Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Análise da distribuição espacial de indicadores sociais e demográficos: uma abordagem baseada em mineração de dados

Analysis of the spatial distribution of social and demographic indicators: a data-mining based approach

Tiago Prudencio Silvano[1], Bryan Maia Correa[2] e Ivanildo Barbosa [3]

1 Instituto Militar de Engenharia, Seção de Ensino de Engenharia Cartográfica, Rio de Janeiro, Brasil. E-mail: tiagoprudencio16@gmail.com
ORCID: https://orcid.org/0000-0002-3287-5907
2 Instituto Militar de Engenharia, Seção de Ensino de Engenharia Cartográfica, Rio de Janeiro, Brasil. E-mail: bryan.maia.c7@gmail.com
ORCID: https://orcid.org/0000-0002-8146-4151
3 Instituto Militar de Engenharia, Seção de Ensino de Engenharia Cartográfica, Rio de Janeiro, Brasil. E-mail: ivanildo@ime.eb.br.
ORCID: https://orcid.org/0000-0003-4164-0875

Recebido: 10.2019 | Aceito: 01.2020

Resumo: Os dados censitários levantados periodicamente permitem retratar o perfil da população em termos étnicos, sociais e econômicos. Com isso, é possível diagnosticar dimensões do desenvolvimento humano que demandam a atuação prioritária do poder público, assim como as similaridades e heterogeneidades entre as diferentes regiões do País. Este trabalho analisa a distribuição espacial em grupos de municípios brasileiros, classificados com base na similaridade de indicadores como a porcentagem de alfabetizados por faixa etária, composição étnica e os componentes do índice de desenvolvimento humano municipal – IDHM. Além de comparar a dependência espacial dentro de cada grupo com a dependência espacial observada em todos os municípios, foram extraídos modelos baseados em árvores de decisão para indicar as regras de formação de cada grupo. Os resultados explicitam a dependência espacial existente em alguns indicadores levantados pelo censo demográfico pelo aumento do índice I de Moran nos grupos quando comparados com o conjunto original. A metodologia proposta pode ser adaptada para outras aplicações especializadas.
Palavras-chave: Desenvolvimento Humano. K-médias. Árvores de Decisão. Índice de Moran.

Abstract: Census data collected periodically depicts the ethnic, social and economic aspects of population's profile. They support the diagnosis of the dimensions of human development that demand priority actions from public agents as well as both similarities and heterogeneities between different regions all through the country. This paper analyzes the spatial distribution of groups of Brazilian municipalities, clustered according the similarities between indices such as the literacy rates by age, ethnical composition and the components of Human Development Index. After the analysis of the spatial dependency within every group, decision trees were modeled aiming at determining the rules that define the clusters. The results pointed the spatially dependent indicators computed by the demographic census when compared to the overall dataset. The proposed methodology is suitable to fit other specialized applications.
Keywords: Human Development. K-means. Decision Trees. Moran's I.

1 INTRODUÇÃO

Os dados censitários levantados periodicamente permitem retratar o perfil da população em termos étnicos, sociais e econômicos. Com isso, é possível identificar necessidades que demandam a atuação prioritária do poder público, assim como as similaridades e heterogeneidades entre as diferentes regiões do País. Januzzi (2005) aborda a escolha de indicadores sociais para uso no processo de formulação e avaliação de políticas públicas, classificação dos indicadores e importância da escolha dos indicadores corretos para cada situação.

Outro conjunto de dados relevante para avaliar a qualidade de vida da população é o que compõe o Índice de Desenvolvimento Humano (IDH), inicialmente proposto pela Organização das Nações Unidas (ONU) em contraponto ao Produto Interno Bruto (PIB) como indicador de desenvolvimento dos países. Esse indicador foi adaptado para avaliar os municípios brasileiros, surgindo o IDHM, calculado com base nos dados levantados dos censos de 1991, 2000 e 2010, variando entre 0 (menos desenvolvido) e 1 (mais desenvolvido). O IDHM é composto por indicadores do desenvolvimento humano (PNUD, 2013b):

a) Longevidade: que considera a esperança de vida ao nascer;
b) Educação: que reflete a escolaridade da população adulta e o fluxo escolar da população jovem; e
c) Renda: que considera a renda per capita da população.

A metodologia de cálculo e a consolidação dos índices referentes aos censos anteriores também podem ser encontrados em (PNUD, 2013b).

Municípios vizinhos podem apresentar um conjunto de indicadores parecidos, tendo em vista que certos fenômenos não são influenciados pelos limites administrativos impostos pelo homem. Em contrapartida, a atuação humana dentro dos limites administrativos municipais pode influenciar os indicadores de modo diferente.

O uso de mapas coropléticos auxilia na interpretação da distribuição espacial de indicadores por meio da atribuição de cores às partições do território de acordo com faixas de valores de cada um desses atributos (ou categorias). Os limites que definem cada categoria são definidos de acordo com a base de conhecimento relacionada ou com a intenção do autor em enfatizar aspectos específicos do fenômeno representado (MARTINELLI; GRAÇA, 2015; MONMONIER, 1991).

O emprego de métodos de agrupamento (ou clusterização) é uma opção para definir categorias a partir de um conjunto (simultâneo) de indicadores não correlacionado. Porém, assim como na elaboração de mapas coropléticos, é necessário definir parâmetros como a quantidade de classes, a métrica de similaridade adotada e a métrica de distância entre pares de grupos, visando à correta interpretação dos fenômenos representados. Entretanto, os grupos criados traduzem somente a minimização das distâncias dos objetos em cada grupo e a maximização das distâncias entre os grupos (GOLDSCHMIDT; PASSOS, 2005).

Isso significa que, enquanto a definição de categorias estava atrelada ao particionamento do domínio de uma variável, as categorias obtidas por clusterização dependem de um conjunto de regras baseadas no particionamento combinado de valores das variáveis empregadas. Para compreender as características inerentes a cada grupo, podem ser empregados métodos de classificação supervisionada tais como árvores de decisão e florestas aleatórias. Os modelos gerados devem ser capazes de descrever de maneira inequívoca os critérios que classificam um objeto em cada uma das categorias analisadas.

