usando a proposta metodológica de aplicação de EEG em dados espaciais em áreas

Primeiramente, ajustou-se um modelo linear generalizado (MLG) considerando a variável resposta e todas as 13 covariáveis usadas no estudo. No processo de modelagem foi usada a distribuição binomial e a função de ligação logit. Em seguida, aplicou-se o procedimento de seleção de covariáveis com o propósito de obter um modelo com maior poder explicativo e com menor número de parâmetros. Nesse procedimento utilizou-se o algorítmo "backward" que inicia com um modelo incluíndo todas as covariáveis e em passos subsequentes são retiradas algumas covariáveis sem influência no modelo até obter-se um modelo com o menor valor de AIC.

Após o ajuste do MLG aplicou-se o índice de Moran nos resíduos do modelo, para evidenciar a presença da autocorrelação espacial residual. Uma vez comprovada essa dependência espacial entre os resíduos do modelo, indicando que a autocorrelação espacial presente na variável resposta não foi incluída no modelo, havendo necessidade de incluí-la no processo de modelagem.

A inclusão da dependência espacial no processo de modelagem foi feita usando as equações de estimação generalizadas (EEG) por meio da definição de uma matriz de correlação espacial de trabalho, o que corresponde à proposta metodológica desta tese.

As estruturas de matrizes de correlação de trabalho apresentas na seção 2.3.1 referem-se aos casos de dados longitudinais em que é estimada uma correlação no tempo. Para o caso de dados de superfície contínua (geoestatística) existem na literatura metodologias para a definição da matriz de correlação de trabalho. Porém, para o caso de dados espaciais em áreas, não constam na literatura metodologias para a construção da referida matriz de

correlação de trabalho e uso das EEG, pelo que no presente trabalho propõe-se uma metodologia de construção da matriz de correlação de trabalho quando são considerados dados espaciais em áreas.

Assim, propõe-se que a matriz de correlação de trabalho seja definida com base no índice global de Moran e na matriz de vizinhança (**W**) construida com base no critério de fronteira. Para facilitar a notação, o índice global de Moran será denominado por ρ. A matriz de correlação de trabalho **R(ρ)** é obtida multiplicando o índice de Moran (ρ) pela matriz de vizinhança (**W**) resultando numa matriz $\rho\mathbf{W}$. Como os elementos da diagonal principal de $\rho\mathbf{W}$ são compostos por zeros, adiciona-se a matriz identidade de ordem *n* para obter-se os elementos da diagonal principal de **R(ρ)** composta por uns. Dessa forma, a matriz **R(ρ),** sendo simétrica, positiva semi definida e composta por uns na diagonal principal, apresenta as características de uma matriz de correlação. Essas características foram avaliadas com base nos autovalores da matriz que devem ser não negativos. Portanto, a matriz de correlação de trabalho em dados de área, proposta neste trabalho é definida por:

$$\mathbf{R}(\rho) = \rho\mathbf{W} + \mathbf{I}, \tag{29}$$

em que **R(ρ)** é a matriz de correlação espacial de trabalho; ρ é o índice de Moran para a variável resposta (proporção dos produtores que usaram semente melhorada de milho); **W** é a matriz de proximidade espacial construida utilizando o critério de fronteira; **I** é a matriz identidade de ordem *n*.

Fazendo uma analogia com as estruturas de matriz de correlação definidas na seção 2.3.1 verifica-se:

- na equação (29), **R(ρ)** define uma matriz de correlação de trabalho do tipo “Toeplitz” ou M-dependent (M=1), indicando que as áreas que fazem fronteira entre si (vizinhança de primeira ordem) estão espacialmente autocorrelacionadas, enquanto que as áreas não vizinhas (sem fronteira entre si) não apresentam autocorrelação espacial.
- **R(ρ)** irá definir uma matriz de correlação AR(1) se forem consideradas matrizes de vizinhança de ordem *m* ($\mathbf{W}^{(m)}$), com *m = 2,3,...h*, indicando a ordem de vizinhança. Nesse caso, serão determinados diferentes índices de Moran ($\rho^{(m)}$), para cada ordem de vizinhança *m*, isto é, será construído um correlograma de Moran com vista a indicar a correlação espacial existente nas diferentes ordens de vizinhança. Deste modo, os

diferentes valores de $\rho^{(m)}$ obtidos pelo correlograma de Moran serão usados para construção da matriz de correlação de trabalho **R($\boldsymbol{\rho}$)**.

- A matriz de correlação de trabalho permutável pode ser obtida ao se considerar que todas as áreas são vizinhas entre si. Portanto, a matriz de vizinhança (**W**) será definida com base no número máximo de vizinhos. Desse modo, o índice de Moran irá indicar que todas as áreas apresentam a mesma correlação espacial, o que na prática é uma pré-suposição pouco realística.

Assim, dada a estimativa do índice de Moran ($\hat{\rho}$), obtida a partir dos dados da proporção de produtores que adotaram sementes melhoradas de milho, a matriz de correlação espacial de trabalho **R($\hat{\rho}$)** definida em (29), foi substituída na equação (17) para definir a matriz de covariâncias espacial da variável resposta. Portanto, $\boldsymbol{\Omega}$ foi definida por:

$$\boldsymbol{\Omega} = \phi^{-1}\mathbf{V}^{\mathbf{1/2}}\mathbf{R}(\hat{\rho})\mathbf{V}^{\mathbf{1/2}}, \tag{30}$$

em que $\mathbf{V} = diag\{\mathrm{V_i}, \dots, \mathrm{V_n}\}$, com $\mathrm{V_i} = \mu_i(1-\mu_i)$ e $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/[1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]$ para i = 1, 2, ... , n definindo o número de áreas na região de estudo; $\boldsymbol{x}_i^T = (1, x_1, \dots, x_p)$ são os valores observados das covariávies descritas no sub capítulo 3.3; $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ é um vetor de parâmetros desconhecidos; ϕ é o parâmetro de dispersão que neste trabalho se assume igual a 1 (ϕ = 1).

As estimativas dos parâmetros do modelo foram obtidas por meio do processo iterativo definido na equação (19), substituíndo $\boldsymbol{\Omega}$ pela equação (30). Para tal, foram consideradas como valores iniciais, as estimativas de $\boldsymbol{\beta}$ obtidas a partir do modelo logístico iteragindo-se o processo pelo método de escore Fisher até a convergência. As expressões das matrizes envolvidas nos cálculos encontram-se descritas no apêndice A.

A inferência sobre os parâmetros do modelo foi feita usando o teste de Wald, recorrendo ao estimador robusto da matriz de covariâncias de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{G}}$ dado na equação (22). A estatística do teste de Wald proposto por Rotnitzky e Jewel (1990) é definida por:

$$\tau = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\boldsymbol{o}}\right)^{\boldsymbol{T}}\widehat{\mathbf{V}}_{\mathbf{G}}^{-\mathbf{1}}\left(\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\boldsymbol{o}}\right) \sim \chi_p^2 \; sob \; Ho, \tag{31}$$

em que:

$\hat{\beta}$ é o estimador dos parâmetros;

β^o é o valor do parâmetro sob Ho;

$\widehat{\mathbf{V}}_{\mathbf{G}}$ é o estimador da matriz de covariâncias dos estimadores dos parâmetros.

O parâmetro de associação das EEG ($\alpha|\hat{\beta}$) foi estimado a partir do modelo de erro espacial usando os resíduos do modelo. O modelo de erro espacial (CAR) é definido por:

$$\boldsymbol{\varepsilon} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{\delta}, \tag{32}$$

em que $\boldsymbol{\varepsilon}$ é o vetor de resíduos do modelo, λ é o coeficiente espacial autoregressivo que irá definir o parâmetro de associação espacial, **W** é a matriz de proximidade espacial e **δ** é o termo aleatório normalmente distribuido, com média zero e variância constante, isto é, $\delta \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$.

De acordo com Bivand et al. (2013), a estimação do parâmetro λ no modelo CAR consiste em explorar a decomposição da matriz |**I** - λ**W**| em termos dos autovalores da matriz **W**. Assim, tem-se:

$$ln(|\mathbf{I} - \lambda \mathbf{W}|) = ln[\textstyle\prod_{i=1}^{n}(1 - \lambda\omega_i)], \tag{33}$$

em que ω_i são os autovalores da matriz **W** e λ é estimado a partir de métodos iterativos. Obtido o $\hat{\lambda}$ de (33), este é substituído na equação (22) para determinar as estimativas das variâncias dos estimadores dos parâmetros de posição.

No processo de modelagem foram consideradas duas classes de modelos: com ausência e presença de dependência espacial. Na presença da depêndência espacial foram considerados dois casos:

i. ajuste do modelo considerando a matriz de correlação espacial de trabalho “toeplitz” m-dependente (m=1), onde a matriz de vizinhança usada é de primeira ordem.
ii. ajuste do modelo considerando a matriz de correlação espacial de trabalho “AR(1)”, em que a matriz de proximidade espacial utilizada é de ordem *m*, isto é, definida com base no correlograma de Moran.

Em cada um dos casos, foram usadas duas abordagens para comparação dos modelos. Na primeira, foi determinada a eficiência relativa do estimador considerando a dependência espacial em relação ao estimador que “ignora” a estrutura de correlação espacial. O cálculo de eficiência relativa foi feito usando a abordagem descrita em 2.3.1. Outra abordagem para comparação dos modelos foi a aplicação dos critérios de seleção de matriz de correlação de trabalho descritas na seção 2.3.2 Foram consideradas como melhores estruturas, aquelas que forneceram os menores valores de QIC, CIC e RJ1.

3.6 Simulação de dados

Para validar a metodologia proposta neste trabalho foi feito um processo de simulação dos dados variando o valor da estatística de Moran. Uma vez que não foram encontradas metodologias na literatura para simulação de dados de área usando o índice de Moran ou outro índice de correlação espacial aplicado a esse tipo de dados, a simulação foi baseada na geração de um conjunto de dados com dependência espacial avaliada com base no índice de Moran para uma região fixa, isto é, foi usado o mapa descrito pela Figura 3.

A geração da variável com dependência espacial foi feita seguindo o procedimento descrito a seguir:

De acordo com Anselin (2005), o índice de Moran corresponde ao parâmetro "*b*" do seguinte modelo de regressão linear:

$$\begin{gathered} \boldsymbol{Y}^* = a\mathbf{1} + b\boldsymbol{Y} + \boldsymbol{\varepsilon}, \\ \text{em que } \varepsilon \sim N(0, \sigma^2) \text{ e } \boldsymbol{Y}^* = \mathbf{W}\boldsymbol{Y}. \end{gathered} \tag{34}$$

Se assumirmos que os valores de "*a*", "*b*" e a distribuição dos erros são conhecidos, assim como a matriz **W**, então o problema consiste em determinar o conjunto de valores que irão constituir o vetor ***Y***. Para uma melhor compreensão, considere a região a seguir com a respectiva matriz de vizinhança normalizada nas linhas (FIGURA 4):

Figura 4 - Exemplo de definição de matriz de proximidade espacial.

	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Fonte: Druck et al. (2004).

Para cada observação, o modelo apresentado na (34) pode ser descrito como:

$$\begin{bmatrix} y_A^* \\ y_B^* \\ y_C^* \\ y_D^* \\ y_E^* \end{bmatrix} = \begin{bmatrix} a + by_A + e_A \\ a + by_B + e_B \\ a + by_C + e_C \\ a + by_D + e_D \\ a + by_E + e_E \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & w_{AB} & 0 & w_{AD} & 0 \\ w_{BA} & 0 & w_{BC} & w_{BD} & w_{BE} \\ 0 & w_{CB} & 0 & 0 & w_{CE} \\ w_{DA} & w_{DB} & 0 & 0 & w_{DE} \\ 0 & w_{EB} & w_{EC} & w_{ED} & 0 \end{bmatrix} \begin{bmatrix} y_A \\ y_B \\ y_C \\ y_D \\ y_E \end{bmatrix} = \begin{bmatrix} a + by_A + e_A \\ a + by_B + e_B \\ a + by_C + e_C \\ a + by_D + e_D \\ a + by_E + e_E \end{bmatrix} \Leftrightarrow$$

$$\Leftrightarrow \begin{bmatrix} w_{AB}y_B + w_{AD}y_D \\ w_{BA}y_A + w_{BC}y_C + w_{BD}y_D + w_{BE}y_E \\ w_{CB}y_B + w_{CE}y_E \\ w_{DA}y_A + w_{DB}y_B + w_{DE}y_E \\ w_{EB}y_B + w_{EC}y_C + w_{ED}y_D \end{bmatrix} = \begin{bmatrix} a + by_A + e_A \\ a + by_B + e_B \\ a + by_C + e_C \\ a + by_D + e_D \\ a + by_E + e_E \end{bmatrix}.$$

Nesse caso temos um sistema de equações lineares com 5 variáveis (y_A, ... ,y_E). Assim, temos:

$$\begin{bmatrix} -by_A & +w_{AB}y_B & +0 & +w_{AD}y_D & +0 \\ w_{BA}y_A & -by_B & +w_{BC}y_C & +w_{BD}y_D & +w_{BE}y_E \\ 0 & +w_{CB}y_B & -by_C & +0 & +w_{CE}y_E \\ w_{DA}y_A & +w_{DB}y_B & +0 & -by_D & +w_{DE}y_E \\ 0 & +w_{EB}y_B & +w_{EC}y_C & +w_{ED}y_D & -by_E \end{bmatrix} = \begin{bmatrix} a + e_A \\ a + e_B \\ a + e_C \\ a + e_D \\ a + e_E \end{bmatrix}.$$

A solução desse sistema de equações é da forma $\boldsymbol{y = A^{-1}e^*}$, em que:

$$A = \begin{bmatrix} -b & w_{AB} & 0 & w_{AD} & 0 \\ w_{BA} & -b & w_{BC} & w_{BD} & w_{BE} \\ 0 & w_{CB} & -b & 0 & w_{CE} \\ w_{DA} & w_{DB} & +0 & -b & w_{DE} \\ 0 & w_{EB} & w_{EC} & w_{ED} & -b \end{bmatrix} \quad ; \quad e^* = \begin{bmatrix} a + e_A \\ a + e_B \\ a + e_C \\ a + e_D \\ a + e_E \end{bmatrix}.$$

É fácil ver que a matrix **A,** é constituída pelos elementos da matriz de vizinhança **W** com os elementos da diagonal principal atualizados com a estimativa simétrica do índice de Moran. Assim, para gerar amostras com dependência espacial para um índice de Moran fixo, construiu-se a matriz **A** acima descrita e gerou-se amostras aleatórias de uma distribuição normal com parâmetros conhecidos. Desse modo, a variável aleatória com dependência espacial fixada para um determinado valor do índice de Moran foi definida por:

$$\boldsymbol{Y = \mathbf{A^{-1}}e^*},$$

em que **A = W – *diag (b)*** e $\boldsymbol{e^* = a\mathbf{1} + \varepsilon}$ conforme a metodologia já descrita.

Para cada valor fixo do índice de Moran, foram geradas 16100 amostras de tamanho n = 128 correspondente ao número de distritos do mapa de Moçambique. A simulação foi feita em quatro etapas, isto é, na primeira etapa foram geradas 100 amostras, na segunda 1000, na terceira 5000 e na quarta etapa 10000 amostras usando o procedimento

anteriormente descrito. Em cada uma das etapas do processo de simulação foi determinada a raíz quadrada do erro quadrático médio que é dada por:

$$REQM = \sqrt{\frac{\sum_{i=1}^{n}(\hat{I}_i - I)^2}{n}},$$

em que I é a verdadeiro valor do índice de Moran; $\hat{I}_i$ é a estimador do índice de Moran em cada amostra e n é o número de amostras simuladas.

As 10000 amostras geradas na última etapa foram também usadas para construir a distribuição empírica para as estimativas do índice de Moran através do processo Monte Carlo.

Além da variável dependente (com correlação espacial), foram geradas aleatoriamente duas covariáveis (uma binária e outra contínua normalmente distribuida) usadas no processo de modelagem. Os valores do índice de Moran considerados estão definidos no intervalo (-1; 1). Para cada valor fixo de autocorrelação espacial, foram geradas 1000 amostras e em cada uma delas foram ajustadas duas classes de modelos (MLG e EEG proposta neste trabalho). Para o caso do modelo baseado em EEG foram utilizadas duas estruturas de matriz de correlação espacial de trabalho, a Toeplitz e a AR(1). Em cada amostra, determinou-se a eficiência relativa do estimador com dependência espacial em relação ao estimador com a estrutura independente, com o propósito de avaliar o ganho ou perda da eficiência utilizando a metodologia proposta.

Uma vez comprovado o ganho da eficiência do estimador dos parâmetros considerando a dependência espacial captada com base no índice de Moran em relação ao estimador que considera a estrutura independente, comparou-se também a eficiência relativa dos estimadores com dependência espacial avaliada com base no índice de Moran em relação ao uso do índice de Geary considerando as estruturas de matriz de correlação espacial de trabalho do tipo Toeplitz e AR(1).

Pelo fato do índice de Geary estar limitado no intervalo [0; 2], fez-se uma modificação na escala com vista a limitá-lo no intervalo [-1; 1]. A modificação foi feita do seguinte modo:

$$c^* = 1 - c,$$

em que c^* é o índice de Geary modificado definido no intervalo [-1;1]; c é a estatística de Geary no intervalo [0; 2].

Para determinar qual dos índices de correlação espacial se traduz numa maior eficiência dos estimadores, foi usado o mesmo procedimento de geração de amostras aleatórias anteriormente descrito. A diferença residiu na definição da matriz de correlação de

trabalho, isto é, um estimador considerou a matriz de correlação de trabalho construída com base no índice de Moran e o outro com base no índice de Geary. Em cada amostra determinou-se a eficiência relativa de um estimador em relação ao outro e determinou-se a eficiência relativa média considerando as mil simulações para cada valor fixo do índice de correlação espacial. Além disso, foram também determinadas as estimativas médias dos critérios de seleção de matriz de correlação de trabalho.

Todas as análises foram feitas utilizando o *software* Geoda (2010) e funções disponíveis e desenvolvidas dentro do ambiente R (R CORE TEAM, 2018), com o auxílio das bibliotecas *spdep* (BIVAND et al., 2011), *maptools* (BIVAND et al., 2013) e *geepack* (HALEKOH et al., 2006).

As rotinas utilizadas para análise dos dados reais de adoção de variedades melhoradas de milho estão descritas no apêndice C enquanto que as funções para gerar amostras com dependência espacial avaliada com base no índice de Moran estão no apêndice D.

