UNIVERSIDADE FEDERAL DE ALFENAS

WALEF MACHADO DE MENDONÇA

CONCENTRAÇÕES REGIONAIS DO SEGURO RURAL NO BRASIL EM 2019

Varginha/MG

2022

WALEF MACHADO DE MENDONÇA

CONCENTRAÇÕES REGIONAIS DO SEGURO RURAL NO BRASIL EM 2019

Trabalho de conclusão de curso apresentado ao Instituto de Ciências Sociais Aplicadas da Universidade Federal de Alfenas como requisito parcial para a obtenção do título de Bacharel em Ciências Econômicas.

Orientadora: Prof[a]. Dr[a]. Patrícia de Siqueira Ramos.

Varginha/MG
2022

Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Campus Varginha

Mendonça, Walef Machado de.
Concentrações regionais do seguro rural no Brasil em 2019 / Walef Machado de Mendonça. - Varginha, MG, 2022.
45 f. : il. -

Orientador(a): Patrícia de Siqueira Ramos.
Trabalho de Conclusão de Curso (Graduação em Ciências Econômicas) - Universidade Federal de Alfenas, Varginha, MG, 2022.
Bibliografia.

1. Seguro rural. 2. Municípios. 3. Regiões. 4. Agrupamentos. 5. Estatística espacial. I. Ramos, Patrícia de Siqueira, orient. II. Título.

Ficha gerada automaticamente com dados fornecidos pelo autor.

RESUMO

O ambiente no qual se desenvolvem as atividades agropecuárias apresenta elevado risco e grande incerteza. Variáveis relacionadas aos mercados agropecuários podem gerar oscilações na renda do setor que devem ser enfrentadas por meio de políticas de apoio à gestão de riscos. Uma das formas mais usuais de gerenciamento de risco é a contratação de seguro rural, o qual possibilita a recuperação da capacidade financeira do produtor na ocorrência de sinistros. Destaca-se, no entanto, que o mercado de seguro rural não se consolida sem a participação do Estado devido a problemas como os elevados investimentos e custos administrativos, a possibilidade de riscos catastróficos, e a forte influência do risco moral e da seleção adversa na formação das carteiras. Como mecanismo de estímulo para o desenvolvimento do seguro rural, o governo brasileiro criou o Programa de Subvenção ao Prêmio do Seguro Rural, que divide os custos de aquisição da apólice entre o governo e os produtores. Nesse sentido, o objetivo deste trabalho é agrupar espacialmente os municípios do Brasil segundo variáveis relativas ao seguro rural com o intuito de subsidiar a tomada de decisões em políticas públicas de estímulo à demanda por produtos de seguro específicos para cada grupo de municípios. Para tanto, é apresentado um procedimento para descobrir e explorar padrões de agrupamento espacial com base na distribuição espacial de dados multivariados. Esse procedimento faz uso do algoritmo de agrupamento não hierárquico das k-médias e incorpora a estrutura espacial dos dados através do uso de medidas locais de autocorrelação espacial. A aplicação do procedimento aponta para a existência de maiores concentrações de apólices de seguro rural nas regiões Sul, Centro-Oeste e Sudeste, no sul do Estado de São Paulo.

Palavras-chave: Política agrícola. Zoneamento agrícola. Riscos agropecuários. Autocorrelação espacial.

ABSTRACT

The environment in which agricultural activities are carried out presents high risk and great uncertainty. Variables related to agricultural markets can generate fluctuations in the sector's income that must be addressed through policies to support risk management. One of the most common forms of risk management is the contracting of rural insurance, which makes it possible to recover the producer's financial capacity in the event of accidents. It is noteworthy, however, that the rural insurance market is not consolidated without the participation of the State due to problems such as high investments and administrative costs, the possibility of catastrophic risks, and the strong influence of moral hazard and adverse selection on formation of portfolios. As a mechanism to encourage the development of rural insurance, the Brazilian government created the Subsidy Program for the Rural Insurance Premium, which divides the costs of purchasing the policy between the government and the producers. In this sense, the objective of this work is to spatially group Brazilian municipalities according to variables related to rural insurance in order to support decision-making in public policies to stimulate demand for specific insurance products for each group of municipalities. Therefore, a procedure to discover and explore patterns of spatial clustering based on the spatial distribution of multivariate data is presented. This procedure uses the non-hierarchical k-means clustering algorithm and incorporates the spatial structure of the data through the use of local measures of spatial autocorrelation. The application of the procedure points to the existence of greater concentrations of rural insurance policies in the South, Midwest and Southeast regions, in the south of the State of São Paulo.

Keywords: Agricultural Policy. Agricultural Zoning. Agricultural risks. Spatial autocorrelation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A contratação de seguro rural é uma das formas mais frequentes de gerenciamento de riscos agropecuários. Essa forma de seguro opera com o objetivo de atenuar as perdas e viabilizar a recuperação da capacidade financeira do produtor na ocorrência de sinistros. O seguro rural também promove um ambiente mais propício ao desenvolvimento das atividades agropecuárias, pois proporciona a garantia do fluxo de renda, contribui para um aumento da área plantada e facilita a obtenção de financiamento. Além disso, ele se apresenta como um instrumento que promove o compartilhamento do risco da agropecuária com outros agentes e setores econômicos da sociedade (BRASIL, 2019a).

É importante destacar que o mercado de seguro rural não se consolida sem a participação do Estado. O ambiente de elevado risco e considerável incerteza no qual se desenvolvem as atividades agropecuárias impõe a existência de fatores que limitam a eficiência da iniciativa privada na oferta de produtos de seguro rural. Destacam-se problemas como os elevados investimentos e custos administrativos, a possibilidade de riscos catastróficos, a forte influência do risco moral e da seleção adversa na formação das carteiras, como fatores que limitam a eficiência da iniciativa privada na oferta do seguro rural.

Os principais riscos agropecuários se devem, principalmente, às instabilidades climáticas e ameaças sanitárias, que podem afetar a produção, ou a razões de mercado, como variações das taxas de câmbio e juros, ou a condições ligadas ao ambiente de negócios, tais como alterações em marcos regulatórios e em políticas públicas. Todas essas variáveis, relacionadas aos mercados agropecuários, geram variações na renda do setor, que devem ser enfrentadas através de políticas de gerenciamento de gestão de riscos (BRASIL, 2018). Nesse sentido, o poder público é demandado a interferir no mercado, seja atuando diretamente como seguradora, seja criando programas que estimulem a oferta e a demanda por produtos de seguro.

Uma das formas utilizadas como mecanismo de estímulo ao desenvolvimento do seguro rural no Brasil, é o Programa de Subvenção ao Prêmio do Seguro Rural (PSR), instituído pela Lei 10.823/2003 e decreto nº 5.121/2004 (BRASIL, 2018). Esta política do governo brasileiro tem como objetivo tornar o seguro rural mais barato para os produtores rurais, dividindo os custos de aquisição da apólice entre o governo e os produtores.

Neste contexto, este trabalho busca identificar grupos de municípios com características semelhantes em relação à adesão ao seguro rural através do agrupamento de dados multivariados com base em medidas locais de autocorrelação espacial. Isso faz com que os agrupamentos obtidos levem em consideração não apenas o valor das variáveis de seguro rural, mas também seu posicionamento geográfico. Dessa forma, busca-se auxiliar na tomada de decisões sobre políticas públicas de estímulo à demanda por produtos de

seguro específicas para cada grupo de municípios.

Para tanto, é apresentado um procedimento para descobrir e explorar padrões de agrupamento espacial com base na distribuição espacial de dados multivariados. Esse procedimento faz uso do algoritmo de agrupamento não hierárquico das k-médias e incorpora a estrutura espacial dos dados através do uso de medidas locais de autocorrelação espacial como o I de Moran local e G de Getis e Ord local.

2 REFERENCIAL TEÓRICO

2.1 SEGURO RURAL

De acordo com o Guia de Seguros Rurais[1], a ocorrência de sinistros devido a eventos climáticos adversos tem provocado prejuízos consideráveis para agricultores, mesmo tendo em conta o elevado nível tecnológico empregado nas atividades agropecuárias (GUIA DE SEGUROS RURAIS, 2020).

A ocorrência desses sinistros se deve ao fato de que as atividades do setor agropecuário são dotadas de especificidades em relação à dimensão dos riscos aos quais estão expostas. Os principais riscos estão associados aos aspectos biológicos da produção e sua interdependência com os fatores climáticos (BURGO, 2005). Também se destacam os riscos associados à volatilidade dos preços, assim como variações das taxas de juros e de câmbio. Dessa forma, os riscos e incertezas inerentes à produção agropecuária podem causar perdas econômicas que têm potencial de afetar não apenas os produtores rurais como toda a sociedade (BRASIL, 2019b).

Este contexto torna evidente a necessidade de medidas de gestão de risco como a contratação do seguro rural. De um modo geral, o seguro caracteriza-se pela transferência das consequências econômicas, da realização de um determinado risco, do segurado para a seguradora. O mesmo ocorre com o seguro rural, cujo papel é reduzir riscos e proteger a renda dos produtores rurais (GUIA DE SEGUROS RURAIS, 2020; BRASIL, 2021a).

Um sistema de seguridade rural eficiente tem como principal vantagem garantir ao produtor a segurança necessária para continuar a investir na produção e se manter competitivo no setor agropecuário mesmo em situações de ocorrência de adversidades climáticas, que ocasionam perda patrimonial ou da safra. Dessa forma, o seguro rural se apresenta como um importante recurso para estabelecer uma proteção à renda dos produtores (GUIA DE SEGUROS RURAIS, 2020).

Além do mais, do ponto de vista dos efeitos agregados, o seguro rural desempenha um papel importante, pois proporciona um bom ambiente de desenvolvimento para a agricultura. O seguro rural auxilia na oferta de financiamento, promove o crescimento da

[1] Elaborado pela Comissão Nacional de Política Agrícola da Confederação da Agricultura e Pecuária do Brasil (CNA) e pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA).

área plantada e é uma ferramenta que pode compartilhar os riscos agrícolas com os demais agentes e setores da economia (BRASIL, 2021a).

Contudo, é importante destacar que o desenvolvimento de um sistema de seguro rural é em si um desafio devido à natureza dos riscos da atividade agropecuária. Por exemplo, há uma dependência espacial na ocorrência de eventos climáticos adversos, o que faz com que o seguro rural contrarie a suposição de que, em um seguro, o risco agregado deve ser menor que o risco individual. Fatores como este fazem com que as seguradoras tenham dificuldade em constituir carteiras que viabilizem a diversificação dos riscos (BARROS *et al*, 2012; FORNAZIER; SOUZA; PONCIANO, 2012)

Associados aos altos custos operacionais e a dificuldades em uma precificação que leve em conta a dependência e heterogeneidade dos sinistros, é importante destacar que elementos como a assimetria de informação e o risco moral, devido à ausência de dados históricos dos produtores, prejudicam a atividade das empresas de seguro rural. Isso acontece porque, diante deste cenário, as seguradoras que atuam no seguro rural acabam cobrando valores altos para os prêmios, o que causa um desincentivo à aderência ao seguro. Dessa forma, ressalta-se que as subvenções concedidas pelo governo desempenham um atribuição essencial, pois, ao tornar viável a cobertura para o produtor rural, buscam corrigir as falhas de mercado e proporcionar a expansão da área segurada (GUIMARÃES; NOGUEIRA, 2009; BARROS *et al*, 2012).

2.1.1 O Seguro Rural no Brasil

As primeiras tentativas de implantar o seguro rural no Brasil se passaram na década de 1930 no estado de São Paulo. A criação, em 1939, de um seguro obrigatório que oferecia proteção contra o granizo na produção de algodão foi uma das iniciativas pioneiras no estado de São Paulo (MAIA; ROITMAN, DE CONTI, 2011). A partir de então, os resultados positivos alcançados pelo seguro para a proteção da lavoura do algodão inspiraram a criação de novos programas como a Carteira de Seguro Agrícola contra Granizo para a Viticultura, instituída em 1948, e a Carteira de Seguro Agrícola contra Geada para Horticultura criada em 1964 (SILVA; TEIXEIRA; SANTOS, 2014).

No final da década de 1940, foi criado no Instituto Rio-Grandense do Arroz (Irga) o seguro para granizo e o seguro para indenizar os produtores de fumo nos estados de Santa Catarina e Rio Grande do Sul, criado pela Associação dos Fumicultores do Brasil (Afubra) e mantido com recursos próprios (SILVA; TEIXEIRA; SANTOS, 2014).