Assumindo como premissa de que todas as coisas estão relacionadas entre si, no entanto, as coisas mais próximas estão mais relacionadas que as distantes (TOBLER, 1970), foi cogitada a hipótese de que os municípios pertencentes ao mesmo cluster e, consequentemente, com maior relação de similaridade, sejam vizinhos entre si. Portanto, espera-se que a dependência espacial global seja maior dentro de cada grupo, em comparação com a dependência espacial calculada com base em todos os municípios.

Neste contexto, este trabalho tem como objetivo analisar a distribuição espacial de grupos de municípios brasileiros definidos pela similaridade de conjuntos de indicadores demográficos, sociais e econômicos. Carvalho et al. (2009) apresentaram uma metodologia para clusterização hierárquica espacial de polígonos contíguos, aplicando-a sobre a malha municipal brasileira de 2000, abordando indicadores como taxa de emprego, percentual da população em áreas urbanas, entre outras. O enfoque da abordagem era, todavia, a formação de clusters mais homogêneos, quantificados com as métricas de formação de clusters ao invés da autocorrelação espacial.

Além de avaliar a dependência espacial dos indicadores dentro de cada grupo, foram extraídos modelos baseados em árvores de decisão para descrever as características mais relevantes de cada grupo.

Cabe ressaltar que a interpretação das causas do aumento ou da redução da dependência espacial foge ao escopo deste trabalho, uma vez que demanda uma análise mais profunda dos aspectos ambientais e antrópicos que podem influenciar cada variável individualmente.

Após esta contextualização, a seção 2 apresenta sucintamente conceitos necessários ao embasamento da metodologia proposta na seção 3. A seção 4 aplica a metodologia a conjuntos de indicadores calculados com base nos levantamentos censitários disponibilizados pelo Sistema IBGE de Recuperação Automática – SIDRA e nos indicadores de Desenvolvimento Humano disponibilizado pelo Portal Atlas Brasil, apresentando e discutindo sucintamente os resultados obtidos. As conclusões obtidas a partir dos resultados obtidos estão organizadas na seção 5, com indicação de possíveis aplicações e oportunidades de melhoria a serem abordadas em trabalhos futuros.

2 REVISÃO CONCEITUAL

2.1 Descoberta de conhecimento nas bases de dados

Alguns autores consideram os termos "descoberta de conhecimento em bases de dados" e "mineração de dados" (descoberta de padrões em dados) sinônimos. Em contrapartida, Fayyad, Piatetsky-Shapiro e Smyth (1996) argumentam que a descoberta de conhecimento (*Knowledge Discovery in Databases* – KDD) se refere a todo processo de descoberta de conhecimento enquanto a mineração de dados deve ser vista como uma atividade do processo.

Para o desenvolvimento do processo KDD é necessário interpretar corretamente os dados, encontrar padrões e similaridades. Essa extração de informações não é um processo trivial devido à grande quantidade de dados normalmente envolvida, demandando a divisão do processo em fases.

Han, Kamber e Pei (2001), propõem a divisão do processo de KDD nas etapas:

a) Limpeza de dados e integração: a integração ocorre com a combinação de diferentes origens dos dados (banco de dados, planilhas, imagens, *etc.*) com objetivo de desenvolver um repositório único e consistente. Porém, essa combinação de dados é passível de inconsistências e valores conflitantes, sendo necessário filtrar, combinar e preencher valores vazios para preparar os dados para aplicação do método proposto.
b) Seleção e transformação: Conforme o objetivo proposto, os dados relevantes são identificados e reunidos, formando um subconjunto de dados que, caso necessário, deve sofrer transformações de forma apropriada para a mineração. São exemplos, a compatibilização de sistemas de referência espacial e a normalização de valores (a fim de dimensionar as variáveis em uma mesma escala).
c) Mineração de dados: o objetivo da mineração de dados é procurar padrões de interesse em um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 83). Entre as principais tarefas de mineração de dados, pode-se citar o agrupamento de dados (ou clusterização) e a classificação supervisionada, empregadas no escopo deste trabalho.
d) Avaliação e apresentação: nessa fase, exige-se a participação de especialistas no domínio considerado, visando obter a confiabilidade do modelo e indicadores para auxiliar a análise dos resultados obtidos. Na apresentação são utilizadas ferramentas gráficas (como mapas coropléticos e dasimétricos) para facilitar a visualização e análise dos resultados.

2.2 Agrupamento

Dentro do processo de mineração de dados, a tarefa de agrupamento ou clusterização (do inglês, clustering) reúne os itens de um banco de dados em grupos definidos pela afinidade entre elementos do mesmo grupo. De acordo com Ochi, Dias e Soares (2004), o problema de clusterização consiste em, dada uma base de dados, agrupar (clusterizar) seus objetos (elementos) de modo que objetos mais similares fiquem no mesmo cluster e objetos menos similares sejam alocados para clusters distintos. Braga (2005) apresenta este método como de classificação não supervisionada, de cunho descritivo, sem menção ao significado de cada classe nem à quantidade de classes esperadas.

Algoritmos de agrupamento dependem da definição de parâmetros como o número de clusters, a métrica de similaridade (entre os itens a serem agrupados) e a métrica de distância entre os grupos formados.

Carvalho et al. (2009) citam os critérios *cubic clustering criterion* (CCC), pseudo-F, pseudo-t2, R^2 e R^2 semiparcial como alternativas para a definição do número de grupos. Podem ser citados, também, a Largura de Silhueta, *Variance Ratio Criterior* (VRC) – Calinski-Harabaz, Davies-Bouldin e Índice de Dunn. Eles também citam como métricas de similaridade a distância euclidiana (norma L_2), norma L_1 (distância de Manhattan), norma L_p (caso mais geral), distância de Mahalanobis e distância euclidiana corrigida pela variância (*variance corrected*). Por fim, descrevem as métricas de distância entre grupos: ligação simples (*single linkage*), ligação completa (*complete linkage* – *unweighted* e *weighted*), associação média (*average linkage*), variância mínima de *Ward* e mediana.

2.3 Classificação

De acordo com Goldschmidt e Passos (2005, p. 13), a tarefa de classificação *"consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes".* O principal uso dessa função é avaliar novas instâncias quanto ao potencial de pertencer a uma ou outra classe. Como se baseia nas características de objetos de classes conhecidas para orientar o aprendizado, alguns autores se referem a esta tarefa como "classificação supervisionada", em contraponto à classificação não supervisionada resultante do processo de agrupamento.