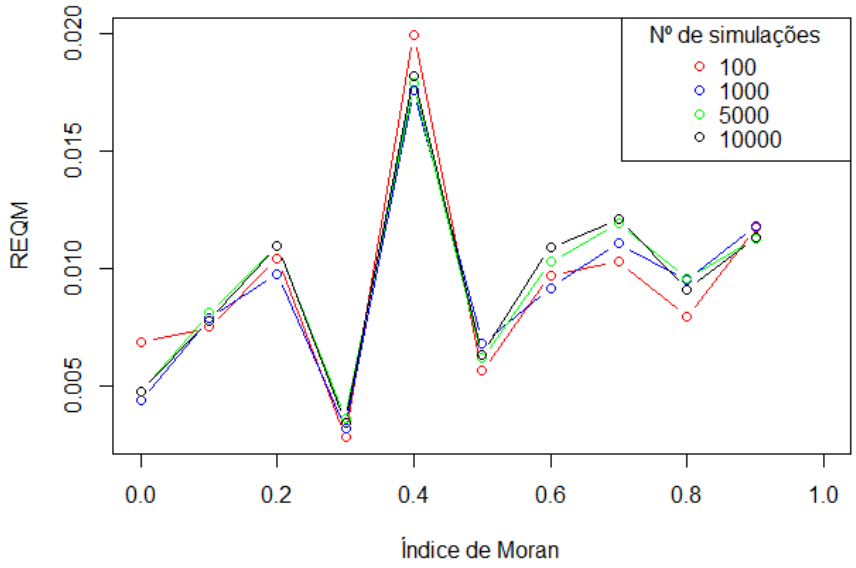
4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados, primeiramente, os resultados do processo de simulação de dados e, em seguida, os resultados referentes aos dados reais. Para o caso dos dados simulados, apresentam-se os resultados da avaliação do método de simulação de dados proposto neste trabalho, incluindo o cálculo da eficiência dos estimadores, assim como os resultados baseados nos critérios de seleção da matriz de correlação de trabalho. Já, para os dados reais, são apresentados resultados das estatísticas descritivas de todas as variáveis consideradas no estudo, assim como a análise da dependência espacial da variável resposta usando as estatísticas de Moran, Geary e Getis e Ord. Posteriormente, seguem os resultados do processo de modelagem.

4.1 Avaliação do método de simulação dos dados

Na Figura 5 descrevem-se os resultados das estimativas da raíz quadrada do erro quadrático médio (REQM) obtidos para valores positivos do índice de Moran nas diferentes etapas do processo de simulação consideradas no estudo, isto é, para o caso de 100, 1000, 5000 e 10000 amostras simuladas.

Figura 5 - Raíz do erro quadrático médio para valores positivos do índice de Moran em diferentes amostras simuladas.

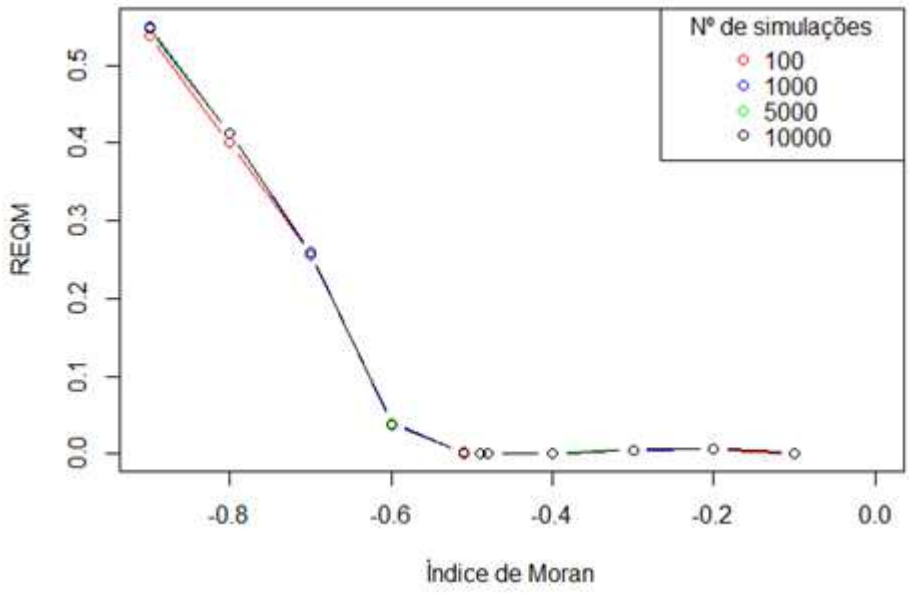


Fonte: Do autor (2019).

De um modo geral, os resultados mostram que a REQM nos diferentes valores do índice de Moran, apresenta o mesmo comportamento nos diversos números de amostras

simuladas, isto é, para um determinado valor do índice de Moran a REQM tende apresentar valores muito próximos independentemente do número de amostras consideradas. A título de exemplo, a maior amplitude dos valores da REQM nos diferentes números de amostras simuladas é observada para o caso em que o índice de Moran é igual a 0,4. Nessa situação os valores mínimo e máximo da REQM são 0,017 e 0,019, respectivamente. Esse comportamento sugere que o método de simulação de dados de área proposto neste trabalho apresenta uma boa consistência. Além disso, verifica-se que os valores da REQM em todos os cenários avaliados são muito baixos, encontrando-se próximos de zero (REQM < 0,02), o que pressupõe uma alta acurácia do método. Esses resultados são similares aos obtidos para o caso dos valores negativos do índice de Moran. Contudo, apesar do método proposto apresentar uma boa constistência também para valores negativos, a acurácia diminui de forma substancial para valores do índice de Moran inferiores a -0,5 (FIGURA 6).

Figura 6 - Raíz do erro quadrático médio para valores negativos do índice de Moran em diferentes amostras simuladas.

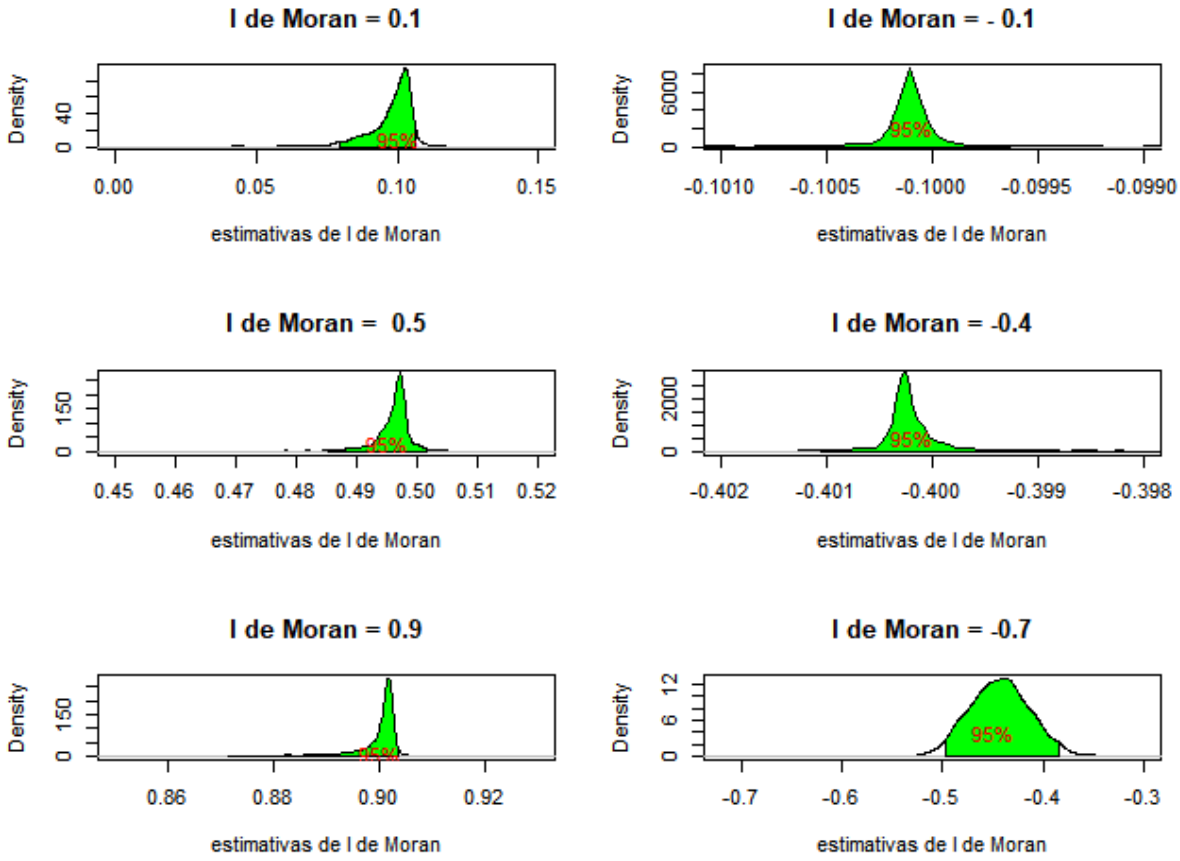


Fonte: Do autor (2019).

A baixa acurácia observada para valores do índice de Moran definidos no intervalo (-1; -0,5) sugere uma limitação para aplicação do método para esse conjunto de valores, isto é, para dados com associação espacial negativamente forte o método proposto neste trabalho apresentou um resultado pouco satisfatório. Porém, vale ressaltar que para o caso de dados espaciais em áreas a ocorrência de uma autocorrelação espacial negativa forte raramente ocorre nos vários estudos envolvendo esse tipo de dados (ROGERSON; YAMADA, 2009).

Analisando os resultados das simulações Monte Carlo para 10000 amostras geradas verificou-se que a distribuição das estimativas do índice de Moran, tanto para valores positivos assim como negativos tem tendência a apresentar-se assimétrica (FIGURA 7).

Figura 7 - Distribuição das estimativas do índice de Moran baseada em 10 mil simulações Monte Carlo.



Fonte: Do autor (2019).

Para o caso de valores positivos, verifica-se que a distribuição das estimativas do índice de Moran apresenta uma ligeira assimetria para o conjunto de valores do índice considerados. Os quantis correspondentes a 2,5% e 97,5%, que delimitam a área de 95% apresentada na Figura 7, são muito próximos aos verdadeiros valores dos índices de Moran em todos os casos considerados (TABELA 3). Isso sugere que a maior parte das amostras geradas a partir do método proposto nesse trabalho, irão estimar um intervalo de confiança de 95% que contém o verdadeiro valor do índice de Moran. Assim, a maior parte das amostras geradas usando o método proposto irão estimar um índice de Moran que se situe dentro da área de 95% nas distribuições apresentadas na Figura 7 mais comumente chamada de “região de não rejeição de Ho”. Esse resultado sugere que a metodologia de simulação de dados de

área baseado no índice de Moran, proposta neste trabalho, produz bons resultados para o caso de ocorrência de autocorrelação espacial positiva.

Tabela 3 - Quantis da distribuição do índice de Moran.

Quantis	**Índice de Moran**							
	-0,9	**-0,7**	**-0,5**	**-0,4**	**-0,1**	**0,1**	**0,5**	**0,9**
2,50%	-0,41	-0,49	-0,51	-0,41	-0,11	0,08	0,49	0,87
97,50%	-0.29	-0,38	-0,49	-0,39	-0,09	0,11	0,51	0,91

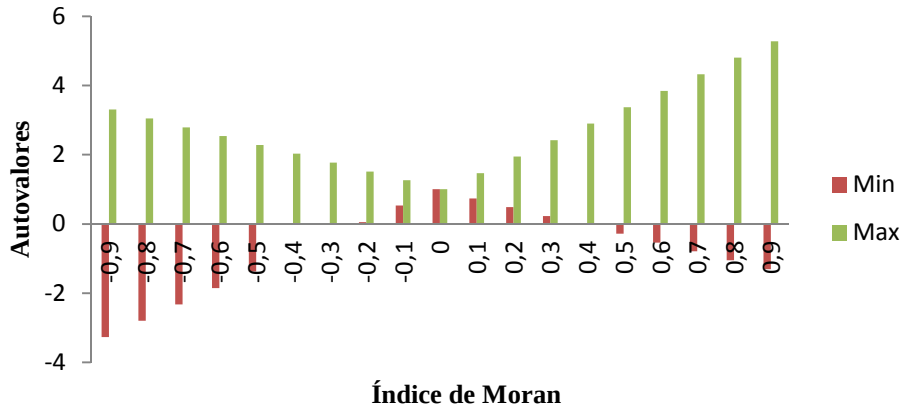
Fonte: Do autor (2019).

Em relação aos valores do índice de Moran negativos, a distribuição das estimativas desse índice apresentou-se menos assimétrica, comparativamente ao caso dos valores positivos. Os quantis 2,5% e 97,5% apresentaram valores próximos ao verdadeiro valor do índice de Moran apenas para valores do índice definidos em [-0,5; 0). Para valores do índice definidos no intervalo (-1; -0,5) os referidos quantis encontram-se mais distantes do verdadeiro valor do índice e este é menor que os dois quantis (TABELA 3). Isto sugere que as amostras obtidas a partir do método proposto neste trabalho irão apresentar bons resultados em todo espaço paramétrico do índice de Moran com a exceção de valores de autocorrelação espacial negativa definidos no intervalo (-1; -0,5). Esse resultado corrobora os resultados obtidos na avaliação do método de simulação de dados de área proposto nesse trabalho através do uso da REQM.

4.2 Determinação da eficiência relativa assintótica

Nas EEG, a estimação das variâncias dos estimadores envolvidos no processo de modelagem possui um papel preponderante para a determinação da eficiência relativa assintótica. Para garantir estimativas fiáveis no cálculo dessa eficiência é necessário que a matriz de covariâncias conserve a propriedade de uma matriz positiva semi-definida. Na Figura 8 descreve-se de forma resumida os resultados do maior e menor autovalores da matriz de correlação espacial de trabalho considerando apenas a vizinhança de primeira ordem, para os diferentes valores do índice de Moran utilizados durante o processo de simulação dos dados. Para os valores do índice de Moran definidos em (-1;-0,5] U [0,5;1) observa-se que a matriz de correlação de trabalho do tipo “Toeplitz” com m=1, apresenta alguns autovalores negativos, quebrando a característica de uma matriz positiva semi-definida, típica de uma matriz de covariâncias.

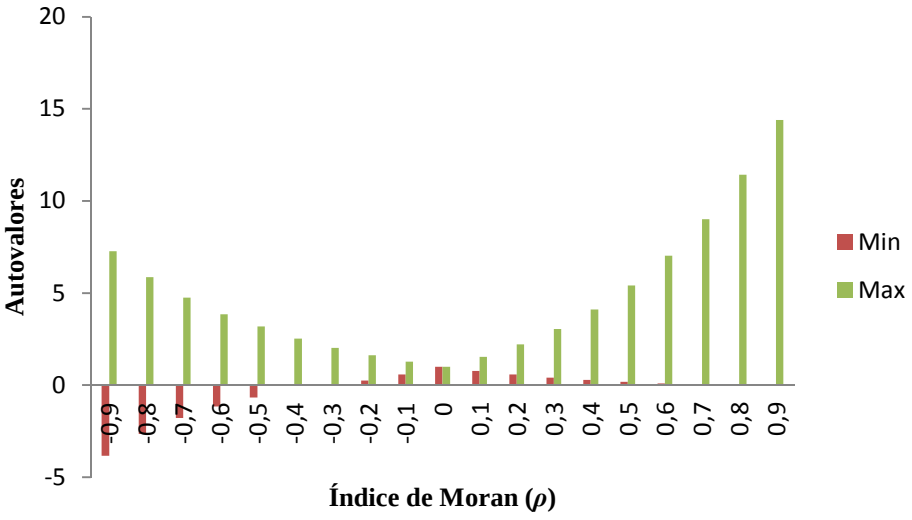
Figura 8 - Autovalores da matriz de correlação espacial de trabalho do tipo Toeplitz de primeira ordem para diferentes valores do índice de Moran.



Fonte: Do autor (2019).

A condição de uma matriz positiva semi definida é observada para o caso dos valores do índice de Moran definidos no intervalo (-0,5; 0,5) quando é usada uma matriz de correlação de trabalho m-dependent de primeira ordem. Já, quando a matriz de correlação de trabalho é construtída usando a estrutura AR(1) na qual são consideradas ordens de vizinhança superiores ($m > 1$) a condição de uma matriz positiva semi-definida é observada para todo espaço parâmetrico do índice de Moran exceto para os casos em que a correlação espacial é negativa e forte, isto é, para os valores do índice definidos em (-1;-0,5] confome indicam os autovalores da matriz de correlação de trabalho que são negativos nesse intervalo (FIGURA 9).

Figura 9 - Autovalores da matriz de correlação espacial de trabalho do tipo AR(1) para diferentes valores do índice de Moran.



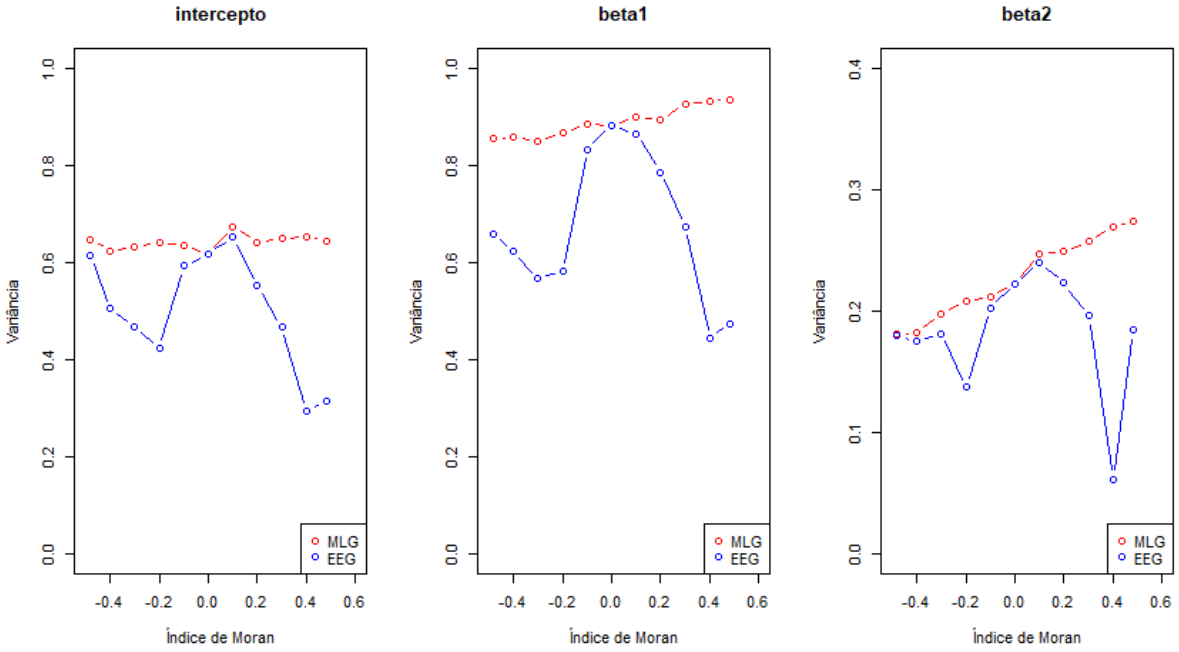
Fonte: Do autor (2019).

Assim a determinação da eficiência relativa assintótica entre o estimador que considera a dependência espacial (EEG) e o estimador que ignora a estrutura espacial (MLG) foi realizada para os casos em que $| \rho | < 0,5$ para a matriz de correlação de trabalho Toeplitz de ordem 1 e $\rho > - 0,5$ para a estrutura AR(1). Autores como Brajendra e Kalyan (1999) e Wang e Carey (2003), em estudos sobre a eficiência relativa assintótica nas equações de estimação generalizadas, também usaram os valores de ρ definidos no intervalo (-0,5; 0,5) para a estrutura Toeplitz enquanto que para a estrutura AR(1) foram usados valores de ρ definidos no intervalo (-1; 1).

Nas Figuras 10 e 11 descrevem-se os resultados correspondente às estimativas médias das variâncias obtidas usando os estimadores EEG e MLG para cada valor do índice de Moran quando são usadas respectivamente, as estruturas Toeplitz (m=1) e AR(1) para um modelo com três parâmetros considerando 1000 simulações.