Além disso, em 1948, foi criado, no âmbito federal, o Instituto de Resseguros do Brasil (IRB), que tinha como objetivo reduzir os prejuízos causados por eventos adversos e assegurar maior estabilidade aos produtores rurais (SILVA; TEIXEIRA; SANTOS, 2014). Também foi fundada em 1954, pelo Governo Federal, a Companhia Nacional de Seguro Agrícola (CNSA) e o Fundo de Estabilidade do Seguro Agrário. Segundo, Maia,

Roitman e de Conti (2011), a estruturação e gestão dos seguros da CNSA ficaram sob responsabilidade do IRB. Contudo, as atividades da CNSA foram finalizadas em 1996, em decorrência, segundo Gemignani (2000), da incapacidade em difundir o seguro rural de forma a tornar possível sua viabilização econômica (MAIA; ROITMAN, DE CONTI, 2011; SILVA; TEIXEIRA; SANTOS, 2014).

Segundo Silva, Teixeira e Santos (2014), o Decreto-Lei nº 73 de 1966 e o Decreto nº 60.459 de 1967 estabelecem as bases legais para as atividades de seguro e a criação do Sistema Nacional de Seguros Privados (SNSP). Com o decreto de 1967 também foi instituído o Fundo de Estabilidade do Seguro Rural (FESR), cujos recursos eram geridos pelo IRB e a principal função era garantir a estabilidade das operações de seguro e proporcionar uma proteção complementar para os riscos de sinistro (SILVA; TEIXEIRA; SANTOS, 2014)

Com um papel importante na definição das modalidades de seguros agrários, a Resolução nº 5 do Conselho Nacional de Seguros Privados foi instituída em 1970. Essa resolução define que o seguro agrícola deve fornecer cobertura contra perdas decorrentes de fenômenos meteorológicos, doenças e pragas. Além disso, no setor pecuário, o seguro deve fornecer cobertura para mortes de animais causadas por doenças ou acidentes, assim como o seguro de benfeitorias e produtos agropecuários. Por fim, a Resolução nº 5 também estabelece o seguro de crédito, que cobre incapacidade de pagamento de compradores dos produtos agropecuários (SILVA; TEIXEIRA; SANTOS, 2014).

Apesar das várias ações implementadas pelo Governo Federal, o seguro rural evoluiu de forma lenta e limitada a uma reduzida parcela da produção. Os problemas do desenvolvimento do seguro rural motivaram a criação do Programa de Garantia da Atividade Agropecuária (Proagro), através da Lei nº 5.969, de 11 de dezembro de 1973 (SILVA; TEIXEIRA; SANTOS, 2014). Inicialmente, o programa ficou sob a responsabilidade do Banco Central, que passou a vincular o seguro rural às operações de crédito agropecuário e utilizou emissões monetárias para pagamentos de sinistros. Segundo Maia, Roitman e De Conti (2011), o sistema de financiamento do Proagro gerou déficits que motivaram diversas modificações no Programa, que ainda continua sendo um dos mais importantes instrumentos para a gestão de riscos na agricultura no Brasil (MAIA; ROITMAN, DE CONTI, 2011).

A partir de 2003, o Governo Federal começou a adotar uma política de subvenção através do Programa de Subvenção ao Prêmio do Seguro Rural (PSR), instituído por meio da Lei nº 10.823 de 19 de dezembro de 2003. Este programa objetiva, assim como ocorre com seguros rurais em países europeus e nos Estados Unidos, conceder subvenção econômica ao valor do prêmio do seguro rural contratado com seguradoras autorizadas (MAIA; ROITMAN, DE CONTI, 2011; SILVA; TEIXEIRA; SANTOS, 2014).

2.1.2 Programas de Subvenção ao Prêmio do Seguro Rural

O PSR foi instituído pelo Decreto nº 5.121 de 2004 e tem a finalidade de subsidiar parte do prêmio do seguro rural, de forma a garantir o papel do seguro rural como mecanismo de estabilidade da renda do produtor, além de propor a aplicação das tecnologias adequadas para os empreendimentos agropecuários (GUIA DE SEGUROS RURAIS, 2020)

A operacionalização do PSR é feita pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Nesse programa, os produtores rurais que devem fazer a contratação do seguro rural diretamente com as seguradoras habilitadas pela Superintendência de Seguros Privados (Susep) e cadastradas no MAPA. Assim, para obter os recursos do programa, os produtores devem contratar a apólice e requerer a subvenção do governo federal. Por sua vez, o governo federal repassa para as seguradoras habilitadas um percentual do prêmio do seguro contratado, resultando na redução do valor a ser pago pelo produtor rural (SILVA; TEIXEIRA; SANTOS, 2014).

As técnicas de execução e prioridades da política do PSR são estabelecidas pelo Comitê Gestor Interministerial do Seguro Rural (CGSR). Com o objetivo de fiscalizar a gestão do programa, o CGSR é formado pelo MAPA, que é responsável pela coordenação, e por representantes do Ministério da Economia (ME) e da Superintendência de Seguros Privados (Susep). Dessa forma, o valor de subvenção ao prêmio do seguro rural leva em conta algumas características como a modalidade do seguro, o tipo de cultura e o tipo de cobertura (BRASIL, 2021a).

É necessário destacar que, conforme apresentado no Guia de Seguros Rurais (2020), o PSR tem como condição o cumprimento dos indicadores do Zoneamento Agrícola de Risco Climático (ZARC) para o recebimento da subvenção. O ZARC é um levantamento do período de plantio das culturas por município e considera características como o tipo de solo e o clima da região. Dessa forma, o ZARC busca evitar que eventos climáticos adversos ocorram durante as fases em que as culturas se encontram mais expostas e, com isso, reduzir as perdas (GUIA DE SEGUROS RURAIS, 2020).

Diante da relevância do setor agrícola para a economia brasileira, é possível entender que o seguro rural e o PSR são importantes formas de gestão de risco e que colaboram para uma agricultura mais eficiente e com riscos reduzidos para o produtor rural. Isso vale especialmente para o PSR, que opera como uma forma de fomentar a aquisição das apólices de seguro rural, objetivando induzir o uso de tecnologias adequadas e modernizando a gestão do empreendimento agropecuário.

2.2 ANÁLISE ESPACIAL

A análise espacial consiste em um conjunto de técnicas destinadas a incorporar o espaço à análise. Ou seja, a análise espacial, constitui-se de métodos capazes de mensurar

propriedades e relações baseados na localização espacial de um determinado fenômeno. Portanto, antes de iniciar a modelagem, muitas vezes é efetuada uma análise exploratória de dados para localizar padrões de dependência espacial no fenômeno em estudo (CÂMARA et al., 2004).

Os principais fenômenos abordados na análise espacial incluem a dependência espacial e a heterogeneidade espacial. O fenômeno da dependência espacial pode ser representado pela primeira lei da geografia ou pela lei de Tobler. O geógrafo suíço Waldo Tobler apresentou a lei no ano de 1970, afirmando que tudo depende de tudo, mas as coisas que estão mais próximas estão mais relacionadas entre si do que coisas mais distantes. Portanto, a dependência espacial estabelece que o valor de uma variável em uma determinada região depende do valor da mesma variável em uma região próxima (ALMEIDA, 2012).

Dessa forma, segundo Almeida (2012), processos temporais apresentam relações diferentes dos processos espaciais. O que quer dizer que, nos processos temporais, o valor de uma variável qualquer no tempo t é influenciado pelo valor da variável no tempo $t-1$, mas o inverso não ocorre. Em contraste a esse tipo de processo, nos processos espaciais têm-se a multidirecionalidade nas relações entre as regiões, ou seja, o valor da variável na região i depende do valor na região j, e a região j depende da região i em relação ao valor da mesma variável.

Além disso, é necessário destacar, entre os objetivos da análise espacial, a investigação da presença dos *outliers* espaciais. Os *outliers* são definidos como valores extremos com relação a suas posições no espaço geográfico. A presença de *outliers* espaciais ocorre levando-se em consideração a diferença entre os valores em relação ao conjunto de valores das regiões vizinhas (HAINING, 2003). Os *outliers* espaciais podem surgir do processo de obtenção e armazenamento dos dados, no entanto, tais valores podem apontar para a existência de valores extremos e representar atributos do fenômeno analisado. Portanto, a presença de *outliers* espaciais impõe a necessidade de uma investigação mais cuidadosa (ALMEIDA, 2012).

2.2.1 Matriz de ponderação espacial

Em uma matriz de ponderação espacial, ou matriz de pesos espaciais, cada conexão entre duas regiões é representada por uma entrada na matriz que é denominada peso espacial. Essa matriz tem como objetivo representar um determinado arranjo espacial das interações resultantes de um fenômeno em estudo (ALMEIDA, 2012). A matriz de ponderação espacial tem dimensão $n \times n$ e é, usualmente, denotada por $\boldsymbol{W}$:

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix},$$

em que n é o número de regiões em análise. Cada elemento w_{ij} retrata uma medida de proximidade entre as áreas i e j (CÂMARA et al., 2004).

É possível enumerar dois problemas com a atribuição da matriz de ponderação espacial a ser utilizada. Primeiramente, é possível que a escolha da matriz $\boldsymbol{W}$ seja arbitrária, uma vez que não existe teste formal para sua escolha. Além disso, há o problema que se relaciona à sensibilidade dos resultados à escolha da matriz. Dessa forma, a escolha da matriz de ponderação espacial possuí fundamental relevância e, portanto, deve ser feita com uma fundamentação teórica (ALMEIDA, 2012).

Os modelos tradicionais comumente utilizam, em sua maioria, características físicas e geográficos para determinar um critério de distâncias entre as regiões. Os critérios de vizinhança, distância geográfica ou de tempo de deslocamento são os mais utilizados. No entanto, é possível que a matriz de pesos espaciais utilize como critério de distâncias entre as regiões os aspectos socioeconômicos (TYSZLER, 2006).

Uma matriz comumente utilizada é a matriz de ponderação espacial por contiguidade. Nessa matriz, são levadas em consideração as fronteiras físicas em comum entre duas regiões. Se duas regiões apresentam relação de vizinhança, é atribuído ao elemento correspondente na matriz o valor 1, caso contrário é atribuído o valor 0. Formalmente, tem-se:

$$w_{ij} = \begin{cases} 1, & \text{se } i \text{ e } j \text{ são contíguos} \\ 0, & \text{se } i \text{ e } j \text{ não são contíguos.} \end{cases}$$

Destaca-se que convencionou-se assumir que $w_{ij} = 0$, para todo $i = j$, ou seja, uma determinada região não é considerada vizinha de si própria. Porém, embora com menor frequência, é possível encontrar na literatura regiões sendo vizinhas de si mesmas.

É necessário ressaltar que a matriz de contiguidade binária é simétrica, de forma que se duas regiões i e j possuem fronteiras físicas em comum, tanto w_{ij} quanto w_{ji} serão iguais a 1. Assim, a influência exercida pela região i em j será observada na relação entre j e i, o que representa formalmente o conceito de multidirecionalidade da influência espacial. Um fato a se destacar é que, normalmente é realizada a normalização da matriz de ponderação espacial, ou seja, divide-se o valor de cada elemento da matriz $\boldsymbol{W}$ pelo valor

da soma dos valores da linha em que ele se encontra. Uma vez realizada a normalização, a matriz resultante não será necessariamente simétrica.

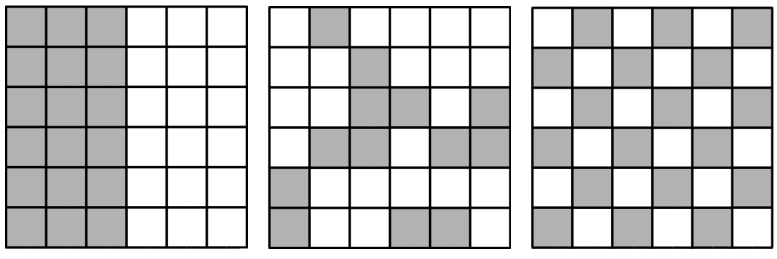
2.2.2 Análise exploratória de dados espaciais

Com o objetivo de descrever como estão distribuídas espacialmente as observações, a análise exploratória de dados espaciais (AEDE) busca compreender os padrões de associação espacial e verificar a existência de *clusters* espaciais. Além disso, essa metodologia objetiva analisar a existência de diferentes regimes espaciais ou outras formas de instabilidade, identificação de observações atípicas, ou como são também denominados, *outliers* espaciais (ALMEIDA, 2012).