Entre os métodos de classificação podem ser citados os métodos baseados em estatísticas, em árvores, em regras e em distâncias, assim como podem ser citadas as redes neurais e máquinas de vetor de suporte (CAMILO; DA SILVA, 2009).

No contexto deste trabalho, as regras obtidas serão empregadas na classificação dos grupos gerados, dado que o critério de formação dos grupos é uma combinação de atributos. Algoritmos de criação de árvores de decisão como o CART – C*lassification and Regression Trees* (KASSAMBARA, 2019a) permitem extrair regras de classificação assim como as estatísticas relacionadas à sua acurácia.

2.4 Dependência Espacial

A análise da dependência espacial tem como objetivo caracterizar a distribuição espacial de um conjunto de valores no espaço. O índice global de Moran I expressa a dependência espacial por meio da autocorrelação espacial, considerando áreas vizinhas (CÂMARA et al, 2004).

O cálculo do índice de Moran I demanda a elaboração de uma matriz de vizinhança (ou de conectividade ou de proximidade), que pode assumir diversas formas, de acordo com as regras adotadas. Podem ser citadas as regras da rainha, da torre, e do bispo, assim como da distância entre centróides e a dos vizinhos mais próximos (CÂMARA et al, 2004; SEFFRIN; DE ARAÚJO; BAZZI, 2018)

Partindo do princípio de que objetos mais próximos são semelhantes, assume-se a premissa de que objetos de um mesmo grupo (semelhantes) sejam próximos entre si. Por entender que as diferentes extensões dos municípios brasileiros influenciariam a análise visual da vizinhança, optou-se pelo cálculo dos índices de correlação espacial global observados entre os objetos dentro de cada grupo. São possíveis os cenários em que a correlação espacial de todo o conjunto seja menor do que dentro de cada grupo, ou seja, os grupos gerados incrementaram a dependência espacial, e os cenários em que o agrupamento retorna valores desfavoráveis no tocante à dependência espacial.

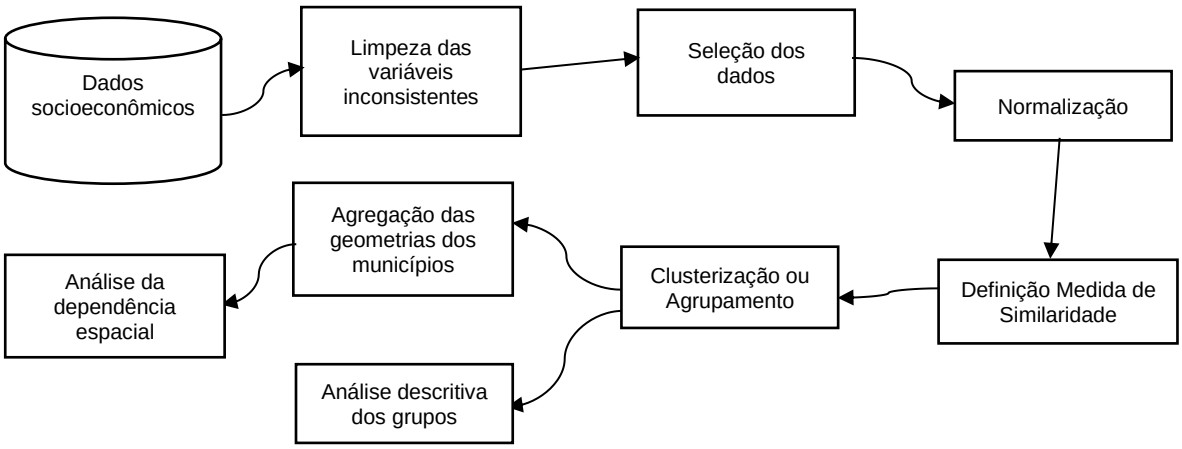
3 METODOLOGIA

A Figura 1 esquematiza as etapas de realização deste trabalho, compreendendo aquisição de dados, limpeza das variáveis inconsistentes, seleção dos dados, normalização dos dados, definição da medida de similaridade, agrupamento, associação com as geometrias dos municípios, análise da dependência espacial global e intragrupos. Faz-se necessária a etapa complementar para identificar as regras de formação do grupo, que pode depender de combinações de valores das variáveis empregadas na formação do grupo.

A aquisição de dados consiste na identificação das variáveis de interesse e na obtenção dos valores consolidados por unidade de referência (no escopo deste trabalho, os municípios). Os dados são disponibilizados em formato de planilhas (XLS), podendo ser convertidos em formato de valores separados por vírgulas (CSV), dependendo do ambiente de processamento.

As demais etapas foram processadas no software RStudio, usando a linguagem R e os conjuntos de bibliotecas que serão apresentadas conforme forem demandadas ao longo do processo.

Figura 1 – Etapas para análise dos indicadores multivariados.



Fonte: Os autores (2020).

A etapa de limpeza de variáveis inconsistentes consiste na verificação dos valores comprovadamente fora do domínio compatível com a variável considerada, além dos campos sem valor que não puderam ser recuperados.

A etapa da seleção de dados resultou na tabela que contém, para cada município, valores válidos e consistentes para processar o agrupamento. Cabe salientar que, conforme aumenta a quantidade de variáveis, os processamentos se tornam mais custosos computacionalmente pois boa parte dos procedimentos envolve comparações e medições entre pares de objetos, ou seja, possuem complexidade $O(n^2)$. No caso de o processamento envolver muitas variáveis, deve-se avaliar a realização de uma etapa extra com a finalidade de reduzir as dimensões do conjunto de dados avaliados (por exemplo, por análise de componentes principais).

A etapa de normalização se aplica quando são selecionadas variáveis cujos valores possuem ordem de grandeza tal que influenciem de forma tendenciosa o processo de medição. Goldschmidt e Passos (2005) citam, como métodos de normalização de dados que podem ser adotados, a normalização linear, por desvio-padrão, pela soma dos elementos, entre outros. A escolha do critério de normalização implica assumir que a distribuição dos valores ocorre de forma linear ou normalizada, por exemplo, o que agrega subjetividade aos valores normalizados. Portanto, é desejável a seleção de indicadores que se enquadrem em uma escala pré-determinada. Um exemplo que pode ser citado são os dados relacionados à composição da população, em que a normalização consiste na divisão de cada parcela pela população total do respectivo município.