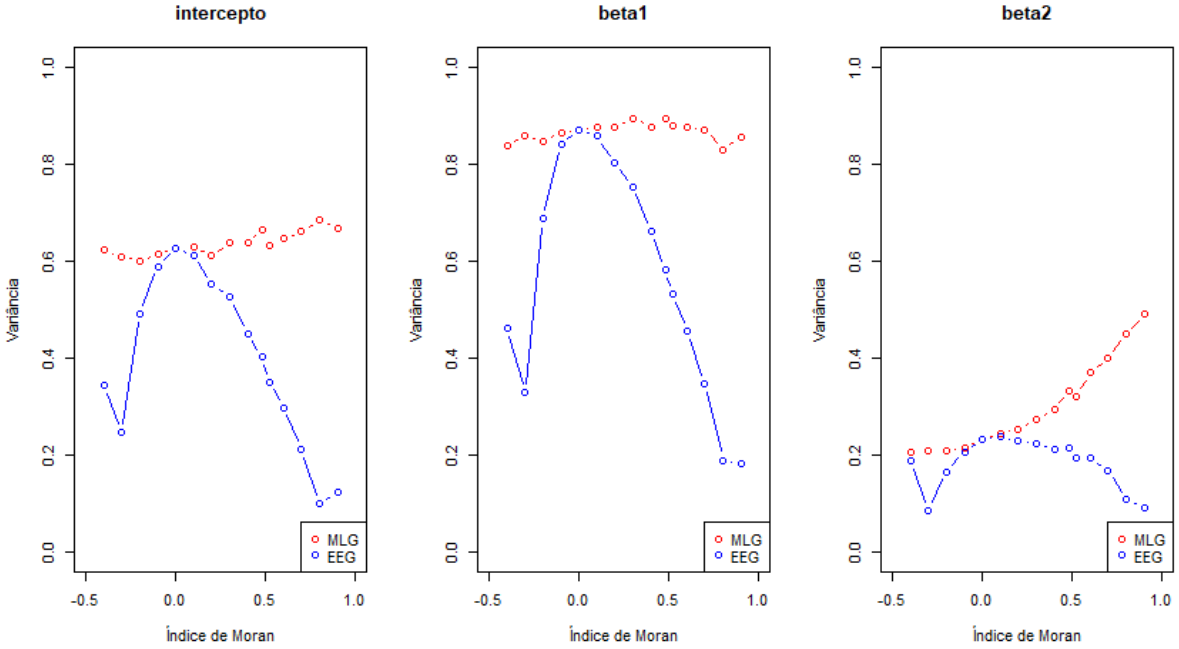
De uma forma geral, observa-se que para ambas as covariáveis, assim como para o intercepto, o estimador utilizando a metodologia EEG proposta neste trabalho apresenta estimativas da variância menores em relação ao estimador assumindo independência (MLG), independentemente da estrutura da matriz de correlação espacial de trabalho usada (Toeplitz ou AR(1)). Para valores de autocorrelação espacial fraca (índice de Moran próximo de zero) as estimativas da variância do estimador EEG e MLG tendem a ser muito próximas uma das outras. Porém, os valores da variância do EEG são sempre ligeiramente menores em relação ao MLG. Isso indica que nessas situações, ambos estimadores podem ser considerados igualmente eficientes. Já, para valores do índice de Moran mais altos, isto é, onde há ocorrência de uma associação espacial positiva, as estimativas da variância do modelo EEG tendem a diminuir com o aumento dos valores do índice de Moran, evidenciando um substancial aumento da eficiência dos estimadores dos parâmetros no processo de estimação, conforme indicam os resultados descritos no apêndice B baseados em 1000 amostras simuladas. Isso é indicativo de que quando a autocorrelação espacial encontra-se presente, ela deve ser considerada no processo de modelagem com vista a obter estimadores ótimos. Esse resultado comprova a hipótese deste estudo em que, obtém-se ganhos na eficiência dos estimadores dos parâmetros no processo de estimação quando há uma correta especificação da estrutura da matriz de correlação espacial de trabalho.

Figura 10 - Estimativa da variância média dos estimadores MLG e EEG para diferentes valores do índice de Moran para a estrutura Toeplitz com m=1.



Fonte: Do autor (2019).

Figura 11 - Estimativa da variância média dos estimadores MLG e EEG para diferentes valores do índice de Moran para a estrutura AR(1).



Fonte: Do autor (2019).

De fato, a estrutura da matriz de correlação de trabalho construída com base no índice de Moran proposta neste trabalho, mostra que existe um ganho na eficiência mesmo para valores baixos do índice de Moran. Na Tabela 4 têm-se as estimativas da eficiência relativa assintótica do estimador EEG em relação ao estimador MLG para o caso em que é usada a matriz de correlação espacial de trabalho AR (1). Ainda na Tabela 4 é comparada a eficiência relativa assintótica entre as duas estruturas de matriz de correlação espacial de trabalho, AR(1) e toeplitz.

Para valores do índice de Moran com correlação espacial fraca, tanto positiva assim como negativa ($|\rho| < 0{,}2$), verifica-se que o estimador EEG apresenta uma eficiência relativa assintótica em relação ao MLG que varia em torno de 80 a 97%, ou seja, mesmo para valores baixos de autocorrelação espacial observa-se um ganho na eficiência dos estimadores quando é aplicada a matriz de correlação espacial de trabalho do tipo AR(1). Para valores de autocorrelação espacial positiva forte, o uso do EEG mostra um ganho considerável na eficiência relativa assintótica chegando a atingir valores abaixo dos 20%, isto é, a variância dos estimadores EEG corresponde a cerca de 20% da variância do MLG. Esse resultado corrobora o resultado de Liang e Zeger (1986) ao afirmar que uma boa especificação da matriz de correlação de trabalho aumenta a eficiência.

Tabela 4 - Eficiência relativa assintótica do estimador EEG usando a estrutura AR(1) em relação ao estimador MLG e em relação a estrutura Toeplitz m=1.

Índice de	**EEG/MLG**			**EEG – AR(1)/EEG - Toeplitz**		
Moran	β_0	β_1	β_2	β_0	β_1	β_2
-0,4	55,38%	55,05%	91,14%	123,27%	132,07%	149,96%
-0,3	40,65%	38,43%	41,41%	34,55%	36,94%	19,49%
-0,2	81,90%	81,10%	79,69%	80,77%	81,80%	77,29%
-0,1	95,90%	97,16%	95,98%	99,92%	99,92%	99,95%
0	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
0,1	97,05%	97,81%	97,11%	99,97%	99,98%	99,96%
0,2	90,43%	91,67%	90,69%	99,27%	99,41%	99,48%
0,3	82,59%	84,25%	81,39%	95,56%	95,52%	96,29%
0,4	70,43%	75,49%	72,38%	59,84%	86,94%	40,97%
0,52	55,43%	60,59%	60,94%			
0,6	46,15%	52,12%	52,43%			
0,7	32,15%	39,78%	42,09%			
0,8	14,52%	22,87%	24,18%			
0,9	18,38%	21,49%	18,93%			

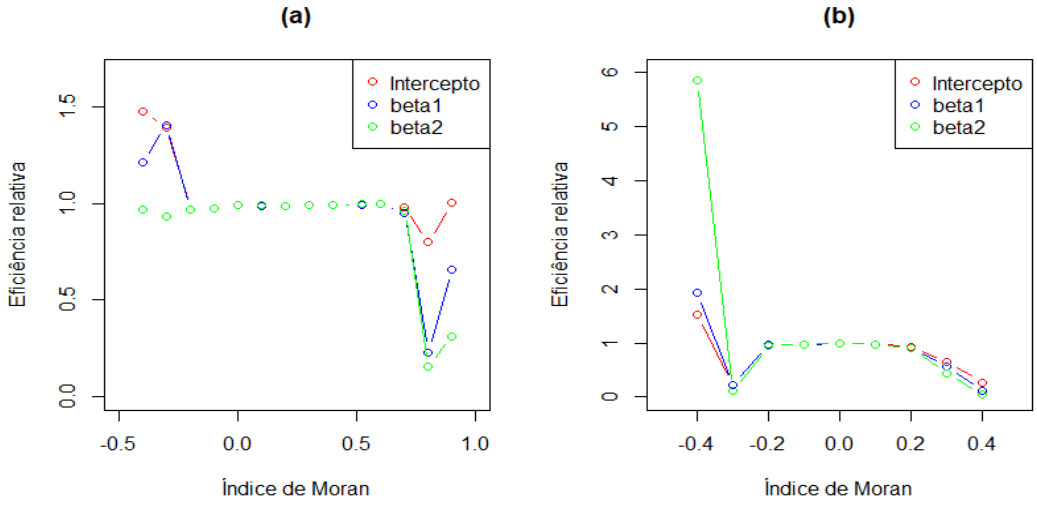
Fonte: Do autor (2019).

Ao comparar as duas estruturas de matriz de correlação de trabalho verifica-se que a estrutura Toeplitz é relativamente mais eficiente em relação a estrutura AR(1) apenas para o caso em que a correlação espacial é igual a -0,4. Para todos os restantes casos, os resultados mostram que ambas estruturas são igualmente eficientes ou a estrutura AR(1) leva vantagem por apresentar menores estimativas para a variância. Esse resultado mostra que a estrutura de matriz de correlação de trabalho do tipo AR(1) é que apresenta melhores resultados comparativamente a estrutura Toeplitz. Wang e Carey (2003), num estudo sobre a performance das equações de estimação generalizadas sob a má especificação da matriz de correlação de trabalho, também chegaram à conclusão de que a matriz de correlação de trabalho do tipo AR(1) apresenta qualitativamente mais robustez, comparativamente às outras estruturas. As estruturas consideradas por esses autores foram “Toeplitz”, AR(1), “Exchangeable” e a Independente.

As análises realizadas anteriormente mostraram que a definição da matriz de correlação de trabalho em dados de área produz bons resultados quando é usado o índice de Moran na definição dessa matriz. Porém, será que essa performance se mantém, quando são considerados outros índices de autocorrelação espacial como é o caso da estatística de Geary?

A resposta a essa pergunta encontra-se descrita na Figura 12, na qual é comparada a eficiência relativa assintótica do estimador EEG definido com base no índice de Moran usando a estrutura AR (1) com o estimador EEG definido com base no índice de Geary usando: (a) a estrutura AR(1); (b) a estrutura Toeplitz.

Figura 12 - Eficiência relativa assintótica do estimador EEG definido com base no índice de Moran usando a estrutura AR (1) em relação ao estimador EEG definido com base no índice de Geary usando: (a) a estrutura AR (1); (b) a estrutura Toeplitz.



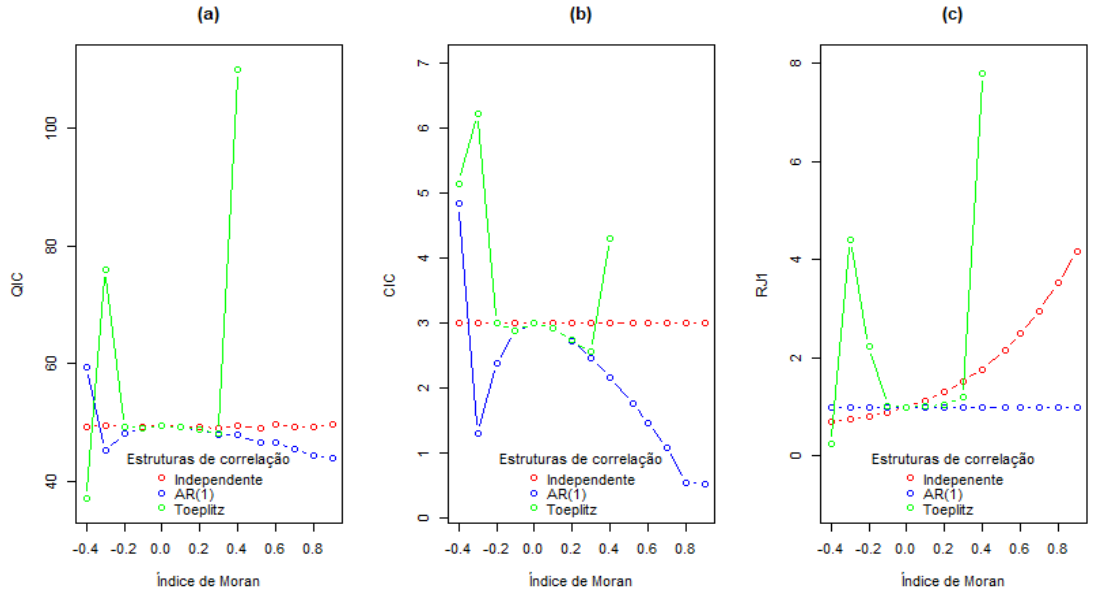
Fonte: Do autor (2019).

Na Figura 12 (a) onde é comparada a eficiência relativa do estimador EEG definido com base no índice de Moran em relação ao estimador EEG definido com base no índice de Geary para a estrutura da matriz de correlação espacial de trabalho do tipo AR(1) em ambos os casos, verifica-se que, para valores de associação espacial mais negativos, o uso do índice de Geary no processo de estimação mostra-se mais eficiente em relação ao índice de Moran. Já, para os casos em que a correlação espacial é negativamente fraca, assim como para as situações de ocorrência de uma associação espacial positiva, ambos estimadores são igualmente eficientes, exceto para os casos em que a correlação espacial é positiva forte. Nessa situação, o uso do índice de Moran na definição da matriz de correlação de trabalho traduz-se numa maior eficiência comparativamente ao uso do índice de Geary. Resultado similar a este é verificado quando se compara a eficiência relativa do estimador EEG usando o índice de Moran para a estrutura AR(1) com o estimador EEG usando o índice de Geary para uma estrutura de matriz de correlação espacial de trabalho do tipo Toeplitz (FIGURA 12b).

4.3 Estimativas dos critérios de seleção de matrizes de correlação

Na Figura 13 têm-se as estimativas médias dos valores dos critérios QIC, CIC e RJC para a seleção da matriz de correlação espacial de trabalho em função dos diferentes valores de índice de Moran, em 1000 amostras simuladas. As matrizes de correlação espacial de trabalho consideradas foram: a estrutura independente, toeplitz e AR(1).

Figura 13 - Critérios de seleção de matriz de correlação de trabalho para diferentes valores do índice de Moran: (a) QIC; (b) CIC; (c) RJ1.



Fonte: Do autor (2019).

De uma forma geral os resultados obtidos utilizando os 3 critérios de seleção são muito similares entre si. Analisando qualquer dos três critérios, para valores do índice de Moran mais próximos de zero ($| \rho | \leq 0,1$), as estimativas obtidas para cada critério são muito próximas entre si quando são comparadas as diferentes matrizes de correlação espacial de trabalho (FIGURA 13). Isso significa que o uso de qualquer das estruturas de matriz de correlação de trabalho no processo de modelagem irá fornecer resultados muito similares principalmente na estimativa da variância dos estimadores. Esse resultado deve-se ao fato de a correlação espacial ser muito baixa, podendo ser considerada desprezível, o que torna as estruturas da matriz de correlação Toeplitz e AR(1) muito parecidas com a estrutura independente, caracterizada pela matriz identidade.

Para valores da autocorrelação espacial positivos ($0,1 < \rho < 1,0$), os três critérios indicam que a estrutura AR(1) apresenta melhores resultados pelo fato das estimativas dos critérios serem mais baixas quando comparadas com as estimativas das outras estruturas de matriz de correlação de trabalho. Esse resultado deve-se ao fato de que a ocorrência de uma associação espacial positiva moderada ou forte possui grande influência no processo de modelagem pelo fato das matrizes de correlação espacial de trabalho que estão sendo comparadas serem numericamente muito diferentes. Porém, vale ressaltar que para valores do índice de Moran definidos em $0,1 < \rho \leq 0,3$, a estrutura Toeplitz é tão eficiente quanto a estrutura AR(1), em qualquer dos critérios usados na análise. Isso significa que a definição da matriz de correlação de trabalho baseada no índice de Moran usando apenas a vizinhança de primeira ordem é eficiente para valores do índice definidos em $0,1 < \rho \leq 0,3$. Quando a autocorrelação espacial presente é superior a 0,3, as matrizes de vizinhança de ordem superior deverão ser levadas em consideração com vista a obter maior eficiência no processo de estimação.

Em termos gerais, a estrutura de correlação espacial AR(1) mostra-se melhor em quase todo domínio do espaço parâmetrico do índice de Moran aqui avaliado.

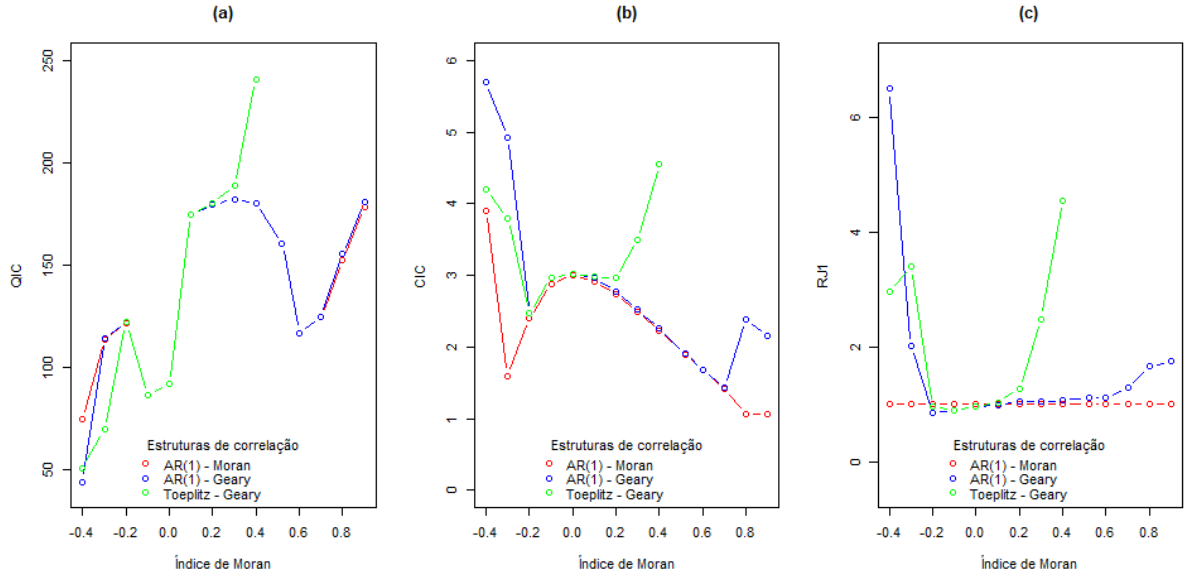
Esses resultados são consistentes com aqueles obtidos por meio da análise da eficiência relativa assintótica, mostrando que a definição da matriz de correlação espacial de trabalho com base no índice de Moran e na matriz de vizinhança traduz-se num maior ganho na eficiência dos estimadores no processo de estimação, principalmente nos casos de ocorrência de uma autocorrelação espacial positiva forte.

Em relação aos valores negativos de correlação espacial aqui considerados, verificou-se que para os valores mais negativos extremos do índice de Moran, os critérios QIC e RJC identificam a estrutura Toeplitz como a melhor. Já, para valores do índice de Moran definidos

em -0,3 ≤ ρ < -0,1 a estrutura AR(1) é identificada como a melhor através dos critérios QIC e CIC.