A necessidade de se fazer uso de metodologias que considerem a distribuição espacial se dá pois a identificação de padrões de associação espacial pela percepção humana tende a criar padrões viesados. Ou seja, é possível que em uma análise menos criteriosa encontrem-se padrões até mesmo em dados distribuídos de forma aleatória (MESSNER, 1999). Através dessas metodologias, podem-se utilizar indicadores de associação espacial globais ou locais com o objetivo de identificar os diferentes regimes espaciais do conjunto de dados.

A Figura 1 apresenta exemplos de padrões de autocorrelação espacial. Pode-se identificar em (a) um padrão de autocorrelação positivo, ou seja, valores semelhantes estão localizados próximos uns dos outros. Em (b) é apresentado um padrão de autocorrelação nula, o que indica que há aleatoriedade na distribuição espacial. Por sua vez, em (c), há um padrão de autocorrelação negativa, assinalando que valores dissimilares localizam-se próximos uns dos outros.

Figura 1 – Padrões de autocorrelação espacial



(a) Positiva (b) Nula (c) Negativa

Fonte: Elaboração própria

Em contraste à Figura 1, que apresenta dados no formato de grades regulares, geralmente as regiões em estudo são polígonos irregulares. Esse fato prejudica a identificação visual dos padrões de autocorrelação espacial (principalmente autocorrelação nula e

negativa). Portanto, é necessário lançar mão de testes estatísticos para a identificação dos padrões de autocorrelação espacial.

2.2.3 Autocorrelação espacial global

Um dos principais indicadores de autocorrelação espacial global é o coeficiente I de Moran (ALMEIDA, 2012). Este coeficiente é definido através de uma medida de autocovariância na forma de produto cruzado. Algebricamente a estatística é dada por:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{1}$$

em que n retrata o número de regiões, z_i indica o valor da variável de interesse padronizada na região i e S_0 representa o somatório de todos os pesos espaciais da matriz $\boldsymbol{W}$ (ALMEIDA, 2012):

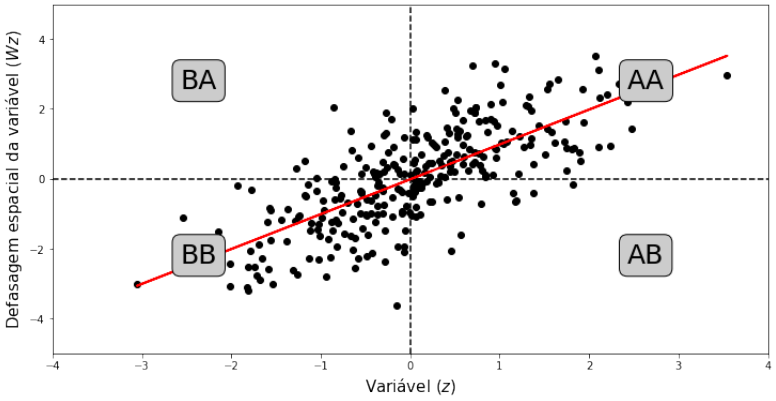
De forma geral, o I de Moran é utilizado para testar a existência de autocorrelação espacial e, nesse contexto, a aleatoriedade espacial é a hipótese nula. Cliff e Ord (CLIFF, 1981) demonstraram que o I de Moran não é centrado em 0, mas sim em seu valor esperado $-[1/(n-1)]$, que é o valor que seria obtido em um cenário de distribuição espacial aleatória. Assim, valores do I de Moran entre o valor esperado e 1 indicam a presença de autocorrelação espacial positiva, valores entre -1 e o valor esperado apontam para a presença de autocorrelação espacial negativa.

A autocorrelação espacial positiva revela que as regiões que apresentam valores altos (acima da média) da variável de interesse tendem a ter ao seu redor regiões que também apresentam valores altos (acima da média) desta variável e regiões com baixos valores (abaixo da média) tendem a estar rodeadas por regiões que também apresentam baixos valores (abaixo da média). Em contraste, a autocorrelação espacial negativa aponta a dissimilaridade entre os valores da variável e a sua localização. Nesse sentido, geralmente regiões com valores altos da variável estão circundadas por regiões com baixos valores e regiões com baixos valores têm em seu entorno regiões com valores altos (ALMEIDA, 2012).

A verificação da significância do valor do índice I de Moran pode ser obtida pela associação do índice a uma distribuição estatística, geralmente a distribuição normal. A forma mais utilizada para tal verificação é o teste de pseudo-significância ou, como também é conhecido, teste de permutação aleatória. Esse teste não possui pressupostos em relação à distribuição e nesta abordagem são geradas diferentes permutações dos valores associados às regiões, produzindo para cada permutação um novo arranjo espacial com os valores redistribuídos entre as áreas. Dado que o cenário real observado corresponde somente a apenas um dos arranjos, tem-se uma distribuição empírica do índice (CÂMARA et al., 2004).

Outra forma de visualizar os regimes espaciais é o diagrama de dispersão de Moran, exemplificado na Figura 2. No eixo horizontal do diagrama de dispersão de Moran, são retratados os valores padronizados da variável de interesse, já o eixo vertical apresenta as médias da variável nos vizinhos das respectivas áreas, também conhecidas como defasagens espaciais.

Figura 2 – Diagrama de dispersão de Moran



Elaboração própria com dados simulados.

2.2.4 Autocorrelação espacial local

Um indicador de concentração com a capacidade de analisar localmente a associação espacial foi inicialmente apresentado por Getis e Ord em 1992 (ALMEIDA, 2012). De acordo com Getis e Ord (1992), essa estatística indica a presença de eventuais agrupamentos localizados de concentração espacial, chamados de *hot spots* e *cool spots*. A estatística denota, para cada observação i, em que medida essa observação está rodeada por valores altos (*hot spot*) ou baixos (*cool spot*). Para uma determinada variável y, a estatística é calculada da seguinte forma:

$$G_i(d) = \frac{\sum_j w_{ij}(d) y_j}{\sum_j y_j}, \quad \text{para } j \neq i. \tag{2}$$

O somatório em j faz com que apenas os valores dos vizinhos próximos da região i sejam utilizados no cálculo da estatística. A matriz de ponderação espacial $\boldsymbol{W}$ pode ou não ser uma matriz de proximidade geográfica, baseada em um raio construído em torno da região i.

A estatística G de Getis e Ord possui duas formulações possíveis. Se não incluir a observação sob consideração i, tem-se a estatística G_i como apresentada da definição 2. Se incluir a observação i no somatório, obtém-se G_i^*, que é expressa como:

$$G^*(d) = \frac{\sum_j w_{ij}(d) y_j}{\sum_j y_j}, \quad \text{para qualquer } j. \tag{3}$$

A média da estatística G_i é dada por

$$\mathbb{E}(G) = \frac{W_i}{(n-1)}$$

em que $W_i = \sum_j x_{ij}(d)$

A variância da estatística G_i é dada por

$$Var(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-1)} \left[\frac{s(i)}{\bar{y}(i)}\right]^2 \tag{4}$$

em que $\bar{y}(i) = \dfrac{\sum_j y_j}{(n-1)}$ e $s^2(i) = \dfrac{\sum_j y_j^2}{(n-1)} - [\bar{y}(i)]^2$.

A interpretação da estatística G é feita com base no sinal de $Z(G_i)$, valores positivos e significativos indicam um *cluster* espacial do tipo *hot spot*, ou seja, com valores altos para a variável de interesse. Por outro lado, um valor negativo e significativo de $Z(G_i)$ indica um *cluster* do tipo *cool spot*, ou seja de baixos valores para a variável de interesse. A inferência a respeito da significância da estatística G_i é baseada na normal padrão, ou seja, $Z(G_i)$ (ALMEIDA, 2012).

Ressalta-se, ainda, que o indicador proposto por Getis e Ord não é capaz de revelar uma situação de correlação negativa, ou seja, um padrão espacial de dispersão da variável. A estatística G_i de Getis e Ord só fornece informação sobre o padrão espacial de concentração, ou seja, *hot spot* ou *cool spot*. Os padrões espaciais do tipo Alto-Baixo (AB) ou Baixo-Alto (BA), comuns nos indicadores locais LISA a serem apresentados, não são identificados por esse indicador. Para mais, o indicador de Getis e Ord não pode ser calculado para valores negativos da variável. Também é importante ressaltar que nessa versão do indicador, a matriz $\boldsymbol{W}$ utilizada tem que ser simétrica e com pesos binários (ALMEIDA, 2012).

Com o objetivo de possibilitar o calculo da estatística G_i com valores negativos e com a possibilidade de incorporação de uma matriz de pesos não simétrica, Getis e Ord (1995) apresentaram uma nova estatística (NG_i), obtida a partir da padronização de G_i. A estatística (NG_i) é expressa como

$$NG_i = \frac{G_i - \mathbb{E}(G_i)}{DP(G_i)} \tag{5}$$

em que $\mathbb{E}(G_i)$ é a média teórica de G_i e $DP(G_i)$ é o desvio padrão de G_i (ALMEIDA, 2012).

O indicador com a capacidade de detectar padrões locais de autocorrelação espacial, estatisticamente significativos, é o I_i de Moran local (ALMEIDA, 2012). Segundo Anselin (1995), um indicador com capacidade de identificar padrões locais de autocorrelação

espacial é denominado *"local indicator of spatial association"*(LISA) e deve atender a duas condições (ANSELIN, 1995):

1. para cada observação, o indicador, deve ser capaz de indicar *clusters* espaciais estatisticamente significativos;
2. o somatório dos indicadores locais, calculados para todas as regiões, deve corresponder ao indicador de autocorrelação espacial global para as mesmas regiões.

O I_i de Moran local apresenta uma decomposição do I de Moran global em quatro grupos: Alto-Alto (AA), Baixo-Baixo (BB), Alto-Baixo (AB) e Baixo-Alto (BA). Esses grupos equivalem aos quadrantes do diagrama de dispersão de Moran, apresentado na Figura 2. Além disso, o I_i de Moran local para a variável padronizada z_i, observada na região i, pode ser expresso como

$$I_i = z_i \sum_j w_{ij} z_j,$$

O valor de I_i é calculado levando-se em consideração apenas as regiões vizinhas de i, definidas através da matriz de pesos espaciais. É obtido um valor de I_i para cada observação n, o que gera uma grande quantidade de informação e prejudica a interpretação. Dessa forma, uma maneira mais eficaz de visualizar o conjunto de estatísticas geradas pelo I_i de Moran local é através de um mapa de *clusters*. Esse mapa combina informação do diagrama de dispersão de Moran (Figura 2) com a significância da medida de associação local I_i (ALMEIDA, 2012).

As duas estatísticas de autocorrelação espacial local apresentadas, o G_i e o I_i, apresentam vantagens e desvantagens. Uma vantagem do G_i, em relação ao I_i, é a capacidade de estabelecer uma definição mais clara de *clusters* com valores altos (*hot spots*) ou *clusters* com valores baixos (*cool spots*). Entretanto, o I_i tem a vantagem de indicar uma associação espacial de valores dissimilares quando apresenta valores negativos. Ou seja, indica a presença de valores baixos da variável, circundados por valores altos ou valores altos rodeados por valores baixos. Estas observações, conhecidas como *outliers* globais, não são captadas pelo G_i (DARMOFAL, 2006).

2.3 ESTATÍSTICA MULTIVARIADA

A estatística multivariada é definida como um conjunto de métodos estatísticos que são empregados com o objetivo de analisar as variáveis como um todo, levando-se em conta, por exemplo, a sua estrutura de correlação. Considera-se que os dados são multivariados quando cada unidade amostral contém diversas variáveis aleatórias. Portanto, os métodos

de estatística multivariada permitem a realização de uma avaliação muito mais abrangente do conjunto de dados, possibilitando a descoberta de padrões que, possivelmente, não seriam revelados ao se analisar cada variável individualmente (MINGOTI, 2005).

Uma das formas de se classificar os métodos multivariados é proposta por Hair et al. (2009). Segundo a classificação dos autores, os métodos multivariados podem ser divididos como métodos de dependência e interdependência ou, *supervised learning* e *unsupervised learning*. Se for possível identificar variáveis com uma estrutura de dependência aconselha-se o uso de técnicas de dependência, tais como regressão múltipla, regressão logística ou análise discriminante. No entanto, se não houver uma distinção entre quais variáveis são dependentes e independentes, aconselha-se que o uso de técnicas de interdependência como análise fatorial e análise de agrupamento.

