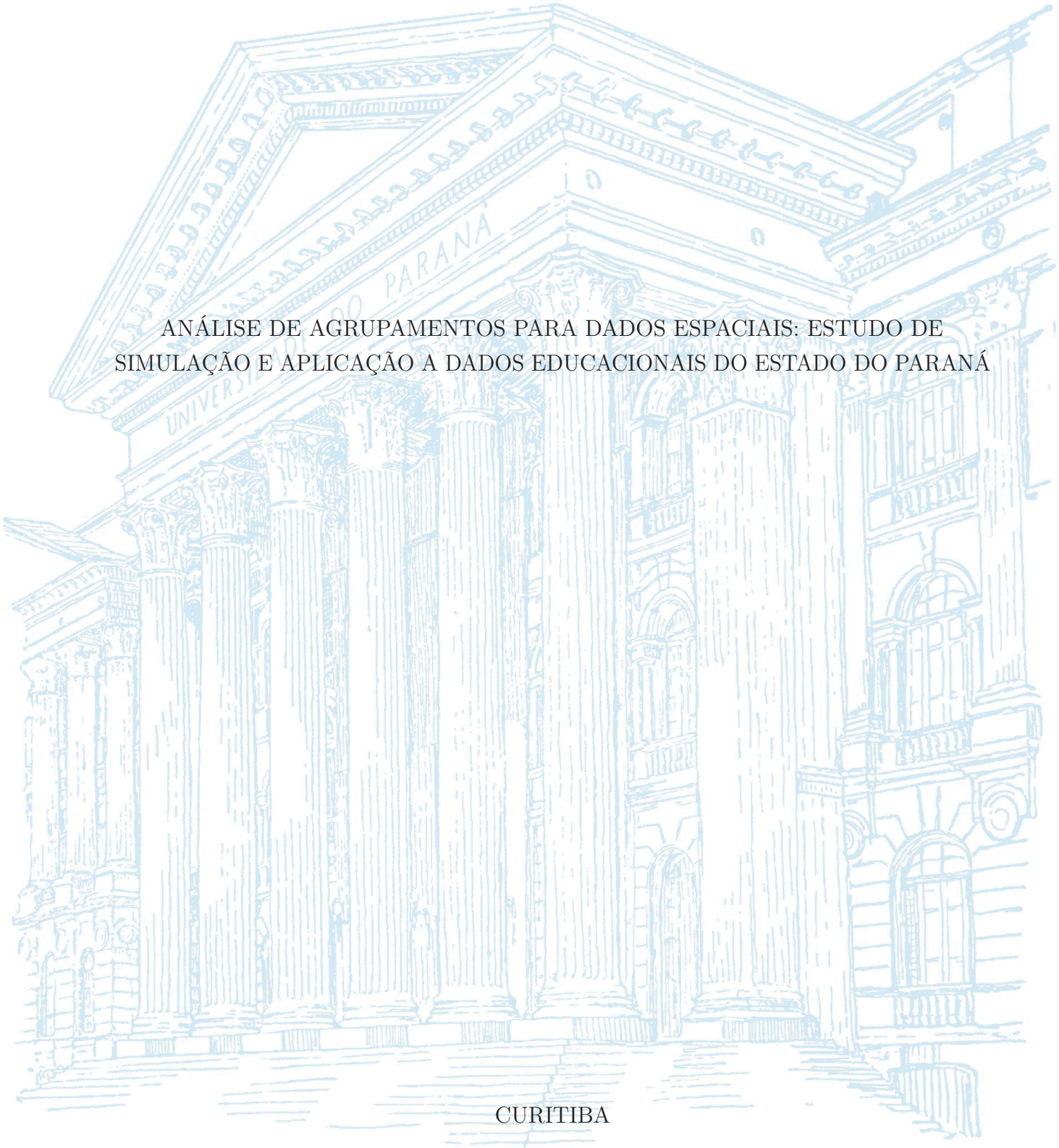
UNIVERSIDADE FEDERAL DO PARANÁ

DAIANE CHITKO DE SOUZA



ANÁLISE DE AGRUPAMENTOS PARA DADOS ESPACIAIS: ESTUDO DE SIMULAÇÃO E APLICAÇÃO A DADOS EDUCACIONAIS DO ESTADO DO PARANÁ

CURITIBA

2021

DAIANE CHITKO DE SOUZA

ANÁLISE DE AGRUPAMENTOS PARA DADOS ESPACIAIS: ESTUDO DE SIMULAÇÃO E APLICAÇÃO A DADOS EDUCACIONAIS DO ESTADO DO PARANÁ

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciências do Curso de Pós-Graduação em Métodos Numéricos em Engenharia na área de concentração em Programação Matemática do Setor de Tecnologia, Departamento de Construção Civil e Setor de Ciências Exatas, Departamento de Matemática da Universidade Federal do Paraná..

Orientador: Prof. Dr. Cesar Augusto Taconeli

CURITIBA

2021

CATALOGAÇÃO NA FONTE – SIBI/UFPR

S729a Souza, Daiane Chitko de

Análise de agrupamentos para dados espaciais: estudo de simulação e aplicação a dados educacionais do estado do Paraná [recurso eletrônico]/ Daiane Chitko de Souza - Curitiba, 2021.

Dissertação (Mestrado) apresentada ao Programa de Pós-graduação em Métodos Numéricos em Engenharia na área de concentração em Programação Matemática do Setor de Tecnologia, Departamento de Construção Civil e Setor de Ciências Exatas, Departamento de Matemática da Universidade Federal do Paraná.
Orientador: Prof. Dr. Cesar Augusto Taconeli

1. Análise matemática. 2. Métodos numéricos. I. Taconeli, Cesar Augusto. II. Título. III. Universidade Federal do Paraná.

CDD 515

Bibliotecária: Vilma Machado CRB9/1563

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **DAIANE CHITKO DE SOUZA** intitulada: **Análise de agrupamentos para dados espaciais: estudo de simulação e aplicação a dados educacionais do Estado do Paraná**, sob orientação do Prof. Dr. CESAR AUGUSTO TACONELI, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 24 de Fevereiro de 2021.

Assinatura Eletrônica
25/02/2021 18:06:59.0
CESAR AUGUSTO TACONELI
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
25/02/2021 10:17:17.0
PAULO JUSTINIANO RIBEIRO JUNIOR
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
25/02/2021 17:38:22.0
IDEMAURO ANTONIO RODRIGUES DE LARA
Avaliador Externo (UNIVERSIDADE DE SÃO PAULO)

AGRADECIMENTOS

Os meus sinceros agradecimentos ao meu orientador professor Dr. Cesar Augusto Taconeli, pelo apoio, confiança e paciência, depositados em mim no decorrer deste trabalho.

À minha família e amigos, pelo apoio incondicional nesta jornada. Obrigada por entenderem minha ausência nos momentos mais preciosos.

Aos meus amigos do programa, por todas as suas contribuições acadêmicas e pessoais e pelos momentos de descontração e amizade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio neste estudo.

A todos que direta ou indiretamente contribuíram, muito obrigada.

A árvore não prova a doçura dos próprios frutos;
o rio não bebe suas próprias ondas;
as nuvens não despejam água sobre si mesmas.
A força dos bons deve ser usada para benefício de todos.
Provérbio Hindu

RESUMO

Este trabalho apresenta dois estudos sob a perspectiva da análise de agrupamentos envolvendo dados espaciais. O primeiro é um estudo de caso com dados educacionais do Estado do Paraná que tem como objetivo identificar padrões baseados em indicadores educacionais e também de constituir grupos de municípios homogêneos quantos aos padrões detectados. O segundo é um estudo por simulação com o objetivo de avaliar a reprodutibilidade de métodos de agrupamentos espaciais e não espaciais de uma partição de grupos predefinida. Foram considerados métodos de agrupamentos não espaciais (K-means, Complete linkage e Ward) e métodos de agrupamentos espaciais (ClustGeo e Skater) que foram comparados, no estudo de caso, por meio de três índices de validação interna (largura média da silhueta, índice de Dunn e índice de conectividade) e, no estudo por simulação, usando dois índices de validação externa (índice de Rand ajustado e variação da informação). Os dois estudos permitiram concluir em um favorecimento aos métodos espaciais como os mais adequados para analisar tanto os indicadores educacionais quanto os dados simulados, em termos da reprodutibilidade da partição predefinida (para o estudo por simulação) e do favorecimento em identificar os padrões educacionais (para o estudo de caso).

Palavras-chaves: Validação de agrupamentos. ClustGeo. Variação da informação. Índice de Rand ajustado. k-means.

ABSTRACT

This work presents two studies from the perspective of cluster analysis involving spatial data.The first is a case study with educational data from the State of Paraná that aims to identify clusters based on educational indicators and to constitute groups of homogeneous municipalities. The second is a simulation study to evaluate the reproducibility of spatial and non-spatial clustering methods of a predefined group partition. Non-spatial clustering methods (K-means, Complete linkage and Ward) and spatial clustering methods (ClustGeo and Skater) were considered, which were compared, in the case study, using three internal validation indexes (average silhouette width, Dunn index and Connectivity index) and, in the simulation study, using two external validation indexes (adjusted Rand index and variation of information). The two studies made it possible to conclude in favor of spatial methods as the most suitable for analyzing both educational indicators and simulated data, in terms of the reproducibility of the predefined partition (in the simulation study) and identifying educational clusterings (in the case study).

Key-words:

Cluster validation. ClustGeo. Variation of information. Adjusted Rand index. K-means.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Analisar dados sempre foi uma das tarefas utilizadas para o desenvolvimento da sociedade. Nos dias atuais, com o avanço da tecnologia e da internet, conseguimos registrar e armazenar mais dados de diferentes áreas do que no passado. Frente à grande quantidade de dados gerados, técnicas de análise aplicadas à extração de informações é de suma importância. Uma das técnicas utilizadas para classificar informações contidas nos dados, e consequentemente analisar essas informações, é a análise de agrupamentos. A análise de agrupamentos é uma técnica de aprendizado não supervisionado. Diferentemente das técnicas de aprendizado supervisionado, no qual já se sabe as informações contidas nos dados podendo assim prever as respostas, na aprendizagem não supervisionada não se têm informações prévias sobre o comportamento ou a resposta contida nos dados.

A análise de agrupamentos tem como objetivo agrupar indivíduos com características semelhantes dentro de grupos (ZHU, 2019). Essa técnica é amplamente utilizada em várias áreas, como na agricultura (TIWARI; MISRA, 2011), criminalidade (AGARWAL; NAGPAL; SEHGAL, 2013), saúde (SIDDIQUI et al., 2020), marketing (PUNJ; STEWART, 1983), entre outras. Por esse motivo, existem diferentes métodos que podem ser aplicados na obtenção dos grupos em uma análise de agrupamentos. Os métodos mais utilizados são os de agrupamentos hierárquicos e não hierárquicos (JAIN, 2010), que são métodos não espaciais. Métodos de agrupamentos que levam em consideração a espacialização dos dados são denominados métodos de agrupamentos espaciais e dados com espacialização recebem o nome de dados espaciais.

Métodos de agrupamentos que levam em conta a espacialização dos dados, vêm apresentando crescente interesse e numerosas aplicações (DISTEFANO; MAMELI; POLI, 2020). Dados espaciais são caracterizados por atributos não espaciais (como variáveis sociais e econômicas), bem como por atributos espaciais (como coordenadas e formas) (DISTEFANO; MAMELI; POLI, 2020). Os métodos de agrupamentos espaciais permitem incorporar a espacialização dos dados e demais atributos para fornecer respostas com informações mais coerentes a partir dos dados utilizados.

Como mencionado anteriormente, existem muitos métodos de análise de agrupamentos e cada método possui características que podem refletir de maneira adequada as informações contidas em diferentes tipos de dados. Uma das questões da análise de agrupamentos é qual método, dentre os existentes, é o mais apropriado para os dados disponíveis (GRUBESIC; WEI; MURRAY, 2014). Adicionalmente, conclusões sobre os agrupamentos apenas com base nos grupos fornecidos por um desses métodos pode não ser eficiente, de maneira que, para avaliar a performance, diferentes índices de validação podem ser utilizados.

Para poder comparar os métodos de agrupamentos existem índices de validação interna (usados para medir a qualidade de uma estrutura de agrupamentos internamente, ou seja, se os grupos formados constituem de boa formação em relação a homogeneidade e separação dos dados) e externa (usados para comparar os grupos constituídos a partir de métodos de agrupamentos com alguma configuração previamente estabelecida de grupos) (HASSANI; SEIDL, 2017). Nesse contexto, avaliar e medir a qualidade das soluções dos métodos de agrupamentos tem-se mostrado tão importante quanto o próprio algoritmo de agrupamento (HASSANI; SEIDL, 2017), (HOEF; WARRENS, 2019).

1.1 OBJETIVOS

Considerando os aspectos mencionados, o objetivo principal do presente trabalho é comparar os métodos de análise de agrupamentos na presença de dados espaciais. Este trabalho é composto por dois estudos: uma aplicação a dados educacionais do Estado do Paraná e um estudo por simulação envolvendo dados espaciais.

O uso da análise de agrupamento em dados educacionais têm o objetivo de analisar padrões de conhecimento, índices de reprovação ou aprovação, interações aluno/universidade, dentre outros (DUTT et al., 2015). Um campo ainda pouco explorado é o de indicadores educacionais no contexto da análise de agrupamentos, levando em consideração a espacialização dos dados (VERICA; VILLWOCK; JOHANN, 2015) e (WEBBER; ZAT; PRADO et al., 2013), que é um dos focos deste trabalho. Nem todos os dados educacionais são dados espaciais e a correta identificação do tipo de dado também interfere na eficácia dos métodos de agrupamentos. Logo, nesse contexto, a aplicação e comparação dos resultados produzidos por diferentes métodos de agrupamentos podem viabilizar a identificação de padrões educacionais no Estado.

Para fomentar a compreensão dos métodos de agrupamentos espaciais e não espaciais apresentados neste trabalho, foi realizado um estudo por simulação que tem como intuito investigar a performance desses métodos quando aplicados a dados espaciais, além de comparar e avaliar os mesmos. A simulação de dados é amplamente usada para estudar e analisar métodos. Como exemplos de uso de simulação para tais fins estão o trabalho de Hassani e Seidl (2017) e Distefano, Mameli e Poli (2020).

Os objetivos específicos do estudo de caso são: de contribuir para a obtenção de um panorama da educação no Estado do Paraná, identificar as similaridades entres os municípios e analisar como os métodos de agrupamentos se comportam com esse tipo de dados. Para o estudo por simulação, como os dados espaciais foram simulados com diferentes parâmetros, os objetivos específicos são tanto de avaliar as performances dos métodos de agrupamentos espaciais e não espaciais usando critérios de validação externa, quanto o impacto dos diferentes parâmetros simulados para os mesmos.

1.2 ESTRUTURA DO TRABALHO

No presente capítulo foi realizada a apresentação da análise de agrupamentos para dados espaciais, definindo-se os objetivos e justificativas deste trabalho. No Capítulo 2, os conceitos sobre os métodos utilizados ao longo do trabalho e os índices de validação são apresentados. O Capítulo 3 é dedicado ao estudo envolvendo os dados educacionais do Estado do Paraná e no Capítulo 4 o estudo por simulação, com as respectivas discussões dos resultados. Finalmente, no Capítulo 5 são apresentadas as conclusões obtidas.

2 CONCEITOS PRELIMINARES

2.1 ANÁLISE DE AGRUPAMENTOS

A análise de agrupamentos tem como objetivo dividir os indivíduos em grupos, de forma que as observações dentro de um mesmo grupo sejam similares entre si, e observações em grupos diferentes sejam dissimilares em relação às mesmas características (MINGOTI, 2005). Nesse sentido, este capítulo apresenta algumas noções sobre cinco métodos de agrupamentos, servindo como base conceitual para o desenvolvimento dos capítulos posteriores.

Primeiramente, define-se o conceito de dissimilaridade, o qual fundamenta os métodos de agrupamentos. Em seguida são apresentados os métodos não espaciais (K-means, Complete linkage e Ward) e os métodos espaciais (ClustGeo e Skater). A Seção seguinte conta com a apresentação de três índices de validação interna (largura média da silhueta, índice de Dunn e índice de conectividade) usados no estudo de caso e de dois índices de validação externa, usados no estudo de simulação (índice de Rand ajustado e variação da informação). Finaliza-se com o índice de Moran usado para analisar a autocorrelação espacial dos dados e as informações do software usado neste trabalho.

2.2 MEDIDAS DE DISSIMILARIDADE

Medidas de dissimilaridade permitem quantificar as diferenças entre pares de indivíduos com base nos valores avaliados para um conjunto de variáveis (JOHNSON; WICHERN, 2002). Essas medidas fundamentam métodos de agrupamentos baseados em distância para agrupar os indivíduos em grupos.