A função *dist* do pacote *stats*, nativo do R, realiza tais cálculos, permitindo os métodos euclidiano (definido como padrão), de Manhattan, Canberra e Minkowski (R CORE TEAM, 2019). O resultado é uma matriz de distâncias que servirá de insumo para os algoritmos de agrupamento disponíveis no ambiente R.

Os parâmetros a serem definidos para a etapa de agrupamento são o número ótimo de clusters, o método de agrupamento e a métrica de distância entre grupos.

Para obter o primeiro parâmetro, emprega-se a biblioteca *factoextra* e a função *fviz_nbclust* (KASSAMBARA; MUNDT, 2019) no software R, fornecendo como entrada uma matriz de dados (variáveis), a função de particionamento (k-médias, k-medoides ou hierárquico) e um método para determinação da quantidade ótima de grupos (silhueta, cotovelo e *gap* estatístico). Neste trabalho, optou-se pelo método da largura das silhuetas, onde os valores mais próximos de 1 indicam melhores resultados (i.e., amostra pertence

ao *cluster* correto) e os valores mais próximos de -1 indicam piores resultados (i.e., amostra provavelmente pertence a outro *cluster*). Kassambara (2019b) apresenta a métrica da largura média das silhuetas dos grupos e os itens com largura de silhueta negativa como indicadores para a qualidade do agrupamento: a primeira, para a formação dos grupos, como um todo; a segunda, a fim de identificar os itens classificados erroneamente.

Quanto ao método de particionamento, optou-se pelo k-médias, com o emprego da função *hkmeans*, do pacote *factoextra* (KASSAMBARA; MUNDT, 2019), que difere da função *kmeans*, do pacote *cluster* (MAECHLER et al., 2019), por adotar uma abordagem híbrida, combinando o método k-médias e o método hierárquico. Além disso, essa função permite a especificação da métrica de distância entre grupos. Neste trabalho, optou-se pela variância mínima de *Ward*. Após a etapa de agrupamento, cada município recebe um rótulo com o número do grupo a que pertence sem qualquer significado semântico senão a similaridade dentro do grupo e a dissimiliaridade com os municípios pertencentes aos demais grupos.

A etapa seguinte consiste em mesclar a tabela de dados com o arquivo vetorial relativo à malha municipal brasileira de 2010 (IBGE, 2016), compatível com os dados censitários. A implementação desta etapa demandou a instalação do pacote *rgdal* (BIVAND; KEITT; ROWLINGSON, 2019), para importação do arquivo vetorial, e do pacote *SpatialEpi* (KIM; WAKEFIELD, 2019), que executou a junção dos arquivos referentes à malha viária e as tabelas que contém os dados censitários. Sobrepondo os *k* grupos formados sobre a malha municipal, pode-se visualizar, de forma imediata, a distribuição dos municípios entre os *k* grupos no mapa.

O principal parâmetro a ser definido para a etapa de análise de dependência espacial é o critério de formação da matriz de vizinhança. Neste trabalho, optou-se pelo critério binário implementado na função *nb2listw*, do pacote *spdep* (BIVAND; WONG, 2018), ou seja, todos os municípios que possuam qualquer interseção com o município de referência são considerados vizinhos, com mesmo peso. A função *nb2listw* se baseia em uma lista de vizinhanças criada pela função *poly2nb*, também do pacote *spdep*, ocasião em que se optou pelo critério *queen*, isto é, considera-se vizinho o polígono que possui qualquer interseção (mesmo pontual) com o polígono de referência (BIVAND, 2019b).

Inicialmente foi calculado o Índice Global de Moran a priori para todas as variáveis de cada classe antes da divisão em grupos, a fim de avaliar os eventuais incrementos nos indicadores de dependência espacial, oriundos da divisão em grupos. A segunda etapa do processo foi calcular o Índice de Moran para cada um dos grupos gerados. Foi empregada a função *moran.test* (BIVAND, 2019a), também do pacote *spdep*, usando como parâmetros os dados de cada indicador e a lista gerada pela função *nb2listw* do mesmo pacote.

A análise descritiva dos grupos é necessária pois os rótulos atribuídos aos grupos não possuem conotação semântica. Dada a lista de municípios classificados em cada grupo, cria-se uma árvore de decisão com base em uma amostra aleatória de 80% dos municípios. A função *rpart.rules*, do pacote *rpart.plot* (MILBORROW, 2019), descreve as regras de partição de uma árvore de decisão gerada pelo algoritmo CART no ambiente RStudio. Muito embora haja métodos mais sofisticados, como as florestas aleatórias, não foi identificada uma forma de descrever o modelo com base em regras, sendo empregado o algoritmo CART. Com base no modelo gerado, realiza-se a predição das classes dos 20% dos municípios restantes. A comparação entre a predição obtida e a classificação real indica a acurácia do modelo.

4 APLICAÇÃO

4.1 Cenário

Os dados utilizados neste estudo referem-se ao censo de 2010, organizados pelos municípios, identificados pelo nome e pelo código oficial do IBGE. Como a metodologia proposta aborda conjuntos de indicadores, o critério de seleção das variáveis baseou-se na capacidade de normalização objetiva (ou seja, os valores são naturalmente representados na escala entre 0 e 1) e no caráter multivariado da análise (ou seja, o mesmo tema é representado por mais de um indicador. Com essas características, foram selecionados os temas:

a) Pessoas de 5 anos ou mais de idade, alfabetizadas, por grupos de idade (IBGE, 2011): Foram selecionados dados da população residente (percentual do total geral, dividido por 100 para

constar entre 0 e 1), todas as faixas etárias (5 a 9, 10 a 14, 15 a 19, 20 a 29, 30 a 39, 40 a 49, 50 a 59 e 60 anos ou mais), dados de 2010;

b) População residente por cor ou raça (IBGE, 2012): Foram selecionados dados da população residente (percentual do total geral, dividido por 100 para constar entre 0 e 1), todas as opções de cor ou raça (branca, preta, amarela, parda, indígena ou sem declaração), dados de 2010;

c) Índice de Desenvolvimento Humano por Município (PNUD, 2019a): IDHM e suas dimensões (Educação, Renda e Longevidade), referentes a 2010. Como os valores originais já constam entre 0 e 1, não foram necessários processamentos adicionais.