Os dados multivariados são representados através de matrizes. Por exemplo, uma amostra aleatória que contenha n observações contendo valores de p variáveis observadas dá origem a uma matriz de dados X com dimensão n (linhas) por p (colunas):

$$X_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \tag{6}$$

em que cada observação é representada por uma linha da matriz de dados X, sendo um vetor com p variáveis, e cada variável é representada por uma coluna de X, sendo um vetor com n elementos, as observações (EVERITT; HOTHORN, 2011).

A exposição de dados multivariados através de uma matriz como exposta na definição (6) pode não ser muito informativa, principalmente se as dimensões n e p foram grandes. Dessa forma, faz-se necessário a utilização de medidas resumo dos dados amostrais, calculando-se a média, mediana, desvio padrão etc. de forma a sintetizar as informações dos dados da amostra (FERREIRA, 2011).

No caso multivariado, a média amostral, muito utilizada como uma medida de tendência central, torna-se o vetor de médias amostral de dimensão $p \times 1$, onde cada elemento $\bar{X}_i$, com $i = 1, 2, \cdots, p$, é a média de uma variável:

$$\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}.$$

No caso multivariado, utiliza-se, como medida de dispersão dos dados, no lugar da variância amostral, a matriz de covariâncias amostral $\boldsymbol{S}$ de dimensão $p \times p$. Esta matriz apresenta, em sua diagonal principal, as variâncias de cada uma das p variáveis, e os elementos fora da diagonal principal são as covariâncias entre as variáveis. A matriz de covariâncias amostral é, portanto, simétrica, ou seja, $S_{ij} = S_{ji}$.

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}.$$

Do mesmo modo, a correlação pode ser utilizada como uma medida de associação entre duas variáveis. Os possíveis valores do coeficiente de correlação se encontram no intervalo entre -1 e 1. Valores próximos de 1 apontam para o fato de que as variáveis estão correlacionadas de forma positiva, ou seja, grandes valores de uma estão associados a grandes valores da outra. Por sua vez, valores próximos de -1 indicam que as variáveis estão correlacionadas de forma negativa, o que sugere que grandes valores de uma estão associados a pequenos valores da outra. A matriz de correlações amostral é dada por

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}.$$

em que $r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$ (FERREIRA, 2011).

Para mais, outras estatísticas descritivas, como a matriz de somas de quadrados e produtos e a matriz de correlações, podem ser consideradas, dependendo do objetivo da pesquisa (FERREIRA, 2011).

As técnicas de estatística multivariada podem ser classificadas em duas categorias, as técnicas exploratórias e as técnicas de inferência estatística. A primeira categoria compreende as técnicas que não dependem do conhecimento da forma matemática da distribuição de probabilidade que gerou a amostra de dados. O uso dessas técnicas permite a identificação de padrões em dados multivariados, o que faz com que estas técnicas possuam considerável apelo prático. É possível citar como exemplos de técnicas exploratórias análise de componentes principais, análise fatorial exploratória, análise de agrupamento. Em contraste, a categoria de técnicas de de inferência estatística têm como objetivo a estimação de parâmetros, testes de hipóteses, análise de regressão multivariada etc. Dessa

forma, as técnicas de inferência fazem uso da amostra para realizar inferências sobre a população de onde essa amostra foi extraída (MINGOTI, 2005).

2.4 ANÁLISE DE AGRUPAMENTO

Também conhecida como classificação ou *cluster analysis*, a análise de agrupamento é uma técnica exploratória que tem como objetivo encontrar partições (ou grupos) dos elementos de uma amostra. Os grupos são obtidos de maneira que as observações de um mesmo grupo sejam similares entre si em relação às variáveis analisadas e que as observações de grupos diferentes sejam heterogêneas em relação a essas mesmas variáveis (MINGOTI, 2005).

Vários métodos, cujo objetivo é encontrar grupos de observações homogêneas, são considerados como "análise de agrupamento". Estes métodos procuram dar um tratamento matemático para o que os seres humanos são capazes de fazer de forma relativamente eficiente em duas ou três dimensões, por meio de diagramas de dispersão, por exemplo (EVERITT; HOTHORN, 2011).

Existem duas possibilidades com relação aos objetivos dos métodos de agrupamento. A primeira delas é agrupar as n observações em um número desconhecido de grupos. Também há a possibilidade de classificar as observações em um conjunto predefinido de grupos. No primeiro caso, a análise de agrupamento deve ser utilizada, no segundo caso, a análise discriminante deve ser empregada. É importante ressaltar que, na análise de agrupamento, geralmente, o número de grupos não é conhecido a princípio e encontrar o melhor agrupamento não é considerado uma tarefa simples (FERREIRA, 2011). Segundo Bartholomew et al. (2008), existem duas etapas fundamentais em qualquer processo de agrupamento. A primeira etapa consiste em obter as distâncias entre todos os pares de pontos para construção da matriz de distâncias. A segunda etapa compreende o desenvolvimento de um algoritmo para formação de grupos baseados nessas distâncias.

As distâncias são definidas segundo medidas de similaridade ou dissimilaridade. As medidas de dissimilaridade correspondem às distâncias, ao passo que as de similaridades complementam as distâncias, assim, quanto maior a medida de similaridade entre dois objetos mais próximos eles serão (FERREIRA, 2011). Algumas considerações a respeito de medidas de distâncias que podem ser utilizadas na análise de agrupamento serão apresentadas na próxima seção.

2.4.1 Distâncias

A utilização de um algoritmo de agrupamento implica que se estabeleça, previamente, uma medida de distância. Há diversas medidas de similaridade ou dissimilaridade entre pares de observações. Entre as medidas de dissimilaridade mais comuns estão: a distância euclidiana, distância euclidiana padronizada, Manhattan, Mahalanobis, etc. É

necessário destacar que, no caso de medidas de dissimilaridade, quanto menores os seus valores, mais próximos ou similares são os objetos comparados. Além disso, a escolha da métrica de distância interfere diretamente no resultado final do agrupamento (MINGOTI, 2005; EVERITT; HOTHORN, 2011).

Uma das métricas de distância mais elementar é a distância euclidiana, que pode ser expressa por

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2},$$

em que d_{ij} é a distância euclidiana entre os elementos i, com os valores X_{i1}, X_{i2}, ..., X_{ip}, e j, com os valores X_{j1}, X_{j2}, ..., X_{jp}.

A distância de Mahalanobis e a euclidiana padronizada são generalizações da distância euclidiana. Sendo assim, a distância generalizada entre dois elementos $X_{i\cdot}$ e $X_{j\cdot}$ pode ser definida como

$$d_{ij} = (X_{i\cdot} - X_{j\cdot})^T A(X_{i\cdot} - X_{j\cdot}). \tag{7}$$

A seleção da matriz A determina a distância a ser calculada. No caso em que $A = I$ obtém-se a distâncias euclidiana, se $A = D^{-1}$ tem-se como resultado a distância euclidiana padronizada. Se $A = S^{-1}$, ou seja, a inversa da matriz de covariâncias dos dados, obtém-se a distância de Mahalanobis (MINGOTI, 2005).

Após a definição da medida de distância a ser utilizada, o passo seguinte da análise consiste em escolher um método de agrupamento. De acordo com Ferreira (2011), os métodos de agrupamento podem ser classificados em não hierárquicos ou hierárquicos (aglomerativos ou divisivos). Nos métodos hierárquicos divisivos, no início, há um único grupo com as n observações e, ao final, haverá n grupos, cada um com uma observação. Nos métodos não hierárquicos é necessário definir o número de grupos (k) previamente de forma que seja possível atribuir as n observações aos k grupos da melhor maneira possível. No início do processo utiliza-se uma alocação arbitrária e, iterativamente, busca-se a alocação ótima. Nos métodos hierárquicos aglomerativos ocorre o processo contrário, o processo de agrupamento se inicia com n grupos, cada um contendo uma observação, no final do método haverá um único grupo com todas as observações. A cada passo do processo iterativo, cada observação ou grupo é unido a outra observação ou grupo. A união se dá através de um critério de similaridade, os objetos mais próximos entre si são alocados para o mesmo grupo, até que, no final, todos estejam em um único grupo (FERREIRA, 2011).

2.4.2 Técnicas hierárquicas aglomerativas

Ao se utilizar o método hierárquico aglomerativo, depois que uma fusão é realizada, ela não será mais desfeita. Assim, quando o método coloca dois elementos em um mesmo grupo, eles não mais aparecerão em grupos diferentes. Dessa forma, para se encontrar uma solução adequada com o número de agrupamentos ótimo, é necessário adotar algum critério de divisão (EVERITT; HOTHORN, 2011).

Para ilustrar as fusões ou divisões realizadas a cada passo do processo de agrupamento, utiliza-se o dendrograma. Este tipo de gráfico tem a capacidade de representar os agrupamentos obtidos a partir de métodos hierárquicos, aglomerativos ou divisivos (EVERITT; HOTHORN, 2011).

Um dos métodos hierárquicos aglomerativos mais frequentemente utilizado é o método de *Ward*, ou método da mínima variância. Este método se fundamenta na mudança de variação entre os grupos e também dentro dos grupos que são formados em cada passo do algoritmo. Assim, a cada passo de agrupamento é calculada a soma de quadrados dentro dos grupos. Esta soma é composta do quadrado da distância euclidiana de cada elemento amostral pertencente ao grupo em relação à seu respectivo vetor de médias. A soma de quadrados dentro do i-ésimo grupo é definida como:

$$SQDG_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

sendo n_i o número de elementos no grupo i, X_{ij} o vetor de observações do j-ésimo elemento amostral pertencente ao i-ésimo agrupamento e $\bar{X}_i$ o centróide do i-ésimo agrupamento, representando a soma de quadrados correspondente a tal grupo.

A distância entre dois grupos quaisquer, G_i e G_l, é definida como:

$$d_{G_i,G_l} = \left(\frac{n_l n_i}{n_l + n_i}\right)(X_{ij} - \bar{X}_i)^2. \tag{8}$$

Em cada iteração do algoritmo, os dois grupos que minimizam a distância em (8) são combinados. Este método geralmente produz agrupamentos com aproximadamente o mesmo número de elementos em cada grupo e é adequado apenas para variáveis quantitativas por se baseado na comparação do vetor de médias amostral (MINGOTI, 2005).

2.4.3 Técnica não hierárquica: K-médias

Outro algoritmo de agrupamento é conhecido como método das k-médias (ou *k-means*). Este método busca por uma partição das n observações em k agrupamentos (G_1, G_2, $\cdots$, G_k), em que G_i indica o conjunto de observações que pertence ao i-ésimo

grupo e k é dado por algum critério numérico de minimização. A implementação mais utilizada do método das k-médias tenta encontrar a partição dos n elementos em k grupos que minimize a soma de quadrados dentro dos grupos ($SQDG$) em relação a todas as variáveis. Esse critério pode ser escrito como

$$SQDG = \sum_{j=1}^{p} \sum_{l=1}^{k} \sum_{i \in G_l} (X_{ij} - \bar{X}_j^{(l)})^2,$$

em que $\bar{X}_j^{(l)} = \dfrac{1}{n_i} \sum_{i \in G_l} X_{ij}$ é a média dos indivíduos no grupo G_l em relação à variável j (EVERITT; HOTHORN, 2011).

Apesar de o problema parecer relativamente simples, ele não é tão elementar. A tarefa de selecionar a partição com a menor soma de quadrados dentro dos grupos se torna complexa pois o número de possíveis partições torna-se muito grande mesmo com um tamanho amostral não tão grande. Por exemplo, para $n = 100$ e $k = 5$, o número de partições é da ordem de 10^{68}. Este fato induziu o desenvolvimento de algoritmos que, embora não garantam encontrar a solução ótima, levam a soluções igualmente aceitáveis.

2.4.4 Número de grupos

Independentemente do algoritmo utilizado, em aplicações de métodos de agrupamento, é necessário que se defina o número adequado de grupos. Isso se deve ao fato de que essa é a última etapa nos métodos de agrupamento hierárquicos aglomerativos e a etapa inicial nos agrupamentos não hierárquicos.

Como exposto, nas técnicas hierárquicas aglomerativas, a cada passo, o algoritmo reúne duas observações ou grupos e ao final, um único grupo com as k observações é obtido. Dessa forma, é necessário estabelecer uma regra de corte para que o número ideal de grupos seja escolhido. No caso de métodos não hierárquicos, como o das k-médias, por definição, o número de grupos seja escolhido a priori (MINGOTI, 2005).