Ao comparar as estruturas de matriz de correlação espacial de Trabalho construídas com base nos índices de Moran e Geary, verifica-se uma similaridade nos resultados quando são aplicados os três critérios de seleção da matriz de correlação de trabalho (FIGURA 14).

Figura 14 - Critérios de seleção de matriz de correlação espacial definidos com base nos índices de Moran e Geary: (a) QIC; (b) CIC; (c) RJ1.



Fonte: Do autor (2019).

Analizando os resultados obtidos com o critério CIC, observa-se que quando os valores do índice de Moran são próximos de zero (| ρ | ≤ 0,1), todas as estruturas de matriz de correlação espacial de trabalho avaliadas, apresentam o mesmo desempenho, isto é, o uso de qualquer dos índices de correlação espacial (Moran e Geary) não se traduz em nenhum ganho, tanto para as estruturas AR(1) como para a Toeplitz.

Para valores de autocorrelação espacial definidos em 0,1 < ρ < 1,0 a estrutura AR(1) mostra-se preferencial em relação a estrutura Toeplitz independentemente do índice de autocorrelação espacial usado na definição da matriz de correlação espacial de trabalho. Contudo, para valores de índice de Moran que indicam uma associação espacial positiva forte, o uso do índice de Geary mostra-se menos eficiente na definição da matriz de correlação espacial de trabalho em relação ao uso do índice de Moran.

Quando a associação espacial é mais fortemente negativa, os critérios CIC e RJ1 identificam a estrutura AR(1) definida com base no índice de Moran como a melhor estrutura.

Porém, o critério QIC apesar de identificar a mesma estrutura de correlação como sendo a melhor, ele indica o uso do índice de Geary para definição da matriz de correlação espacial de trabalho.

A grande similaridade que se observou nos resultados obtidos em quaisquer dos critérios, principalmente nos cenários de ocorrência de associação espacial positiva, ao se comparar as estruturas “AR(1) – Moran” e “AR(1) – Geary” deveu-se ao fato das estimativas dos dois índices serem muito similares entre si devido a modificação na escala efetuada no índice de Geary para permitir seu uso na construção da matriz de correlação.

Os resultados aqui obtidos são similares aos resultados apresentados na Figura 12 baseados no cálculo da eficiência relativa.

4.4 Análise dos dados reais de adoção de variedades melhoradas de milho em Moçambique

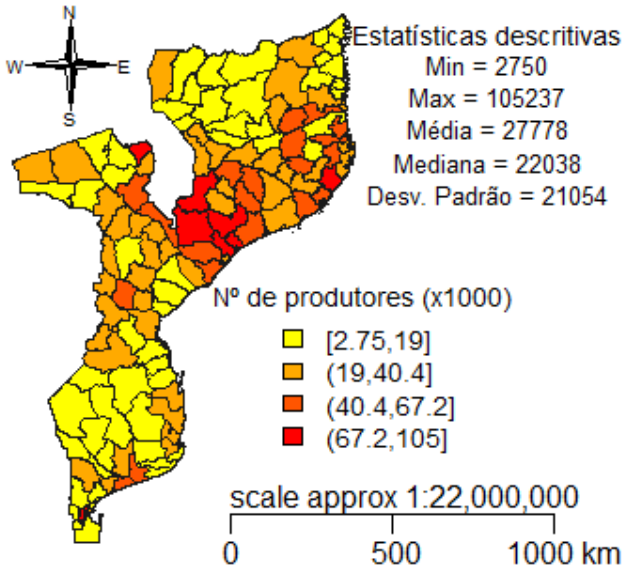
A análise de dados reais referentes ao uso de sementes melhoradas de milho foi precedida de uma análise exploratória para um conjunto de 13 variáveis independentes seguida da análise descritiva da variável resposta e posterior processo de modelagem com aplicação das equações de estimação generalizadas.

4.4.1 Análise descritiva das variáveis independentes

Nas Figuras 15 a 20, estão representados os mapas da distribuição espacial das covariáveis correspondentes aos fatores sócio-demográficos usados no presente estudo, assim como algumas estatísticas descritivas. Quanto ao perfil demográfico do setor agrário em termos do número de produtores por distrito, observa-se que este varia de 2750 a pouco mais de 105 mil produtores com uma média de aproximadamente 28 mil produtores por distrito (FIGURA 15). As regiões norte e centro (principalmente as províncias da Zambézia e Nampula) concentram a maior parte dos pequenos e médios produtores o que coincide com as regiões mais populosas de Moçambique.

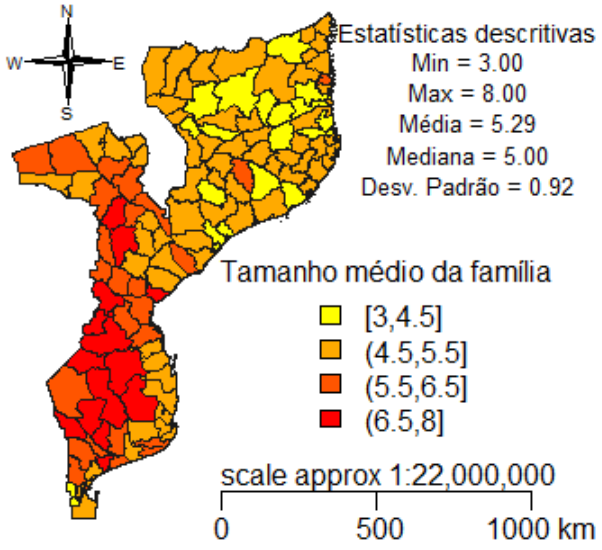
Para o caso do tamanho médio do agregado familiar, verifica-se que as famílias apresentam entre 3 a 8 membros com uma média por distrito de 5 indivíduos, que coincide com a mediana (FIGURA 16). Essa variável possui um papel preponderante na disponibilidade de mão de obra familiar. Portanto, espera-se que a mesma tenha um efeito direto sobre a decisão de adoção de variedades melhoradas de milho, devido à demanda nos tratos que a cultura possui ao longo do seu ciclo.

Figura 15 - Distribuição espacial do número de produtores por distrito.



Fonte: Do autor (2019).

Figura 16 - Distribuição espacial do tamanho médio da família.

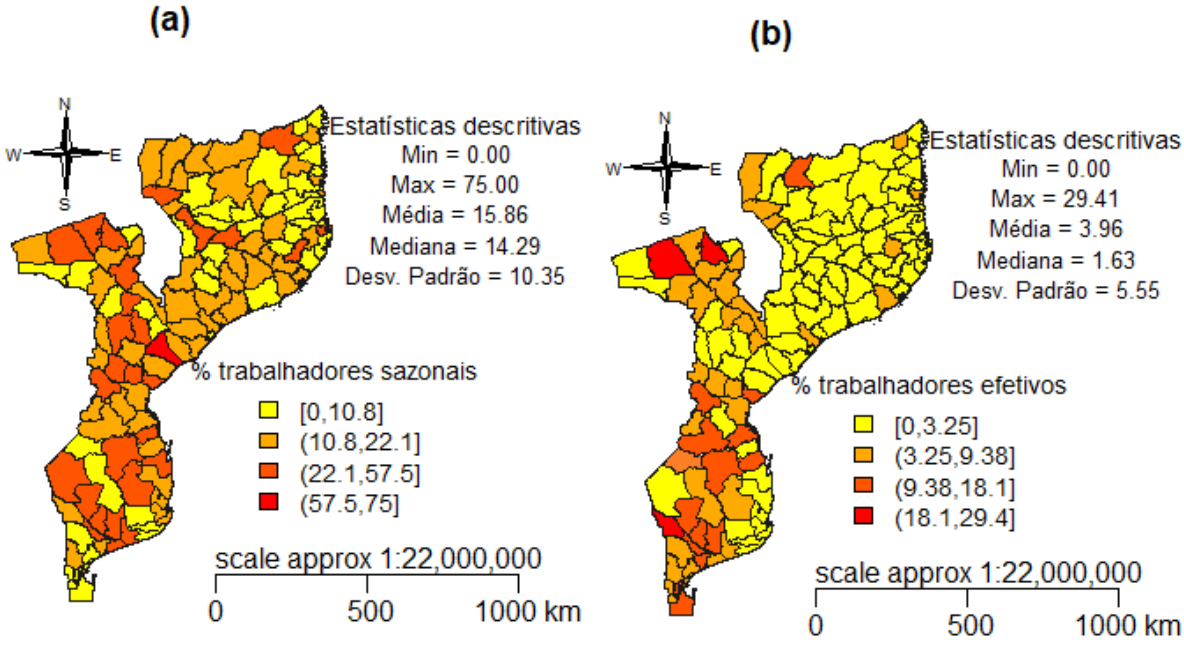


Fonte: Do autor (2019).

Quanto ao número de trabalhadores efetivos e sazonais, a distribuição espacial dessas variáveis é apresentada na Figura 17. Em geral os produtores usam mais trabalhadores sazonais do que efetivos, isto é, cerca de 4% dos produtores usam em média trabalhadores efetivos, enquanto que para o caso dos trabalhadores sazonais, tem-se em média

aproximadamente 16% de produtores recorrendo a esse tipo de mão de obra. Existem distritos cujo percentual de produtores que contrata mão de obra sazonal, atinge um percentual de 75%. Já, para o caso da mão de obra efetiva, o percentual de produtores que recorre a esse tipo atinge no máximo 29%. Ambas as fontes de mão de obra, juntamente com a mão de obra familiar, possuem uma relação direta com a decisão de adotar o uso de variedade melhoradas de milho devido aos tratos culturais que essas variedades demandam.

Figura 17 - Distribuição percentual de tipo de mão de obra: (a) Trabalhadores sazonais; (b) Trabalhadores efetivos.

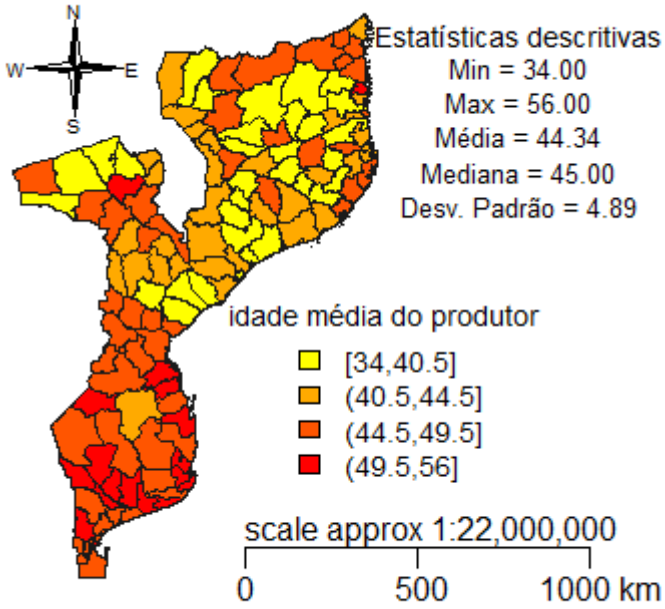


Fonte: Do autor (2019).

A idade do produtor é um dos fatores sócio-demográficos que está ligado aos anos de experiência dentro da atividade agrícola, embora os produtores mais jovens mostrem-se mais flexiveis no processo de tomada de decisão em relação aos mais antigos. De uma forma geral, os produtores possuem uma idade média de 44 anos, sendo que 50% dos distritos apresentam produtores com mais de 45 anos de idade. A zona sul do país parece concentrar a maior parte dos produtores com esse perfil (FIGURA 18). Produtores com uma idade mais avançada são caracterizados por possuirem mais anos de experiência e, portanto, maior conhecimento acumulado das técnicas de produção agrícola e mais cautelosos no processo de tomada de decisão. Por outro lado, os produtores mais novos são mais flexíveis no processo de tomada

de decisão. Por essa razão, para essa variável, a expectativa do seu efeito na decisão do uso de sementes melhoradas de milho é indeterminável.

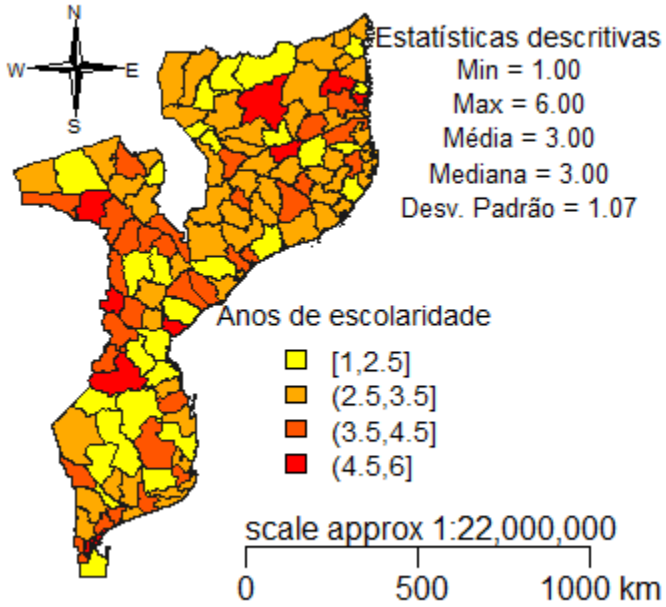
Figura 18 - Distribuição espacial da idade média dos produtores por distrito.



Fonte: Do autor (2019).

Na Figura 19 descreve-se a distribuição espacial do nível de educação média do produtor em termos de anos de escolaridade. O perfil educacional varia de 1 ano de escolaridade até 6 anos com uma média de 3 anos de escolaridade completa, que coincide com a mediana indicando que cerca de 50% dos distritos tem produtores que possuem uma escolaridade de até os 3 anos completos. Em geral, o perfil educacional dos produtores é bastante baixo. Mwuangi e Kariuki (2015) afirmam que muitos estudos de adoção de tecnlogias agrária têm apontado que o nível de educação do produtor tem uma relação positiva com a decisão de adotar uma determinada tecnologia. Além disso, Mignouna et al. (2011) salientam que um alto nível de educação do produtor aumenta sua habilidade de obter e processar informações relevante para adoção de novas tecnologias. Já, autores como Samiee et al. (2009) e Uematsu e Mishra (2010) reportam ausência de efeito ou um efeito negativo da educação formal na adoção de tecnologias agrárias. Assim, a expectativa da direção do efeito da educação na adoção de variedades melhoradas de milho é indeterminável.

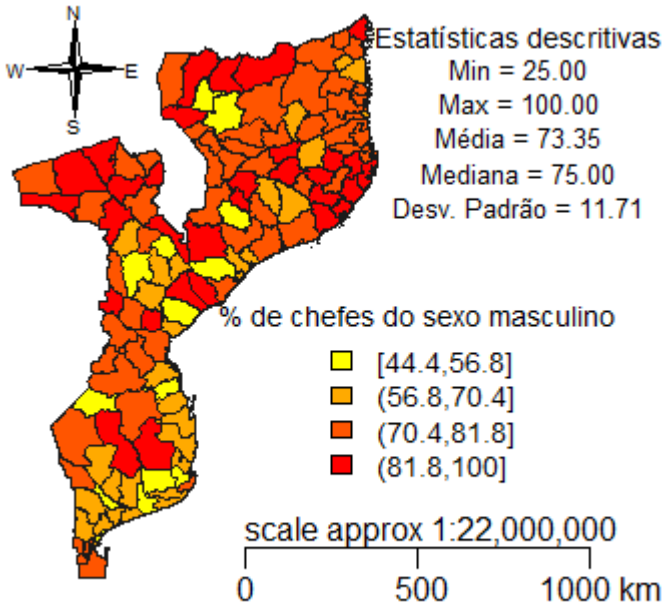
Figura 19 - Distribuição espacial do perfil educacional dos produtores em anos de escolaridade.



Fonte: Do autor (2019).

Na Figura 20 esta descrita a distribuição espacial do percentual dos chefes de família do sexo masculino em cada distrito. Embora a grande maioria dos chefes de família seja do sexo masculino existem alguns agregados familiares cujo chefe de família é do sexo feminino devido a existência de sociedades matrilineares em algumas regiões do país, mas também pelo fato de algumas serem viúvas. Tem-se em média que cerca de 73% dos chefes de família são do sexo masculino. Segundo Bonana–Wabbi (2002) as questões de gênero nos estudos de adoção de tecnologias agrárias têm mostrado diferentes evidências no que diz respeito aos papeis dos homens e das mulheres no processo de adoção de tecnologias. Morris e Doss (2001) não encontraram efeito do gênero no estudo de adoção de variedades melhoradas de milho no Gana. Porém, Mesfin (2005) e Omonona et al. (2005) verificaram que existe um efeito significativo do gênero no processo de adoção de tecnologias agrárias pelo fato dos homens possuirem maior controle e acesso dos recurso de produção em relação às mulheres, devido a aspetos sócio-culturais.

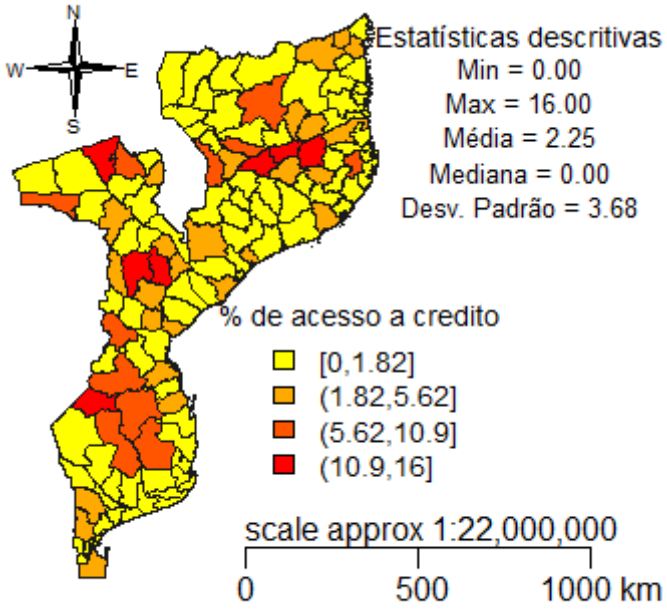
Figura 20 - Distribuição espacial do percentual de chefes de família do sexo masculino.



Fonte: Do autor (2019).

Quanto às covariáveis relacionadas aos fatores econômicos, elas encontram-se descritas nas Figuras 21 a 23. O acesso ao crédito por parte dos pequenos e médios produtores em Moçambique é citado por autores como Uaiene (2009) como uma das variáveis limitantes para a prática da produção agrícola. Na Figura 21 está descrita a distribuição espacial do percentual de produtores que têm acesso ao crédito. De uma forma geral observa-se que o acesso a esse fator ainda é bastante limitado para a maior parte dos produtores. Apenas 2,25% de produtores têm em média acesso ao crédito e só um número ínfimo de distritos apresentam um percentual de acesso ao crédito acima de 10%, mas que não ultrapassa os 16%. O baixo percentual de acesso ao crédito está aliado à fraca rede de instituições financeiras no meio rural mas também às altas taxas de juros praticadas por essas instituições, que são pouco atrativas aos produtores. O acesso ao crédito está diretamente ligado a capacidade financeira que o produtor possui para poder adquirir os fatores de produção necessários para o aumento da produção e produtividade agrícolas. Assim, espera-se que o acesso ao crédito influencie positivamente na decisão do uso de sementes melhoradas de milho.