Aplicando o algoritmo de definição de número ótimo de grupos, verificou-se que o número que torna máxima a largura média de silhueta foi de 2 grupos para todos os grupos de variáveis, o que sugere uma maior homogeneidade na distribuição dos indicadores analisados.

4.2 Alfabetização por faixa etária

A Tabela 1 sintetiza os resultados obtidos após a divisão em grupos das variáveis relacionadas à alfabetização por faixas etárias. O grupo 1 contém 2782 municípios enquanto o grupo 2 abrange 2783 municípios. A largura média das silhuetas foi de 0,46, o que indica que os grupos não estão muito bem definidos. Cem municípios (cerca de 1,8%) apresentaram silhueta negativa, o que indica que foram classificados equivocadamente.

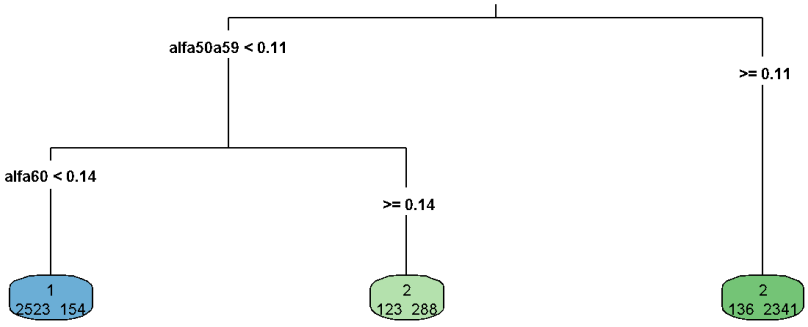
A árvore de decisão gerada com base nessas variáveis, ilustrada na Figura 2, indica um modelo capaz de descrever o grupo 1 com 92% de acurácia e outro para descrever o grupo 2, com 94% de acurácia. Mesmo que sejam avaliadas 8 variáveis, as categorias podem ser descritas por regras que consideram apenas duas variáveis, apontadas como as de maior importância.

Tabela 1– Resultados obtidos após fase de agrupamento (alfabetização por faixas etárias).

Faixa etária	**Regras grupo 1**	**Regras grupo 2**	**Moran *a priori***	**Moran grupo 1**	**Moran grupo 2**
5 a 9 anos	50 a 59 anos < 0,11 & 60+ < 0,14 50 a 59 anos >= 0,11 & 60+ < 0,11	50 a 59 anos < 0,11 & 60+ >= 0,14 50 a 59 anos >= 0,11 & 60+ >= 0,11	0,7889	0,8834	0,9436
10 a 14 anos			0,7912	0,7232	0,7355
15 a 19 anos			0,7470	0,7183	0,7579
20 a 29 anos			0,4952	0,6973	0,7173
30 a 39 anos			0,5558	0,4343	0,5023
40 a 49 anos			0,8039	0,5296	0,5352
50 a 59 anos			0,7754	0,7365	0,7837
60+ anos			0,6476	0,6991	0,7582

Fonte: Os autores (2020).

Figura 2 – Árvore de Decisão baseada nos dados do tema Alfabetização por faixa etária.



Fonte: Os autores (2020).

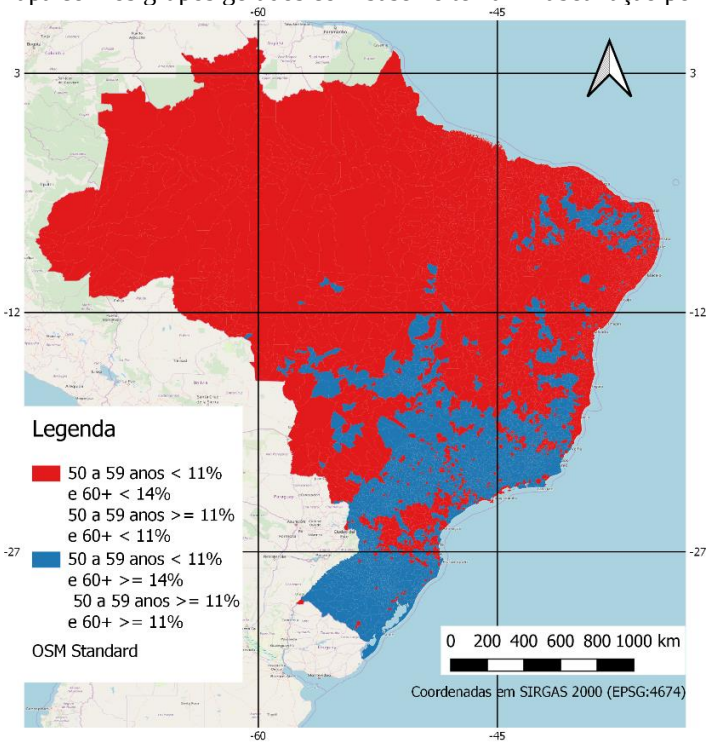
No que diz respeito à análise da dependência espacial, houve incremento nos índices dos dois grupos nas faixas etárias de 5 a 9 anos, de 20 a 29 anos e maiores de 60 anos, o que indica que a divisão adotada permite realçar a dependência espacial observada nesses indicadores. Nas faixas etárias de 15 a 19 anos e de

50 a 59 anos, houve incremento em apenas um dos grupos. Nos demais grupos, houve perda nos indicadores de dependência espacial em ambos os grupos formados, especialmente na faixa de 30 a 49 anos.

A Figura 3 ilustra a distribuição dos municípios conforme os clusters gerados, com base na malha municipal de 2010 compilada (IBGE, 2016). As diferentes extensões dos municípios podem confundir a análise visual e causar estranheza quando comparados com os índices de Moran, porém é possível identificar alguns padrões na distribuição dos municípios.

Pode-se observar que a distribuição espacial dos municípios não é homogênea (exceto na Região Norte, onde prevalecem os municípios do grupo 1). Entretanto, ainda assim é possível verificar a formação de microrregiões de destaque na sua vizinhança, como os municípios do grupo 2 identificados no interior do Nordeste do Centro-Oeste e municípios do grupo1 no interior da Região Sul.