Alguns autores indicam a aplicação de uma combinação dos métodos hierárquicos e das k-médias. Nessa abordagem, primeiramente uma técnica hierárquica é aplicada para identificar o número de grupos e, em seguida, é aplicado o método das k-médias para classificar as observações (HAIR, 2009; MINGOTI, 2005). No entanto, em face a esta profusão de possíveis abordagens e, por não haver um critério adequado em todas as situações, é razoável levar em conta considerações de ordem práticas. Em algumas situações, é possível haver alguma informação a priori ou uma teoria que sugira uma estrutura nos dados. Ademais, a interpretabilidade e o significado prático e útil dos resultados deve ser o principal direcionador da escolha do número de grupos (EVERITT; HOTHORN, 2011).

3 MATERIAL E MÉTODOS

O objetivo desta seção é apresentar os dados utilizados no trabalho, quais as variáveis selecionadas, assim como descrever a metodologia usada na presente análise.

3.1 Dados

Este trabalho utiliza dados referentes a apólices de seguro rural dos municípios brasileiros no ano de 2019. Os dados sobre seguro rural estão disponíveis no endereço eletrônico do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Também foram utilizados dados que contém atributos geográficos, como a posição e o formato, do território brasileiro. Esses dados estão disponíveis no endereço eletrônico do Instituto Brasileiro de Geografia e Estatística (IBGE, 2020). As variáveis utilizadas, bem como suas siglas e descrições, são apresentadas na Tabela 1.

Tabela 1 – Descrição das variáveis utilizadas.

Variável	Sigla
Total de apólices contratadas	TAC
Soma da importância segurada (R$ milhão)	SIS
Soma dos prêmios (R$ milhão)	SPR
Total de subvenção (R$ milhão)	TSB
Soma das indenizações pagas (R$ milhão)	SIP
Taxa média aplicada às apólices	TMA
Número de apólices indenizadas	NAI

Fonte: Elaboração própria

Como as variáveis estão em diferentes escalas e, com isto possuem diferentes variâncias, realizou-se a padronização das variáveis para que tais fatores não interferissem na análise. Tal padronização foi efetuada subtraindo-se de cada observação a média de sua respectiva variável e dividindo-se posteriormente pelo desvio padrão da variável.

3.2 Metodologia

Após a escolha das variáveis, foi realizada uma análise exploratória dos dados em que são apresentadas a distribuição espacial das variáveis, assim como a distribuição espacial das estatísticas de autocorrelação espacial I de Moran e G de Getis e Ord. Em seguida, é aplicado o procedimento, descrito a seguir, para descobrir e explorar padrões de agrupamento espacial com base na distribuição espacial de dados multivariados.

O procedimento para a criação de agrupamentos com baseados medidas locais de autocorrelação espacial pode ser descrito da seguinte forma:

1. Dada uma matriz de pesos espaciais W, calcule para cada variável a estatística G de Getis-Ord local. Seja $z(G_j(x_i))$ a estatística resultante de (2) para a $j-$ésima variável no $i-$ésimo município;

2. Reúna os valores do passo anterior em uma matriz Z de dimensão $(n \times p)$. Cada coluna de Z expressa o padrão de autocorrelação local para uma variável, enquanto cada linha de Z fornece o perfil de agrupamento em torno de cada unidade local;

3. Aplique o método de *Ward* no conjunto Z de novas variáveis para especificar o número de grupos a ser utilizado no método das k-médias;

4. Aplique o algoritmo k-médias no conjunto Z. Esta etapa permite agrupar observações com base em seus perfis espaciais multivariados que contêm informações de localização e das variáveis.

O procedimento também foi empregado utilizando-se, como medida de autocorrelação espacial local, o I de Moran local.

Optou-se pelo método não hierárquico das k-médias para a etapa final, pois este método apresenta um desempenho superior aos métodos hierárquicos (MINGOTI, 2005). No entanto, como o método das k-médias impõe que seja inicialmente definido o número de grupos desejado, nessa abordagem, inicialmente foi utilizada uma técnica hierárquica para identificar este número[2]. Em seguida, o método das k-médias foi aplicado para classificar as observações. Nesse trabalho, optou-se pelo método de *Ward* para especificar o número de grupos pretendido. Além de ser um método com desempenho superior aos demais, ele é considerado indicado para dados quantitativos contínuos (EVERITT; HOTHORN, 2011).

A definição do número de grupos da partição final da aplicação do método de *Ward* foi realizada no final do processo de agrupamento, decidindo-se o ponto de corte no dendrograma. Dessa forma, foram utilizados os critérios da maior diferença no nível de fusão e da interpretabilidade. O primeiro critério faz referência à maior distância no dendrograma, que indica que o grupo pode se tornar menos homogêneo internamente com essa união (EVERITT; HOTHORN, 2011). O segundo critério concerne à solução que proporciona uma interpretação que corresponda aos objetivos da análise.

3.3 Recursos computacionais

Esse estudo foi realizado com a linguagem de programação *Python* (PYTHON, 2017), utilizando-se a interface *Jupyter* (JUPYTER, 2017), (PEREZ; GRANGER, 2007) (KLUYVER, 2016). Além disso, as seguintes bibliotecas foram utilizadas: *Pandas* (MCKINNEY, 2010), para a manipulação de dados, *NumPy* (WALT; COLBERT; VAROQUAUX, 2011), que possibilita computação numérica com *Python*, *Matplotlib* (HUNTER, 2007) e *Seaborn* (WASKOM, 2014), que são bibliotecas para a criação de gráficos. A análise de componentes principais foi realizada através da biblioteca *sklearn* e as bibliotecas

[2] Alguns critérios para escolha do número de grupos foram apresentados na seção 2.4.4, dentre eles o uso de uma técnica hierárquica aglomerativa.

Geopandas (JORDAHL, 2014) e *PySAL* (REY; ANSELIN, 2007) possibilitaram a análise espacial. Os códigos utilizados na análise estão disponíveis no *GitHub* e os link estão no Apêndice B.

4 RESULTADOS E DISCUSSÃO

4.1 Distribuição espacial

A análise se inicia com a visualização da distribuição espacial das variáveis por meio dos mapas temáticos. Busca-se, através desses mapas, identificar visualmente se existem padrões na distribuição espacial das variáveis do seguro rural.

O primeiro grupo de mapas, apresentado na Figura 3, exibe a distribuição espacial de cada uma das variáveis analisadas no ano de 2019. Pela análise da Figura 3, é possível constatar que a distribuição espacial das variáveis é bastante semelhante. Embora com distribuições espaciais muito similares entre si, as variáveis SIP (soma das indenizações pagas) e NAI (número de apólices indenizadas) diferem das demais variáveis. Este fato pode ser causado, entre outros motivos, pelo fato que estas variáveis se relacionam à ocorrência de sinistros. Além disso, é possível destacar que há indícios de concentrações espaciais de todas as variáveis nas regiões do Extremo Oeste Baiano (BA), Sudoeste de Mato Grosso do Sul, Sul Goiano (GO), na região Sudeste e no sul do Estado de São Paulo.

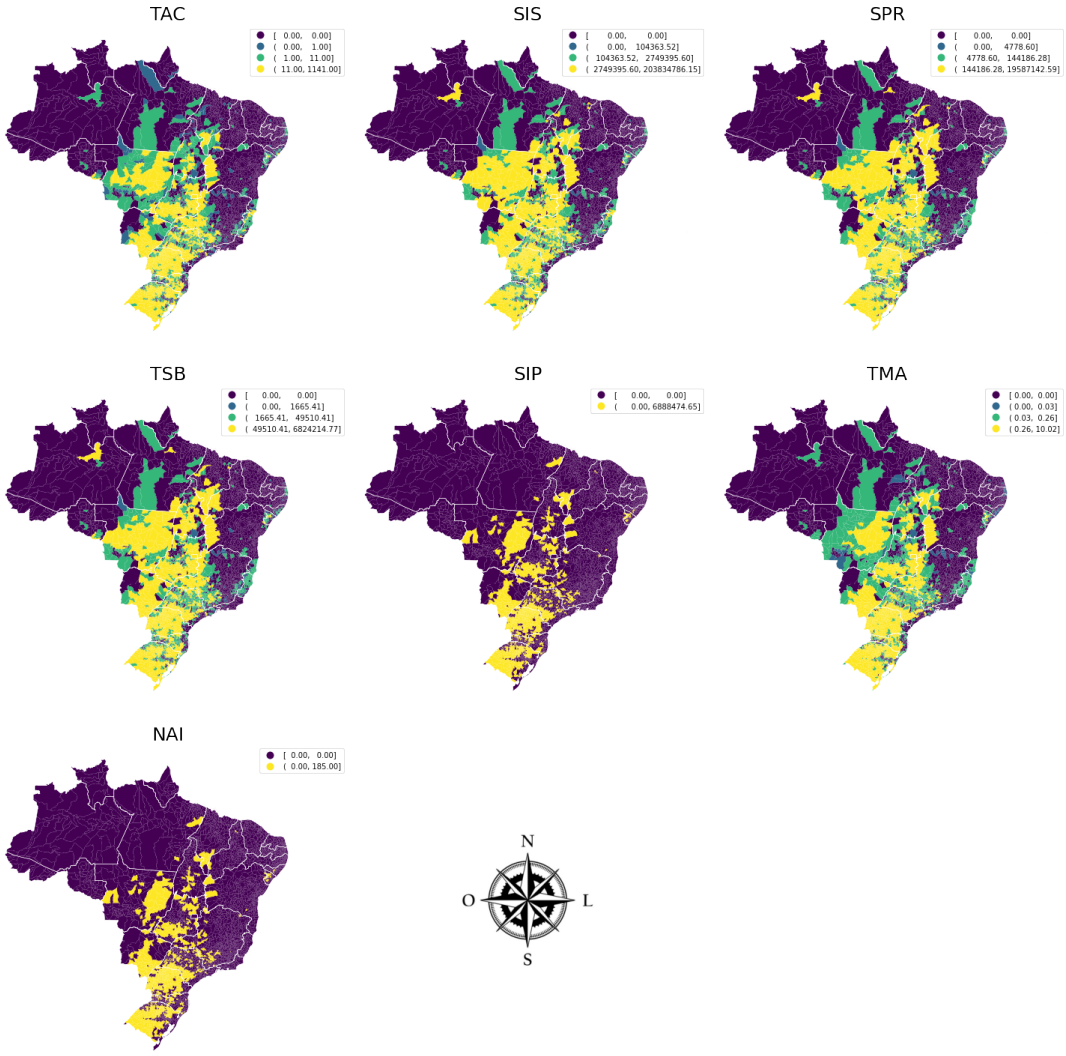
A concentração regional do número de apólices, do valor de subvenção e, consequentemente, das demais variáveis do seguro rural acontece devido a motivos históricos e climáticos. De acordo com Santos e Silva (2017), isso ocorre pela presença de maiores riscos de intempéries nos estados da Região Sul, em São Paulo e Minas Gerais. Além disso, estados como Mato Grosso do Sul, Mato Grosso, Goiás, e a região formada pelos estados do Maranhão, Tocantins, Piauí e Bahia, mais recentemente, também têm aderido aos contratos de seguro rural devido a fatores climáticos (SANTOS; SILVA, 2017).

4.2 Autocorrelação espacial

O próximo passo da análise da distribuição espacial consistiu na construção dos mapas que representam a autocorrelação espacial local. Os mapas *LISA* apresentam quatro grupos (Alto-Alto (AA), Baixo-Baixo (BB), Alto-Baixo (AB) e Baixo-Alto (BA)) com características distintas de associação espacial estatisticamente significativas. Os mapas *LISA* apresentados na Figura 4 apresentam os pontos que foram estatisticamente significativos no ano de 2019.

Na Figura 4 é possível identificar que a variável TAC apresenta concentrações de municípios do tipo AA nos estados de Goiás, Minas Gerais, Mato Grosso do Sul, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul. A variável SIS, apresenta concentrações de municípios do tipo AA nos estados do Tocantins, Bahia, Mato Grosso, Mato Grosso

Figura 3 – Distribuição espacial das variáveis de seguro rural.