Figura 21 - Distribuição espacial do percentual de produtores com acesso ao crédito.



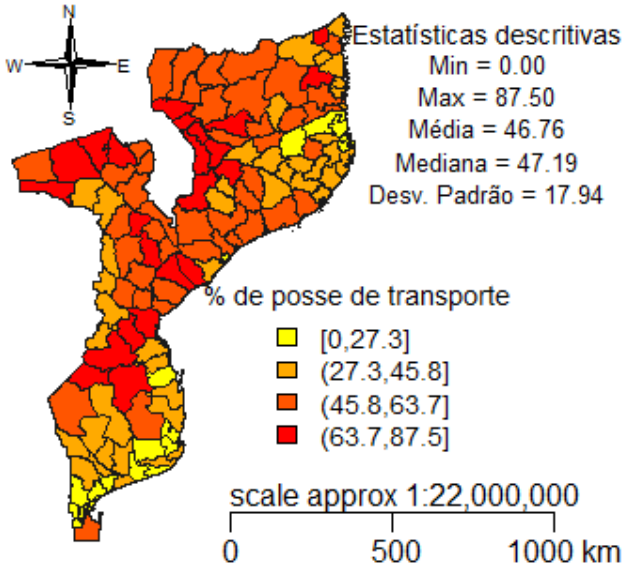
Fonte: Do autor (2019).

A bicicleta é o meio de transporte mais usado no meio rural Moçambicano. Ela não só é usada como um meio para se locomover, mas também para transportar os excedentes de produção agrícolas para comercialização nos mercados locais, como no transporte dos insumos de produção. Na Figura 22 tem-se a distribuição espacial do percentual de produtores que usam esse meio de transporte. Embora existam alguns distritos que não recorrem a este meio de transporte, a maior parte dos produtores usam este tipo de transporte nas suas atividades. A posse de meio de transporte facilita o deslocamento tanto para a aquisição dos insumos agrícolas como para o escoamento da produção. Nesse sentido, espera-se que os produtores que possuem acesso a esse meio de transporte estejam mais propensos a usar variedades melhoradas de milho.

A disponibilidade de celeiros melhorados é um fator que garante um sistema de armazenamento dos excedentes da produção de cereais e de culturas do grão. Ela não só permite que o produtor possa produzir em larga escala mas também permite que ele possa armazenar o seu produto e comercializá-lo nos períodos de fraca oferta, nos quais os preços são mais atrativos. A Figura 23 descreve a distribuição espacial do percentual de produtores que possuem celeiros melhorados que indica uma média de cerca de 13% de produtores com esse sistema de armazenamento. O acesso aos celeiros mostra-se espacialmente distribuído de forma assimétrica, evidenciado alguns distritos em que os produtores não possuem acesso ao

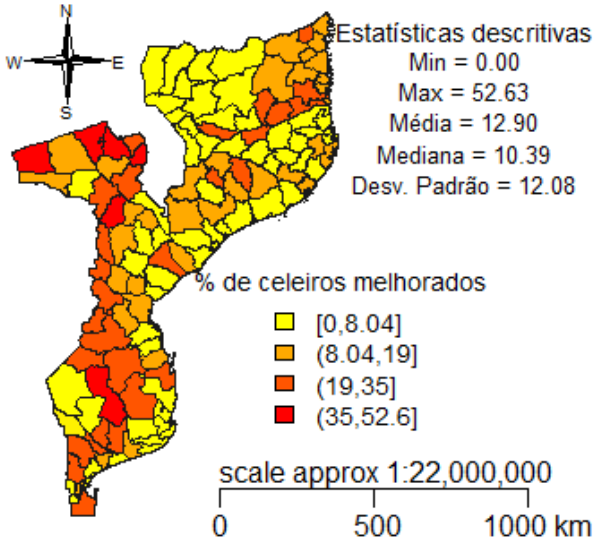
sistema de armazenamento mas também indica que existem distritos em que o percentual de acesso atinge cerca de 53%. Assim, espera-se que os produtores que possuem acesso aos sistemas de armazenamento estejam mais inclinados a adotar o uso de sementes melhoradas devido ao seu alto potencial de produtividade.

Figura 22 - Distribuição espacial do percentual de produtores com posse de meio de transporte (bicicleta).



Fonte: Do autor (2019).

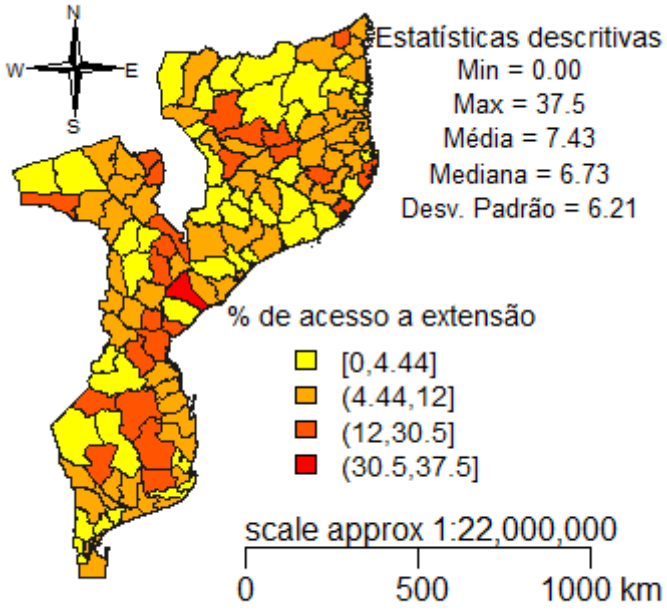
Figura 23 - Distribuição espacial do percentual de acesso a celeiros melhorados.



Fonte: Do autor (2019).

No concernente às variáveis relacionadas com os fatores institucionais, a descrição dos mesmos encontra-se apresentada pelas Figuras 24 a 26. Para o caso do acesso aos serviços de extensão agrária, verifica-se que em geral, seu acesso é bastante baixo, existindo distritos que não possuem acesso a estes serviços e outros com um percentual de acesso até 37,5% (FIGURA 24). Em termos médios, só cerca de 7,5% dos pequenos e médios produtores é que possuem acesso aos serviços de extensão. O acesso a esses serviços é feito pelos agentes de extensão responsáveis pela disseminação das técnicas de produção ao nível dos produtores. Isso confere a possibilidade dos produtores adotarem novas tecnologias agrárias que preconizam o aumento da produção e produtividade das culturas de forma sustentável. Nesse sentido, espera-se que essa variável possua uma relação direta na decisão do uso de sementes melhoradas de milho.

Figura 24 - Distribuição espacial do percentual de acesso aos serviços de extensão agrária.

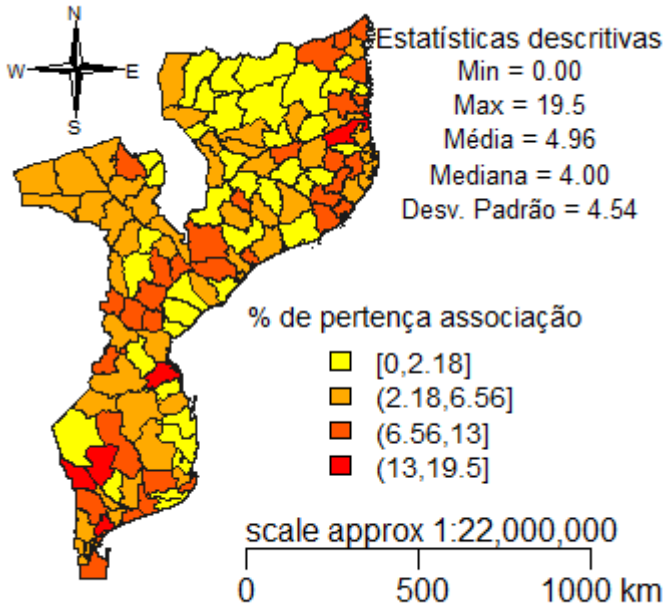


Fonte: Do autor (2019).

O associativismo por parte dos pequenos e médios produtores, é uma das variáveis de carácter institucional que constitui uma das estratégias preconizadas pelo Plano Estratégico do Desenvolvimento do Setor Agrário (PEDSA) para o aumento da produção e comercialização de produtos agrícolas em larga escala (MINAG, 2011). Na Figura 25 descreve-se a distribuição espacial do percentual de produtores que pertencem a alguma associação agrícola. Em termos gerais, verifica-se que a maior parte dos distritos possui menos de 5%

dos produtores afiliados a alguma associação agrícola, embora existam alguns distritos que têm até 19,5% dos produtores inscritos nalguma associação. O associativismo permite que os produtores tenham a possibilidade de trocar experiências sobre as técnicas de produção por eles adotadas e daí influenciar outros membros da associação a aplicarem tais técnicas. Katungi e Akankwasa (2010) e Uaiene (2009) encontraram que produtores pertencentes a alguma associação eram mais propensos em adotar novas tecnologias agrárias em relação aos produtores que não faziam parte de nenhum grupo comunitário. Assim, espera-se que o fato de pertencer a alguma associação agrícola influencie diretamente na decisão da adoção de variedades melhoradas de milho.

Figura 25 - Distribuição espacial do percentual de produtores que pertencem a alguma associação agrícola.

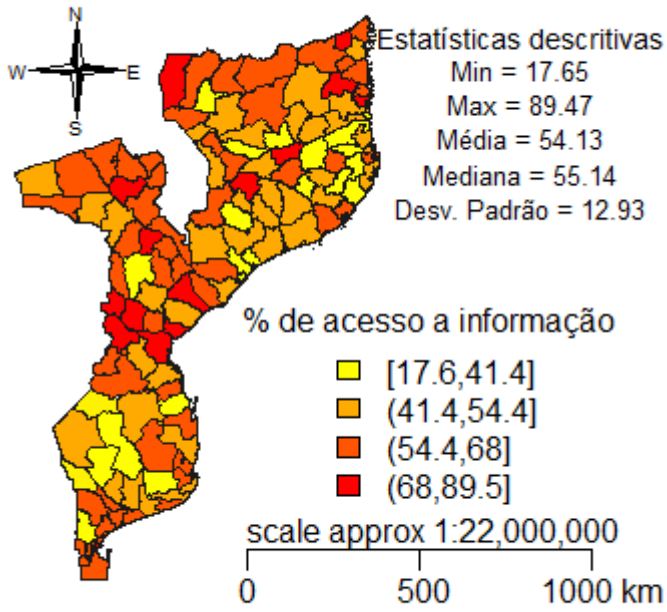


Fonte: Do autor (2019).

O acesso a informação sobre as novas técnicas de produção agrícola também é um dos fatores catalizadores para incrementar a produção e produtividade agrícola. As fontes de informação podem ser de natureza diversa, como rádio, tv, agentes de extensão entre outros. Nesse trabalho considera-se o acesso a informação a todos os produtores que possuem rádio, pois este é o canal de informação mais disponível e utilizado no meio rural. Na Figura 26 apresenta-se a distribuição espacial do percentual de produtores que têm acesso a informação (possuem rádio). Em média cerca de 54% dos produtores possuem acesso a este meio de informação, havendo distritos com um percentual de produtores com acesso ao rádio de cerca

de 90%. Igualmente ao caso de outras variáveis anteriormente descritas, o acesso a informação tem um efeito direto na decisão de adoção de sementes melhoradas de milho.

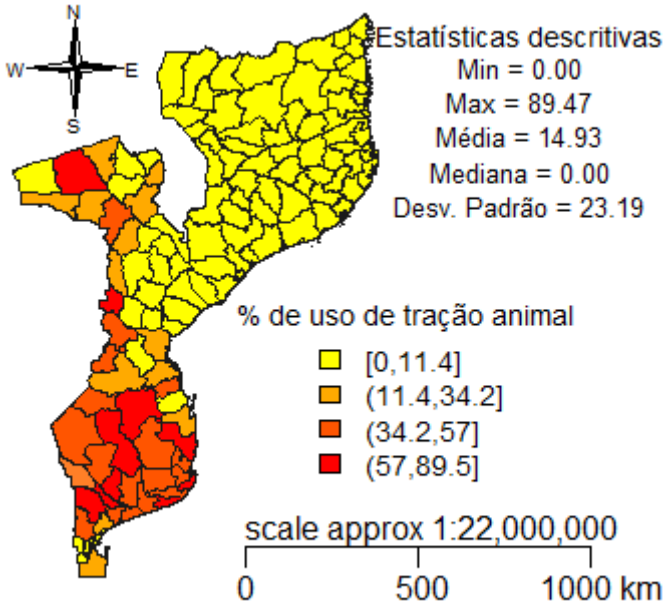
Figura 26 - Distribuição epacial do percentual de produtores com acesso a informação.



Fonte: Do autor (2019).

O uso de tração animal foi a única variável considerada como fator tecnológico. Essa tecnologia empregada no processo de preparação do solo, tem sido uma alternativa para muitos produtores que não têm acesso à maquinas agrícolas como tratores, charruas, grades, etc. Além disso, a tração animal mostra-se economicamente mais viável pelo fato de algumas regiões do país (região sul e uma parte do região centro) apresentarem um bom efetivo de gado bovino, o que facilita a disponibilidade dessa tecnologia. Porém, o uso dessa tecnologia ao longo de todo país apresenta uma assimetria acentuada, conforme descrito na Figura 27. Observa-se que mais de 50% dos distritos não usam a tração animal. Porém, existem distritos em que aproximadamente 90% dos seus produtores aplicam essa tecnologia. Em termos médios, só cerca de 15% dos produtores é que usam essa tecnologia. O uso de tração animal permite que o produtor tenha maior possibilidade de lavrar áreas maiores e assim aumentar sua produção. Dessa forma, é esperado que os produtores que usam a tração animal estejam mais propensos a adotar o uso de variedade melhoradas de milho que possuem maiores rendimentos.

Figura 27 - Distribuição espacial do percentual de produtores que usam tração animal.

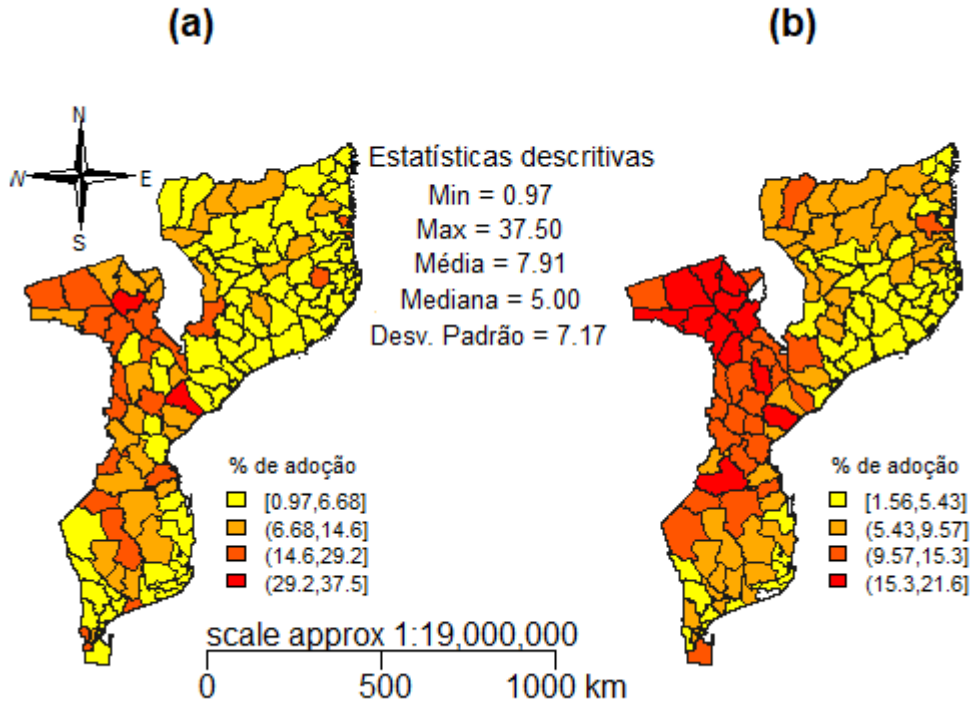


Fonte: Do autor (2019).

4.4.2 Análise exploratória da proporção de produtores que usaram variedades melhoradas de milho

Na Figura 28 descreve-se o mapa temático da proporção de produtores que usaram variedades melhoradas de milho para os 128 distritos em Moçambique, como a superfície suave dessa variável estimada com base na média móvel espacial. De uma forma geral verifica-se que a proporção de adoção de variedades melhoradas de milho varia de menos de 5% até um máximo de 37,5% com uma média por distrito estimada em aproximadamente 8%. Observa-se que a maior parte dos distritos possui um baixo nível de adoção de variedades melhoradas, isto é, cerca de 50% dos distritos apresentam um nível de adoção de variedades de milho inferior a 5%. Os distritos com maior percentual de adoção são pertecentes a zona centro do país com maior enfoque às províncias de Tete e Manica. Isto é facilmente visualizado quando se observa o mapa de suavização espacial indicando que os distritos pertecentes a estas duas províncias tendem a apresentar os maiores percentuais de adoção, ou seja, existe uma tendência de redução da adoção de sementes melhoradas no sentido oeste leste, mas também observa-se que os distritos na região sul e norte do país têm tendência a apresentar níveis baixos de adoção de sementes melhoradas quando comparados com os distritos da região central.

Figura 28 - Mapa temático para proporção de produtores que usaram semente melhorada de milho: (a) Dados originais; (b) Média móvel espacial.



Fonte: Do autor (2019).

Na Tabela 5, encontram-se as estimativas das estatísticas global de Moran, Geary e Getis e Ord e a respectiva avaliação de sua significância baseada no teste de permutação. Para todos os casos verifica-se que os índices indicam a ocorrência de uma associação espacial positiva entre os valores da proporção de produtores que usaram variedades melhoradas de milho com os seus vizinhos.

Tabela 5 - Estimativas das estatísticas globais de Moran, Geary e Getis.

Índices	**Estimativa**	**Valor Esperado**	**Variância**	**Valor p**
Moran	0,307	-0,008	0,0030	0,00000002
Geary	0,696	1,000	0,0042	0,00000149
Getis	0,047	0,038	0,0006	0,00009439

Fonte: Do autor (2019).