Figura 3 – Mapa com os grupos gerados com base no tema Alfabetização por faixa etária.



Fonte: Os autores (2020).

4.3 Cor ou Raça

A Tabela 2 sintetiza os resultados obtidos após a divisão em grupos das variáveis relacionadas à composição étnica da população. O grupo 1 contém 3294 municípios enquanto o grupo 2 abrange 2271 municípios. A largura média das silhuetas foi de 0,62, o que indica que os grupos estão bem definidos. Apenas 0,18% dos municípios apresentaram silhueta negativa, o que indica que foram classificados equivocadamente.

Tabela 2 – Resultados obtidos após fase de agrupamento (composição étnica).

Indicador	**Regras grupo 1**	**Regras grupo 2**	**Moran *a priori***	**Moran grupo 1**	**Moran grupo 2**
Branca	Branca < 0,51	Branca >= 0,51	0,91980	0,85865	0,95265
Preta			0,70068	0,81797	0,89660
Amarela			0,35751	0,58237	0,68655
Parda			0,90565	0,30557	0,35899
Indígena			0,27531	0,80667	0,89182

Fonte: Os autores (2020).

A árvore de decisão gerada com base nessas variáveis, ilustradas na Figura 4, indica um modelo para descrever o grupo 1 com 98,6% de acurácia e outro para descrever o grupo 2, com 97,3% de acurácia. Mesmo que sejam avaliadas 5 variáveis, as categorias podem ser descritas por regras que consideram apenas uma variável, apontadas como a de maior importância.

Figura 4 – Árvore de Decisão baseada nos dados do tema Cor / Raça.

Fonte: Os autores (2020).

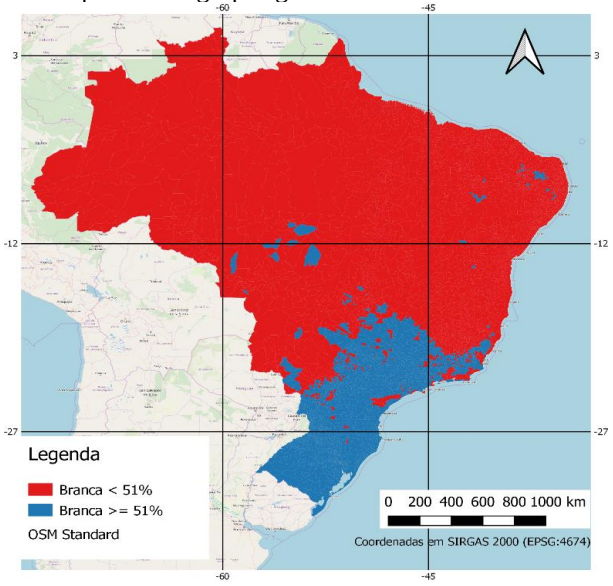
No que diz respeito à análise da dependência espacial, houve incremento nos índices dos dois grupos nas porcentagens de pretos, amarelos e indígenas, demonstrando a correlação entre as regras adotadas e a dependência espacial inerente a esses indicadores. Quanto à proporção de brancos, houve incremento em apenas um dos grupos, enquanto no tocante à proporção de pardos, houve perda nos indicadores de dependência espacial em ambos os grupos formados.

A Figura 5 ilustra a distribuição dos municípios conforme os clusters gerados. Observa-se a homogeneidade da distribuição dos municípios dentro de cada grupo, especialmente nas Regiões Norte, onde predomina o grupo 1, e na Região Sul, onde predomina o grupo 2.

4.4 Índice de Desenvolvimento Humano Municipal

A Tabela 3 sintetiza os resultados obtidos após a divisão em grupos das variáveis relacionadas ao IDHM calculado com base nos dados de 2010 e seus componentes. O grupo 1 contém 3012 municípios enquanto o grupo 2 abrange 2553 municípios. A largura média das silhuetas foi de 0,54, o que indica que os grupos estão bem definidos. Vinte e quatro municípios (cerca de 0,45%) apresentaram silhueta negativa, o que indica que foram classificados equivocadamente.

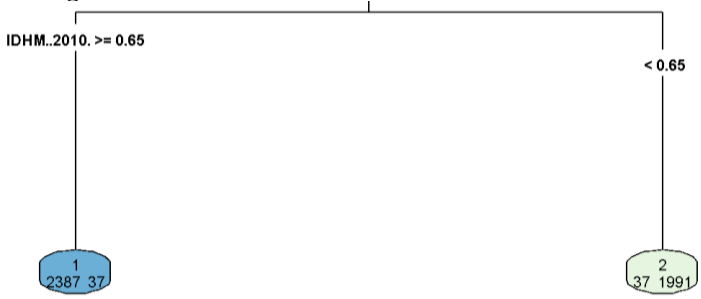
Figura 5 – Mapa com os grupos gerados com base no tema Cor / Raça.



Fonte: Os autores (2020).

A árvore de decisão gerada com base nessas variáveis, ilustrada na Figura 6, indica um modelo para descrever o grupo 1 com 99,8% de acurácia e outro para descrever o grupo 2, com 99,4% de acurácia. O valor que divide as categorias pode ser comparado com os valores de referência estabelecidos pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) em que os municípios englobados no cluster 1 apresentam valores de IDHM alto ou muito alto, enquanto os municípios englobados no cluster 2 apresentam valores médio, baixo e muito baixo (PNUD, 2013b).

Figura 6 – Árvore de Decisão baseada nos dados de IDHM.



Fonte: Os autores (2020).

No que diz respeito à análise da dependência espacial, houve incremento nos valores referentes ao IDHM, e aos componentes Longevidade e Educação, demonstrando a correlação entre a regra adotada na classificação e a dependência espacial inerente a esses indicadores. Quanto ao valor referente ao componente Renda, houve perda, mesmo que sutil, nos indicadores de dependência espacial em ambos os grupos formados. Isso sugere que, em geral, o desenvolvimento humano não está sujeito às delimitações administrativas municipais, mas desenvolve-se regionalmente.