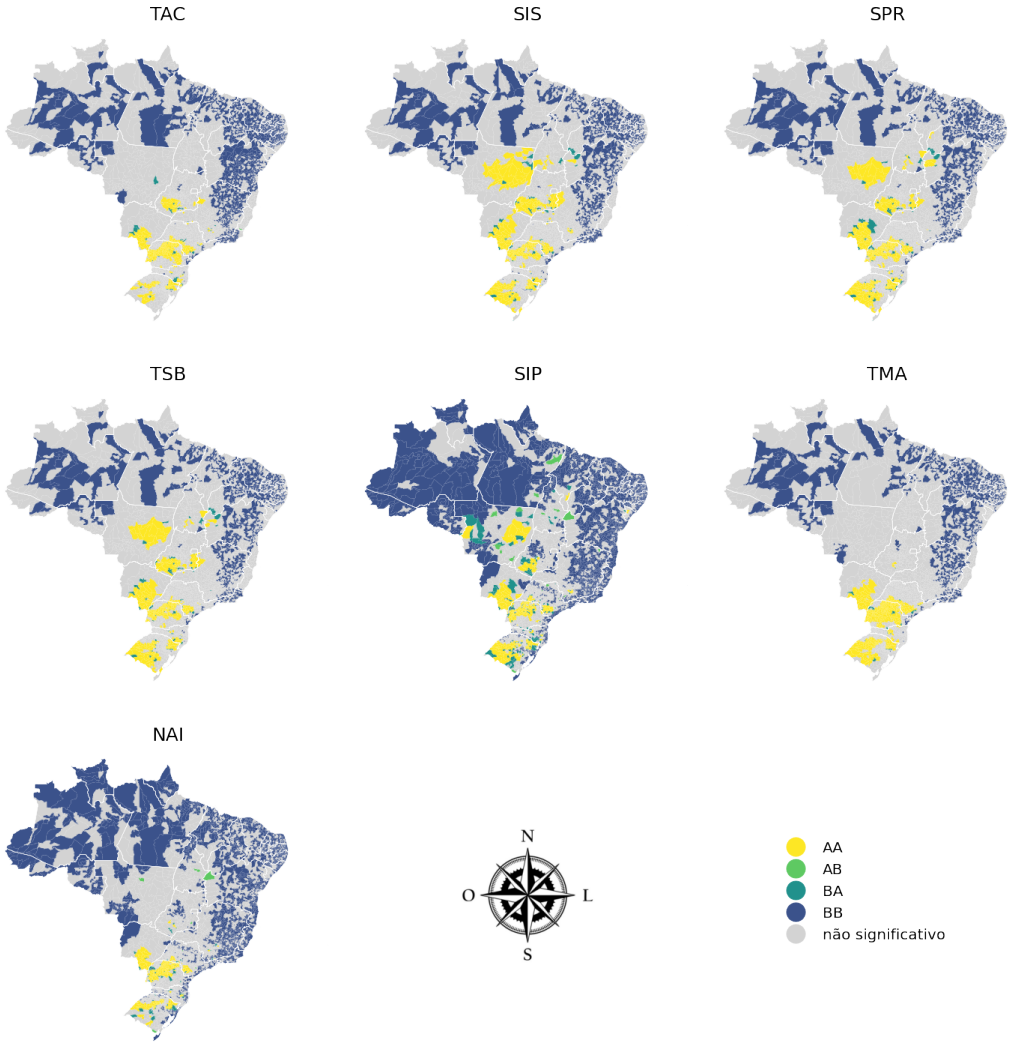
Fonte: Elaboração própria

do Sul, Goiás, Minas Gerais, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul. Distribuições semelhantes também são observadas nas variáveis SPR, TSB e SIP, que também apresentam municípios do tipo AA e AB no estado do Piauí. Os municípios do tipo AA das variáveis TMA e NAI se concentram principalmente na região Sul do país e no sul do estado de São Paulo. Por sua vez, os agrupamentos de municípios do tipo BB se localizam, principalmente, nas regiões Norte e Nordeste.

Os mapas apresentados na Figura 5 apresentam os agrupamentos das variáveis que foram estatisticamente significativos no ano de 2019 considerando-se a estatística G de Getis e Ord. Ou seja, apresenta agrupamentos localizados de concentração espacial (*hot spots* e *cool spots*).

É possível observar nos mapas da Figura 5 que os municípios do tipo *hot spot*, ou seja, concentrações com valores altos da variável, se localizam, principalmente, nos estados da região Centro-Oeste, Sudeste e Sul. Mais especificamente, as variáveis TAI, TMA E NAI,

Figura 4 – Distribuição espacial do I de Moran local para as variáveis de seguro rural.



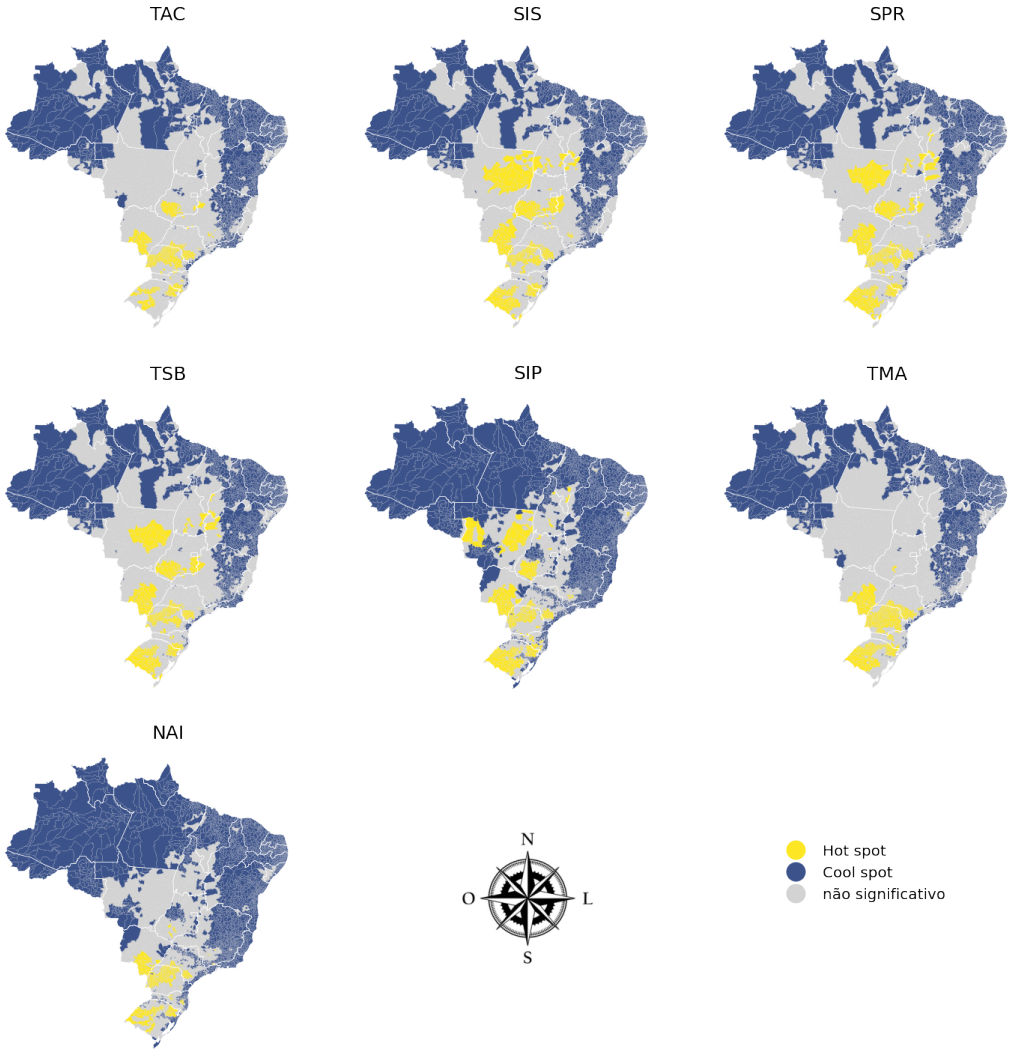
Fonte: Elaboração própria

apresentam concentrações de valores altos nos estados da região Sul, no sul do estado de São Paulo, em alguns municípios de Goiás e no estado do Mato Grosso do Sul. Os municípios do tipo *hot spot* das variáveis SIS, SPR, TSB e SIP apresentam distribuição espacial semelhante àquela dos municípios do tipo AA. Essas variáveis apresentam concentrações de municípios do tipo *hot spot* nos estados do Tocantins, Bahia, Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul.

4.3 Identificação dos agrupamentos

Em seguida, foi obtido o dendrograma da análise de agrupamento pelo método de *Ward*, apresentado na Figura 6. O corte foi realizado na altura que classifica as observações em cinco grupos, para o caso do agrupamento realizado com o I de Moran local (Figura 6(a)), e quatro grupos, quando se considera o G de Getis e Ord (Figura 6(b)). Também foram examinadas as partições com dois grupos, no entanto, ao se considerar a

Figura 5 – Distribuição espacial do G de Getis e Ord local para as variáveis de seguro rural.

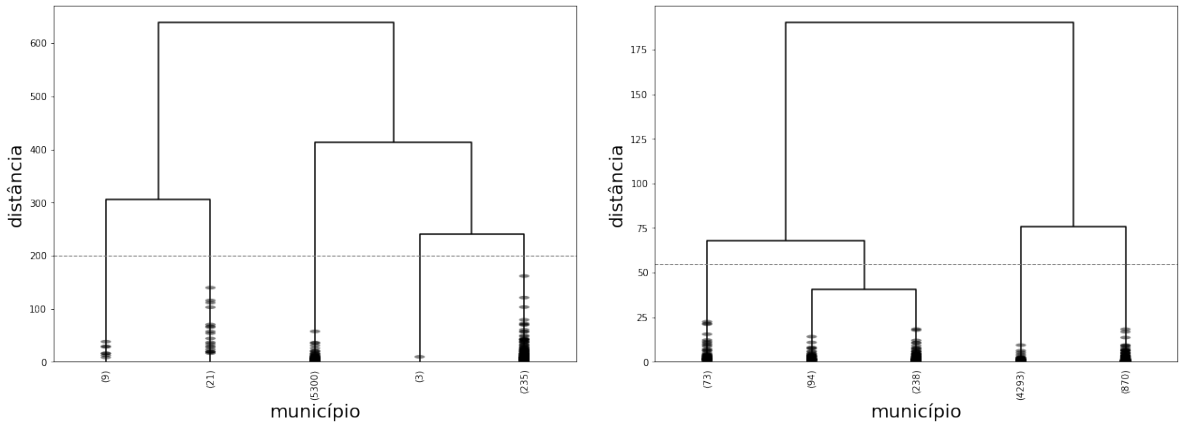


Fonte: Elaboração própria

interpretabilidade dos resultados, não foram encontradas evidências de que seriam mais adequadas. Dessa forma, o número predefinido de grupos usado para aplicação do método das k-médias foi cinco para os agrupamentos formados com o I de Moran local, e quatro para os agrupamentos formados com o G de Getis e Ord.

O grupo de mapas apresentado na Figura 7 exibe a distribuição espacial dos agrupamentos formados pelo método de *Ward* com I de Moran (Figura 7(a)) e G de Getis e Ord (Figura 7(b)). Pela análise da Figura 7(a), é possível observar que os municípios do grupo 1 se localizam nos estados do Mato Grosso, Mato Grosso do Sul, Goiás, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul. Os municípios do grupo 2 se encontram no sul do estado de São Paulo, no estado do Paraná e no Nordeste Rio grandense. Por sua vez, os municípios pertencentes ao grupo 3 localizam-se, principalmente, no Sudoeste do estado do Mato Grosso do Sul e Nordeste Rio grandense e região Serrana se Santa Catarina. Por

Figura 6 – Dendrogramas

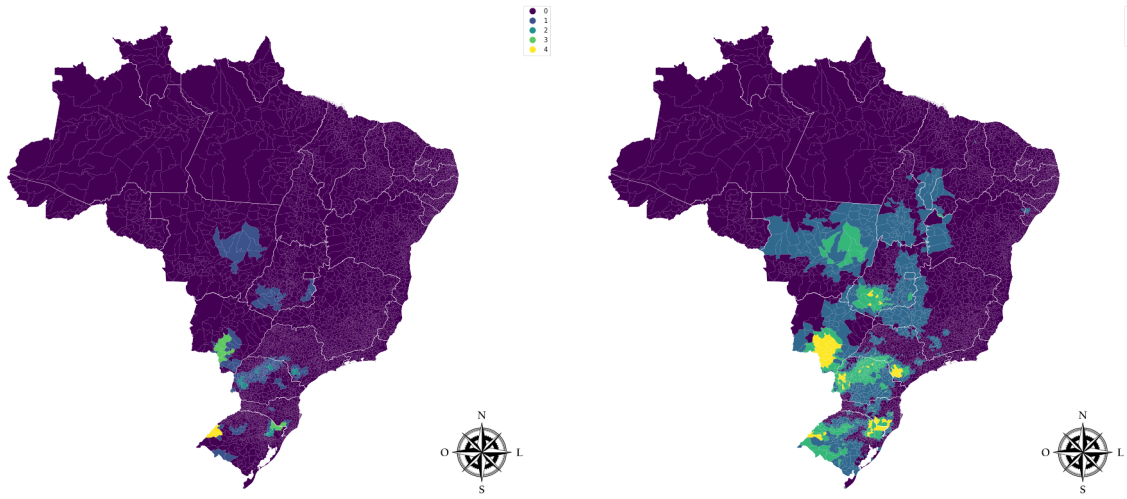


(a) I de Moran (b) G de Getis e Ord

Fonte: Elaboração própria

fim, os municípios do grupo 4 se localizam na mesorregião Sudoeste Rio-Grandense no estado do Rio Grande do Sul.