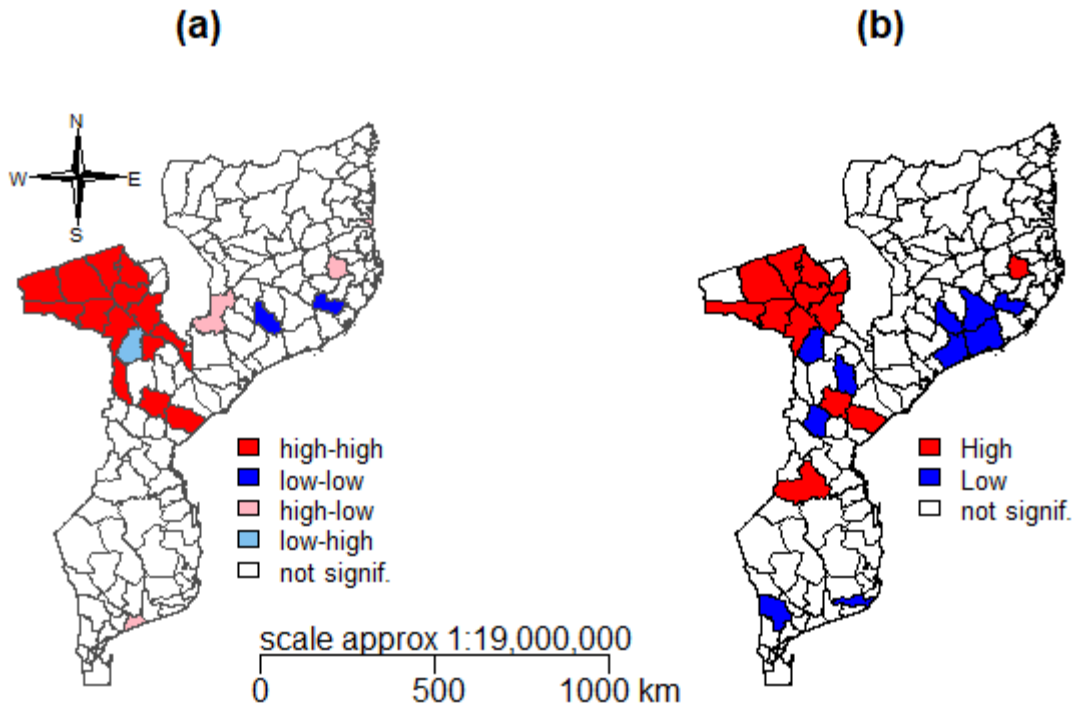
Os indicadores descritos na Tabela 5 são globais e não são capazes de indicar os locais ou distritos que apresentam associação espacial local entre si. Na Figura 29 descreve-se o "LISA cluster map", bem como o mapa para a estatística local G que visa evidenciar os locais de ocorrência do tipo de associação espacial ("*hot spot*" e "*cold spot*").

O "LISA cluster map" indica ocorrência de áreas com denominação "high - high" e "low-low" que são os locais com associação espacial positiva, podendo formar agrupamentos

na região do estudo. Além disso, esse mapa identifica os locais com denominação “low – high” e “high-low” que correspondem às regiões com associação espacial negativa mostrando dissimilaridade entre vizinhos no que diz respeito aos valores da proporção dos produtores que usaram variedades melhoradas de milho.

Pelas estimativas da estatística local *G*, verifica-se que são obtidos resultados similares quando comparados com o “LISA cluster map”. Contudo, ocorre uma pequena diferença pelo fato da estatística *G* não indicar locais de associação espacial negativa (GETIS; ORD, 1992). A ocorrência dos “*hot spots*” (áreas com classificação “high”), por um lado, corresponde aos locais cujo agrupamento ocorre para maiores valores da proporção de produtores que usaram sementes melhoradas de milho. Por outro lado, as áreas com classificação “low” evidenciam os locais onde há ocorrência dos “*cold spot*” que correspondem aos distritos cuja formação de grupos é evidenciada para valores baixos da adoção de variedades melhoradas de milho.

Figura 29 - Autocorrelação espacial local: (a) LISA cluster map; (b) Estatística G local.



Fonte: Do autor (2019).

Nesse sentido, verifica-se que a proporção de produtores que adotou sementes melhoradas de milho constitui uma variável espacialmente autocorrelacionada e, esse grau de dependência espacial deve ser levado em consideração durante o processo de modelagem uma

vez que a não inclusão da correlação espacial presente na variável resposta pode conduzir a conclusões erradas.

4.4.3 Modelagem com aplicação das EEG em dados espaciais em áreas

Na Tabela 6 têm-se os resultados das estimativas dos parâmetros do modelo linear generalizado com a função de ligação logit (modelo logístico) após a aplicação do procedimento se selecção de covariáveis entre a proporção dos produtores que usaram sementes melhoradas de milho e as diferentes covariáveis usadas no estudo. Das 13 covariáveis consideradas nas análises, obteve-se um modelo que inclui nove covariáveis das quais quatro correspondem aos fatores sócio demográficos, duas a fatores econômicos e institucionais e uma ao fator tecnológico. As variáveis destacadas nos fatores sócio-demográficos são: o tamanho médio da família, a idade média do produtor, o uso de trabalhadores sazonais e efetivos. O acesso ao crédito e a posse de celeiros melhorados correspondem às variáveis descriminadas nos fatores econômicos. Já, o acesso a informação e os serviços de extensão são as variáveis selecionadas nos fatores institucionais. A variável de âmbito tecnológico (tração animal) também foi destacada no modelo (TABELA 6).

Tabela 6 - Resultados para o ajuste do MLG (modelo logístico).

Covariáveis	**Descrição**	**Estimativa**	**Erro padrão**	**Z**
Intercepto		-2,82	0,70	-4,05*
Tamanho da família	Contínua	0,31	0,07	4,41*
Idade do produtor	Contínua	-0,06	0,01	-4,17*
Trabalhadores sazonais	1 - se usa; 0 – c.c.	0,78	0,37	2,12*
Trabalhadores efetivos	1 - se usa; 0 – c.c.	0,59	0,14	4,24*
Acesso ao crédito	1 - com acesso; 0 – c.c.	0,19	0,10	$1,83^{ns}$
Celeiros melhorados	1 - se tem acesso; 0 – c.c.	0,34	0,21	$1,64^{ns}$
Acesso a extensão	1 - com acesso; 0 – c.c.	0,36	0,15	2,35*
Acesso a informação	1- maior que 50%; 0 – c.c.	0,22	0,12	$1,82^{ns}$
Tração animal	1 - se usa; 0 – c.c.	0,32	0,16	2,03*
QIC = 89,96; CIC = 10,0; RJC = 1,36				

* siginificativo a 5%; ^{ns} não significativo a 5%; c.c. – caso contrário.
Fonte: Do autor (2019).

Os sinais das estimativas dos parâmetros para todas as covariáveis são positivos com a exceção da idade do produtor cuja estimativa apresentou um valor negativo. Esses sinais coincidem com os resultados esperados para todas as covariáveis. Além disso, quase todas as

covariáveis possuem um efeito significativo no modelo, com a exceção do acesso ao crédito, o acesso a informação e o acesso aos celeiros melhorados (p > 0,05).

A análise dos resíduos do modelo ajustado pela aplicação do índice de Moran, mostrou que os mesmos apresentam autocorrelação espacial (I_{res} *= 0,14, valor p = 0,004*). Isso comprova que a autocorrelação espacial presente na variável resposta não foi absorvida pelas covariáveis do modelo e, quando a mesma é ignorada, pode conduzir a conclusões erradas. Assim, a inclusão da dependência espacial presente na variável resposta foi feita com a aplicação das equações de estimação generalizadas (EEG) na qual é definida uma matriz de correlação espacial de trabalho conforme proposta apresentada na descrição metodológica nesse trabalho.

Na Tabela 7 têm-se os resultados do processo de estimação com aplicação das EEG na qual é utilizada a matriz de correlação espacial de trabalho definida com base no índice de Moran da variável resposta usando as mesmas covariáveis apresentadas no ajuste do modelo logístico.

Tabela 7 - Estimativas dos parâmetros do modelo com aplicação das EEG.

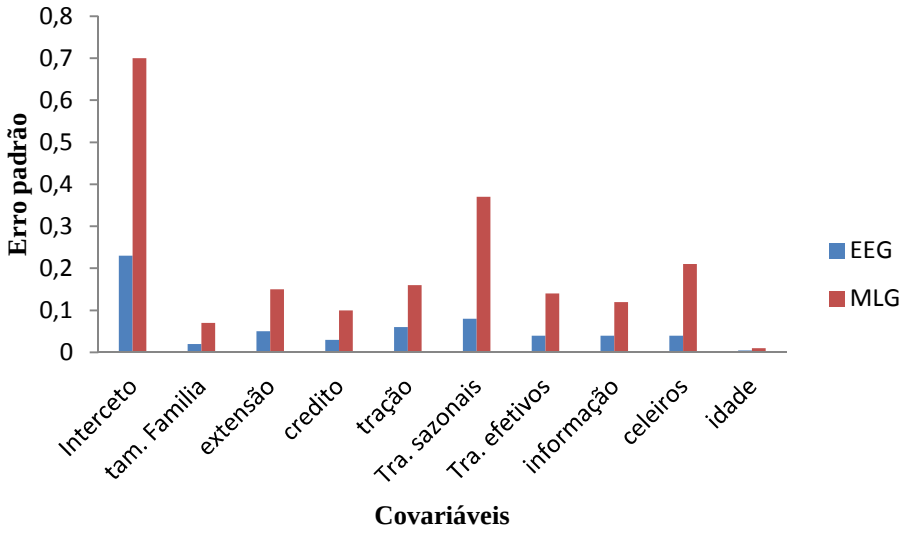
Covariáveis	**Descrição**	**Estimativa**	**Erro padrão**	**Z**
Intercepto		-4,14	0,19	-21,80*
Tamanho da família	Contínua	0,30	0,02	18,05*
Idade do produtor	Contínua	-0,03	0,004	-9,21*
Trabalhadores sazonais	1 - se usa; 0 – c.c.	1,10	0,07	14,90*
Trabalhadores efetivos	1 - se usa; 0 – c.c.	0,69	0,03	20,73*
Acesso ao crédito	1 - com acesso; 0 – c.c.	0,16	0,02	7,32*
Celeiros melhorados	1 - se tem acesso; 0 – c.c.	0,46	0,05	8,76*
Acesso a extensão	1 - com acesso; 0 – c.c.	0,48	0,03	14,86*
Acesso a informação	1- maior que 50%; 0 – c.c.	0,41	0,03	14,69*
Tração animal	1 - se usa; 0 – c.c.	0,14	0,04	3,36*
Parâmetro de associação (λ)		0,33	0,11	3,00*
QIC = 82,94; CIC = 6,34; RJC = 1,0				

* Significativo a 5%; cc – caso contrário.
Fonte: Do autor (2019).

Igualmente ao caso do modelo linear generalizado, as estimativas dos parâmetros das EEG apresentaram sinais positivos com exceção da estimativa para a idade do produtor que foi negativa. Além disso, os valores das estimativas em ambos modelos são muito similares pelo fato dos estimadores serem consistentes (LIANG; ZEGER, 1986). De uma forma geral, o sinal das estimativas corrobora o resultado esperado para todas as covariáveis.

Na Figura 30 tem-se as estimativas do erro padrão obtidas para ambos estimadores (MLG e EEG) em cada covariável do modelo. Vê-se claramente que para todas as covariáveis, os estimadores baseados na aplicação das EEG apresentam estimativas do erro padrão inferiores ao MLG.

Figura 30 - Estimativas do erro padrão dos estimadores MLG e EEG.



Fonte: Do autor (2019).

Com aplicação das EEG na estimação dos parâmetros, observa-se que todas as covariáveis do modelo apresentam efeito significativo devido ao ganho na eficiência dos estimadores o que também influenciou no poder dos testes. A título de exemplo, o erro padrão do estimador para a covariável acesso ao crédito passa de uma estimativa de 0,1 quando é ignorada a estrutura de correlação (MLG) para uma estimativa de 0,02 quando a autocorrelação espacial é levada em consideração por meio do uso do matriz de correlação espacial de trabalho usando a estrutura Toeplitz. Além disso, os critérios QIC, CIC e RJC usados para seleção da matriz de correlação de trabalho também indicam que a matriz de correlação de trabalho construída com base no índice de Moran mostra-se melhor em relação a matriz que ignora a estrutura espacial pelo fato dos três critérios apresentarem os menores valores para a estrutura das EEG.

Esses resultados estão coerentes com os resultados obtidos pelo processo de simulação dos dados os quais indicaram uma diminuição da variância dos estimadores dos parâmetros devido a boa especificação da matriz de correlação espacial de trabalho.

Na Tabela 8 descrevem-se os resultados do efeito marginal de cada covariável usada no estudo dada pela razão de chances. Essa medida descreve a chance do produtor adotar o uso de semente melhorada de milho dado o "incremento" em 1 unidade na covariável em análise.

Tabela 8. Estimativas de razão de chances para as covariáveis do modelo ajustado.

Covariáveis	**Descrição**	**Estimativa($\widehat{\boldsymbol{\beta}}$)**	**Razão de chances (exp($\widehat{\boldsymbol{\beta}}$))**
Tamanho da família	Contínua	0,30	1,35
Idade do produtor	Contínua	-0,03	0,97
Trabalhadores sazonais	1 - se usa; 0 – c.c.	1,10	3,00
Trabalhadores efetivos	1 - se usa; 0 – c.c.	0,69	1,99
Acesso ao crédito	1 - com acesso; 0 – c.c.	0,16	1,17
Celeiros melhorados	1 - se possui; 0 – c.c.	0,46	1,58
Acesso a extensão	1 - com acesso; 0 – c.c.	0,48	1,62
Acesso a informação	1- maior que 50%; 0 – c.c.	0,41	1,51
Tração animal	1 - se usa; 0 – c.c.	0,14	1,15
Parâmetro de associação (λ)		0,33	1,39

c.c – caso contrário.
Fonte: Do autor (2019).

Para o caso das variáveis sócio-demográficas relacionadas com a disponibilidade de mão de obra, isto é, o tamanho médio da família, uso de trabalhadores sazonais e efetivos, verifica-se que as mesmas possuem um efeito positivo na decisão de adoção de sementes melhoradas de milho. Os produtores que possuem famílias com mais membros, espera-se que a chance dos mesmos adotarem o uso de sementes melhoradas do milho aumente em 35% em relação aos produtores com famílias menos numerosas. Já, para o caso dos trabalhadores efetivos e sazonais, a contratação desse tipo de mão de obra irá influenciar na decisão de usar variedade melhoradas de milho em 99% e 200%, para os trabalhadores efetivos e sazonais, respectivamente. De fato a mão de obra baseada no uso de trabalhadores sazonais é que mais caracteriza o perfil da agricultura moçambicana, a qual é essencialmente familiar e cuja força de trabalho depende na sua maioria do capital humano. Autores como Mignouna et al. (2011) e Bonana–Wabbi (2002) também mostraram que a mão de obra tem um impacto positivo no uso de tecnologias melhoradas pelo fato de sua disponibilidade em larga escala constituir um fator preponderante durante a fase de introdução de novas tecnologias agrárias.

Outro fator sócio-demográfico que apresentou um efeito significativo no modelo é a idade do produtor. Do conjunto de covariáveis com efeito significativo no modelo, a idade média do produtor foi a única variável que apresentou um impacto negativo na decisão de

adoptar o uso de semente melhorada de milho. Os resultados sugerem que os produtores mais velhos são menos propensos em adotar o uso de sementes melhoradas em relação aos produtores mais jovens, isto é, o fato do produtor possuir uma idade mais avançada, reduz em 3% a chance de ele usar uma variedade melhorada de milho, comparativamente aos produtores mais novos. Resultados similares são reportados por diversos autores como Zavale (2005), Mauceri et al. (2005) e Adesina e Zinnah (1993), os quais afirmam que à medida que o produtor fica mais velho, existe uma tendência de ele se expor menos aos riscos, mas também ocorre uma diminuição do investimento na atividade agrária a longo prazo. Por outro lado, os produtores mais jovens têm tendência a maior exposição ao risco e sempre aptos em experimentar novas tecnologias. Contudo, Kariyasa e Dewi (2011) e Mwuangi e Kariuki (2015), reportam que, o fato do produtor possuir mais tempo dentro da atividade agrária, é sinônimo de acúmulo de conhecimento e experiência o que lhe permite avaliar melhor as informações referentes às novas tecnologias, comparativamente aos produtores mais jovens.

Em relação aos fatores econômicos, nomeadamente a posse de celeiros melhorados e o acesso ao crédito, verificou-se que estes contribuiram positivamente para adoção de sementes melhoradas do milho. Para o caso do primeiro fator, verifica-se que, se o produtor possui acesso a um celeiro melhorado, então a chance que ele tem em usar variedades melhoradas de milho é de 58% a mais do que o produtor que não tem disponibilidade desse sistema de armazenamento de cereais e grãos. De fato a disponibilidade e acesso a celeiros melhorados garante um ótimo sistema de armazenamento do milho em larga escala. Isso tem também influência no acesso ao mercado e na decisão sobre o período no qual o milho será comercializado. Preferencialmente, os produtores com acesso a celeiros melhorados, escolhem comercializar seus produtos em períodos de maior demanda, nos quais os preços são mais atrativos e conferem maior lucratividade. Já, para o segundo fator econômico, os resultados sugerem que, se o produtor tem acesso ao crédito, a chance que o mesmo possui em adotar o uso de sementes melhoradas de milho aumenta em 17%, comparativamente aos produtores sem acesso a fontes de financiamento. Vários estudos empíricos têm mostrado que o acesso ao crédito influencia positivamente no uso de novas tecnologias pelo fato de permitir ao produtor uma maior capacidade em obter os fatores de produção necessário para garantir uma boa produtividade. Uaiene (2009) também encontrou um efeito positivo do acesso ao crédito no estudo da adoção de tecnologias agrárias em Moçambique. Contrariamente, Zavale et al. (2005) num estudo sobre adoção de sementes melhoradas de milho encontraram uma relação inversa entre o acesso ao crédito e o uso de sementes melhoradas. Os autores afirmam que a justificativa para tal relação, está aliada ao fato das instituições financeiras concederem

crédito apenas aos produtores que possuem outras fontes de renda fora da atividade agrária. Nesse caso, esses produtores têm tendência em canalizar seus investimentos em outras culturas de rendimento como o tabaco e algodão ou em outras atividades que se mostram mais rentáveis.

No que se refere aos fatores institucionais, isto é, o acesso a informação e aos serviços de extensão, verifica-se que essas covariáveis também possuem um impacto positivo na decisão de adoção de uso de sementes melhoradas de milho. Dado que o produtor tem acesso a informação e aos serviços de extensão, a chance do mesmo usar variedades melhoradas de milho aumenta em 51% e 62%, respectivamente. Uaiene (2009) num estudo sobre adoção de tecnologias agrárias em Moçambique encontrou que os serviços de extensão possuem efeito significativo apenas no uso de tração animal. Para outras tecnologias como o uso de sementes melhoradas, fertilizantes, pesticidas e mecanização, o acesso a extensão não mostrou influência significativa nos modelos ajustados. Porém, vários autores como Mignouna et al. (2011), Sserunkuuma (2005) e Akudugu et al. (2012) são unânimes em destacar os serviços de extensão como uma força motriz no processo de adoção de tecnologias agrárias. Além disso, Mwuangi e Kariuki (2015) afirmam que a influência dos serviços de extensão, na decisão de adoção de tecnologias agrárias, pode suprir o efeito do baixo nível de escolaridade que caracteriza muitos produtores nos países em desenvolvimento.

Igualmente ao caso das outras covariáveis, o uso de tração animal, identificado como o único fator tecnológico, apresentou uma relação positiva sobre a decisão de adoção de sementes melhoradas de milho. Contudo, seu impacto foi relativamente menor quando comparado com as outras covariáveis. Ora vejamos, se o produtor usa tração animal, a chance que o mesmo possui em adotar variedade melhoradas de milho é de 15% a mais comparativamente ao produtor que não usa a tecnologia. Em todas as variáveis analisadas com efeito positivo na adoção de sementes melhoradas de milho, o uso de tração animal é que apresentou a menor razão de chances.