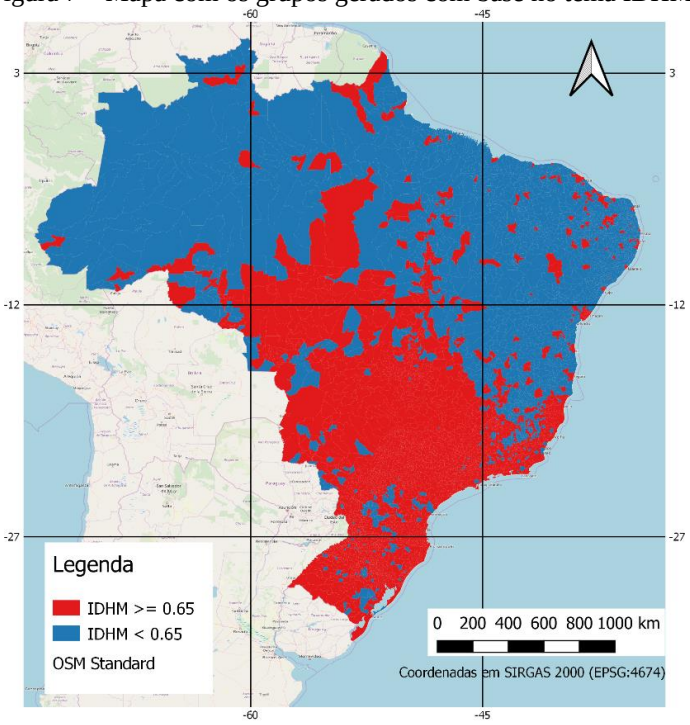
A Figura 7 ilustra a distribuição dos municípios conforme os clusters gerados. Pode-se observar que a distribuição espacial dos municípios não é homogênea, com muitos municípios que destoam de sua vizinhança. Entretanto, ainda assim é possível verificar a concentração de municípios do grupo 1 entre o litoral de São Paulo, sul de Goiás e Mato Grosso do Sul, e do grupo 2 nos estados do Acre, Amazonas, norte do Pará e Maranhão.

Tabela 3 – Resultados obtidos após fase de agrupamento (IDHM).

Indicador	**Regras formadoras do grupo 1**	**Regras formadoras do grupo 2**	**Moran *a priori***	**Moran grupo 1**	**Moran grupo 2**
IDHM	IDHM >= 0.65	IDHM < 0.65	0,79354	0,93936	0,90899
Renda			0,80574	0,73024	0,78908
Longevidade			0,74035	0,75289	0,79891
Educação			0,70214	0,70246	0,71419

Fonte: Os autores (2020).

Figura 7 – Mapa com os grupos gerados com base no tema IDHM.



Fonte: Os autores (2020)

5 CONCLUSÕES

O objetivo proposto para este trabalho foi analisar a distribuição espacial de grupos de municípios brasileiros definidos pela similaridade de indicadores demográficos, sociais e econômicos. Além de avaliar a dependência espacial dos indicadores dentro de cada grupo, foram extraídos modelos baseados em árvores de decisão para descrever as características mais relevantes de cada grupo.

A metodologia proposta compreende a aquisição de dados, limpeza das variáveis inconsistentes, seleção dos dados, normalização dos dados, definição da medida de similaridade, agrupamento, associação com as geometrias dos municípios, análise da dependência espacial global e intragrupos. Após a divisão em grupos, faz-se necessária a etapa complementar para identificar as regras de formação do grupo, que pode depender de combinações de valores das variáveis empregadas na formação do grupo.

A metodologia proposta foi aplicada a conjuntos de dados relacionados à proporção de população alfabetizada por faixa etária, composição étnica da população, IDHM e seus componentes.

O primeiro resultado apresentado no trabalho foi a possibilidade de elaborar mapas com os municípios brasileiros classificados com base nas similaridades entre os conjuntos de indicadores (análise multivariada).

O segundo resultado apresentado foi a homogeneidade na distribuição dos indicadores, caracterizada pela divisão dos municípios em apenas dois grupos. Cabe relembrar que a última etapa do processo de descoberta de conhecimento em bancos de dados (Seção 2.1) exige a participação de especialistas no domínio considerado, visando obter a confiabilidade do modelo e indicadores para auxiliar a análise dos resultados obtidos. Consequentemente, os padrões identificados ao longo deste trabalho podem fomentar discussões especializadas quanto à distribuição desses padrões e de outros potencialmente identificáveis com o emprego da metodologia proposta.

A terceira conclusão baseada no que foi abordado no texto foi que, apesar de não serem consideradas as vizinhanças geográficas na formação dos grupos, a ocorrência de extensas regiões formadas por municípios contíguos indica a continuidade da distribuição dos valores de alguns indicadores (ver Figuras 3, 5 e 7). Contudo, é possível identificar municípios e regiões que se destacam dos municípios no seu entorno, o que impacta os índices de Moran calculados. Esse fato está relacionado à hipótese proposta, indicando a aderência de alguns indicadores à Lei de Tobler, caracterizada pelo incremento nos valores da autocorrelação espacial entre os municípios pertencentes ao mesmo grupo. Das variáveis analisadas, cabe destacar os incrementos dos

índices de Moran referentes às variáveis *Amarela* e *Indígena*, ao mesmo tempo em que se destacam os decréscimos observados nos índices de Moran referentes à variável *Parda*.

A metodologia proposta pode ser empregada para avaliar outros conjuntos de dados que atendam às especificações adotadas na Seção 3. A limitação do número de variáveis está associada ao método de agrupamento, demandando a substituição do método ou a redução das dimensões, se necessário.

Outras discussões que podem ser avaliadas futuramente são sobre alternativas aos parâmetros empregados na formação dos grupos (número de grupos, métodos de clusterização e métricas de similaridade), na avaliação da dependência espacial (critérios de vizinhança e o emprego do Indicador Local de Associação Espacial – LISA, para quantificar a dependência espacial nos grupos), e o particionamento das análises por Regiões e Unidades da Federação.

Contribuição dos autores

Silvano, T. P. e Correa, B. M. contribuíram coletando e organizando a base de dados (curadoria dos dados), calculando os valores de dependência espacial e redigindo a minuta do artigo. Barbosa, I. participou na conceptualização, na supervisão e no desenvolvimento da versão preliminar dos códigos relacionados à clusterização e à classificação (*software*). Todos contribuíram conjuntamente na análise formal, no aperfeiçoamento dos códigos (*software*) e da redação.