Uma medida de dissimilaridade é tal que, quanto maior seu valor para um par de indivíduos, menos parecidas (mais dissimilares) são estes indivíduos. Existem várias medidas de dissimilaridades como, por exemplo, distância de Mahalanobis, distância Euclidiana, distância de Minkowsky, entre outras (SHIRKHORSHIDI; AGHABOZORGI; WAH, 2015). Realizada as devidas verificações, a distância Euclidiana foi escolhida e usada para os métodos apresentados neste trabalho.

Considere $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ e $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ dois indivíduos descritos por p atributos numéricos. A distância Euclidiana para esse par de indivíduos, i e j, denotada por $d(i, j)$, fica definida por:

$$d(i,j) = \sqrt{\sum_{i=1}^{p} (x_{ip} - x_{jp})^2} \tag{2.1}$$

Para cada par de indivíduos temos um valor para a medida de dissimilaridade (em particular, para a distância Euclidiana) e a partir disso pode-se construir uma matriz de dissimilaridades. A matriz de dissimilaridades (2.2) é definida como a matriz simétrica de distâncias nxn, $\boldsymbol{D} = [d_{ij}]_{\text{nxn}}$, onde d_{ij} é dissimilaridade entre o indivíduo i e o indivíduo j.

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{bmatrix} \tag{2.2}$$

2.3 MÉTODOS DE AGRUPAMENTOS

Os métodos de agrupamentos usados neste trabalho estão divididos em métodos espaciais e não espaciais. Dentre os métodos de agrupamentos não espaciais foram considerados no presente trabalho o método não hierárquico K-means e dois métodos hierárquicos, complete linkage e Ward, usando como referência o trabalho dos autores Johnson e Wichern (2002). Os métodos de agrupamentos espaciais utilizados foram o método ClustGeo, usando como referências Chavent et al. (2018) e Distefano, Mameli e Poli (2020), e o algoritmo Skater, usando o trabalho de Assunção et al. (2006) para embasamento teórico. Esses cinco métodos foram escolhidos por terem características de agrupamentos sem espacialização (métodos não espaciais), com espacialização não fortemente restritiva (ClustGeo) e com espacialização mais restritiva (Skater).

2.3.1 Métodos de agrupamentos não espaciais

Os métodos de agrupamentos não espaciais podem ser divididos em métodos hierárquicos e não hierárquicos. Os métodos de agrupamentos não hierárquicos objetivam encontrar uma partição de n indivíduos em k grupos, tal que o número de grupos deve ser definido a priori (MINGOTI, 2005). Os métodos de agrupamentos hierárquicos, respeitam uma hierarquia na formação dos grupos e não é necessária a definição inicial do número de grupos. Os métodos hierárquicos são classificados em aglomerativos (os grupos são criados a partir dos indivíduos isolados) e divisivos (inicialmente, todos os indivíduos fazem parte de um único grupo e posteriormente são divididos).

Alguns exemplos de métodos de agrupamentos hierárquico são: single linkage, average linkage, complete linkage e Ward. Usamos os métodos de agrupamentos aglomerativos, especificamente o método Ward e o complete linkage, escolhidos por preferencia pessoal. Entre os métodos não hierárquicos estão os métodos K-means, Fuzzy e K-medoids, sendo que usamos o método K-means neste trabalho, por ser o mais utilizado e estudado.

2.3.1.1 Método K-means

O algoritmo K-means aloca cada indivíduo ao grupo cujo centroide (vetor de médias do grupo) tem a menor distância (MINGOTI, 2005). O método é composto por quatro passos:

1. Define-se o número grupos (k) e os centroides (que são escolhidos aleatoriamente dentre os indivíduos e correspondem ao mesmo número de grupos).
2. Calcula-se a distância de cada indivíduo ao centróide de cada grupo, alocando-os aos centróides mais próximos. Após os n indivíduos serem alocados, têm-se os k grupos formados.
3. Calculam-se os novos valores dos centroides e se houver alterações dos centroides aplica-se novamente o passo 2.
4. Repetir os passos 2 e 3 até que não haja alterações nos centroides.

2.3.1.2 Método complete linkage

Os métodos de agrupamentos hierárquicos aglomerativos, método complete linkage e Ward, seguem os seguintes passos para a formação dos grupos:

1. Defina cada indivíduo como um grupo.
2. Busque o par de grupos mais similares, de acordo com uma medida de dissimilaridade escolhida.
3. Combine o par encontrado em 2, formando assim um novo grupo e recalcule a dissimilaridade deste grupo para os demais grupos.
4. Repita os passos 2 e 3 até sobrar um único grupo que contenha todos os indivíduos.

O resultado dos métodos hierárquicos é um dendrograma. Um dendrograma representa a sequência de agrupamentos, com as ligações entre os elementos e a dissimilaridade das ligações no eixo vertical. Para o método complete linkage a dissimilaridade entre dois grupos, C e C', é a maior distância de um indivíduo de C para algum indivíduo de C', sendo definida por:

$$d(C, C') = \max\{d(\boldsymbol{i}, \boldsymbol{j})\}, \text{ para } \boldsymbol{i} \in C, \boldsymbol{j} \in C'. \tag{2.3}$$

2.3.1.3 Método Ward

Sejam C e C' dois grupos quaisquer que serão agrupados, $SQE_{CC'}$ representa a soma dos erros quadrados antes da junção, SQE_C e $SQE_{C'}$ as somas de quadrados dos erros dos grupos que serão combinados, no qual x, y e z são os indivíduos respectivamente do grupos C, C' e da juncão dos grupos C e C'.

$$SQE_{CC'} = \sum_{i=1}^{n_{CC'}} (z_i - \bar{z}_{CC'})'(z_i - \bar{z}_{CC'}) \tag{2.4}$$

$$SQE_C = \sum_{i=1}^{n_C} (x_i - \bar{x}_C)'(x_i - \bar{x}_C) \tag{2.5}$$

$$SQE_{C'} = \sum_{i=1}^{n'_C} (y_i - \bar{y}'_C)'(y_i - \bar{y}_{C'}) \tag{2.6}$$

O método Ward agrupa dois grupos distintos, C e C', tais que o aumento da soma dos quadrados dos erros (SQE), dado pela equação (2.7), seja mínimo.

$$\delta(C, C') = SQE_{CC'} - (SQE_C + SQE_{C'}) \tag{2.7}$$

Em cada passo os grupos que minimizam a SQE são combinados. Este método tende a resultar em grupos com tamanhos mais semelhantes devido à minimização da variação interna feita na junção dos grupos (HAIR et al., 2010)

2.3.2 Métodos de agrupamentos espaciais

Os métodos de agrupamentos espaciais usados neste trabalho são os métodos ClustGeo e Skater. O algoritmo Skater conta com imposições espaciais mais restritivas em sua formulação e assim os grupos retornados são mais homogêneos. Por outro lado, o método ClustGeo permite incorporar a espacialização dos dados, mas, diferentemente do Skater, agregam grupos que não sejam necessariamente formetemente homogêneos.

2.3.2.1 Método ClustGeo (CG)

O método ClustGeo consiste em combinar duas matrizes de dissimilaridades, $D_0 = [d_{0(ij)}]$ (que representa a matriz de dissimilaridades obtida com base nos valores das variáveis não espaciais) e $D_1 = [d_{1(ij)}]$ (uma matriz de dissimilaridades associada às unidades espaciais que são baseadas em distâncias geográficas e/ou contiguidade), usando um parâmetro de mistura $\alpha \in [0, 1]$, o qual determinará qual das matrizes terá maior importância na determinação dos grupos.

Seja um conjunto de n indivíduos, w_i é o peso do i-ésimo indivíduo para $i = 1, \ldots, n$. Suponha que os n indivíduos são particionados em k grupos, denotados por C_k^α, com $k = 1, \ldots, n$, formando a partição $\mathcal{P}_k^\alpha = \{C_1^\alpha, \ldots, C_n^\alpha\}$. O método ClustGeo, baseado no método de agrupamento Ward (2.3.1.3), implica na minimização da inércia dentro de um grupo C_k^α da partição $\mathcal{P}_k^\alpha$ que é definida como:

$$W(\mathcal{P}_k^\alpha) = \sum_{k=1}^{n} I_\alpha(C_k^\alpha), \tag{2.8}$$

onde $I_\alpha(C_k^\alpha)$ é a inércia de C_k^α, definida como:

$$I_\alpha(C_k^\alpha) = (1-\alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i wj}{2\mu_k^\alpha} (d^2_{0(ij)}) + (\alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i wj}{2\mu_k^\alpha} (d^2_{1(ij)}), \tag{2.9}$$

onde $\mu_k^\alpha = \sum_{i \in C_k^\alpha} w_i$ é o peso de C_k^α.

Quando $\alpha = 0$, a equação (2.9) baseia-se apenas nas dissimilaridades da matriz D_0 e quando $\alpha = 1$ apenas a matriz de dissimilaridades D_1. Para o método ClustGeo ter bom desempenho é necessário encontrar um valor adequado para α no intervalo [0,1], já que os valores 0 e 1 são os extremos. Para determinar o valor adequado do parâmetro de mistura α deve-se levar em conta um valor de α que aumente a homogeneidade geográfica dos grupos sem deteriorar a homogeneidade relativa aos valores das variáveis.

O efeito de α nas matrizes D_0 e D_1 é baseado na proporção de inércia para a partição $\mathcal{P}_K^\alpha$, com $\beta \in [0, 1]$, definida por:

$$Q_\beta(P_k^\alpha) = 1 - \frac{W_\beta(P_k^\alpha)}{W_\beta(P_1).} \tag{2.10}$$

A escolha do parâmetro de mistura é feita por meio da plotagem dos valores de Q_β para os diferentes valores de α, tanto para a matriz D_0 quanto D_1. Essas duas curvas, representadas na mesma figura, permitem que se escolha o valor de α mais conveniente. Quando $\beta = 0$, valores elevados de Q_β indicam maior homogeneidade do grupo em termos dos atributos não espaciais. Quando $\beta = 1$, elevados valores de Q_β indicam maior homogeneidade do grupo em termos dos atributos espaciais.

Uma outra forma de construir a matriz D_1 é com base em vizinhanças. Essa matriz é criada pelo critério de contiguidade (C) dado pela equação $C = (c_{ij})_{\text{nxn}}$, onde $c_{ij} = 1$, se o i-ésimo e o j-ésimo objeto são contíguos e 0 caso contrário. A matriz de adjacência A, que é criada pelo critério de contiguidade, é a base para a construção da matriz de vizinhança/contiguidade:

$$D_1 = P - A, \tag{2.11}$$

onde $P = [p_{ij}]_{\text{nxn}} = 1,\ \forall$ i e j, $A = [a_{ij}]_{\text{nxn}}$ será igual a 1 se os indivíduos i e j são vizinhos e 0 caso contrário, e a diagonal $a_{ii} = 1$ por convenção.

Para esse tipo de matriz de dissimilaridades, a coesão geográfica, quando se tem poucos grupos, é frequentemente pequena, ou seja, $W_1(P_k)$ pode ser muito pequeno e $Q_1(P_k)$ assumirá valores muito maiores que os obtidos por $Q_0(P_k)$, uma vez que pares de indivíduos próximos, mas que não possui vizinhança, recebem o mesmo valor que indivíduos muito distantes. Por exemplo, se pensarmos no mapa do Brasil para a configuração de vizinhança/contiguidade, o par de Estados Paraná-Rio Grande do Sul recebe o valor 0, da mesma forma que o par de Estados Paraná-Acre, em que a distância relativa ao primeiro par é significantemente menor se comparado ao segundo par. Para minimizar a desigualdade nas duas matrizes usadas pelo método, Q_β precisa ser normalizado, conforme descrito em (2.12). Deste modo, garante-se escalas similares para Q_0 e Q_1, e uma análise mais coerente baseada em tais quantidades.

$$Q_{\beta_{\text{norm}}}(P_k^\alpha) = Q_\beta(P_k^\alpha) Q_\beta(P_k^\beta) \tag{2.12}$$

A análise e escolha do parâmetro de mistura poderá ser realizada após fixado o número de grupos. Para determinar o número de grupos pode ser usado o dendrograma, construído com base na matriz de dissimilaridades D_0, pelo método Ward. Para os seguintes capítulos, por simplicidade, as abreviações ClustGeo (CG), ClustGeo matriz de distâncias geográficas (CGG) e ClustGeo matriz de vizinhança/contiguidade (CGV) são usadas.

2.3.2.2 Algoritmo SKATER

O algoritmo Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) baseia-se na transformação de um mapa (conjunto de objetos geográficos contíguos) em um grafo, em que os nós (vértices) representam os municípios e as arestas a relação de vizinhanças (fronteiras comuns). A partir do grafo uma árvore geradora mínima (AGM) é produzida e esta será particionada iterativamente gerando os grupos desejados. A proposta do algoritmo Skater é limitar a complexidade do grafo, cortando as arestas com alta dissimilaridade.

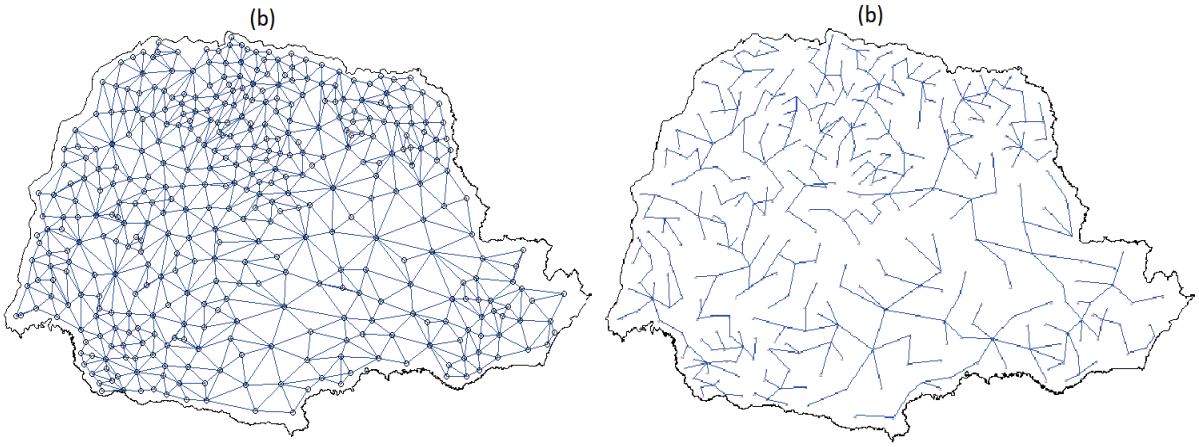
O funcionamento do algoritmo Skater pode ser descrito em duas etapas. A primeira etapa corresponde à produção da AGM que representa um gráfico de vizinhança. Dado um grafo de conectividade $G = (V, L)$ com um conjunto de vértices *(V)* e um conjunto de arestas *(L)*, o algoritmo inicia com uma árvore T_1, contendo apenas um vértice. A cada iteração, uma nova aresta e um novo vértice são adicionados à árvore. Na iteração n, a árvore T_n contém todos os n vértices de V e um subconjunto, L_m de L, com $n-1$ arestas. O custo de cada aresta é proporcional à dissimilaridade do par de indivíduos. A soma dos custos (que é a calculada pela dissimilaridade dos indivíduos) associados às arestas em L_m é mínima. Nos ANEXOS é apresentada a ilustração das etapas para geração do AGM

conforme apresentado a seguir:

1. Escolha qualquer vértice v_i no conjunto completo de vértices *(V)* ,$v_i \in V$, configurando $T_k = T_1 = (\{v_i\}, \varnothing)$
2. Encontre a aresta de menor custo *(l')* $\in L$ que conecta qualquer vértice de T_k a outro vértice, v_j, pertencente a V, mas não a T_k. Se existir mais de uma aresta com essa propriedade, escolha uma delas arbitrariamente
3. Acrecente o vértice v_j e a aresta l' à árvore T_k, criando uma nova árvore T_{k+1}.
4. Repita 2 e 3 até que todos os vértices tenham sido incluídos na árvore (T_n).

A FIGURA 1 apresenta o grafo (a), que é a etapa 1 da do algoritmo Skater, e a AGM (b) que é a finalização da etapa 4, como descrita acima.

FIGURA 1 – Grafo (a) e AGM (b) do Estado do Paraná construídas pelo algoritmo Skater



FONTE: A autora (2020)

Após concluída a construção da AGM, inicia-se a segunda etapa do algoritmo. Esta etapa consiste em particionar a AGM para obter os grupos, usando uma estratégia de divisão hierárquica.