Quanto ao parâmetro de associação espacial, que quantifica a medida de influência entre áreas vizinhas, este também apresentou um impacto positivo sobre a decisão de uso de sementes melhoradas de milho, isto é, o fato dos produtores se encontrarem em áreas (distritos) mais próximas entre si, aumenta em 39% a chance desses adotarem o uso de variedade melhoradas de milho, comparativamente ao caso em que é ignorada a associação espacial. Comparativamente aos produtores mais distantes, os produtores mais próximos têm a tendência em apresentar similaridade quanto ao grau de adoção de sementes melhoradas de milho pelo fato dos mesmos interagirem entre si. Essa interação, produz um efeito de

dependência espacial entre as proporções de adoção de sementes melhoradas de milho, comprovada pelas estatísticas local de Getis e Ord e de Moran (FIGURA 29). De fato, a inclusão da dependência espacial nas análises atráves do uso das EEG com base na metodologia proposta nesse trabalho permitiu ilustrar o impacto da dependência espacial na decisão de adotar semente melhorada de milho através da razão de chances do parâmetro de associação.

4.5 Considerações finais

A definição da matriz de correlação espacial de trabalho baseada no índice de Moran, mostrou um ganho na eficiência dos estimadores em relação ao índice de Geary e a estrutura independente.

De uma forma geral, a estrutura de matriz de correlação espacial de trabalho AR(1) mostrou-se eficiente em relação a estrutura Toeplitz embora para valores de índice de Moran inferiores a 0,3 ambas estruturas sejam igualmente eficientes.

Os dados da proporção de produtores que adotaram sementes melhoradas de milho em Moçambique no ano de 2012, apresentam uma autocorrelacão espacial positiva significativa.

A aplicação das EEG no estudo da adoção de sementes melhoradas de milho em Moçambique, permitiu observar que os fatores determinantes do uso dessa tecnologia são de ordem sócio-demográfico, econômicos, tecnológicos e institucionais. São eles: a disponibilidade de mão de obra, a idade do produtor, o acesso ao crédito, a posse de celeiros melhorados, o uso de tração animal, o acesso à informação e aos serviços de extensão. Além disso, a interação entre os produtores avaliada através da dependência espacial foi também identificada como um dos fatores determinantes do uso de sementes melhoradas de milho.

A modelagem de dados espaciais em áreas usando as EEG baseadas na metodologia proposta nesta tese pode ser aplicada não somente para dados de tipo proporções mas também para outros tipos de dados, como contagens, respostas contínuas e dados binários.

Seria interessante aplicar as EEG usando a metodologia aqui proposta em bancos de dados de outras áreas de conhecimento como por exemplo em estudos epidemiológicos.

Para futuros trabalhos pode-se propor uma extensão da metodologia aqui proposta na análise espaço-tempo em dados de área.

Ainda para futuros trabalhos pode-se propor metodologias para análise de resíduos em EEG em dados espaciais em áreas. Além disso, seria importante incluir a estimação do parâmetro de dispersão nas análises usando o resíduo de Pearson ou outro tipo de resíduo.

5 CONCLUSÃO

Os resultados obtidos nessa tese permitiram afirmar que apesar das equações de estimação generalizadas estarem amplamente desenvolvidas para dados longitudinais, sua aplicação na análise de dados espaciais em áreas mostra-se eficaz na modelagem de dados cuja variável resposta constitui uma proporção que se apresente espacialmente autocorrelacionada.

A inclusão do índice de Moran na definição da matriz de correlação de trabalho aumentou a eficiência dos estimadores dos parâmetros do modelo.

O método de simulação de dados de área proposto nessa tese, baseado no índice de Moran, mostrou-se eficaz para os valores do índice definidos em quase todo espaço paramétrico com a exceção do intervalo (-1;-0,5).

A aplicação das equações de estimação generalizadas no estudo de adoção de variedades melhoradas de milho em Moçambique constitui uma alternativa à modelagem quando a autocorrelação espacial encontra-se presente nos dados. Além disso, com o uso das EEG foi possível quantificar o impacto da dependência espacial na decisão de adoção de variedades melhoradas de milho.

Os resultados obtidos nesta tese, juntos à uma pesquisa mais rigorosa, apresentam um grande potencial de auxílio às políticas agrárias mais direcionadas ao agronegócio do país.

REFERÊNCIAS

ADESINA, A.; ZINNAH, M. Technology characteristics, farmers' perceptions and adoption decisions: a tobit model analysis in Sierra Leone. **Agricultural Economics**. v. 9, n. 4, p. 297-311, 1993.

AKUDUGU, M.; GUO, E.; DADZIE, S. Adoption of modern agricultural production technologies by farm households in Ghana: what factors influence their decisions? **Journal of Biology, Agriculture and Healthcare**, v. 2, n. 3, p. 1 – 14, 2012.

ALBERT, P. S.; MCSHANET, L. M. Generalized estimating equation approach for spatially correlated binary data: application to the analysis of neuroimaging data. **Biometrics**, v. 51, n. 2, p. 627 – 638, Jun. 1995.

ALMEIDA, M. C. D. et al. Dinâmica intra-urbana da epidemia de dengue em Belo Horizonte, Minas Gerais, Brazil, 1996-2002. **Cadernos de Saúde Publica**, Rio de Janeiro, v. 24, n. 10, p. 2385-2395, out. 2008.

ANSELIN, L. **Exploring spatial data with GeoDaTM: a workbook.** Urbana: Spatial Analysis Laboratory Department of Geography. University of Illinois, 2005. 245 p.

ANSELIN, L. **Local indicators of spatial association - LISA**. Geographical Analysis, 1995. 115 p.

ASSUNÇÃO, R. M. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: Universidade Federal de São Carlos, 2001. 136 p.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. London: Longman, 1995. 413 p.

BIVAND, R. S.; PEBESMA, E. J.; GOMES, V. R. **Applied spatial data analysis with R**. 1st ed. New York: Springer, 2011. 378 p.

BIVAND, R. S.; PEBESMA, E. J.; GOMES, V. R. **Applied spatial data analysis with R**. 2nd ed. New York: Springer, 2013. 413 p.

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.

BRAJENDRA, C.S.; KALYAN, D. On the efficiency of regression estimators in generalised linear models for longitudinal data. **Biometrika**, v. 86, n. 2, p. 459-465, 1999.

CAVANE, E.; CUNGUARA, B.; JORGE, A. **Adopção de tecnologias agrárias em Moçambique**: **revisão, interpretação e síntese de estudos feitos**. OMR, 2013.

CAVANE, E.; DONAVAN, C. Determinants of adoption of improved maize varieties and chemical fertilizers in Mozambique. **Journal of International Agricultural and Extension Education**, v. 18, n. 3, p. 1-17, 2011.

CENACARTA. **Cartografia da divisão administrativa de Moçambique.** 1997. Disponível em < http:www.cenacarta.com>. Acesso em Abril de 2017.

CLIFF, A. D.; ORD, K. **Spatial processes**: **models and applications**. London: Pion, 1981.

CRESSIE, N. A. C. **Statistics for spatial data**. New York: J. Wiley, 1993. 900 p.

DIAS, P. **Analysis of incentives and disincentives for maize in Mozambique**. Technical notes series, MAFAP, FAO: Rome, 2013.

DONOVAN, C.; TOSTÃO, E. **Staple food prices in Mozambique.** Paper prepared for the Comesa policy seminar on "variation in staple food prices: Causes, consequence, and policy options", Maputo, Mozambique, Jan. 2010. 19 p.

DRUCK, S. et al. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004. 209 p.

FENG, G.F. et al. Border is better than distance? Contagious corruption in one belt one road economies. **Qual Quant.** DOI 10.1007/s11135-017-0579-3, Sep. 2017.

FISHER, W. D. On grouping for maximum homogeneity. **Journal of the American Statistical Association**, v. 53, p. 789–798, 1958.

FLAHAUT, B. et al. The local spatial autocorrelation and the kernel method for identifying black zones - A comparative approach. Accident analysis and prevention. **Elmsford,** v. 35, n. 6, p. 991-1004, Nov. 2002.

GEODA **Center for geospatial analysis and computation.** Version 9.5. 2010. Disponível em: <http://www.geodacenter.asu. edu>. Acesso em: Fev. 2017.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. **Geographical Analysis**, v. 24. n. 3, p. 189 – 206, 1992.

GRIFFITH, D. A. The Moran coefficient for non-normal data. **Journal of Statistical Planning and Inference**, Amsterdam, v. 140, n. 11, p. 2980-2990, Nov. 2010.

HALEKOH, U.; HOJSGAARD, S.; YAN, J. The R package geepack for generalized estimating equations. **Journal of Statistical Software**, v. 15, n. 2, p. 1 – 11, Jan. 2006.

HARDIN, J. W.; HILBE, J. M. **Generalized estimating equations.** 2nd ed. Florida: Taylor and Francis/CRC, 2013. 255 p.

HIN, L.Y.; WANG, Y.G. Working correlation structure identification in generalized estimating equations. **Stat. Med**, v. 28, p. 642–658, 2009.

HOWARD, J. et al. **Comparing yields and profitability in MADER's high- and low-Input maize programs 1997/98**. Survey Results and Analysis. Research Report nr. 39. MADER, 2000.

INAN, G.; LATIF, M. A. H. M.; PREISSER, J. A prediction criterion for working correlation structure selection in GEE. **Stat.ME**, arXiv:1803.06383, p. 1 – 29, 2018.

INE – Instituto Naciona de Estatística (Moçambique). **Anuário estatístico 2016.** Moçambique, 2017. 112 p.

INE - Instituto Naciona de Estatística (Moçambique). **Censo agropecuário 2009 – 2010**. Maputo: Moçambique, 2011. 115 p.

JANG, M. J. **Working correlation selection in generalized estimating equation**s. 2011. 259 p. Doctorate Thesis. University of Iowa. Iowa, 2011.

KARIYASA, K.; DEWI, A. Analysis of factors affecting adoption of integrated crop management farmer field school (Icm-Ffs) in swampy areas. **International Journal of Food and Agricultural Economics,** v. 1, n. 2, p. 29-38, 2011.

KATUNGI, E.; AKANKWASA, K. Community-based organizations and their effect on the adoption of agricultural technologies in Uganda: a study of banana (Musa spp.). Pest Management Technology. **Acta Horticulturae**, v. 879, p. 719 – 726, 2010.

LIANG, K. Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, v. 73, p. 13-22, 1986.

LIN, P.S.; CLAYTON, M. K. Analysis of binary spatial data by quasi-likelihood estimating equations. **The Annals of Statistics**. v. 33, n. 2, p. 542-555, 2005.

MACULUVE, T.V. **Improving dryland water productivity of maize through cultivar selection and planting date optimization in Mozambique**. 2011. 76 p. MSc Thesis Faculty of Natural and Agricultural Sciences, University of Pretoria. Pretoria, 2011.

MANUEL, L. et al. Spatial Linear Regression models in the infant Mortality analysis. **Multi-science Journal**, v. 13, p. 39-44, 2018.

MAUCERI, M. et al. **Adoption of integrated pest management technologies: A case study of potato farmers in Carchi, Ecuador:** Selected paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Providence, Rhode Island, Jul. 2005. 27 p.

MCCULLAGH, P. E; NELDER, J. A. **Generalized linear models**. 2nd ed. London: Chapman and Hall, 1989. 511 p.

MESFIN, A. **Analysis of factors influencing adoption of triticale and its impact**. The case of Farta District. 2005. 112 p. Msc. Thesis. School of graduate studies of Haramaya University. Ethiopia, 2005.

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.

MINAG - MINISTÉRIO DE AGRICULTURA (Moçambique). **Plano estratégico do desenvolvimento do sector Agrário 2011-2010**. Maputo: Moçambique, 2011. 76 p.

MONDINI, A.; CHIARAVALLOTI, F. N. Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. **Science of the**

Total Environment, Amsterdam, v. 393, n. 3, p. 241- 248, Apr. 2008.

MORRIS, M.; DOSS, C. How does gender affect the adoption of agricultural innovations? The case of improved maize technology in Ghana. Elsevier. **Agricultural Economics**, v. 25, p. 27 – 39, 2001.

MUDEMA, J. A.; SITOLE, R. F.; MLAY, G. **Rentabilidade da cultura do milho na zona sul de Moçambique: Estudo de caso do distrito de Boane**. Relatório Preliminar de Pesquisa No. 3P. IIAM, Out. 2012.

MWANGI, M.; KARIUKI, S. Factors determining adoption of new agricultural technology by smallholder farmers in developing countries. **Journal of Economics and Sustainable Development**, v. 6, n. 5, p. 208 – 217, 2015.

OMONONA, B.; ONI, O.; UWAGBOE, O. Adoption of improved cassava varieties and its impact on rural farming households in Edo State, Nigeria. **Journal of Agriculture and Food Information,** v. 7, n. 1, p. 40-45, 2005.

PAN, W. Akaike's information criterion in generalized estimating equations. **Biometrics**, v. 57, n. 1, p. 120 – 125, 2001.

POULIOU, T.; ELLIOTT, S. J. An exploratory spatial analysis of overweight and obesity in Canada. **Preventive Medicine**, San Diego, v. 48, n. 4, p. 362-367, Apr. 2009.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: <http://www.r-project.org>. Acesso em Jul. 2018.

RENAPRI. **Unfolding agricultural transformation in Africa**: Strategies for sustainable development. Cape Town: South Africa, 2017.

ROGERSON, P.; YAMADA, I. **Statistical detection and surveillance geographic clusters**. U.S: Chapman & Hall/CRC, 2009. 302 p.

ROTNITZKY, A.; JEWELL, N.P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. **Biometrika**, v. 77, n. 3, p. 485-97, 1990.

SAMIEE, A.; REZVANFAR, A.; FAHAM, E. Factors affecting adoption of integrated pest management by wheat growers in Varamin County, Iran: **African Journal of Agricultural Research**, v. 4, n. 5, p. 491-497, 2009.

SANTOS, S. M.; SOUSA, W. V. **Introdução à estatística espacial para a saúde pública.** Brasília, DF: Ministério da Saúde, Fundação Oswaldo Cruz, 2007. 123 p.

SEBER, G.A.F. **A Matrix handbook for statisticians**. New Jersey: John Wiley and Sons, 2008. 593p.

SSERUNKUUMA, D. The adoption and impact of improved maize and land management technologies in Uganda. **Electronic Journal of Agricultural and Development Economics**, v. 2, n. 1, p. 67 – 84, 2005.

UAIENE, R. **Introdução de novas tecnologias agrícolas e estratégias de comercialização no centro de Moçambique**. IIAM. Relatório de Pesquisa No. 2P, 2006.

UAIENE, R. N.; ARNDT. C.; MASTERS, W. A. **Determinant of agricultural technology adoption in Mozambique**. Ministry of Planning and Development Republic of Mozambique. Discussion papers No. 67E, 2009.

UEMATSU, H.; MISHRA, A. **Can education be a barrier to technology adoption?** Selected Paper prepared for presentation at the Agricultural & Applied Economics Association 2010 AAEA,CAES, & WAEA Joint Annual Meeting, Denver, Colorado, Jul. 2010. 38 p.

WALLER, L. A.; GOTWAY, C. A. **Applied spatial statistics for public health data**. Hoboken: J. Wiley, 2004. 518 p.

WANG, Y. G.; CAREY, V. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. **Biometrika**, v. 90, n. 1, p. 29 – 41, 2003.

WANG, Y. G.; LIN, X. Effects of variance-function misspecification in analysis of longitudinal data. **Biometrics**, v. 61, n. 2, p. 413-421, 2005.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. **Biometrika**, v. 61, n. 3, p. 439-447, 1974.

WERNECK, G. L. Georeferenced data in epidemiologic research. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 13, n. 6, p. 1753-1766, 2008.

ZAVALE, H.; MABAYA, E.; CHRISTY, R. **Adoption of improved maize Seed by smallholder farmers in Mozambique**. Department of Applied Economics and Management. Cornell University. Ithaca: New York, 2005. 24 p.

ZHANG, Y.J.; HAO, J. F.; SONG, J. The CO_2 emission efficiency, reduction potential and spatial clustering in China´s industry: Evidence from the regional level. Elsevier. **Applied Energy,** v. 174, p. 213-223, 2016.

APÊNDICE A – Matrizes envolvidas no processo de estimação

Matriz de correlação espacial de trabalho considerando apenas os distritos da província de Maputo para a estrutura Toeplitz m=1.

$$R(\rho)=\begin{bmatrix} 1 & 0 & 0 & \rho & 0 & \rho & \rho & \rho \\ 0 & 1 & \rho & 0 & 0 & 0 & \rho & 0 \\ 0 & \rho & 1 & 0 & \rho & 0 & \rho & 0 \\ \rho & 0 & 0 & 1 & \rho & \rho & \rho & 0 \\ 0 & 0 & \rho & \rho & 1 & 0 & \rho & 0 \\ \rho & 0 & 0 & \rho & 0 & 1 & 0 & \rho \\ \rho & \rho & \rho & \rho & \rho & 0 & 1 & \rho \\ \rho & 0 & 0 & 0 & 0 & \rho & \rho & 1 \end{bmatrix}.$$

Matriz de Derivadas $\mathbf{D_i} = \partial\mu_i/\partial\beta_j$, com i = 1, 2, ... 8 definindo o número de áreas e j = 0, 1, 2, isto é, considerando apenas duas covariáveis.

$$\frac{\partial\mu_i}{\partial\beta_0}=\frac{exp\{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}\}}{[1+exp\{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}\}]^2},$$

$$\frac{\partial\mu_i}{\partial\beta_1}=\frac{exp\{\beta_0+\beta_1x_{i1}+\beta_1x_{i2}\}x_{i1}}{[1+exp\{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}\}]^2}$$

e

$$\frac{\partial\mu_i}{\partial\beta_2}=\frac{exp\{\beta_0+\beta_1x_{i1}+\beta_1x_{i2}\}x_{i2}}{[1+exp\{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}\}]^2}.$$

Portanto a matriz **D** fica definida por:

$$\mathbf{D}=\begin{bmatrix} \frac{\partial\mu_1}{\beta_0} & \frac{\partial\mu_1}{\beta_1} & \frac{\partial\mu_1}{\beta_2} \\ \vdots & \vdots & \vdots \\ \frac{\partial\mu_8}{\beta_0} & \frac{\partial\mu_8}{\beta_1} & \frac{\partial\mu_8}{\beta_2} \end{bmatrix}.$$

Raíz quadrada da matriz da função de variância:

$$\mathbf{V}^{1/2}=\begin{bmatrix} \sqrt{\mu_1(1-\mu_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\mu_8(1-\mu_8)} \end{bmatrix}.$$