Conflitos de Interesse

Os autores declaram que não há conflitos de interesse.

Referências

BIVAND, R.; **moran.test**. Disponível em: <https://www.rdocumentation.org/packages/spdep/versions/1.1-3/topics/moran.test>. Acesso em: 9 set. de 2019a.

BIVAND, R.; **poly2nb**. Disponível em: <https://www.rdocumentation.org/packages/spdep/versions/1.1-2/topics/poly2nb>. Acesso em: 9 set. de 2019b.

BIVAND R; WONG D. W. S. Comparing implementations of global and local indicators of spatial association. **TEST**, v. 27, n. 3, p. 716–748, Sep 2018. DOI. 10.1007/s11749-018-0599-x

BIVAND, R.; KEITT, T.; ROWLINGSON, B. **rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.3-9.** Disponível em: <https://CRAN.R-project.org/package=rgdal>. Acesso em: 9 set. de 2019.

BRAGA, L. P. V. **Introdução à Mineração de Dados**. Editora E-papers, 2005.

CÂMARA, G.; CARVALHO, M.S.; GONÇALVES CRUZ, O.; CORREA, V. Análise Espacial de Áreas. In: DRUCK, S. (ed.); CARVALHO, M. S. (ed.); CÂMARA, G. (ed.); MONTEIRO, A.V.M. (ed.). **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004. pp. 1-44.

CAMILO, C. O.; DA SILVA, J. C. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**: Relatório Técnico RT-INF_001-09, Goiânia: Universidade Federal de Goiás, 2009.

CARVALHO, A. X. Y.; ALBUQUERQUE, P. H. M.; ALMEIDA JUNIOR, G. R.; GUIMARÃES, R. D. **Clusterização Hierárquica Espacial**: Texto para Discussão n° 1427. Brasília, 2009. Disponível em: <http://www. cnps.embrapa.br/search/unids/rtec97/rtec97.html>. Acesso em: 22 ago. 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining To Knowledge Discovery In Databases. **AI Magazine**. v. 17, n. 03, p. 37-54, Fall 1996. DOI. 10.1609/aimag.v17i3.1230.

GOLDSCHMIDT, R.; PASSOS, E.; **Data mining:** um guia prático. Gulf Professional Publishing, 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. São Francisco: Morgan Kaufmann, 2001.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Malhas Municipais – Brasil**. Rio de Janeiro, 2016. Disponível em:

<ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2010>. Acesso em: 9 set. de 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Tabela 2093 - População residente por cor ou raça, sexo, situação do domicílio e grupos de idade - Amostra - Características Gerais da População.** Rio de Janeiro, 2012. Disponível em: <https://sidra.ibge.gov.br/tabela/2093>. Acesso em: 9 set. de 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Tabela 3150 - Pessoas de 5 anos ou mais de idade, total e as alfabetizadas, por grupos de idade - Resultados Preliminares do Universo.** Rio de Janeiro, 2011. Disponível em: <https://sidra.ibge.gov.br/tabela/3150>. Acesso em: 9 set. de 2019.

JANNUZZI, P. M. Indicadores para diagnóstico, monitoramento e avaliação de programas sociais no Brasil. **Revista Do Serviço Público**. Brasília, v. 56, n. 2, p. 137-160, 2005. DOI. 10.21874/rsp.v56i2.222

KASSAMBARA, A. **CART Model: Decision Tree Essentials**. Disponível em: <http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials>. Acesso em: 9 set. de 2019a.

KASSAMBARA, A. **Visualize Silhouette Information from Clustering.** Disponível em: <https://rpkgs.datanovia.com/factoextra/reference/fviz_silhouette.html#examples>. Acesso em: 9 set. de 2019b.

KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5.** Disponível em: <https://CRAN.R-project.org/package=factoextra>. Acesso em: 9 set. de 2019.

KIM, A. Y.; WAKEFIELD, J. **SpatialEpi: Methods and Data for Spatial Epidemiology.** R package version 1.2.3. Disponível em: <https://CRAN.R-project.org/package=SpatialEpi>. Acesso em: 9 set. de 2019.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. **cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.** Disponível em: <https://CRAN.R-project.org/package=cluster>. Acesso em: 9 set. de 2019.

MARTINELLI, M.; GRAÇA, A. J. S. Cartografia Temática: Uma Breve História Repleta de Inovações. **Revista Brasileira de Cartografia,** v. 67, n. 4, p. 913-928, Jul 2015.

MILBORROW, S. **Package 'rpart.plot', Version 3.0.8**, Disponível em: <https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf>. Acesso em: 9 set. de 2019.

MONMONIER, M. **How to lie with maps**, Chicago: University of Chicago Press, 1996.

OCHI L. S.; DIAS C. R.; SOARES S. S. F. **Clusterização em Mineração de Dados.** Disponível em: <http://www2.ic.uff.br/~satoru/conteudo/artigos/ERI-Minicurso-SATORU.pdf>. Acesso em: 19 ago. 2019.

PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO (PNUD). **Consulta**. Brasília, 2013. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/consulta>. Acesso em: 2 dez. de 2019a.

PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO (PNUD). **O IDHM**. Brasília, 2013. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/>. Acesso em: 2 dez. de 2019b.

SEFFRIN, R.; DE ARAÚJO, E. C.; BAZZI C. L. Análise espacial de área aplicada a produtividade de soja na região oeste do Paraná utilizando o software R. **Revista Brasileira de Geomática**, v. 6, n. 1, p. 23-43, 2018. DOI: 10.3895/rbgeo.v6n1.5912.

R CORE TEAM, **Distance Matrix Computation**. Disponível em: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html>. Acesso em: 9 set. de 2019.

TOBLER, W. R. "A Computer Movie Simulating Urban Growth in the Detroit Region." **Economic Geography**. v. 46, n.1, p. 234-240, Jun 1970. DOI. 10.2307/143141.

Biografia do autor principal

Tiago Prudencio Silvano, nasceu em 1989 na cidade do Rio de Janeiro-RJ. Bacharel em Ciências Militares pela Academia Militar das Agulhas Negras, atualmente é aluno do Curso de Graduação em Engenharia Cartográfica do Instituto Militar de Engenharia.