Inicialmente, o grafo G^* contém apenas uma árvore (que pode ser interpretado como um único grupo), que é a AGM, a cada iteração uma aresta é removida, particionando a árvore examinada T_i em duas árvores T_{i_1} e T_{i_2} (que pode ser interpretado como a divisão de um grupo em outros dois). Feita a divisão da AGM em duas árvores, as demais divisões devem analisar cada uma das árvores formadas na etapa anterior. Para isso, as melhores soluções para cada uma das árvores $T_1, \ldots, T_n$ são comparadas através de uma função objetivo (2.13). A solução que maximiza a função objetivo é a que melhor subdivide uma

árvore T em duas novas árvores, também chamada de solução ideal S_*^R da função objetivo. Este processo é repetido até atingir o número de grupos estipulados a priori.

$$f_1(S_l^T) = SSD_{T_i} - (SSD_{T_{i_1}} + SSD_{T_{i_2}}) \tag{2.13}$$

em que o desvio quadrado dentro do grupo (SSD), equação (2.14), é uma medida da dispersão dos valores dos indivíduos em um mesmo grupo,

$$SSD_k = \sum_{j=1}^{k} \sum_{i=1}^{n_k} (x_{ij} - \bar{x}_j)^2, \tag{2.14}$$

onde n_k é o número de objetos espaciais na árvore k; x_{ij} é o j-ésimo atributo do objeto espacial i; m é o número de atributos considerados na análise e $\bar{x}_j$ é o valor médio do j-ésimo atributo (variável) para todos os objetos na árvore k.

A cada iteração do algoritmo, a árvore T_i, que tem o maior valor da função objetivo $f_1(S_*^{T_i})$ comparada as outras árvores, é dividida. O objetivo desse corte é maximizar a qualidade em cada etapa. A qualidade de uma partição é dada por,

$$Q(\sqcap) = \sum_{i=1}^{k} SSD_i, \tag{2.15}$$

onde $\sqcap$ é uma partição dos objetos espaciais em k árvores. Assim, quanto menor o $Q(\sqcap)$, melhor a partição, pois regiões homogêneas produzem menores valores de SSD.

Nos ANEXOS é apresentada a ilustração das etapas para o particionamento da AGM. A construção dos grupos é descrita por cinco etapas:

1. Comece com o grafo $G = (T_0)$, em que $T_0 =$ AGM;
2. Identifique a aresta que possui a $f_1(S_*^{T_0})$ mais alta;
3. Enquanto $\#(G^* \prec k)$, (k é o número desejado de grupos) faça:
   - 3.1 Para todas as árvores em G^*, selecione a T_i responsável por maximizar $f_1(S_*^{T_i})$;
   - 3.2 Divida T_i em duas novas subárvores e atualize G^*.

2.4 ÍNDICES DE VALIDAÇÃO INTERNA E EXTERNA PARA ANÁLISE DE AGRUPAMENTOS

Existem diferentes métodos de agrupamentos, cada um deles produz resultados distintos para um mesmo conjunto de dados. Como não é possível apontar um método que se sobressai aos outros o processo de validação é determinante para se avaliar a qualidade

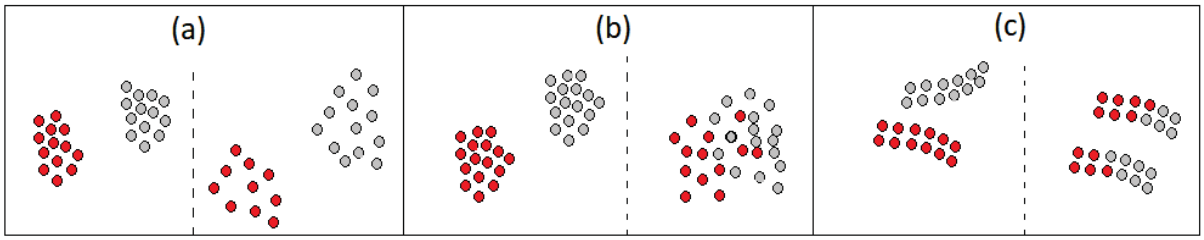
dos agrupamentos (HOEF; WARRENS, 2019). No presente trabalho, para avaliar os agrupamentos, os índices de validação interna foram usados para os dados educacionais do Estado do Paraná e os índices de validação externa para o estudo baseado em dados simulados.

2.4.1 Índices de validação interna

A validação interna permite a avaliação dos agrupamentos internamente, ou seja, avalia os agrupamentos de uma mesma partição. Esse tipo de validação geralmente é usado para avaliar agrupamentos que não possuem referências presumidas das partições (DISTEFANO; MAMELI; POLI, 2020), pois, diferentemente da validação externa, esses índices não comparam métodos de agrupamento e sim a estrutura da partição e dos grupos formados.

A validação interna leva em consideração se os indivíduos pertencentes a um mesmo grupo são homogêneos e se são heterogêneos aos indivíduos que compõem os demais grupos que formam a partição. Neste trabalho três índices de validação interna são usados para avaliar: a separação (largura média da silhueta), a compactação (índice de Dunn) e a conectividade (índice de Conectividade) dos dados, como exemplificado na FIGURA 2. A validação interna é baseada apenas nas informações intrínsecas dos dados e seus respectivos agrupamentos, logo os dois principais conceitos são a separação e a compacidade, já que a conectividade é o oposto da separação.

FIGURA 2 – Exemplos da comparação de conjuntos de dados representado grupos com melhor (à esquerda) e pior (à direita) compactação (a), separação (b) e conectividade espacial (c).



FONTE: A autora (2020)

Os índices apresentados a seguir, bem como as interpretações e explicações, foram extraídos de Distefano, Mameli e Poli (2020) e Hassani e Seidl (2017).

2.4.1.1 Largura média da silhueta

A largura média da silhueta mede o quão semelhante um indivíduo é ao seu próprio grupo (compactação/homogeneidade) em relação com os outros grupos (separação). Esse

índice fornece uma representação (gráfica) assertiva de quão bem cada indivíduo foi classificado. A equação da largura média da silhueta é apresentada da seguinte forma:

$$\mathrm{S} = \frac{1}{k}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in C_k}\frac{b(i)-a(i)}{\max[a(i),b(i)]}, \tag{2.16}$$

onde a(i) $= \frac{1}{n_k-1}\sum_{j\in C_k, j\neq k} d(i,j)$ a distância dos pares entre todos os indivíduos em um grupo, b(i) $= \min_{1\leq l\leq K, l\neq k}[\frac{1}{n_l}\sum_{j\in C_l} d(i,j)]$ é distância média dos indivíduos de um grupo ao segundo grupo mais próximo, n_k denota o número de indíviduos do grupo C_k e $d(i,j)$ a distância entre o i-ésimo e o j-ésimo indivíduo.

A largura média da silhueta assume valores no intervalo [-1,1]. Indivíduos bem agrupados tendem a ter valores próximos a 1 e indivíduos mal agrupados tendem a ter valores próximos a -1.

2.4.1.2 Índice de Dunn

O índice de Dunn (2.17) tem como o objetivo identificar se os grupos de uma partição são compactos, ou seja, se os indivíduos de um mesmo grupo são homogêneos entre si e heterogêneos em relação aos indivíduos que compõem os demais grupos formados.

$$Dunn = \frac{\min_{1\leq l,k\leq K, l\neq k}(\min_{i\in C_k, j\in C_l} d(i,j))}{\max_{1\leq k\leq K}(\max_{i,j\in C_k, i\neq j} d(i,j))}, \tag{2.17}$$

onde $d(i,j)$ é a distância entre o i-ésimo e o j-ésimo indivíduos. O índice de Dunn assume valores no intervalo [0,∞). Valores positivos, e quanto mais altos, indicam grupos homogêneos.

2.4.1.3 Índice de conectividade

O índice de conectividade (2.18) avalia se os indivíduos vizinhos compartilham o mesmo grupo. Esse índice indica o grau de conectividade dos grupos, conforme determinado pelos k-vizinhos mais próximos.

$$\mathrm{Conn} = \sum_{i=1}^{n}\sum_{j=i}^{L}\delta(i, nn_{i(j)}), \tag{2.18}$$

onde L fornece o número de vizinhos mais próximos do indivíduo (i) avaliado, $nn_{i(j)}$ representa um desses vizinhos mais próximos, para $j = 1, 2, ..., L$. Em que, $\delta(i, nn_{i(j)})$ é igual a 0 se o i-ésimo e j-ésimo indivíduo pertencem ao mesmo grupo e $\delta(i, nn_{i(j)})$ é igual a $\frac{1}{j}$ se os indivíduos pertencem a grupos difrentes. O índice de conectividade produz valores no intervalo [0, ∞) e valores proxímos de zero indicam uma partição do conjunto de dados com elevada conectividade.

2.4.2 Índices de validação externa

Neste trabalho foram considerados dois índices de validação externa. Esses índices são geralmente usados para avaliar a similaridade entre agrupamentos, ou seja, quanto os agrupamentos são semelhantes dado uma mesma base de dados, mas obtidos por métodos de agrupamentos diferentes (PFITZNER; LEIBBRANDT; POWERS, 2009). Em nosso estudo de simulação, os índices de validação externa foram aplicados na comparação dos agrupamentos fornecidos pelos métodos de agrupamentos com aqueles predefinidos para os municípios. O objetivo é avaliar a habilidade dos métodos de agrupamentos (espaciais e não espaciais) em reproduzir uma estrutura preenxistente (na prática, conhecida) de grupos. As definições a seguir têm como referência Meilă (2003), Distefano, Mameli e Poli (2020) e Haslbeck e Wulff (2020).

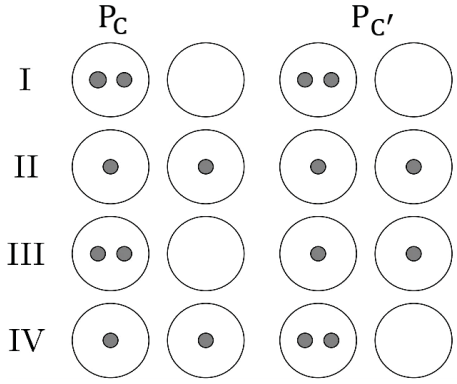
2.4.2.1 Índice de Rand ajustado

O índice de Rand ajustado é um ajuste do índice de Rand. O índice de Rand é baseado na abordagem de contagem de pares de pontos. São analisados pares de pontos que simultaneamente pertencem (ou não pertencem) a duas partições distintas de grupos.

Sejam $\mathcal{P}_C = \{C_1, C_2, \ldots, C_k\}$ e $\mathcal{P}_{C'} = \{C'_1, C'_2, \ldots, C'_{k'}\}$ duas partições de um mesmo conjunto de dados D em subconjuntos (grupos) C_i e C'_j, com $i = 1, \ldots, k$ e $j = 1, \ldots, k'$ e como as partições são diferentes, o tamanho dos grupos também pode variar de partição para partição.

Para um par de indivíduos do conjunto de dados, os componentes para o cálculo dos índices de Rand são ilustrados (FIGURA 3) e definidos abaixo.

FIGURA 3 – Ilustração esquemática das quatro configurações possíveis para dois indivíduos em dois agrupamentos



FONTE: Reproduzido de Haslbeck e Wulff (2020)

I. N_{11} = Número de indivíduos que pertencem ao mesmo grupo em ambos $\mathcal{P}_C$ e $\mathcal{P}_{C'}$.

II. N_{00} = Número de indivíduos que pertencem a diferentes grupos em ambos $\mathcal{P}_C$ e $\mathcal{P}_{C'}$.

III. N_{10} = Número de indivíduos que pertencem ao mesmo grupo na $\mathcal{P}_C$, mas não em $\mathcal{P}_{C'}$.

IV. N_{01} = Número de indivíduos que pertencem ao mesmo grupo na $\mathcal{P}_{C'}$, mas não em $\mathcal{P}_C$.

Segue que:

$$N_{11} + N_{00} + N_{10} + N_{01} = n(n-1)/2 \tag{2.19}$$

O índice de Rand (R) produz valores no intervalo [0,1] sendo definido por:

$$R(\mathcal{P}_C, \mathcal{P}_{C'}) = \frac{N_{11} + N_{10}}{n(n-1)/2}, \tag{2.20}$$

que corresponde à proporção de pontos que são alocados ao mesmo grupo ou a grupos distintos em ambos P_C e $P_{C'}$

A sobreposição de $\mathcal{P}_C$ e $\mathcal{P}_{C'}$ pode ser resumida em uma tabela de contingência $KxK'[n_{ij}]$, FIGURA (4), onde cada entrada n_{ij} denota o número de indivíduos pertencentes a cada combinação de C_i e C'_j, $n_{ij} = |C_i \cap C'_j|$.

FIGURA 4 – Tabela de contingência

$\mathcal{P}_C \diagdown \mathcal{P}'_C$	C'_1	C'_2	$\cdots$	$C'_{k'}$	somas
C_1	n_{11}	n_{12}	$\cdots$	$n_{1k'}$	n_1
C_2	n_{21}	n_{22}	$\cdots$	$n_{2k'}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_k	n_{k1}	n_{k2}	$\cdots$	$n_{kk'}$	n_k
somas	n'_1	n'_2	$\cdots$	$n'_{k'}$	

FONTE: A autora (2020)

O índice de Rand ajustado assume valores no intervalo [-1,1] sendo definido por:

$$ARI(\mathcal{P}_C, \mathcal{P}_{C'}) = \frac{N_{11} - \frac{(N_{11}+N_{10})(N_{11}+N_{00})}{n(n-1/2)}}{\frac{(N_{11}+N_{10})+(N_{11}+N_{00})}{2} - \frac{(N_{11}+N_{10})(N_{11}+N_{00})}{n(n-1/2)}} \tag{2.21}$$

O índice de Rand ajustado é a versão corrigida pelo acaso do índice Rand e valores mais próximos a 1 indicam maior concordância de comparação de $\mathcal{P}_C$ e $\mathcal{P}_{C'}$. Outra

maneira de definir esse índice é apresentada na equação (2.22), em que E[R] é a esperança estatística do índice de Rand.

$$\mathrm{ARI}(\mathcal{P}_C, \mathcal{P}_{C'}) = \frac{R(\mathcal{P}_C, \mathcal{P}_{C'}) - E[R]}{1 - E[R]} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_i}{2} - \sum_j \binom{n'_j}{2}] - [\sum_i \binom{n_i}{2} \sum_j \binom{n'_j}{2}]/\binom{n}{2}} \tag{2.22}$$

onde n_{ij}, n_i, n'_j são valores extraídos da FIGURA 4.

2.4.2.2 Variação da informação

A variação da informação faz parte do conceito de teoria da informação. Índices teóricos da informação têm sua construção baseada em teroremas, por esse motivo a variação da informação não é apenas um índice, mas sim uma métrica. Os índices teóricos da informação quantificam a diferença na informação compartilhada entre duas partições (HOEF; WARRENS, 2019).

A variação da informação estabelece quanta informação existe em cada um dos agrupamentos e quanta informação um agrupamento fornece sobre o outro através da entropia e informação mútua. A incerteza (entropia) associada ao agrupamento $\mathcal{P}_C$ é dada pela equação:

$$H(\mathcal{P}_C) = -\sum_{i=1}^{k} P(i) \times \log P(i), \tag{2.23}$$

onde $P(i) = \frac{n_k}{n}$ é a probabilidade de um indivíduo escolhido ao acaso pertencer ao grupo C_i.

Seja $P(i, j)$ a probabilidade de um ponto pertencente a C_i em $\mathcal{P}_C$ e a $\mathcal{P}_{C'}$ em C'_j:

$$P(i, j) = \frac{|C_i \cap C'_j|}{n} \tag{2.24}$$

A informação mútua para um par de agrupamentos é dada pela seguinte expressão:

$$I(\mathcal{P}_C, \mathcal{P}_{C'}) = \sum_{i=1}^{k}\sum_{j=1}^{k} P(i, j).log\frac{P(i, j)}{P(i)P(j)} \tag{2.25}$$

A informação mútua para dois indivíduos é sempre não negativa e não excede o total de incerteza do agrupamento. Definidas essas duas medidas, a variação da informação é dada pela equação:

$$VI(\mathcal{P}_C, \mathcal{P}_{C'}) = H(\mathcal{P}_C) + H(\mathcal{P}_{C'}) - 2I(\mathcal{P}_C, \mathcal{P}_{C'}). \tag{2.26}$$

A variação da informação assume valores no intervalo [0, 2.ln(k)], onde k representa o número de grupos e pode ser reescrita como uma soma de dois termos positivos, representada por:

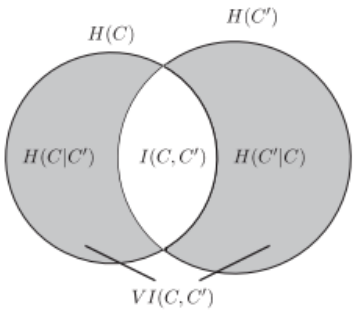
$$VI(\mathcal{P}_C, \mathcal{P}_{C'}) = [H(\mathcal{P}_C) - I(\mathcal{P}_C, \mathcal{P}_{C'})] + [H(\mathcal{P}_{C'}) - I(\mathcal{P}_C, \mathcal{P}_{C'})], \tag{2.27}$$

onde $[H(\mathcal{P}_C) - I(\mathcal{P}_C, \mathcal{P}_{C'})] = H(\mathcal{P}_C|\mathcal{P}_{C'})$ e $[H(\mathcal{P}_{C'}) - I(\mathcal{P}_C, \mathcal{P}_{C'})] = H(\mathcal{P}_{C'}|\mathcal{P}_C)$ que representam as entropias condicionais. O primeiro termo mede a quantidade de informação sobre C que não é preservada, enquanto o segundo mede a quantidade de informação preservada, quando vamos do agrupamento $\mathcal{P}_C$ para o agrupamento $\mathcal{P}_{C'}$. Uma expressão equivalente para VI é dada por:

$$VI(\mathcal{P}_C, \mathcal{P}_{C'}) = H(\mathcal{P}_C|\mathcal{P}_{C'}) + H(\mathcal{P}_{C'}|\mathcal{P}_C). \tag{2.28}$$

A FIGURA 5 exemplifica a expressão 2.28.