Figura 7 – Agrupamentos formados pelo método de *Ward* com I de Moran e G de Getis e Ord



(a) Grupos I de Moran (b) Grupos G de Getis e Ord

Fonte: Elaboração própria

Nos agrupamentos formados utilizando a estatística G de Getis e Ord (Figura 7(b)) os municípios do grupo 1 estão localizados nos estados do Maranhão, Piauí, Bahia e Sergipe na região Nordeste, nos estados da região Centro-Oeste, nos estados de Minas Gerais e São Paulo, no Sudeste, e nos estados da região Sul. Os municípios do grupo 2 localizam-se, predominantemente, nas regiões Centro-Oeste e Sul, com exceção de alguns

municípios localizados no Sul do estado de São Paulo. Por fim, os municípios pertencentes ao grupo 3 encontram-se no Sul Goiano, nas mesorregiões Sudoeste e Centro-Norte do Mato Grosso do Sul, na mesorregião de Itapetininga no estado de São Paulo, no Oeste Paranaense, na região Serrana de Santa Catarina e no Sudoeste e Nordeste Rio-Grandense.

Para auxiliar na análise dos resultados a média das estatísticas G_i e do I de Moran nos grupos formados pelos métodos de *Ward* e das $k-$médias são apresentados nas tabelas 2 e 5, respectivamente. A partir desses resultados, foi possível encontrar o perfil de autocorrelação espacial de cada agrupamento resultante.

Tabela 2 – Média da estatística G_i nos grupos formados pelos métodos de *Ward* e das $k-$médias

Método de agrupamento	Grupos	Variáveis						
		TAC	SIS	SPR	TSB	SIP	TMA	NAI
Ward	0	-0,25	-0,25	-0,24	-0,24	-0,19	-0,31	-0,20
	1	0,34	0,56	0,42	0,42	0,31	0,54	0,18
	2	2,19	1,82	1,80	1,81	1,57	2,63	2,04
	3	4,29	4,56	4,90	4,91	3,70	3,37	3,13
$k-$médias	0	-0,23	-0,23	-0,22	-0,22	-0,18	-0,29	-0,19
	1	0,47	0,73	0,57	0,57	0,40	0,72	0,27
	2	4,43	4,22	4,40	4,41	3,61	3,65	3,60
	3	2,14	1,74	1,81	1,81	1,59	2,65	1,98

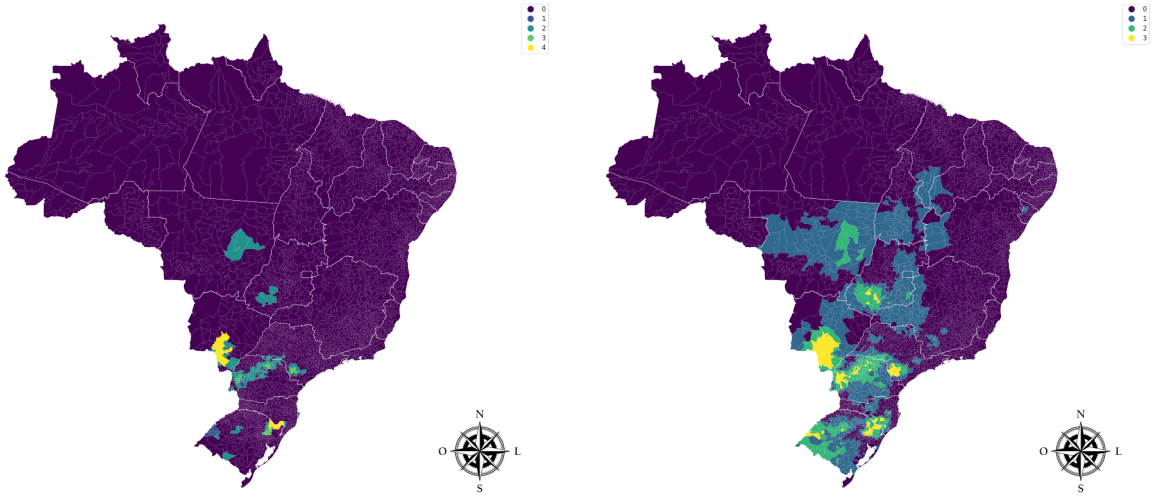
Fonte: Elaboração própria.

Os grupos formados pelo método de *Ward* apontam para uma caracterização de grupos em que o grupo 0 é formado por concentrações do tipo *cool spot*. Ou seja, o grupo 0 é constituído por municípios com valores baixos de todas as variáveis. O grupo 1 caracteriza-se por concentrações do tipo *hot spot* com valores da estatística G_i entre $0,56$ na variável SIS e $0,18$ na variável NAI. Por sua vez, o grupo 2 também é constituído de municípios classificados como *hot spot* e o valor da estatística G_i varia de $1,57$ na variável SIP a $2,36$ na variável TMA. Além disso, observa-se na tabela 2 que a utilização do método das $k-$médias apresenta uma divisão de grupos com características muito semelhantes àquela feita pelo método de *Ward*. Ou seja, o grupo 0 é o único constituído de municípios com característica de concentração de valores baixos das variáveis (*cool spots*) e os demais grupos apresentam concentrações de valores altos das variáveis (*hot spots*).

Os mapas apresentados na Figura 8 apresentam a distribuição espacial dos agrupamentos formados pelo método das $k-$médias com I de Moran (Figura 8(a)) e G de Getis e Ord (Figura 8(b)). Pela análise da Figura 8(a), observa-se que os municípios do grupo 1 localizam-se nos estados do Mato Grosso, Mato Grosso do Sul, Goiás, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul. Por sua vez, os municípios pertencentes ao grupo 2 se encontram no sul do estado se São Paulo, no estado do Paraná e no Nordeste Rio grandense. Já os municípios do grupo 3 se localizam, principalmente, no Sudoeste

do estado do Mato Grosso do Sul e Nordeste Rio grandense e região Serrana de Santa Catarina. Por último, os municípios pertencentes ao grupo 4 localizam-se na mesorregião Sudoeste Rio-Grandense no estado do Rio Grande do Sul.

Figura 8 – Agrupamentos formados pelo método das $k-$médias com I de Moran e G de Getis e Ord



(a) Grupos I de Moran

(b) Grupos G de Getis e Ord

Fonte: Elaboração própria

A Figura 8(b) apresenta os agrupamentos formados utilizando a estatística G de Getis e Ord. Observa-se que os municípios pertencentes ao grupo 1 estão localizados nos estados do Maranhão, Piauí, Bahia e Sergipe na região Nordeste, nos estados da região Centro-Oeste, nos estados de Minas Gerais e São Paulo, no Sudeste, e nos estados da região Sul. Por sua vez, os municípios do grupo 2 se localizam nas regiões Centro-Oeste e Sul, com exceção de alguns municípios localizados no Sul do estado de São Paulo. Por fim, os municípios classificados como pertencentes ao grupo 3, encontram-se no Sul Goiano, nas mesorregiões Sudoeste e Centro-Norte do Mato Grosso do Sul, na mesorregião de Itapetininga no estado de São Paulo, no Oeste Paranaense, na região Serrana de Santa Catarina e no Sudoeste e Nordeste Rio-Grandense.

É possível observar na Tabela 5 que o grupo 0 é constituído por municípios com valores mais baixos da estatística I de Moran local em ambos os métodos, $k-$médias e *Ward*. O grupo 1 é o segundo grupo com maiores valores do I de Moran local. Os grupos 2 e 3 são os grupos com maiores valores de I de Moran local.

Diferentes razões podem considerados como determinantes da concentração do mercado de seguro rural em algumas regiões brasileiras. De acordo com Santos, Sousa e Alvarenga (2013), os principais fatores que devem ser considerados são: o pequeno número e as disparidades no porte das seguradoras que oferecem o seguro rural, dificuldades das

Tabela 3 – Média da estatística I de Moran nos grupos formados pelos métodos de *Ward* e das $k-$médias

Método de agrupamento	Grupos	Variáveis						
		TAC	SIS	SPR	TSB	SIP	TMA	NAI
Ward	0	0,11	0,12	0,09	0,09	0,08	0,26	0,11
	1	7,50	6,18	5,43	5,48	3,91	11,15	5,68
	2	55,71	21,25	23,71	23,66	16,12	25,49	43,24
	3	14,16	44,68	62,54	64,17	11,47	7,50	4,27
	4	1,33	10,35	15,23	13,95	101,29	6,81	5,06
$k-$médias	0	0,14	0,17	0,13	0,13	0,09	0,32	0,12
	1	6,23	8,65	12,17	11,43	82,02	14,88	25,84
	2	10,29	7,06	6,39	6,46	5,15	13,60	8,27
	3	65,42	25,53	29,04	28,90	13,88	28,56	45,68
	4	13,62	43,16	59,62	61,16	13,41	7,54	4,45

Fonte: Elaboração própria.

instituições bancárias com operações no meio rural, levando a informações imprecisas e à elevação de riscos e preços, parcerias governamentais com operadores, como por exemplo, programas de crédito oficial operados pelo Banco do Brasil, que também é o controlador da maior seguradora agrícola, o grau de oportunidade avaliado como pequeno pelas seguradoras, e por fim, o pequeno peso de parcerias e de avaliação das oportunidades envolvendo o segmento de corretagem.

5 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi encontrar agrupamentos de municípios com características semelhantes em relação à adesão ao seguro rural no ano de 2019. Os agrupamentos foram obtidos de forma a levar em consideração não apenas o valor das variáveis de seguro rural, mas também seu posicionamento geográfico.

Os resultados da AEDE apontam que existe dependência espacial em todas as variáveis. Ou seja, existem padrões de associação espacial estatisticamente significativos. Também foi possível identificar a presença de *clusters* espaciais significativos. Tal resultado é observado em todas as variáveis de seguro rural analisadas.

Em geral, identificou-se que as maiores concentrações de apólices de seguro rural estão situadas nas regiões Sul, Centro-Oeste e Sudeste, no sul do Estado de São Paulo. Ou seja, municípios que possuem uma maior adesão ao seguro rural tendem a ser geografica mente próximos de municípios que também têm maior número de apólices de seguro rural contratadas. Além disso, ressalta-se que, embora exista uma concentração do número de apólices contratadas, número de apólices indenizadas, valores de subvenção, indenização e prêmio, atualmente há expansão da demanda por seguros agrícolas.

Apesar de os dados indicarem uma concentração geográfica da adesão ao sistema de seguro rural no Brasil, é necessário levar em consideração outros sistemas, como o Proagro e o programa Garantia Safra. É necessário, ainda, ressaltar que a adesão dos produtores deve ocorrer como uma resposta à percepção do risco das atividades agropecuárias, ou seja, o seguro deve difundir-se com base na compreensão dos riscos e das vantagens de sua contratação.

REFERÊNCIAS

ALMEIDA, E. **Econometria Espacial Aplicada**. Campinas-SP: Alínea, 2012.

ANSELIN, Luc. Local indicators of spatial association – LISA. **Geographical analysis**, v. 27, n. 2, p. 93-115, 1995.

BARROS, A. et al. Seguro Agrícola no Brasil: uma visão estratégica de sua importância para a economia brasileira. **MB Agro**, 2012.