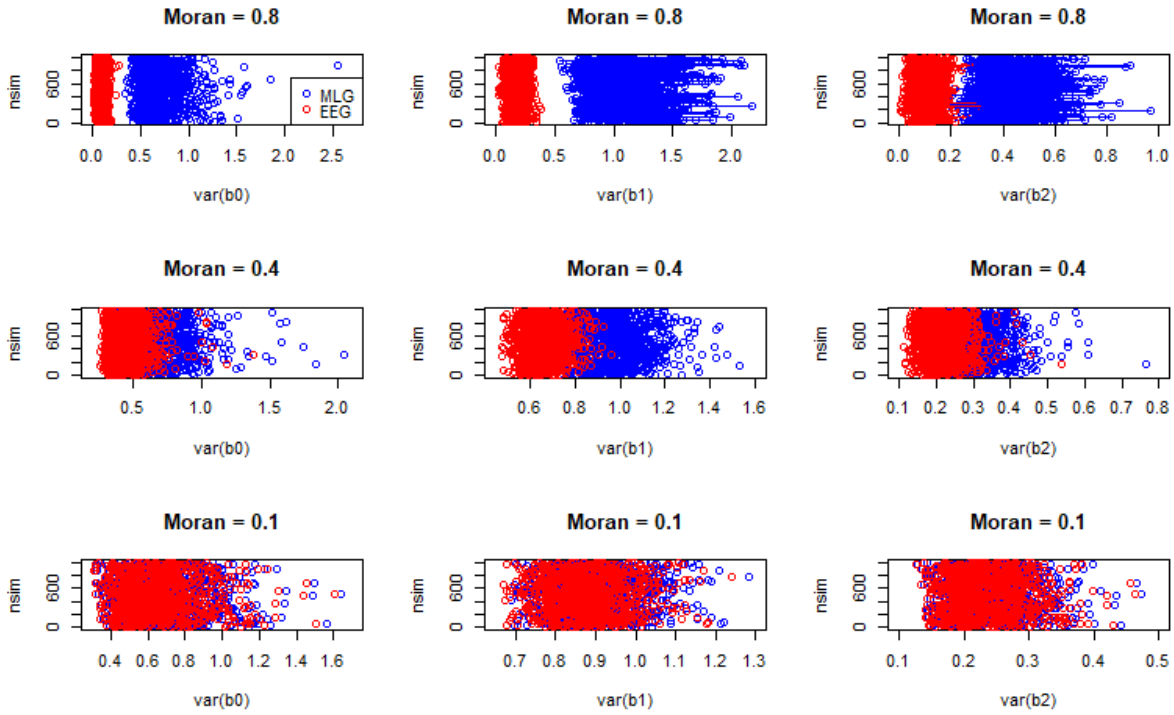
Matriz de covariância espacial de trabalho (**Ω**):

$$\Omega = \begin{bmatrix} \sqrt{\mu_1(1-\mu_1)} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sqrt{\mu_8(1-\mu_8)} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \rho & 0 & \rho & \rho & \rho \\ 0 & 1 & \rho & 0 & 0 & 0 & \rho & 0 \\ 0 & \rho & 1 & 0 & \rho & 0 & \rho & 0 \\ \rho & 0 & 0 & 1 & \rho & \rho & \rho & 0 \\ 0 & 0 & \rho & \rho & 1 & 0 & \rho & 0 \\ \rho & 0 & 0 & \rho & 0 & 1 & 0 & \rho \\ \rho & \rho & \rho & \rho & \rho & 0 & 1 & \rho \\ \rho & 0 & 0 & 0 & 0 & \rho & \rho & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\mu_1(1-\mu_1)} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sqrt{\mu_8(1-\mu_8)} \end{bmatrix}$$

$$\Omega = \begin{bmatrix} V_1 & 0 & 0 & \rho\sqrt{V_1V_4} & 0 & \rho\sqrt{V_1V_6} & \rho\sqrt{V_1V_7} & \rho\sqrt{V_1V_8} \\ 0 & V_2 & \rho\sqrt{V_2V_3} & 0 & 0 & 0 & \rho\sqrt{V_2V_7} & 0 \\ 0 & \rho\sqrt{V_2V_3} & V_3 & 0 & \rho\sqrt{V_3V_5} & 0 & \rho\sqrt{V_3V_7} & 0 \\ \rho\sqrt{V_1V_4} & 0 & 0 & V_4 & \rho\sqrt{V_4V_5} & \rho\sqrt{V_4V_6} & \rho\sqrt{V_4V_7} & 0 \\ 0 & 0 & \rho\sqrt{V_3V_5} & \rho\sqrt{V_4V_5} & V_5 & 0 & \rho\sqrt{V_5V_7} & 0 \\ \rho\sqrt{V_1V_6} & 0 & 0 & \rho\sqrt{V_4V_6} & 0 & V_6 & 0 & \rho\sqrt{V_6V_8} \\ \rho\sqrt{V_1V_7} & \rho\sqrt{V_2V_7} & \rho\sqrt{V_3V_7} & \rho\sqrt{V_4V_7} & \rho\sqrt{V_5V_7} & 0 & V_7 & \rho\sqrt{V_7V_8} \\ \rho\sqrt{V_1V_8} & 0 & 0 & 0 & 0 & \rho\sqrt{V_6V_8} & \rho\sqrt{V_7V_8} & V_8 \end{bmatrix}$$

em que $V_i = \mu_i(1 - \mu_i)$, com i = 1, 2, ..., 8.

APÊNDICE D - Estimativas das variâncias dos estimadores em 1000 simulações para diferentes valores do índice de Moran



APÊNDICE C – Rotina para análise de dados reais da adoção de sementes melhoradas de milho em Moçambique

```
##### Carregando os pacotes
library(geepack)
library(spdep)
library(maptools)
library(Matrix)
library(SpatialEpi)
library(maps)
library(classInt)

####       Importação do shapefile dos distritos de Moçambique    ####################

moz < - readShapeSpatial("MOZ_adm2.SHP", ID="ID_2")
coords<-coordinates(moz)

## criando o data.frame do shapefile
moz.data < - data.frame(moz)
attach(moz.data)

####   Importação do banco de dados
dados=read.table("dadosfinal.txt",h=T)
head(dados)
attach(dados)
## estatísticas descritivas das variáveis do estudo

Summary(dados)

### Criação da matriz de vizinhança

w.geral = poly2nb(moz,queen = T)

### matriz de vizinhança normalizada nas linhas
w.geral.stand<- nb2mat(w.geral, style="W"); w.geral.stand

### Acessando a variável resposta (proporção de produtores que usou semente)
resp = 100*(improve/af)

### Estimando o índice de Moran

moran(resp, w1.stand, 128, 128, zero.policy=NULL, NAOK=FALSE)

#testando o índice de moran

### teste de permutação
moran.test (resp, w.geral.stand, randomisation=TRUE, zero.policy=NULL,
                  alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL,
                  adjust.n=TRUE)
```

```
#### teste Monte Carlo
moran.mc (resp, w.geral.stand, 999, zero.policy=NULL, alternative="greater",
        na.action=na.fail, spChk=NULL, return_boot=FALSE, adjust.n=TRUE)
```

testando o índice de Geary

```
geary (resp, w.geral.stand, 128, 127, 128, zero.policy=NULL)

## teste de permutação
geary.test (resp, w1.stand, randomisation=TRUE, zero.policy=NULL,
         alternative="greater", spChk=NULL, adjust.n=TRUE)


### teste Monte Carlo
geary.mc(resp, w.geral.stand, 999, zero.policy=NULL, alternative="greater",
        spChk=NULL, adjust.n=TRUE, return_boot=FALSE)
```

#estatistica G

```
globalG.test(resp, B1.stand, zero.policy=NULL,
         alternative="greater", spChk=NULL, adjust.n=TRUE,B1correct = TRUE)
```

Mantel statistic

```
sp.mantel.mc (resp, w.geral.stand, 999, type = "sokal", zero.policy = NULL,
          alternative = "greater", spChk=NULL, return_boot=FALSE)

#correlograma
sp.correlogram (w.geral, resp, order = 5, method = "I",
            style = "W", randomisation = TRUE, zero.policy = NULL, spChk=NULL)
```

Índice de Moran local (LISA)

```
lisa <- localmoran(resp,nb2listw(w.geral,style = "W"), zero.policy = T,alternative = "greater")

### centralizar a variável

cDV <- ( resp – mean (resp) )
mI <-  lisa[, 1]
pval = pnorm (q = abs(lisa[,4]), lower.tail = F)

quadrant  <-  vector (mode = "numeric" , length = nrow (lisa) )
quadrant [ cDV > 0 & mI > 0 ]  <-  1
quadrant [ cDV < 0 & mI > 0 ]  <-  2
quadrant[ cDV > 0 & mI < 0 ] <-  3
quadrant[ cDV < 0 & mI < 0] <- 4
```

```
#### set a statistical significance level for the local Moran's
signif <- 0.05

### places non-significant Moran's in the category "5"
Quadrant [pval >  signif ] <-  5

colors <- c("red", "blue", "lightpink", "skyblue2","white")

par (mfrow = c (1,2) )

#### LISA significance map
plot (moz, border = gray(0.3), col=colors[quadrant],
    axes = F,main = "(a)")

legend (36,-19, legend = c ("high-high", "low-low", "high-low", "low-high", "not signif."),
      fill = colors, bty="n", cex=0.8, y.intersp = 1, x.intersp = 1)

compassRose(32,-12,cex = 0.6)
```

estatística G local

```
glocal = localG (resp,nb2listw(w.geral,style = "W",zero.policy = T))

p = pnorm(q = abs(glocal),lower.tail = F)

pvalue <- vector(mode="numeric",length=length(glocal))

pvalue[cDV > 0 & p < 0.05] <- 1   #hotspot
pvalue[cDV < 0 & p < 0.05 ] <- 2  #lowspot


### places non-significant G in the category "3"
pvalue [ p  > 0.05 ] <- 3

colorss <-  c ("red", "blue", "white")


#### G local map (hot and cold spots)

plot(moz, border="black", col=colorss[pvalue],
    axes=F,main = "(b)")

legend(36,-19,legend=c("High","Low","not signif."),
      fill=colorss,bty="n",cex=0.8,y.intersp=1,x.intersp=1)
```

Suavização espacial (media móvel espacial)

```
resp.smoth=w.geral.stand%*%resp  ## estima a média móvel espacial.
```

mapeamento

spatial smoothing

intervalos de classes – usa o método de fisher para definir as classes
intervalo = classIntervals (round(resp.smoth,2), n=4, style = "fisher")

corte = cut (resp.smoth, intervalo$brks, include.lowest = TRUE)

niveis=levels(corte)

(niveis=levels(corte))

cores = palette (c (" #FFFF00FF" , " #FFAA00FF" ," #FF5500FF" , "#FF0000FF "))

levels(corte)=cores

plot(moz, border = gray(0.1),lwd=.1,axes=F,las=1,col=as.character(corte),main="(b)")

legend(36,-20,niveis,fill = cores,bty = "n",title = "% de adoção",cex = 0.7)

incluir a rosa dos ventos
compassRose(32,-11.5,cex = 0.7)

Mapeamento da variável tamanho médio da família
hhsize

intervalos de classes
intervalo = classIntervals(round(hhsize,2),n=4,style = "fisher")
corte=cut(hhsize,intervalo$brks,include.lowest=TRUE)
niveis=levels(corte)
(niveis=levels(corte))
cores = palette(c("#FFFF00FF","#FFAA00FF","#FF5500FF","#FF0000FF"))
levels(corte)=cores

par(mfrow=c(1,1))

plot(moz,border=gray(0.1),lwd=.1,axes=F,las=1,col=as.character(corte))

#text(coords,labels=as.character(pct),cex=.8)
legend(36,-18,niveis,fill = cores,bty = "n",title = "Tamanho médio da família",cex = 1)

summary(hhsize)
sqrt(var(hhsize))

text(45,-11,labels = "Estatísticas descritivas",cex=0.9)
text(45,-12,labels = "Min = 3.00",cex=.8)

```
text(45,-13,labels = "Max = 8.00",cex=.8)
text(45,-14,labels = "Média = 5.29",cex=.8)
text(45,-15,labels = "Mediana = 5.00",cex=.8)
text(45,-16,labels = "Desv. Padrão = 0.92",cex=.8)

### incluir a rosa dos ventos
compassRose(32,-12,cex = 0.6)
map.scale(x=37,y = -26,relwidth = 0.2)
```

Modelagem com uso das EEG

```
### define a matriz de vizinhança binária
wbin = nb2mat(w.geral,style = "B")

### covariáveis do modelo usadas como variáveis dummy

extension.1=as.numeric(extension==0)
association.1=as.numeric(association>0)
credit.1 = as.numeric(credit==0)
tracao.1=as.numeric(tracao>0)
fullworker.1=as.numeric(fullworker>0)
partworkers.1 = as.numeric(partworkers==0)
radio.1=as.numeric(radio/af>0.5)
bike.1=as.numeric(bike/af>0)
cellphone.1=as.numeric(cellphone>0)
celeiro.1=as.numeric(celeiro>0)
hhsex.1=as.numeric(hhsex/af>0.5)

### Ajustando o glm

mod.glm = glm(cbind(improv, af-improv) ~ hhsize + extension.1 +
                association.1 + credit.1 + tracao.1 + partworkers.1+
                fullworker.1+radio.1+bike.1+celeiro.1+cellphone.1+
                hhage+education+hhsex.1,family = "binomial")

summary(mod.glm)

#### indice de moran dos residuos
moran.test (mod.glm$residuals, w.geral.stand, randomisation=TRUE, zero.policy=NULL,
                alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL,
                adjust.n=TRUE)

#### step wise regression
 reg = step (mod.glm, direction = "both")


summary(reg)
```

```
## moran of residuals

moran.test (reg$residuals, w.geral.stand, randomisation=TRUE, zero.policy=NULL,
            alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL,
            adjust.n=TRUE)

### matriz de delineamento

x  = cbind(rep(1,128),hhsize,extension.1,credit.1,
           tracao.1,partworkers.1,fullworker.1,radio.1,celeiro.1,
           hhage)

####### GEE  ###################

####   funcao para estrutura correlacionada AR(1) #####

### Moran da resposta

im=moran.test(improv/af, w.geral.stand, randomisation=TRUE, zero.policy=NULL,
            alternative="greater", rank = FALSE, na.action=na.fail, spChk=NULL,
adjust.n=TRUE)

coef=im$estimate[1]

r.ar=coef*wbin

diag(r.ar) = 1

zco = r.ar[lower.tri(r.ar)]

mod.ar = geeglm(cbind(improv, af-improv) ~ hhsize + extension.1 +
                  credit.1 + tracao.1 + partworkers.1+
                  fullworker.1+radio.1+celeiro.1+hhage,id=rep(1,128),family = binomial(link =
"logit"),scale.fix = TRUE,scale.value = c(rep(1,128)),zcor = zco,corstr = "fixed")

beta1=as.matrix(mod.ar$coefficients)

############

#######     Estimador Sanduiche    #####

mi.ar = exp(x%*%beta1)/(1+exp(x%*%beta1))

peso.meio.ar = Diagonal(128)

peso.ar = Diagonal(128)

diag(peso.ar) = mi.ar*(1-mi.ar)

diag(peso.meio.ar) = sqrt(mi.ar*(1-mi.ar))
```

```
mi.glm = exp(x%*%beta)/(1+exp(x%*%beta))

peso.glm = Diagonal(128)

diag(peso.glm) = mi.glm*(1-mi.glm)

peso.meio.glm = Diagonal(128)

diag(peso.meio.glm) = sqrt(mi.glm*(1-mi.glm))

h1.ar = t(x)%*%(peso.meio.ar)%*%solve(r.ar)%*%(peso.meio.ar)%*%x

h1.glm = t(x)%*%peso.glm%*%x

h2.glm = t(x)%*%peso.meio.glm%*%r.ar%*%peso.meio.glm%*%x

vglm = solve(h1.glm) %*%h2.glm%*%solve(h1.glm)

Vot = mod.ar$geese$vbeta.naiv

Diag.ar = diag(Vot)

sder=sqrt(Diag.ar)

z=beta1/sder

pvalue = pnorm(q = abs(z),lower.tail = F)

table=cbind(beta1,sder,z,round(pvalue,3))
table

res=(improv/af-mi.ar)/sqrt(mi.ar*(1-mi.ar))

as.data.frame(res)

rho = lagsarlm(res~1,listw = nb2listw(w.geral,style = "W"))
summary(rho)

## ##### determinação do QIC ###

# ######### termo referente ao traço  ##########

AIinverse <- solve(vglm)

V.msR <- solve(h1.ar) #Vot  # modelo sob investigacao

trace.term <- sum(diag(AIinverse%*%V.msR))
```

```
 # ######## estima a média e os valores observados #########

mu.R <- mod.ar$fitted.values

y <- improv/af

# ######## scale for binary data ########
scale <- 1

quasi.R <- sum(y*log(mu.R/(1-mu.R))+log(1-mu.R))/scale

QIC <- (-2)*quasi.R + 2*trace.term
QIC

CIC = trace.term
CIC

Rj = sum(diag(solve(h1.ar)%*%h1.ar))/10
Rj

##### determinação do QIC for glm ###

 # ######### termo referente ao traço  ##########

AIinverse <- solve(vglm)

V.msR <- vglm #Vot  # modelo sob investigacao

trace.term <- sum(diag(AIinverse%*%V.msR))

# ######## estima a média e os valores observados #########

mu.R <- reg$fitted.values

y <- improv/af

# ######## scale for binary data ########

scale <- 1

quasi.R <- sum(y*log(mu.R/(1-mu.R))+log(1-mu.R))/scale

QIC <- (-2)*quasi.R + 2*trace.term
QIC

CIC = trace.term
CIC

Rj = sum(diag(h1.glm%*%vglm))/10
Rj
```

APÊNDICE D – Rotina para gerar amostras com dependência espacial com base no índice de Moran

```
setwd("C:/Users/Samsung/Dropbox/TIA2012_march")

#### carregando pacotes
library(maptools)
gpclibPermit()
library(maptools)
library(spdep)
library(Matrix)
library(SpatialEpi)

##### Importação do shapefile dos distritos de MOz ####################

moz <- readShapeSpatial("MOZ_adm2.SHP", ID="ID_2")
coords<-coordinates(moz)

## criando o data.frame do shapefile
moz.data<-data.frame(moz)
attach(moz.data)

### matriz de vizinhança
w.geral=poly2nb(moz,queen = T)

### W normalizada
w.geral.stand<- nb2listw(w.geral, style="W"); w.geral.stand


#### considere as seguintes normais para obter valores da resposta no intervalo (0,1)

# moran(-0.1,-0.4,-0.48) rnorm(10,0.001)+0.5
# moran (-0.2,-0.3) rnorm(20,0.1)+0.5
# moran (0.1 a 0.3) rnorm(33,0.05)
# moran (0.4 a 0.6) rnorm(33,0.1)
# moran (0.7 a 0.8) rnorm(5,0.05)
# moran (0.9) rnorm(5,0.1)


mini = matrix(0, 1, 1)
maxi = matrix(0, 1, 1)
rho=matrix(0,1,1)
gear=matrix(0,1,1)

 f=function(nsim,coef,md,sd)
 {
   for(i in 1:nsim)
   {
  y=rnorm(128,md,sd)+0.5 #0.05 para moran = 0.9
  a=nb2mat(w.geral,style = "W")
```

```
 diag(a)=-coef
 y0=solve(a)%*%y

 im= moran(y0, w.geral.stand, 128, 128, zero.policy=NULL, NAOK=FALSE)

 c=geary.test(as.vector(y0), w.geral.stand, randomisation=TRUE, zero.policy=NULL,
          alternative="greater", spChk=NULL, adjust.n=TRUE)


yes2=max(y0)

 yes = min(y0)

mini=rbind(mini,as.matrix(yes))

maxi=rbind(maxi,as.matrix(yes2))

rho=rbind(rho,as.matrix(im$I[1]))

gear=rbind(gear,as.matrix(c$estimate[1]))

}


   return(list(min=mini,max=maxi,indice=rho,igear=gear))
}

nsim=1000
coef=0.7
md = 5
sd = 0.05
fu=f(nsim,coef,md,sd)
```