FIGURA 5 – Variação da informação



FONTE: Marina Meila(2003)

2.5 ÍNDICE GLOBAL DE MORAN (I)

Os testes mais utilizados para analisar a autocorrelação espacial global são o I de Moran e a estatística de Geary (GEARY, 1954). O Índice Global de Moran (I), que se baseia em uma medida de autocovariância na forma de produto cruzado, definido pela equação (2.29) foi usado neste trabalho. Este teste também é conhecido como I de Moran ou estatística de Moran.

Para avaliar a existência de autocorrelação espacial a hipótese nula do teste de I de Moran é sempre a de que não há associação espacial, caso tenha associação espacial a hipótese nula é rejeitada. O trabalho dos autores Bivand, Pebesma e Rubio (2008) serviu de

embasamento teórico para definir e interpretar os resultados da análise de autocorrelação espacial.

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.29}$$

onde n corresponde ao número de áreas; y_i é o valor do atributo na i-ésima observação, idem para y_j; $\bar{y}$ é a média da variável de interesse e w_{ij} é o peso espacial do vínculo entre as áreas i e j. O I de Moran resulta em valores no intervalo [-1,1], de maneira que

- $-1 \leq I < 0$: Autocorrelação espacial negativa ou inversa.
- $I = 0$: Não há autocorrelação (padrão espacial aleatório).
- $0 < I \leq 1$: Autocorrelação espacial positiva ou direta.

Existem diferentes estilos de pesos, como, por exemplo, padronização por linha (W), codificação binária básica (B) e padronização global (C). A construção de vizinhança também pode ser do tipo *queen* (rainha) ou *rook* (torre), que é construída semelhante ao movimento dessas duas peças no jogo de xadrez, ou outros, como a contiguidade da vizinhança por distância, k vizinhos mais próximos, dente outras possibilidades de definição de estruturas de vizinhaça. Neste trabalho usamos o estilo de peso W, no qual os pesos são a unidade dividida pelo maior e o menor número de vizinhos. Este tipo de pesos espaciais pode ser interpretado como um estilo que permite o cálculo de valores médios entre vizinhos. Para a definição das vizinhanças usamos o tipo *queen*, que é definida por cidades que compartilham pelo menos um vértice.

2.6 SOFTWARE

Todas as análises deste trabalho foram realizadas no software R Core Team (2020) e seus pacotes disponíveis.

3 ESTUDO DE CASO - DADOS EDUCACIONAIS

Neste capítulo é apresentado o estudo de caso contemplando dados educacionais do ensino básico do Estado do Paraná, estando dividido em quatro seções. A Seção 3.1 apresenta a descrição da base de dados utilizada. As Seções 3.2 e 3.3 apresentam as análises de correlação e autocorrelação espacial, respectivamente. Na Seção 3.4 apresenta-se as análises dos agrupamentos, dividida pela definição dos parâmetros para os métodos de agrupamentos e pelos resultados do estudo de caso, juntamente com as discusões e validações realizadas.

3.1 BASE DE DADOS

A base de dados considerada no presente trabalho é formada por dados educacionais públicos da educação básica (ensino fundamental 1, ensino fundamental 2 e ensino médio) referentes aos 399 municípios do Estado do Paraná. Esses dados foram extraídos dos sites oficiais do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e do Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES) . A base de dados é composta pelas seguintes seis variáveis (indicadores):

i. **APROV**: Taxa de aprovação (INEP, 2018)

   A taxa de aprovação no ensino básico é a razão entre a quantidade de matrículas com a situação de aprovado em uma etapa de ensino pelo total de matrículas nessa mesma etapa.

ii. **IDEB**: Média da nota do Ideb (INEP, 2017)

   O Índice de Desenvolvimento da Educação Básica (Ideb) é o principal indicador da qualidade da educação básica no Brasil. O Ideb é calculado a partir dos dados sobre aprovação escolar, obtidos no Censo Escolar realizado todos os anos, e das médias de desempenho no Sistema Nacional de Avaliação da Educação Básica (Saeb) aplicadas aos alunos do 5º e 9º ano do ensino fundamental e 3º ano do ensino médio.

iii. **IDHME**: Índice de Desenvolvimento Humano Municipal-Educação (IBGE/IPARDES, 2010)

   Segundo o Atlas de desenvolvimento humano do Brasil de 2013 , o IDHM - Educação é definido por meio de dois índices. O primeiro, com peso 1, refere-se à escolaridade da população adulta, que é o percentual de pessoas com 18 anos (ou mais) de idade com ensino fundamental completo. O segundo, com peso 2, refere-se ao fluxo escolar da população jovem, que corresponde à média aritmética (i) do percentual de crianças

de 5 a 6 anos frequentando a escola, (ii) do percentual de jovens de 11 a 13 anos frequentando os anos finais do ensino fundamental, (iii) do percentual de jovens de 15 a 17 anos com ensino fundamental completo e (iv) do percentual de jovens de 18 a 20 anos com ensino médio completo. A média geométrica desses dois componentes resulta no IDHM-Educação.

iv. **MAT**: Média de alunos por turma (INEP, 2019)

Refere-se ao tamanho médio das turmas no ensino básico.

v. **DIS**: Distorção de idade-série (INEP, 2019)

Corresponde à proporção de alunos com dois ou mais anos de atraso escolar na educação básica.

vi. **DCS**: Docentes com Curso Superior (INEP, 2019)

Percentual de docentes com nível superior completo lecionando na educação básica.

Para a variável IDEB alguns municípios do Estado do Paraná apresentaram dados faltantes em uma ou mais etapas de ensino. O QUADRO 1 apresenta a relação de municípios com dados indisponíveis.

QUADRO 1 - Municípios com dados faltantes para a variável IDEB

Ens. Fund. 1	**Ens. Fund. 2**	**Ens. Médio**
Porto Rico	Ariranha do Ivaí	Arapuã
Santa Cruz de Monte Castelo	Santa Cruz de Monte Castelo	Cafelândia
	Jataizinho	Jataizinho
	Querência do Norte	Querência do Norte
	São Sebastião da Amoreira	Formosa do Oeste
	Três Barras do Paraná	Três Barras do Paraná
		Itambé
		Catanduvas
		Lunardelli
		Nova Aliança do Ivaí
		Pérola d'Oeste
		Florestópolis
		Sapopema
		Teixeira Soares
		Indianópolis

FONTE: A autora (2020)

No total houve 2% de dados faltantes, sendo 0,5% para o ensino fundamental 1, 1,5% para o ensino fundamental 2 e 4% para o ensino médio. Os dados indisponíveis representam menos de 5% por etapa de ensino. Nessas circunstâncias, segundo Graham(2009)[1]

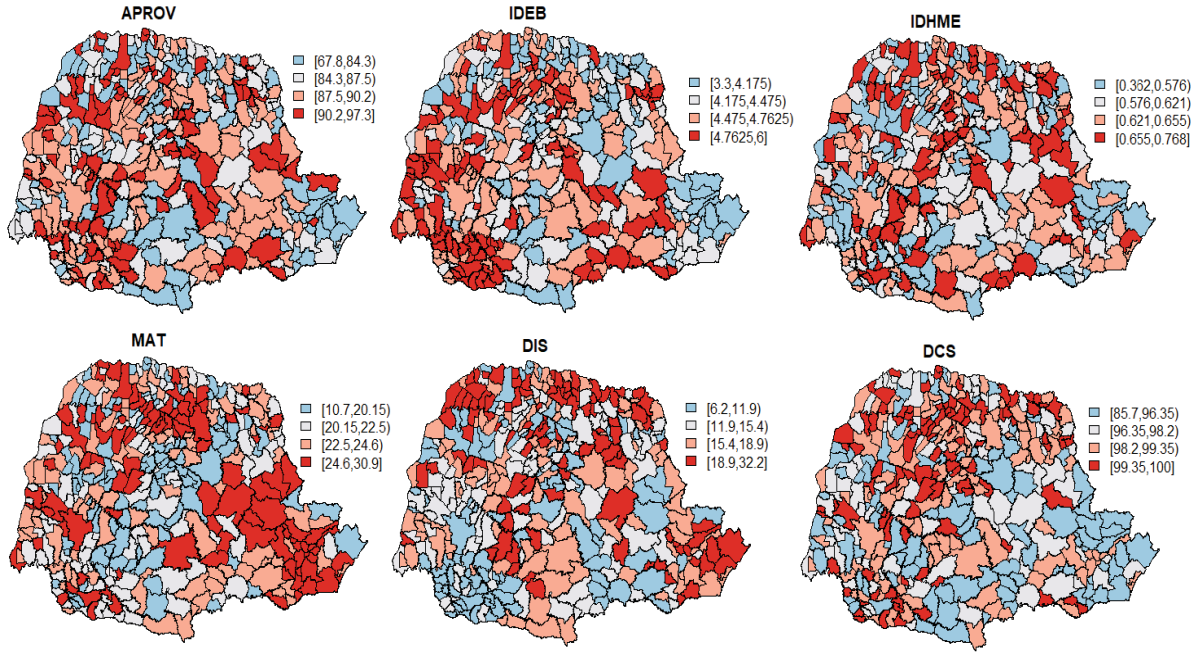
[1] Graham J.W. (2009) Missing data analysis: making it work in the real world. Annual Review of Psychology, 60, 549–576, DOI: 10.1146/annurev.psych.58.110405.085530

Schafer(1999) [2]; *apud* Azur et al. (2011) a utilização de dados imputados, no lugar dos dados faltantes, é uma abordagem aceitável.

A imputação dos dados faltantes foi realizada por meio do pacote Mice (VAN BUUREN; GROOTHUIS-OUDSHOORN, 2011), com o método de imputação de correspondência média preditiva (pmm). Após a execusão, a função retorna cinco valores gerados como resultado do processo de imputação para cada observação faltante e a média desses cinco valores para cada município foi escolhida para preencher os dados ausentes.

Na FIGURA 6 os diferentes mapas representam a distribuição espacial de cada uma das seis variáveis. Os valores das variáveis foram categorizados de acordo com os respectivos quartis, para uma melhor visualização dos resultados.

FIGURA 6 – Distribuição espacial dos indicadores educacionais para os municípios do Estado do Paraná



FONTE: A autora (2020)

Observando a FIGURA 6 podemos notar que na região sudeste do mapa para a variável docentes com curso superior (DCS), os municípios compartilham de menores valores desse indicador, o que para a mesma região no mapa da variável média de alunos por turma (MAT) os municípios compartilham dos maiores valores desse indicador. Para o mapa da variável da média da nota do Ideb (IDEB), ao sudoeste, esse indicador possui uma concetração de valores elevados, o que não ocorre para a variável distorção idade-série

[2] Schafer J.L. (1999) Multiple imputation: a primer. Statistical Methods in Medical Research, 8, 3–15, DOI: 10.1177/096228029900800102

(DIS) já que esse indicador possui uma concentração de valores baixos. Pode-se observar também que as escalas das seis variáveis analisadas são expressas em diferentes escalas, o que pode distorcer os resultados da análise de agrupamentos, de maneira que os dados necessitaram ser normalizados. As variáveis foram normalizadas da seguinte maneira:

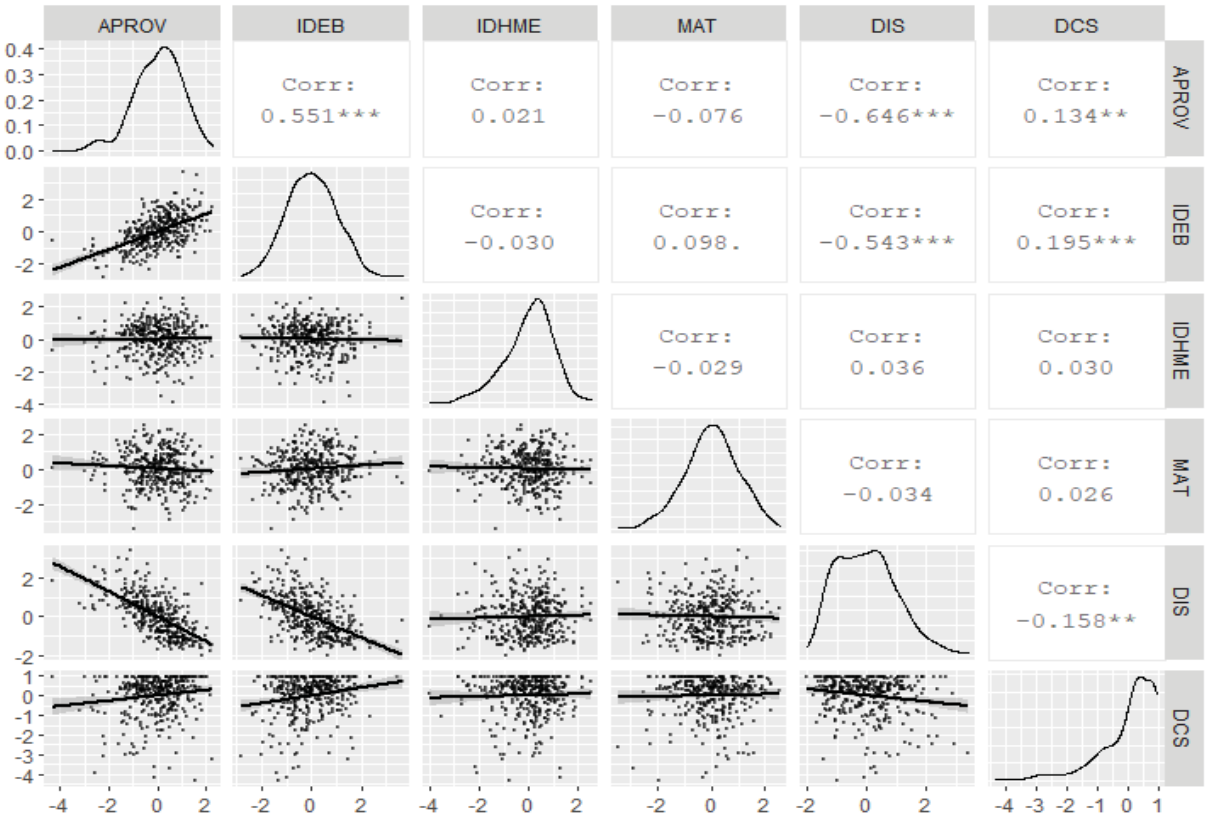
$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.1}$$

onde X representa uma particular variável, X_{min} e X_{max} o menor e o maior valor amostrado de X, respectivamente, e X_{norm} é a variável normalizada.

3.2 ANÁLISE DAS CORRELAÇÕES

Para analisar as correlações entre as variáveis usa-se a matriz de gráficos de correlações apresentada na FIGURA 7. Na parte superior estão representados os valores do coeficiente de correlação de Pearson e na parte inferior estão os gráficos de dispersão para cada par de variáveis. Os gráficos localizados na diagonal permitem avaliar a distribuição de cada indicador da educação básica para os 399 municípios do Estado do Paraná.

FIGURA 7 – Matriz de gráficos de correlações para os indicadores educacionais dos municípios do Estado do Paraná



FONTE: A autora (2020)

Analisando a FIGURA 7, observa-se que a correlação positiva de maior magnitude (0.551) ocorre para a taxa de aprovação (APROV) e para a média da nota do Ideb (IDEB). As correlações negativas numericamente mais expressivas ocorrem para taxa de aprovação (APROV) e distorção idade série (DIS) com -0.646, e para a média da nota do Ideb (IDEB) e distorção idade-série (DIS) com -0.5545, indicando que elas estão inversamente associadas.