BRASIL. **Agropecuária Brasileira em Números**. Brasília, DF: Brasília: Ministério da Agricultura, Pecuária e Abastecimento (MAPA), 2019a

BRASIL. **Programa de Subvenção ao Prêmio do Seguro Rural: Relatório de Resultado 2018**. Brasília, DF: Brasília: Ministério da Agricultura, Pecuária e Abastecimento (MAPA), 2019b.

BRASIL. **Programa de Subvenção ao Prêmio do Seguro Rural: Relatório de Resultado 2020**. Brasília, DF: Brasília: Ministério da Agricultura, Pecuária e Abastecimento (MAPA), 2021a.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. (2021). **Atlas do seguro rural**. Disponível em <http://indicadores.agricultura.gov.br/atlasdoseguro/index.htm>. Acesso em: 26 jun 2021b.

BURGO, M. N. **Caracterização espacial de riscos na agricultura e implicações para o desenvolvimento de instrumentos para seu gerenciamento**. Dissertação de Mestrado, Escola Superior de Agricultura "Luiz de Queiroz", ESALQ/USP, 2005.

CÂMARA, G. *et al.* Análise espacial e geoprocessamento. In: DRUCK, S. *et al.* (Ed.). **Análise espacial de dados geográficos.** 2. ed. Brasília: Embrapa, 2004. cap. 1, p. 21–52.

CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA (CEPEA) **PIB do Agronegócio Brasileiro**. Disponível em: <https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx>. Acesso em: 26 jun 2021.

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

DARMOFAL, D. Spatial Econometrics and Political Science. In **Annual Meeting of the Southern Political Science Association.** Atlanta, GA, USA: The Society for Political Methodology, 2006. Disponível em: <http://web.cenet.org.cn/upfile/103632.pdf>. Acesso em: 28 jul. 2020.

EVERITT, B.; HOTHORN, T. **An introduction to applied multivariate analysis with R.** Nova York: Springer-Verlag, 2011.

FERREIRA, D. F. **Estatística multivariada**. 2. ed. Lavras: Editora UFLA, 2011.

FERREIRA, A. L. C. J.; FERREIRA, L. da R. Experiências internacionais de seguro rural: as novas perspectivas de política agrícola para o Brasil. **Econômica**, Rio de janeiro, v. 11, n. 1, p. 131-156, 2009.

FORNAZIER, A.; SOUZA, P. M.; PONCIANO, N. J. A importância do seguro rural na redução de riscos da agropecuária. Revista de Estudos Sociais, v. 14, n. 28, p. 39-52, 2012

GEMIGNANI, A. S. **Seguro rural**. Brasília, DF: Fundação Escola Nacional de Seguros, 2000

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

GRIFFITH, D. A. The boundary value problem in spatial statistical analysis. **Journal of regional science**, v. 23, n. 3, p. 377-387, 1983.

Guia de Seguros Rurais 2020. Disponível em: <https://www.cnabrasil.org.br/documentos-tecnicos/guia-de-seguros-rurais-2020>. Acesso em: 5 nov. 2021.

GUIMARÃES, M. F.; NOGUEIRA, J. M.. A experiência norte-americana com o seguro agrícola: lições ao Brasil?. **Revista de Economia e Sociologia Rural**, v. 47, n. 1, p. 27-58, 2009.

HAINING, R. **Spatial data analysis: theory and practice.** Cambrige: Cambridge university press, 2003.

HAIR, J. F. et al. **Análise multivariada de dados.** 6. ed. Porto Alegre: Bookman, 2009

HUNTER, J. D. **Matplotlib**: A 2D graphics environment. **Computing In Science & Engineering**, v. 9, n. 3, p. 90-95, 2007

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo Agropecuário 2017**. Rio de Janeiro, v. 8, p.1-105, 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Malha Municipal**. Disponível em: <https://ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15774-malhas.html?=&t=o-que-e> Acesso em: 18 out. 2020.

JENKS, G. Optimal data classification for choropleth maps. **Department of Geography, University of Kansas Occasional Paper**, 1977.

JORDAHL, K. **GeoPandas**: Python tools for geographic data. 2014. Disponível em: <github.com/geopandas/geopandas>, Acesso em: 28 jul. 2020.

JUPYTER. **Jupyter**: a computational environment. Disponível em: <github.com/jupyter/notebook> Acesso em: 18 jul. 2017.

KLUYVER, T. *et al.* Jupyter Notebooks: a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, p. 87–90, 2016.

MACEDO, L. O. B.; PACHECO, A. B.; DO ESPÍRITO SANTO, E. S. A evolução do Programa de Subvenção do Prêmio do Seguro Rural: uma avaliação do período 2006-10. **Indicadores Econômicos FEE**, v. 40, n. 4, 2013.

MAIA, G. B. S.; ROITMAN, F. B.; DE CONTI, B. M. Instrumentos de gestão do risco agrícola: o caso do Brasil. **Informativo Técnico SEAGRI**, Rio de Janeiro, n.1 , p. 1-16, jan. 2011.

MCKINNEY, W. Data Structures for Statistical Computing in Python. **Proceedings of the 9th Python in Science Conference**, v. 1697900, n. Scipy, p. 51–56, 2010. Disponível em: <https://conference.scipy.org/proceedings/scipy2010/mckinney.html>. Acesso em: 28 jul. 2020.

MESSNER, S. F. *et al.* The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. **Journal of Quantitative criminology**, Springer, v. 15, n. 4, p. 423-450, 1999.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte: Editora UFMG, 2005.

ORD, J. K.; GETIS, A. Local spatial autocorrelation statistics: distributional issues and an application. **Geographical analysis**, v. 27, n. 4, p. 286-306, 1995.

OZAKI, V. A. Uma digressão sobre o Programa de Subvenção ao Prêmio do Seguro Rural e as implicações para o futuro deste mercado. **Revista de Economia e Sociologia Rural**, v. 48, n. 4, p. 495-514, 2010.

OZAKI, V. A. Avanços no programa de seguro agrícola norte-americano: novos produtos, aumento da participação e dos subsídios. **Revista Brasileira de Risco e Seguro**, Rio de Janeiro, v. 2, n. 1, p. 23-48, 2006.

OZAKI, V. A. Análise espacial da produtividade agrícola no Estado do Paraná: implicações para o seguro agrícola. **Revista de Economia e Sociologia Rural**, v. 46, n. 3, p. 869-886, 2008.

PÉREZ, F.; GRANGER, B. E. IPython: a system for interactive scientific computing. **Computing in science & engineering**, v. 9, n. 3, p. 21-29, 2007.

PYTHON. **The Python programming language**. Disponível em: <github.com/python/cpython> Acesso em: 18 jul. 2017.

REY, S. J. Spatial empirics for economic growth and convergence. **Geographical Analysis**, Ohio State University Press, v. 33, n. 3, p. 196-2014, 2001.

REY, S. J.; ANSELIN, L. PySAL: A Python library of spatial analytical methods.**Review of Regional Studies**, v. 37, n. 1, p. 5–27, 2007.

ROSSETTI, L. A. Zoneamento agrícola em aplicações de crédito e securidade rural no Brasil: aspectos atuariais e de política agrícola. **Revista Brasileira de Agrometeorologia**, v. 9, n. 3, p. 386-399, 2001.

SANTOS, G. R.; SOUSA, A. G.; ALVARENGA, G. **Seguro agrícola no Brasil e o desenvolvimento do Programa de Subvenção ao Prêmio.** Brasília: Ipea, 2013. (Texto para Discussão, n. 1910). Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td_1910.pdf> Acesso em: 14 jun. 2021.

SANTOS, G. R.; SILVA, F. C. **Dez anos do Programa de Subvenção ao Prêmio de Seguro Agrícola:** proposta de índice técnico para análise do gasto público e ampliação do seguro. Rio de Janeiro: Ipea, 2017. Disponível em: <http://repositorio.ipea.gov.br/bitstream/11058/7718/1/td_2290.pdf>. Acesso em: 14 jun. 2021.

SANTOS, W. G. dos; MARTINS, J. I. F. O Zoneamento Agrícola de Risco Climático e sua contribuição à agricultura brasileira. **Revista de Política Agrícola**Revista de Política Agrícola, v. 25, n. 3, p. 73-94, dez. 2016

SILVA, J. A. da; TEIXEIRA, M. do S. G.; SANTOS, V. G. dos. Avaliação do Programa de Subvenção ao Prêmio do Seguro Rural–2005 a 2012. **Revista de Política Agrícola**, v. 23, n. 1, p. 105-118, 2014.

TYSZLER, M. **Econometria Espacial: Discutindo Medidas para a Matriz de Ponderação Espacial**. 2006. 155 p. Dissertação (Mestrado em Administração Pública e Governo) - Fundação Getúlio Vargas. São Paulo, 2006.

WALT, S. van der; COLBERT, S.; VAROQUAUX, G. The NumPy Array: A Structurefor Efficient Numerical Computation. **Computing in Science Engineering**, v. 13, n. 2, p.22–30, 2011.

WASKOM, M. *et al.* **Seaborn**: statistical data visualization. 2014. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 28 jul. 2020.

APÊNDICE A – **TABELAS**

Tabela 4 – Média das variáveis de seguro rural nos grupos formados pelos métodos de *Ward* e das $k-$médias utilizando o I de Moran

Método de agrupamento	Grupos	Variáveis						
		TAC	SIS	SPR	TSB	SIP	TMA	NAI
Ward	0	7,68	1.892.193,38	105.276,81	36.532,61	29.693,40	0,15	0,68
	1	173,36	32.377.896,41	2.152.607,72	748.847,23	556.841,71	2,31	16,93
	2	526,00	69.400.787,04	4.947.731,92	1.716.262,10	1.427.141,41	3,86	64,84
	3	344,89	124.095.719,64	11.129.426,06	3.870.037,04	1.407.619,53	2,36	20,22
	4	983,50	95.669.588,24	7.877.683,44	2.756.971,44	1.715.536,03	3,94	105,50
	5	117,67	58.773.616,53	4.980.581,99	1.655.144,56	4.453.559,68	2,28	26,00
$k-$médias	0	9,00	2.210.399,50	125.163,46	43.435,23	32.044,65	0,17	0,75
	1	192,60	53.892.341,69	4.446.303,68	1.499.628,28	4.243.046,92	3,00	61,20
	2	205,80	34.675.760,99	2.351.545,91	818.622,42	690.872,13	2,60	22,08
	3	640,93	80.999.520,67	5.967.192,65	2.070.337,33	1.313.862,44	4,26	69,60
	4	336,20	121.956.455,51	10.746.772,22	3.739.577,62	1.600.186,85	2,37	21,20

Fonte: Elaboração própria.

Tabela 5 – Média das variáveis de seguro rural nos grupos formados pelos métodos de *Ward* e das $k-$médias utilizando o G_i de Getis e Ord

Método de agrupamento	Grupos	Variáveis						
		TAC	SIS	SPR	TSB	SIP	TMA	NAI
Ward	0	2,03	575.862,61	26.371,01	9.127,21	6.397,54	0,05	0,11
	1	31,34	8.223.079,03	467.073,47	161.725,15	121.009,70	0,53	2,53
	2	131,03	22.334.516,01	1.570.846,99	545.795,72	465.570,70	1,82	14,83
	3	210,66	41.553.730,58	3.055.409,76	1.066.913,45	700.965,69	1,99	20,14
$k-$médias	0	2.86	841,211.68	38,991.52	13,541.56	10,018.64	0.06	0.17
	1	38.21	9,766,021.38	566,793.66	196,009.24	152,086.38	0.64	3.39
	2	226.97	40,993,363.61	2,967,590.76	1,037,155.87	678,955.97	2.22	21.02
	3	124.95	20,869,352.28	1,529,173.71	530,743.28	454,040.21	1.81	14.53

Fonte: Elaboração própria.

APÊNDICE B – **CÓDIGOS**

- Código correspondente ao tratamento dos dados. Disponível online em: <https://github.com/walefmachado/seguro_rural_espacial/blob/main/scripts/seguro_rural_dados.py>
- Código correspondente aos agrupamentos com o G de Getis e Ord. Disponível online em: <https://github.com/walefmachado/tcc_econ_walef/blob/main/codigos/matriz_moran_brasil.py>
- Código correspondente aos agrupamentos com o I de Moran local. Disponível online em: <https://github.com/walefmachado/tcc_econ_walef/blob/main/codigos/matriz_moran_brasil.py>