3.3 ANÁLISE DA AUTOCORRELAÇÃO ESPACIAL

Com o objetivo de se investigar a existência de padrão espacial não aleatório, procedeu-se o teste de autocorrelação espacial dos dados educacionais. Os resultados para o teste I de Moran, para as seis variáveis, são apresentados na TABELA 1. As três primeiras colunas contém o valor observado de I, a esperança estatística (E(I)) e variância $\left(\frac{-1}{(n-1)}\right)$ sob a hipótese nula. Na quarta coluna está a diferença padronizada $\left(\frac{(I-E(I))}{\sqrt{var(I)}}\right)$ e, na última coluna, o p-valor do teste da hipótese de ausência de autocorrelação espacial.

TABELA 1 – Análise da autocorrelação espacial para os indicadores educacionais

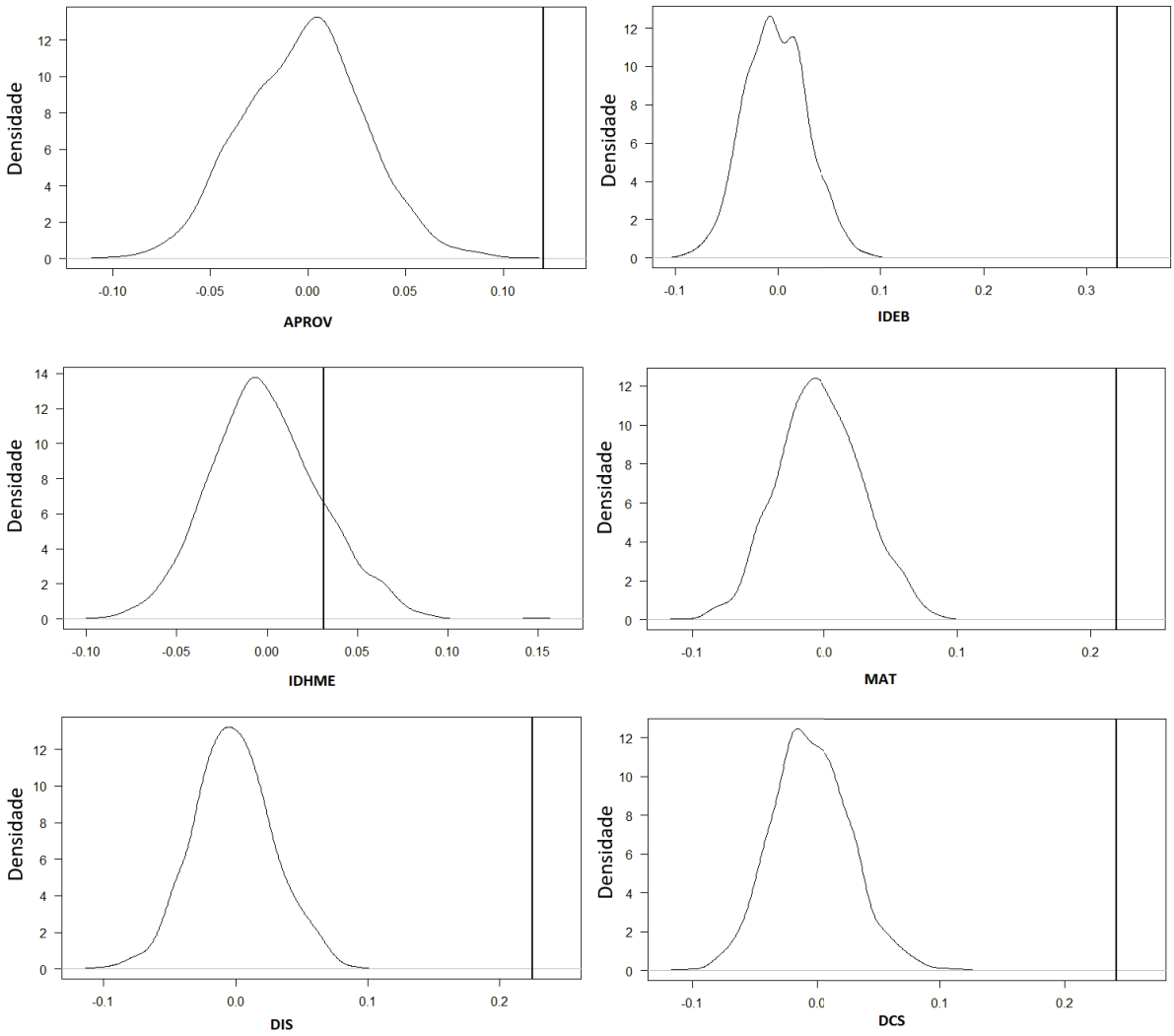
Variáveis	**I de Moran**	**E(I)**	**var(I)**	$\left(\frac{(I-E(I))}{\sqrt{var(I)}}\right)$	**p-valor**
APROV	0,120	-0,002	$9{,}37\text{x}10^{-4}$	4,014	<0,0001
IDEB	0,329	-0,002	$9{,}39\text{x}10^{-4}$	10,843	<0,0001
IDHME	0,031	-0,002	$9{,}37\text{x}10^{-4}$	1,104	0,1348
MAT	0,219	-0,002	$9{,}38\text{x}10^{-4}$	7,250	<0.0001
DIS	0,225	-0,002	$9{,}39\text{x}10^{-4}$	7,420	<0,0001
DCS	0,241	-0,002	$9{,}32\text{x}10^{-4}$	8,005	<0,0001

FONTE: A autora (2020)

O coeficiente I de Moran é positivo para todos os indicadores, ou seja, à medida que o valor da variável para um referido polígono aumenta, o mesmo tende a ocorrer com os polígonos vizinhos. Já analisando os p-valores, não há evidência, ao nível de significância de 5%, contra a hipótese nula para o indicador IDHME. Para os outros indicadores há evidência significativa contra a hipótese nula, confirmando a existência de autocorrelação espacial.

Como explanado na Seção 2.5, os resultados do índice de Moran podem depender das especificações feitas, como os pesos espaciais e as estruturas de vizinhança. Para confirmar a existência de autocorrelação espacial, usa-se a simulação de Monte Carlo, com n = 1000 simulações. A FIGURA 8 apresenta as distribuições obtidas por simulação de Monte Carlo. As curvas representam o caso em que as variáveis não apresentassem autocorrelação espacial e as linhas verticais os valores para o índice I de Moran calculados com base nos valores observados para os indicadores educacionais.

FIGURA 8 – Gráficos da densidade estimada via simulação Monte Carlo para o índice I de Moran na ausência de autocorrelação espacial para cada uma das seis variáveis



FONTE: A autora (2020)

Observa-se, a partir da FIGURA 8, que o índice I de Moran observado para as variáveis APROV, IDEB, MAT, DIS e DCS ficam à direita das distribuições dos valores simulados sob a hipótese de ausência de autocorrelação espacial. Para a variável IDHME, com I = 0.031, a simulação aponta a não rejeição da hipótese nula. A TABELA 2 sumariza os resultados da simulação de Monte Carlo. Analisando os p-valores, observamos que o indicador IDHME apresentou $p > 0{,}05$. Para os outros cinco indicadores, o $p < 0{,}05$ e, portanto, deste modo pode-se concluir que os resultados do índice I de Moran não ocorreram ao acaso nesses indicadores. Logo, pode-se concluir que os resultados do teste I de Moran são confirmados e assim reconfirma-se a autocorrelação espacial para os indicadores.

TABELA 2 – Resultados da simulação de Monte Carlo do I de Moran dos indicadores

Variáveis	**I de Moran**	$\#I_{\mathbf{sim}} < I$	**p-valor**
APROV	0,1203	1000	0,001
IDEB	0,3296	1000	0,001
IDHME	0,0312	845	0,155
MAT	0,2196	1000	0,001
DIS	0,2248	1000	0,001
DCS	0,2418	1000	0,001

FONTE: A autora (2020)

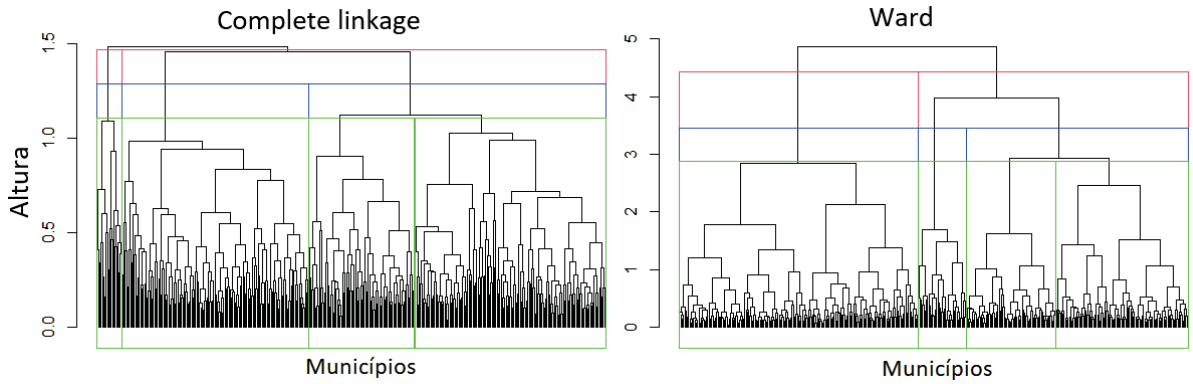
3.4 ANÁLISE DE AGRUPAMENTOS - ESTUDO DE CASO

Os métodos de agrupamentos necessitam de algumas definições antes de serem executados. As definições são: o número de grupos, a determinação do valor de α para as duas variações do método CG e uma restrição nos tamanhos mínimos dos grupos para o método Skater. Posteriormente a essas definições são apresentados os resultados retornados pelos métodos de agrupamentos para o estudo de caso envolvendo os indicadores educacionais do Estado do Paraná.

3.4.1 Definições dos parâmetros

Determinar o número de grupos (k) é um dos pontos essenciais para a aplicação dos métodos de agrupamentos. Para determinar o número adequado de grupos, foram usados os dendrogramas dos métodos complete linkage e Ward, apresentados na FIGURA 9. No entanto, como a determinação dos grupos por esse método é por meio da visualização, precisamos de outros critérios de determinação de k para validar a escolha.

FIGURA 9 – Dendrogramas com partições para dois (vermelho), três (azul) e quatro (verde) grupos dos métodos complete linkage e Ward aplicado aos dados educacionais



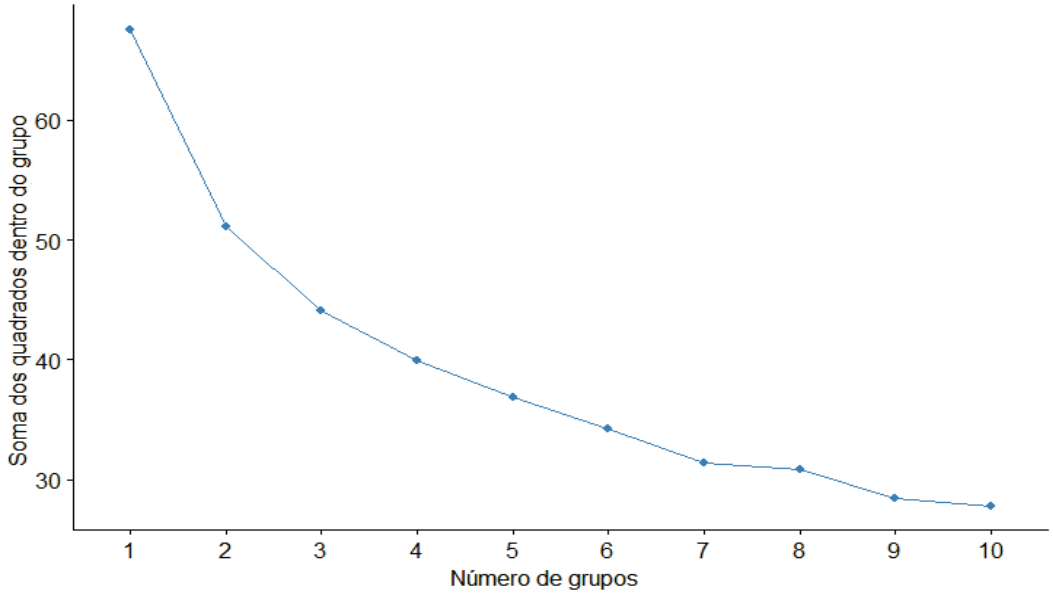
FONTE: A autora (2020)

A verificação do número adequado de grupos é feita buscando identificar onde o

dendrograma apresenta maiores segmentos verticais, indicando haver maior dissimilaridade na conjunção dos grupos. Logo, os dendrogramas parecem sugerir k = 2, 3 ou 4 como especificações a serem consideradas.

Para auxiliar na escolha de k a FIGURA 10 foi utilizada. Essa Figura representa o gráfico do cotovelo executado para o método K-means, em um intervalo k de 1 a 10, e para cada um desses dez valores é calculada a soma das distâncias quadradas intragrupos. Quando plotados os resultados no gráfico do cotovelo é possível determinar, visualmente, o valor de k tal que a soma das distâncias quadradas intragrupos começa a diminuir menos rapidamente.

FIGURA 10 – Gráfico do cotovelo para o método K-means



FONTE: A autora (2020)

Observando a FIGURA 10 podemos perceber uma substancial redução na soma de quadrados de k = 2 para k = 3. Além disso, a partir de k = 3, os valores de k produzem somas de quadrados mais próximas, e portanto, este gráfico sugerire a escolha de k = 3.

Os dois métodos usados anteriormente para a determinação de k são essencialmente visuais envolvendo, portanto, subjetividade. Como terceiro método para determinar k usamos o pacote NbClust (CHARRAD et al., 2014), que segundo os autores fornece 30 critérios objetivos para determinar o melhor número de grupos. Os 30 critérios foram aplicados aos dados educacionais, e, a partir disso, a função retornou as frequências com que os critérios indicaram os diferentes números de grupos. O número de grupos apontado pelo maior número de critérios foi k = 3, tanto para o método K-means, quanto complete linkage e Ward como apresentado na TABELA 3.

TABELA 3 – Distribuição de frequências para os números de grupos mais indicados pelos 30 critérios implementados na função `NbClust`

Nº de grupos (k)	**Frequências**		
	K-means	**Complete linkage**	**Ward**
2	7	8	6
3	8	11	10
4	1	1	1
5	0	0	1
6	0	1	0
7	0	0	2
9	0	0	1
10	1	0	0
11	0	1	0
12	3	2	2
18	1	0	0
19	2	0	0

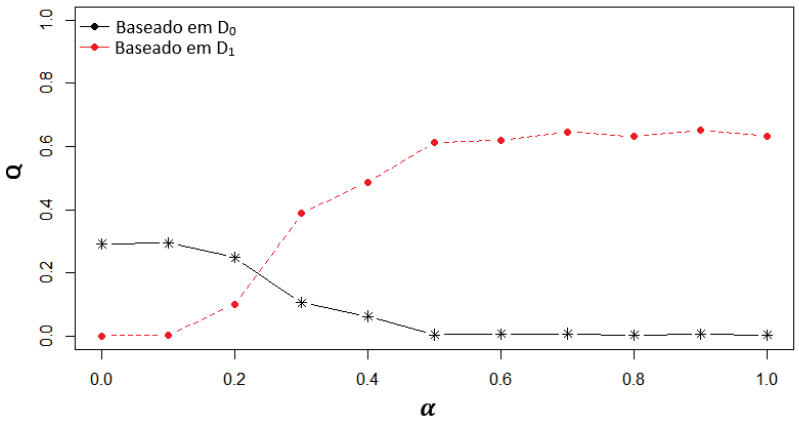
FONTE: A autora (2020)

Portanto, com ajuda de vários métodos de determinação do número k de grupos, definimos k = 3 como o número de grupos usado para todos os métodos de agrupamentos.

A determinação do parâmetro de mistura (α) para o método CG deve considerar que à medida que se aumenta o valor de α, os grupos são mais homogêneos espacialmente, mas menos homogêneos quanto aos valores das variáveis (indicadores). Nas FIGURAS 11 e 12 a linha preta corresponde ao efeito dos diferentes valores de α na matriz D_0 (indicadores) e a linha vermelha corresponde ao efeito dos diferentes valores de α atribuído à matriz D_1 (critério geográfico escolhido) e o peso de α é representado no eixo horizontal.

Para o método CGG, em que a matriz D_1 é construída por meio das distâncias geográficas, usamos o gráfico obtido a partir da proporção de inércia (Q),denominada de Q_1 para a matriz D_1 e para a matriz D_0 denominada de Q_0, e os valores do parâmetro α. Para o método CGV, que é referente à matriz D_1 construída por meio de vizinhanças, usamos o gráfico obtido a partir da proporção de inércia normalizada (Q_{norm}), denominada de $Q_{1\text{norm}}$ para a matriz D_1 e para a matriz D_0 denominada de $Q_{0\text{norm}}$, e dos valores do parâmetro α.

FIGURA 11 – Gráfico e valores para a escolha de α para o método ClustGeo baseado nas distâncias geográficas

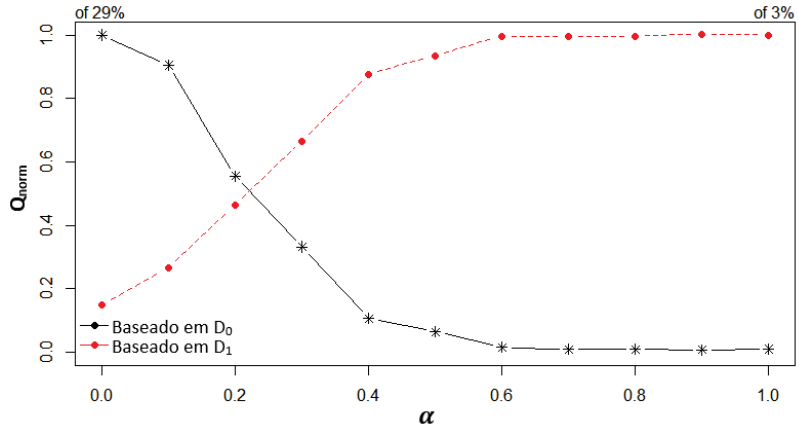


α	Q_0	Q_1
0.0	0.2925	0.0023
0.1	0.2942	0.0053
0.2	0.2495	0.1006
0.3	0.1069	0.3896
0.4	0.0636	0.4858
0.5	0.0077	0.6195
0.6	0.0077	0.6195
0.7	0.0086	0.6460
0.8	0.0049	0.6317
0.9	0.0076	0.6531
1.0	0.0050	0.6330

FONTE: A autora (2020)

Analisando os resultados dispostos na FIGURA 11, aparentemente as melhores escolhas são $\alpha = 0{,}2$ ou 0,3, pois as linhas se cruzam em algum valor entre esses dois pontos. Para confirmar a escolha, analisamos os valores de Q_0 e Q_1 no gráfico. Para $\alpha = 0{,}2$ temos uma perda de aproximadamente 15% na homogeneidade dos indicadores e um ganho de aproximadamente 16% na espacialização. Para $\alpha = 0{,}3$ temos uma perda de homogeneidade dos indicadores de aproximadamente 64%, e ganho de espacialização de aproximadamente 62%. Como a perda de homogeneidade dos indicadores é expressiva de 0,2 para 0,3 o valor escolhido é $\alpha = 0{,}2$.

FIGURA 12 – Gráfico e valores para escolha de α para o método ClustGeo baseado em vizinhaças



α	Q_{0norm}	Q_{1norm}
0.0	1.0000	0.1493
0.1	0.9044	0.2675
0.2	0.3326	0.6663
0.3	0.3326	0.6663
0.4	0.1068	0.8763
0.5	0.0647	0.9345
0.6	0.0158	0.9969
0.7	0.0099	0.9965
0.8	0.0099	0.9965
0.9	0.0084	1.0022
1.0	0.0100	1.0000

FONTE: A autora (2020)

Analisando os resultados dispostos na FIGURA 12, as duas linhas se cruzam em um ponto próximo a $\alpha = 0{,}2$. Analisando os valores de $Q_{0\text{norm}}$ e $Q_{1\text{norm}}$, os valores de $\alpha = 0{,}2$ e 0,3 são iguais. Logo, por meio do gráfico e da tabela, para o método CGV o valor escolhido é, também neste caso, $\alpha = 0{,}2$.

A priori os métodos K-means, Ward e Complete linkage podem ser executados sem necessidade de outras definições além do número de grupos. Para o método CG, uma vez definidos os parâmetros de mistura para suas duas variações, a análise de agrupamentos pode ser executada sem outras definições. No entanto, algumas especificações adicionais relativas ao método Skater precisam ser feitas.

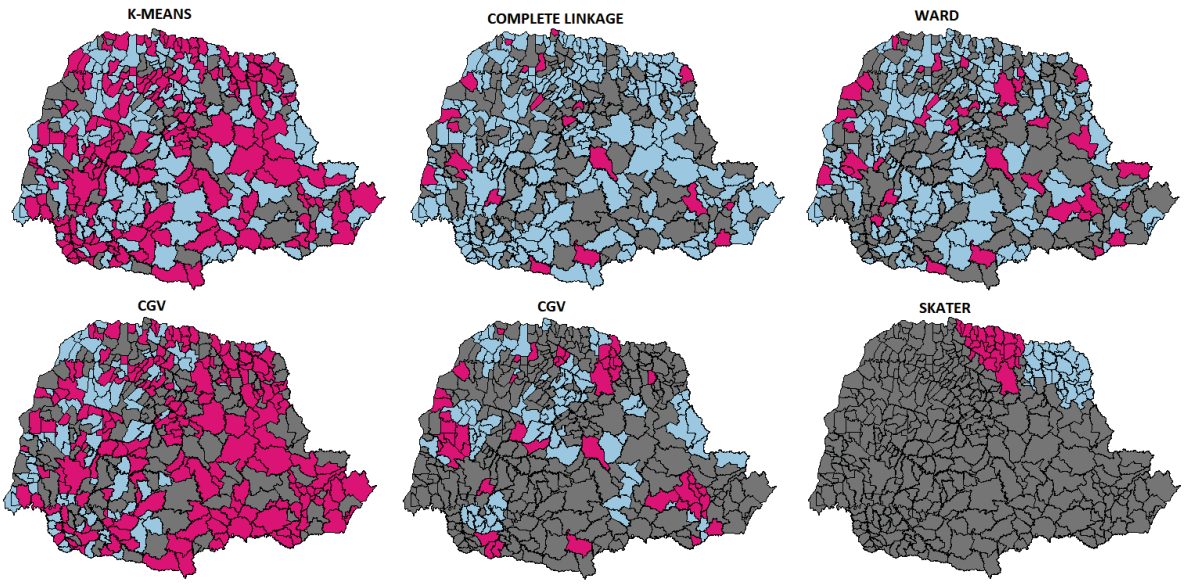
O método Skater, pelas caracteristicas do algoritmo, tende, muitas vezes, a compor um grande grupo, englobando a quase totalidade dos municípios, em detrimento a grupos pequenos, eventualmente unitários. Neste sentido, foi aplicada uma restrição de fixar um minímo de 20 municípios por grupo, o que corresponde a um mínimo de 5% de municípios em cada grupo.

Concluídas as definições de números de grupos, parâmetros de mistura para as duas variações do método CG e a questão particular do método Skater, os resultados de cada um dos métodos de agrupamentos para os indicadores educacionais do Estado do Paraná são apresentados na seção a seguir.

3.4.2 Resultados

Nesta seção os resultados provenientes dos métodos de agrupamentos aplicados a base de dados, que é composta pelos indicadores educacionais dos municípios do Estado do Paraná, são apresentados e avaliados. As análises são baseadas em mapas, na análise descritiva e exploratória dos grupos constituídos e nos índices de validação interna. Na FIGURA 13 os mapas dos resultados da aplicação dos métodos de agrupamentos aos indicadores educacionais, anteriormente discutidos, são apresentados. Os grupos são representados por cores distintas em cada mapa e não há conexão entre as cores usadas nos diferentes mapas.

FIGURA 13 – Mapas contendo os grupos resultantes dos diferentes métodos de agrupamentos aplicados aos indicadores educacionais dos municípios do Estado do Paraná



FONTE: A autora (2020)

Pode-se observar que o método Skater produz grupos espacialmente homogêneos, compostos por municípios vizinhos, em comparação aos outros métodos de agrupamentos, mas o número de municípios que compõem o maior grupo é muito elevado quando comparado aos dois outros grupos que formam a partição. O método CGG tem o grupos mais homogêneos para a metade direita do mapa, correspondendo as regiões leste, nordeste e sudeste do Estado. O método CGV tem um grupo, correspondente a cor cinza, mais homogêneo que os demais grupos. Os métodos complete linkage, Ward e K-means têm seus grupos organizados de forma mais heterogênea.

A TABELA 4 apresenta, para cada método, o número de grupos e as respectivas porcentagens de municípios por grupo. Ressaltando que os rótulos 1, 2 e 3 são arbitrários e não comparáveis.

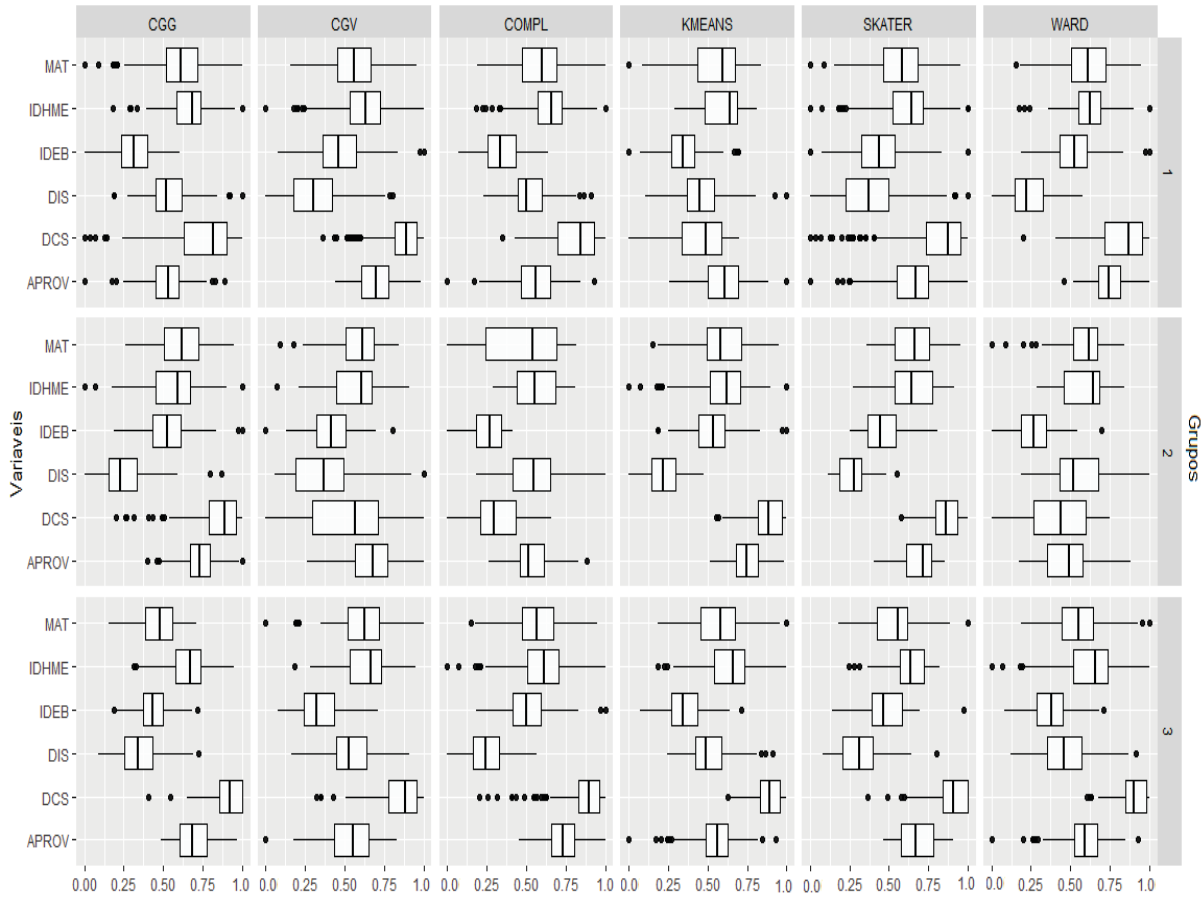
TABELA 4 – Resultados dos métodos de agrupamentos para o número de municípios do Estado do Paraná contidos em cada grupo

Método	**Grupos**		
	1	*2*	*3*
k-means	55	190	154
	13,7%	49,6%	38,6%
Complete linkage	146	20	233
	36,6%	5%	58,4%
Ward	187	38	147
	46,8%	9,5%	36,8%
CGG	130	182	87
	32,6%	45,6%	21,8%
CGV	264	47	85
	66,2%	11,8%	21,3%
Skater	340	26	33
	85,2%	6,5%	8,3%

FONTE: A autora (2020)

De acordo com a TABELA 4 o método CGG produziu grupos com tamanhos mais homogêneos. O método Skater produziu grupos com tamanhos mais heterogêneos, sendo que em apenas um dos grupos do método estão mais de 80% dos municípios, que também é o maior grupo retornado (340 municípios) entre os métodos análisados. O menor grupo, com 20 municípios, é retornado pelo método complete linkage e representa 5% do total de municípios do Estado do Paraná. O método CGV retorna o segundo maior grupo, com 264 municípios, que representa um total mais de 66% dos munícipios totais. Os métodos Ward e K-means apresentam resultados semelhantes, de certa forma, pois os grupos construídos apresentam tamanhos próximos. Na FIGURA 14 são apresentados os box plots para os seis indicadores educacionais divididos por métodos representados nas colunas e grupos indicados à direita. As variáveis estão representadas à esquerda. Vale ressaltar que os rótulos dos grupos são usados apenas para distinguir os grupos dentro dos métodos de agrupamento.

FIGURA 14 – Box plots dos resultados dos métodos de agrupamentos aplicados aos dados educacionais divididos por indicadores em cada grupo.



FONTE: A autora (2020)

Para o método K-means verificam-se maiores valores para a variável docentes com curso superior (DCS) nos grupo 1 e 2, enquanto menores valores para a variável distorção idade-série (DIS), para o grupo 2, e média da nota do Ideb (IDEB), para grupos 1 e 3, podem ser observados para os municípios que compõem esses grupos. O grupo 1 apresenta as variáveis mais dispersas ao ser comparado com os grupos 2 e 3, também esse grupo é o mais heterogêneo dos três grupos, que pode ser verificado na FIGURA 13. Para o método Complete linkage verificam-se maiores valores para as variáveis docentes com curso superior (DCS) e taxa de aprovação (APROV) no grupo 2, enquanto menores valores para a variável média da nota do Ideb (IDEB) podem ser observados para os municípios que compõem os grupos 2 e 3. O grupo 2, que é o grupo mais heterogêneo, apresentam as variáveis com as medianas mais próximas, se compararmos aos grupos 1 e 3 que são mais homogêneos que o grupo 2. Para o método Ward verificam-se maiores valores para a variável docentes com curso superior (DCS) nos grupo 1 e 3, enquanto menores valores para a variável média da nota do Ideb (IDEB) podem ser observados para os municípios que compõem os grupos 2 e 3. O grupo 2 apresenta as variáveis menos dispersas ao ser

comparado com os grupos 1 e 3, também esse grupo é o mais heterogêneo dos três grupos.

Para o método CGG verificam-se maiores valores para as variáveis docentes com curso superior (DCS), para os três grupos, e taxa de aprovação (APROV) nos grupos 2 e 3, enquanto menores valores para a variável distorção idade-série (DIS) podem ser observados para os municípios que compõem os grupos 2 e 3. Os grupos 1 e 2, que são os grupos mais homogêneos, apresentam semelhanças das variáveis média de alunos por turma (MAT), docentes com curso superior (DCS) e IDHME. Pelo método CGV verificam-se maiores valores para a variável docentes com curso superior (DCS) nos grupos 1 e 3, enquanto menores valores para a variável distorção idade-série (DIS) podem ser observados para os municípios que compõem os grupos 1 e 2. O grupo 1, que é o grupo mais homogêneo, apresentam as variáveis docentes com curso superior (DCS) e taxa de aprovação (APROV) com maiores valores para os munícipios deste grupo. Para o método Skater verificam-se maiores valores para as variáveis docentes com curso superior (DCS) e taxa de aprovação (APROV) nos três grupos, enquanto menores valores para a variável distorção idade-série (DIS) e média da nota do Ideb (IDEB) podem ser observados para os municípios que compõem os três grupos. O grupos, que são todos muito homogêneos, apresentam as variáveis com comportamentos com muita semelhança.

Os métodos não espaciais em que os grupos são mais heterogêneos, para K-means o grupo 1, Complete linkage e Ward o grupo 2, são caracterizados de forma semelhante pelos valores, tanto maiores quanto menores, das variáveis que os compõem. Os métodos espaciais apresentam maior homogeneidade espacial para CGG grupo 2, CGV grupo 1 e Skater todos os grupos e as variáveis desses grupos são caracterizadas dos menores para os maiores valores (respectivamente DIS, IDEB, IDHME, MAT, APROV,DCS) de forma semelhante, mostrando que os grupos com mais homogeneidade são similarmente construídos.

Por fim, a TABELA 5 apresenta os resultados dos índices de validação interna.

TABELA 5 – Resultados dos índices de validação interna para os métodos de agrupamentos construídos por meio da base de dados educacionais do Estado do Paraná

Método	**Silhueta média**	**Índice de Dunn**	**Conectividade**
K-means	-0,0289	$2{,}79\text{x}10^{-4}$	713
Complete linkage	-0,0265	$2.80\text{x}10^{-4}$	601
Ward	-0,0269	$2{,}97\text{x}10^{-4}$	701
CGG	-0,0684	$3{,}01\text{x}10^{-4}$	717
CGV	-0,0667	$2{,}96\text{x}10^{-4}$	572
Skater	-0,0351	$5{,}62\text{x}10^{-4}$	266

FONTE: A autora (2020)

No que diz respeito ao índice de largura da silhueta média que descreve a estutura dos grupos de uma mesma partição, podemos notar que todos os métodos de agrupamen-

tos produziram valores negativos, indicando que os cinco métodos, em alguma medida, resultaram em municípios mal agrupados. Em termos do índice de Dunn, que mede a homogeneidade entre os indivíduos de um mesmo grupo, temos o método Skater com o maior valor para esse índice seguido do método CGG indicando uma maior compactação dos grupos desses métodos. O índice de conectividade que avalia se os indivíduos vizinhos compartilham o mesmo grupo, apresentou menores valores para o método Skater seguido do método CGV. Esses resultados indicam que o método Skater pode identificar de forma precisa a estrutura do agrupamento, seguido do método CG.

Portanto, analisando o conjunto de resultados (mapas, box plots, número de municípios por grupo e índices de validação interna) os métodos espaciais tendem a reproduzir melhor os dados educacionais do Estado do Paraná. Os métodos espaciais apresentam grupos mais homogêneos e resultados superiores nos índices de validação interna. No entanto, vale ressaltar que as características e finalidades da utilização dos dados educacionais devem ser levadas em consideração na escolha dos métodos, pois nenhum dos métodos de agrupamentos têm resultados muito superiores aos demais no conjunto de análises feitas neste trabalho.

4 ESTUDO DE SIMULAÇÃO

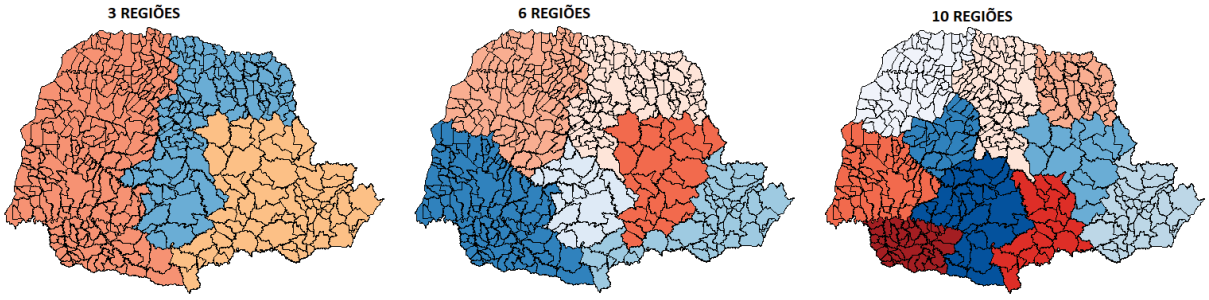
O objetivo deste estudo por simulação de dados espaciais é comparar o desempenho dos métodos espaciais e não espaciais. Essa comparação é feita por meio da reprodutibilidade dos métodos de agrupamentos a uma configuração predefinida de grupos, a partir da qual os dados foram simulados, denominada "gabarito". Foram considerados diversos cenários simulados, contemplando diferentes números de grupos e variáveis simuladas com diferentes níveis de variação espacial e intergrupos para avaliar o impacto desses fatores na performance dos métodos de agrupamentos espaciais e não espaciais.

Este capítulo está dividido em duas seções, sendo que na Seção 4.1 é descrita a metodologia do estudo por simulação, enquanto na Seção 4.2 são apresentados os respectivos resultados.

4.1 METODOLOGIA DO ESTUDO POR SIMULAÇÃO

Para este estudo por simulação foi adotado o mapa da divisão municipal do Estado do Paraná. Os municípios foram agrupados com base na divisão do Estado em regiões e mesorregiões. Com essas duas configurações foram formandos k = 10 grupos (mesorregiões - norte pioneiro paranaense, região metropolitana de Curitiba, norte central paranaense, sudoeste paranaense, oeste paranaense, centro ocidental paranaense, noroeste paranaense, centro oriental paranaense, centro-sul paranaense, sudeste paranaense), k = 6 grupos (regiões – Ponta Grossa, Curitiba, Londrina, Guarapuava, Cascavel, Maringá) e k = 3 grupos (junção de duas regiões – Ponta Grossa e Curitiba, Londrina e Guarapuava, Cascavel e Maringá) como apresentado na FIGURA 15.

FIGURA 15 – Mapas das divisões



FONTE: A autora (2020)

Para cada um dos 399 municípios, sem perda de generalidade, foi simulada uma

única observação (resposta univariada dada por y_{ij}) da seguinte forma:

$$y_{ij} = b_i + s_{ij}, \tag{4.1}$$

em que $b_i \sim N(0, \sigma_b^2)$ é o efeito do i-ésimo grupo, enquanto s_{ij} foi simulado de um modelo autorregressivo espacial (SAR acrônimo, em inglês, para *Spatial Auto Regressive*) tendo como referência o trabalho dos autores Bivand, Pebesma e Rubio (2008), com distribuição normal, parâmetro de autocorrelação ρ e variância $\sigma_s^2 = 4$.

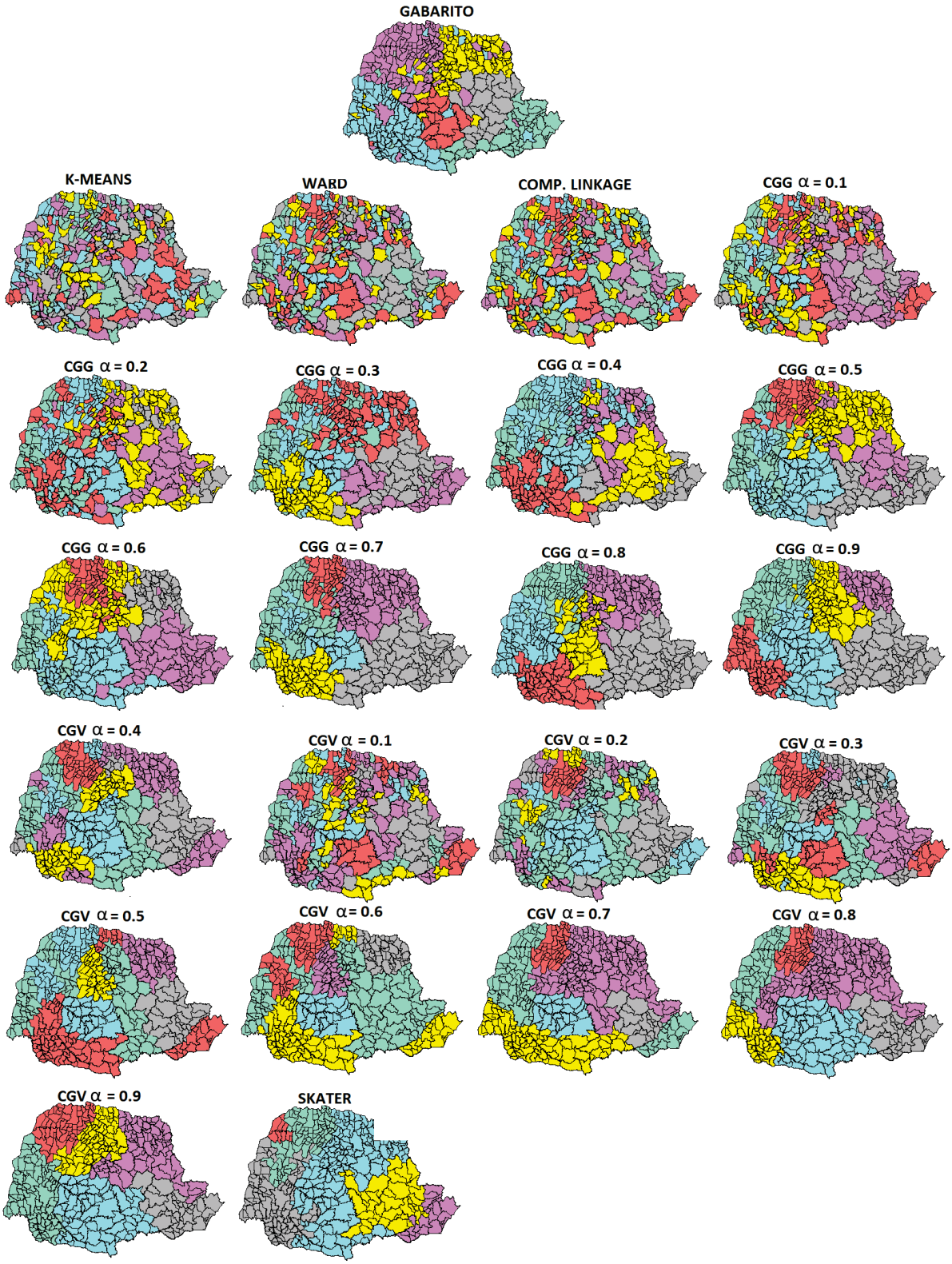
Realizada a simulação, uma porcentagem p dos municípios teve sua resposta permutada com municípios de grupos diferentes. O gabarito final é estabelecido após a permutação. Com esses parâmetros quanto maior o valor de σ_b^2, maior a heterogeneidade entre grupos; quanto maior o valor de ρ, maior a associação espacial em todo o Estado; e quanto maior p, maior a perturbação (ruído) introduzida na configuração regional de grupos.

Os seguintes parâmetros foram adotados para o estudo por simulação:

- $\sigma_b^2 = 1$, 4 e 16.
- $\rho =$ 0,5, 0,7 e 0,9.
- p = 5%, 10%, 20% e 33%.

Combinados os parâmetros e o número de grupos, resultam em um total de 108 cenários simulados e para cada um desses cenários foram realizadas 1.000 simulações. Para cada simulação resultante foram empregados os seguintes métodos de agrupamentos: K-means, complete linkage, Ward, CGG, CGV e Skater. Com base na solução produzida por um particular método e no gabarito, foram calculados os valores do índice de Rand ajustado e da variação da informação. Desta forma, para cada método e cenário simulado dispõe-se de 1.000 valores dos dois índices, que foram utilizados para avaliar as performances dos diferentes métodos. A FIGURA 16 mostra os resultados que os diferentes métodos de agrupamentos retornaram para um específico cenário simulado além do gabarito, a título de ilustração.

FIGURA 16 – Exemplo dos resultados dos métodos de agrupamentos para um cenário simulado com parâmetros $\rho = 0{,}7$, $\sigma_b^2 = 4$, p = 10% e k = 6.



FONTE: A autora (2020)

Com base nos resultados simulados os índices de validação externa foram calculados

para avaliar como cada um dos métodos de agrupamentos consegue reproduzir o gabarito.

4.2 RESULTADOS DO ESTUDO POR SIMULAÇÃO

Os resultados das simulações foram analisados com base em cinco aspectos:

1. Métodos não espaciais *vs* Métodos espaciais;
2. CG *versus* Skater;
3. CGG *versus* CGV;
4. Efeitos de α nos resultados produzidos pelas duas varições do método CG.
5. Efeito de σ_b^2, ρ e p.

Cada um desses aspectos foi analisado para os três diferentes números de grupos e para cada um dos índices de validação separadamente. Os resultados mais próximos de um, para o índice de Rand ajustado, configuram uma maior concordância do gabarito e o método de agrupamentos analisado. Para o índice de variação da informação, resultados mais próximos de zero implicam em maior concordância entre o gabarito e o método de agrupamentos analisado. Todas as figuras seguem o mesmo padrão de cores e, por simplicidade, as abreviações K-means (KMS), Ward (WD), complete linkage (CPT) e Skater (SKT), além das já usadas ao longo do trabalho CGG e CGV, foram adotadas.

4.2.1 Resultados do estudo por simulação para k = 3

As FIGURAS 17, 18 e 19 referem-se aos resultados do estudo por simulação para o índice de Rand ajustado, considerando k = 3 grupos. As análises de reprodutibilidade dos gabaritos pelos métodos de agrupamentos, medidas por este índice de validação externa, são apresentadas na sequência:

i. p = 5%: Independentemente do valor de ρ com $\sigma_b^2 = 1$ e 4 os métodos espaciais apresentam melhores resultados em reproduzir os gabaritos do que os métodos não espaciais, sendo que CG tem desempenho superior que Skater e CGG tem resultados melhores comparados ao método CGV. Para as mesmas configurações só que com $\sigma_b^2 = 16$ os métodos espaciais também têm desempenho melhor que os não espaciais e, ainda, o método Skater tem melhor desempenho que CG. Finalmente, verifica-se que CGG produziu melhores resultados que CGV neste cenário.

ii. p = 10%: Independentemente do valor de ρ com $\sigma_b^2 = 1$ os métodos espaciais têm melhores resultados do que os métodos não espaciais, ou seja, os resultados dos métodos espaciais estão mais próximos de 1 (valor máximo para o índice de

Rand ajustado) que os métodos não espaciais. O método CG apresenta melhor reprodutibilidade dos gabaritos do que o método Skater. O método CGG tem desempenho superior ao CGV em termos de reprodutibilidade dos gabaritos. Para ρ = 0,5 com $\sigma_b^2 = 4$, ρ = 0,7 e 0,9 com $\sigma_b^2 = 4$ e 16 os métodos espaciais apresentam melhores desempenhos do que os métodos não espaciais e CGG tem melhores resultados em reproduzir os gabaritos que o Skater, embora CGV tenha resultados piores comparado ao Skater. Para $\rho = 0{,}5$ com $\sigma_b^2 = 16$ os métodos espaciais também apresentam melhores resultados do que os métodos não espaciais, sendo que o Skater apresenta resultados superiores ao CG e com CGG também superior ao CGV.

iii. p = 20%: Para $\rho = 0{,}5$ e 0,7 com $\sigma_b^2 = 1$ e 4 os métodos espaciais apresentam melhores resultados em termos de reprodutibilidade dos gabaritos em comparação aos métodos não espaciais, CG tem melhores resultados comparado ao método Skater e CGG apresenta desempenho melhor que CGV. Para $\rho = 0{,}5$ e 0,7 com $\sigma_b^2 = 16$ os métodos espaciais também possuem resultados melhores do que os métodos não espaciais, CGG com melhores resultados comparados ao método Skater. No entanto, para CGV ocorre ao contrário, já que este método apresenta valores menores em termos de reprodutibilidade dos gabaritos comparado ao método Skater e, consequentemente, CGG é melhor que o CGV. Para $\rho = 0{,}9$ com $\sigma_b^2 = 1$ o método CG apresenta resultados superiores aos métodos não espaciais e ao Skater. O método K-means tem resultados superiores ao método espacial Skater e entre as variações do método CG, CGG apresenta resultados superiores ao CGV. Com $\sigma_b^2 = 4$ o resultado é o mesmo que do $\sigma_b^2 = 1$, mas em vez do método K-means os resultados são melhores para o método Ward. Com $\sigma_b^2 = 16$ o método CG apresenta resultados melhores que os outros métodos de agrupamentos, CGG tem resultados melhores em comparação com CGV e Skater. O método não espacial K-means, neste cenário, apresenta desempenho superior ao método Skater.

iv. p = 33%: com $\sigma_b^2 = 1$, para todos os valores de ρ, o método Ward possui resultados melhores se comparados com o método Skater, enquanto que o método CG apresenta resultados superiores aos métodos de agrupamentos não espaciais, e também CGG é melhor que CGV. Para todos os valores de ρ com $\sigma_b^2 = 4$ e 16, os métodos não espaciais K-means e Ward são melhores que CGV e Skater, mas piores que CGG.

Na sequência são discutidos os efeitos de ρ, σ_b^2 e p, respectivamente, nos resultados da validação externa e os efeitos de α nos resultados dos métodos CGG e CGV:

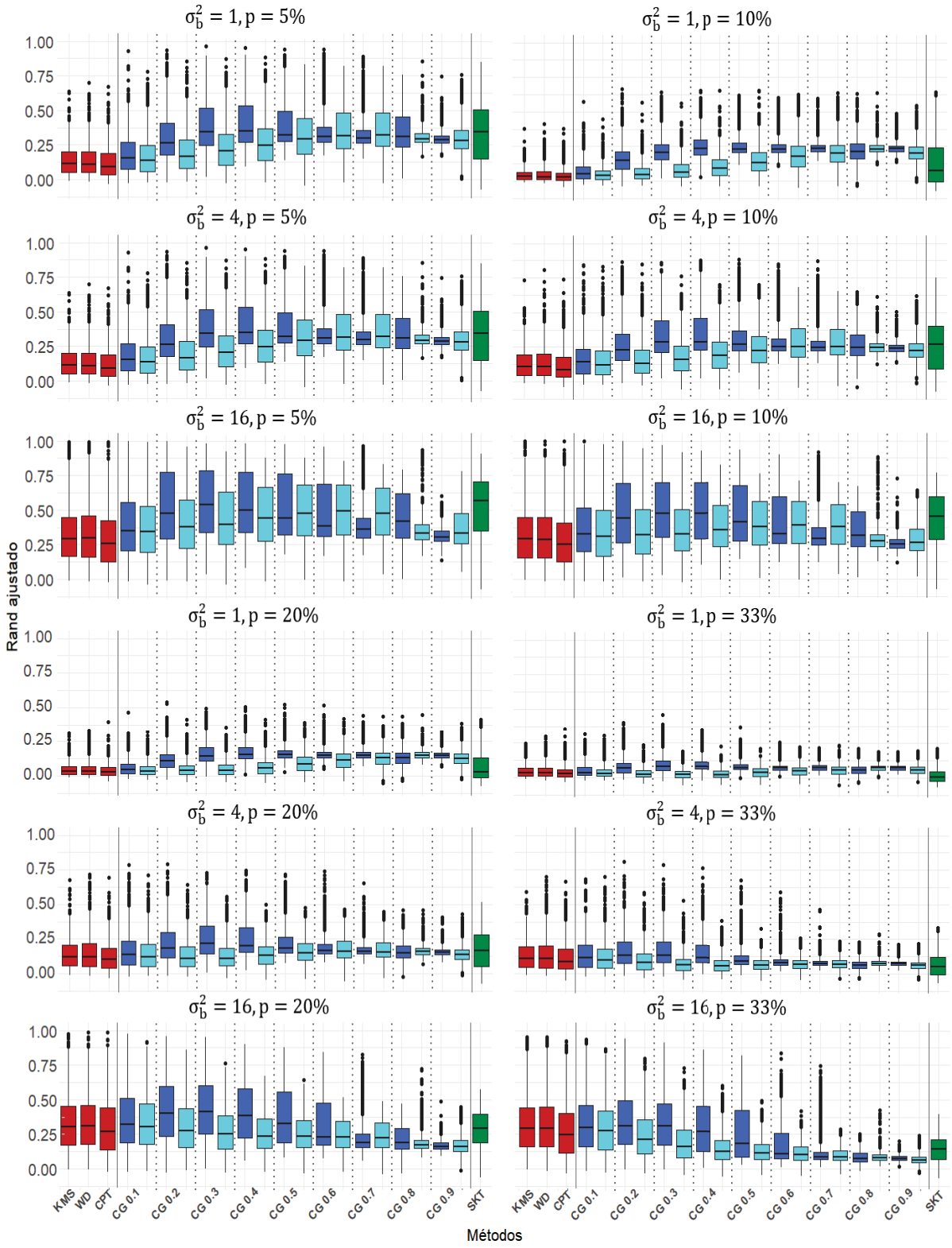
i. Os resultados simulados não apresentam grande variação para os diferentes valores de ρ.

ii. De maneira geral as variações para os diferentes valores de σ_b^2, apresentam mudanças não muito significativas em termo de resultados. Entretanto, pode-se notar que conforme o σ_b^2 aumenta os métodos não espaciais apresentam aumento de heterogeneidade intergrupos se comparados ao valores menores de σ_b^2.

iii. Para os valores até p = 20%, os resultados de reprodutibilidade do gabarito pelos métodos de agrupamentos não apresenta variações expressivas, sendo que os métodos espaciais se saem melhores do que os não espaciais. Entretanto para p = 33% métodos não espaciais apresentam melhores desempenhos ao serem comparados com o método Skater, mas não ao serem comparados com o método CG.

iv. O parâmetro α, que confere maior ou menor peso à matriz de distâncias baseadas no espaço geográfico no método CG, produz, na maior parte das vezes, melhores resultados (maiores valores do índice de Rand ajustado) para CGG quando α está em torno de 0,3, que seria o valor apropriado. O mesmo ocorre, na maior parte das vezes, para valores de α em torno de 0,6 no método CGV.

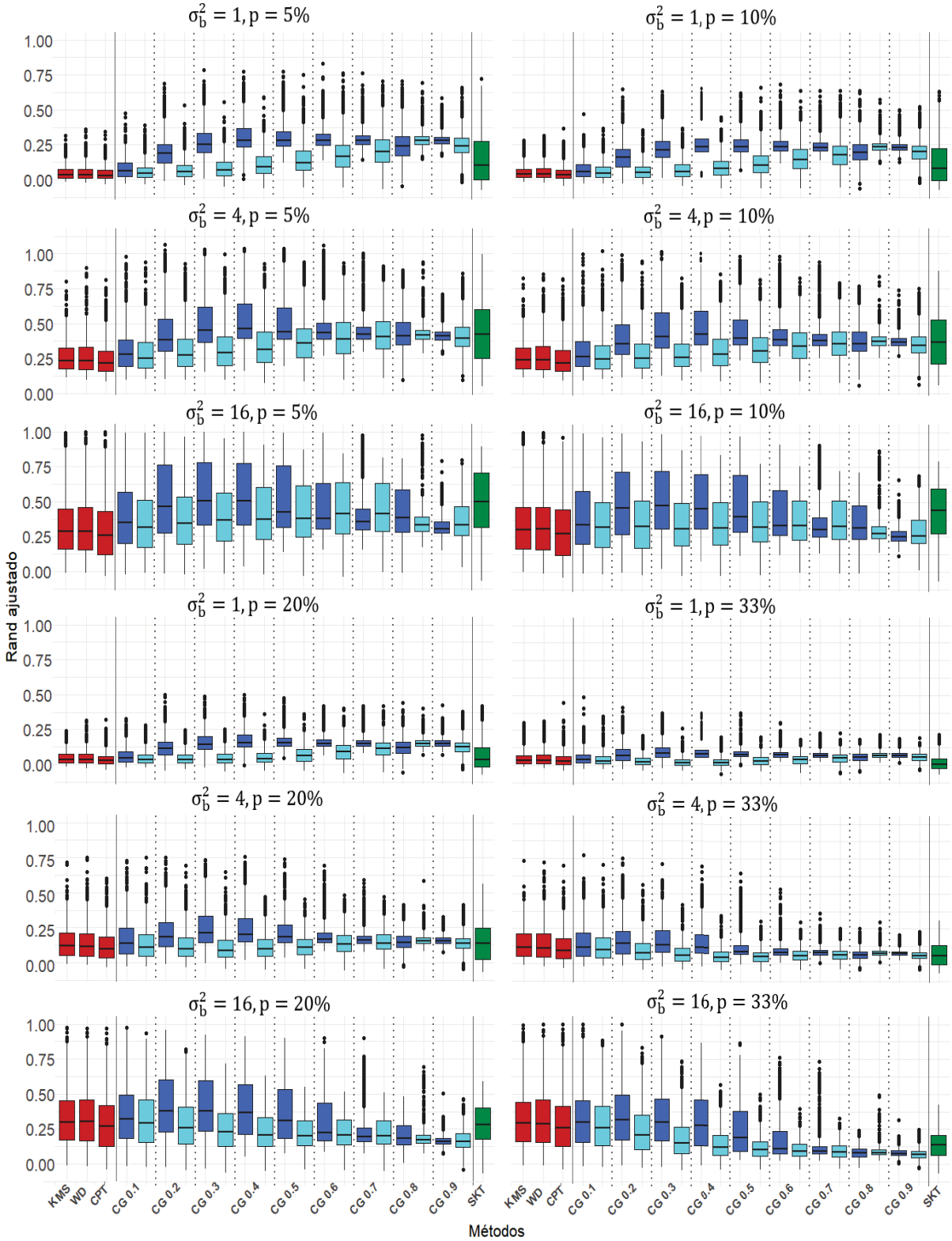
FIGURA 17 – Resultados do estudo por simulação para o índice de Rand ajustado com k=3 e $\rho = 0,5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

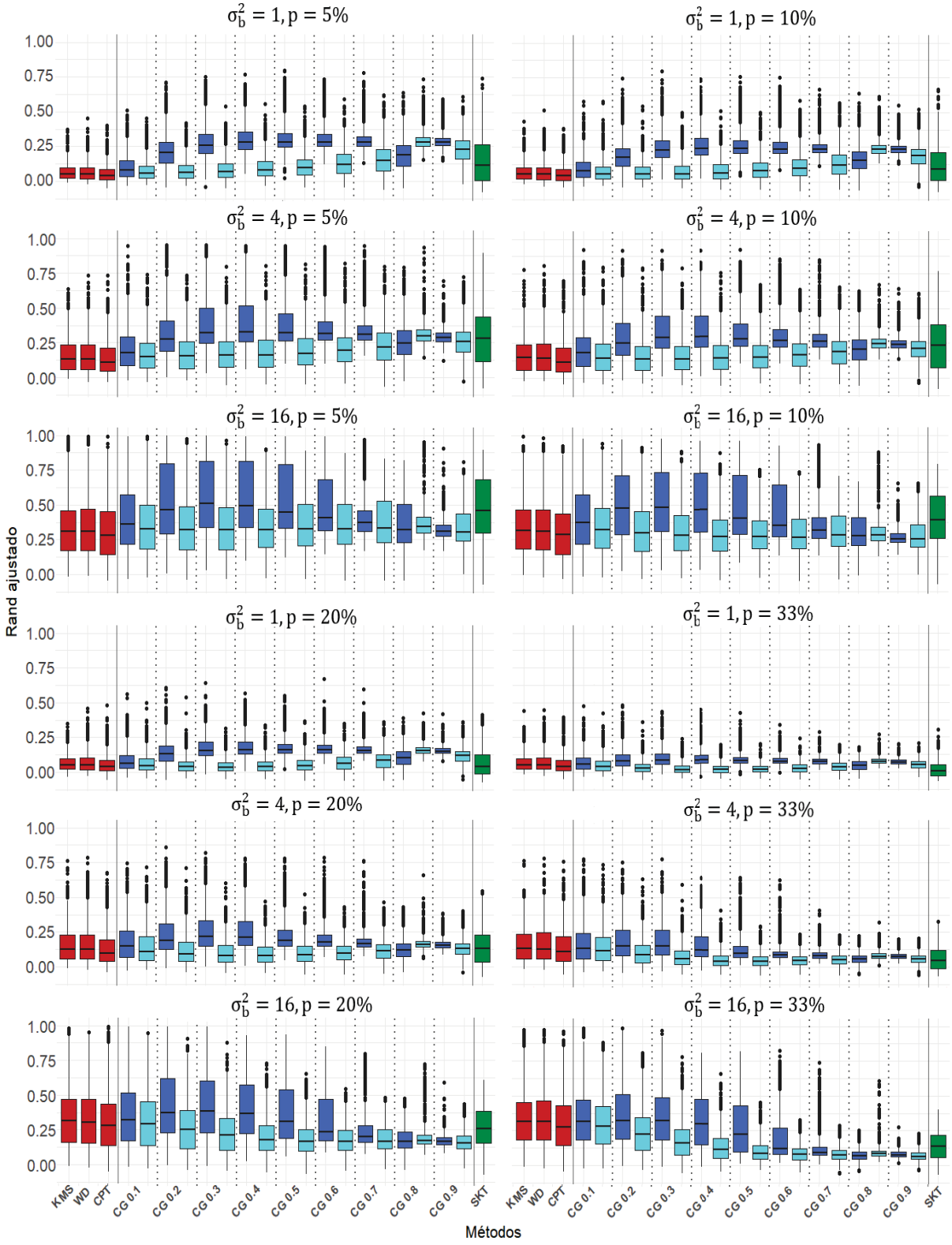
FIGURA 18 – Resultados do estudo por simulação para o índice de Rand ajustado com k=3 e $\rho = 0,7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

FIGURA 19 – Resultados do estudo por simulação para o índice de Rand ajustado com k=3 e $\rho = 0,9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

As FIGURAS 20, 21 e 22 referem-se aos resultados do estudo por simulação para o índice de variação da informação, considerando k = 3 grupos. As análises de reprodutibilidade dos gabaritos pelos métodos de agrupamentos, medidas por esse índice de validação, são apresentadas na sequência:

i. p = 5%: Os resultados dos métodos espaciais sempre apresentam performances superiores do que os métodos de agrupamentos não espaciais, ou seja, os resultados dos métodos de agrupamentos espaciais têm valores mais próximos de zero, indicando maior concordância desses métodos com o gabarito. Para $\rho = 0{,}5$ e 0,7, com todos os valores de σ_b^2 o método Skater apresenta resultados mais compatíveis com o gabarito que CG e, comparando CGG e CGV, os resultados do CGG são mais próximos do gabarito do que CGV. Para $\rho = 0{,}9$ com $\sigma_b^2 = 1$, CG apresenta resultados melhores que o Skater e CGG também tem desempenho superior ao CGV na reprodutibilidade do gabarito. Com $\sigma_b^2 = 4$ e 16, o método Skater apresenta pior desempenho em reproduzir o gabarito do que o método CGG, mas seu desempenho ainda é melhor se comparado com CGV.

ii. p = 10%: Para todos os cenários ao menos um dos métodos espaciais sempre apresentou resultados com maior proximidade ao gabarito, se comparado aos métodos não espaciais. Para $\rho = 0{,}5$ para todas as combinações de valores de σ_b^2 e para $\rho = 0{,}7$ e 0,9 com $\sigma_b^2 = 1$ e 4, o método Skater tem maior desempenho na reprodutibilidade do gabarito do que o método CG. Analisando apenas os métodos CG, CGG apresenta resultados melhores ao ser comparado com CGV. Para $\rho = 0{,}7$ e 0,9 com $\sigma_b^2 = 16$ os resultados apresentam o método complete linkage como mais eficaz que CGV para reproduzir o gabarito. Novamente, o desempenho dos resultados do método Skater são melhores que o CG e CGG tem melhor reprodutibilidade do gabarito ao ser comparado com CGV.

iii. p = 20%: Para $\rho = 0{,}5$ com $\sigma_b^2 = 16$ e $\rho = 0{,}7$ com $\sigma_b^2 = 1$ e 4 os métodos espaciais conseguem reproduzir melhor o gabarito do que os métodos não espaciais, com ressalva para o método complete linkage que apresenta melhor performance nos resultados para a variação da informação do que o método espacial CGV. Já para o método Skater, seus resultados apresentam-se melhores do que os resultados dos métodos CG. Comparando apenas CG, CGG tem resultados de reprodutibilidade superiores em relação a CGV. Para $\rho = 0{,}5$ com $\sigma_b^2 = 4$ e para $\rho = 0{,}9$ com $\sigma_b^2 = 1$ e 4 os resultados são semelhantes aos dos cenários anteriores, com exceção que o método complete linkage é superior em reprodutibilidade do gabarito que as duas variações do método CG e não apenas ao CGV como no cenário anterior. Para $\rho = 0{,}5$ com $\sigma_b^2 = 1$ o método Skater tem resultado melhor que CG e CGG com resultados superiores ao CGV na reprodutibilidade do gabarito. Para $\rho = 0{,}7$ com

$\sigma_b^2 = 16$ o método complete linkage apresenta resultados superiores comparado ao CGV. Os resultados do método CGG têm melhores performances ao ser comparados com os resultados métodos Skater e CGV. Para $\rho = 0{,}9$ com $\sigma_b^2 = 16$ os resultados do método Ward são melhores que CGV, em relação à reprodutibilidade do gabarito. Igualmente ao cenário anterior, o método CGG apresenta resultados superiores ao Skater e CGV.

iv. p = 33%: Para todas as combinações de ρ e σ_b^2, com exceção de $\sigma_b^2 = 16$ e $\rho = 0{,}5$, os resultados do método Skater são melhores, em termos de reprodutibilidade do gabarito do que os resultados dos métodos não espaciais e do método CG. Também, para esses cenários, o método complete linkage apresenta resultados melhores do que o método CG. A comparação apenas das variações do método CG resultam com CGG obtendo desempenho superior aos resultados do CGV. Para $\rho = 0{,}5$ com $\sigma_b^2 = 16$ o método CGV tem pior desempenho em reproduzir o gabarito que o método não espacial complete linkage, já para o método Skater os resultados retornados por esse método são superiores a CG. Ao comparar CGG e CGV o método CGV tem melhores resultados em reproduzir o gabarito.

Na sequência são discutidos os efeitos de ρ, σ_b^2 e p, respectivamente, nos resultados da validação externa e os efeitos de α nos resultados dos métodos CGG e CGV:

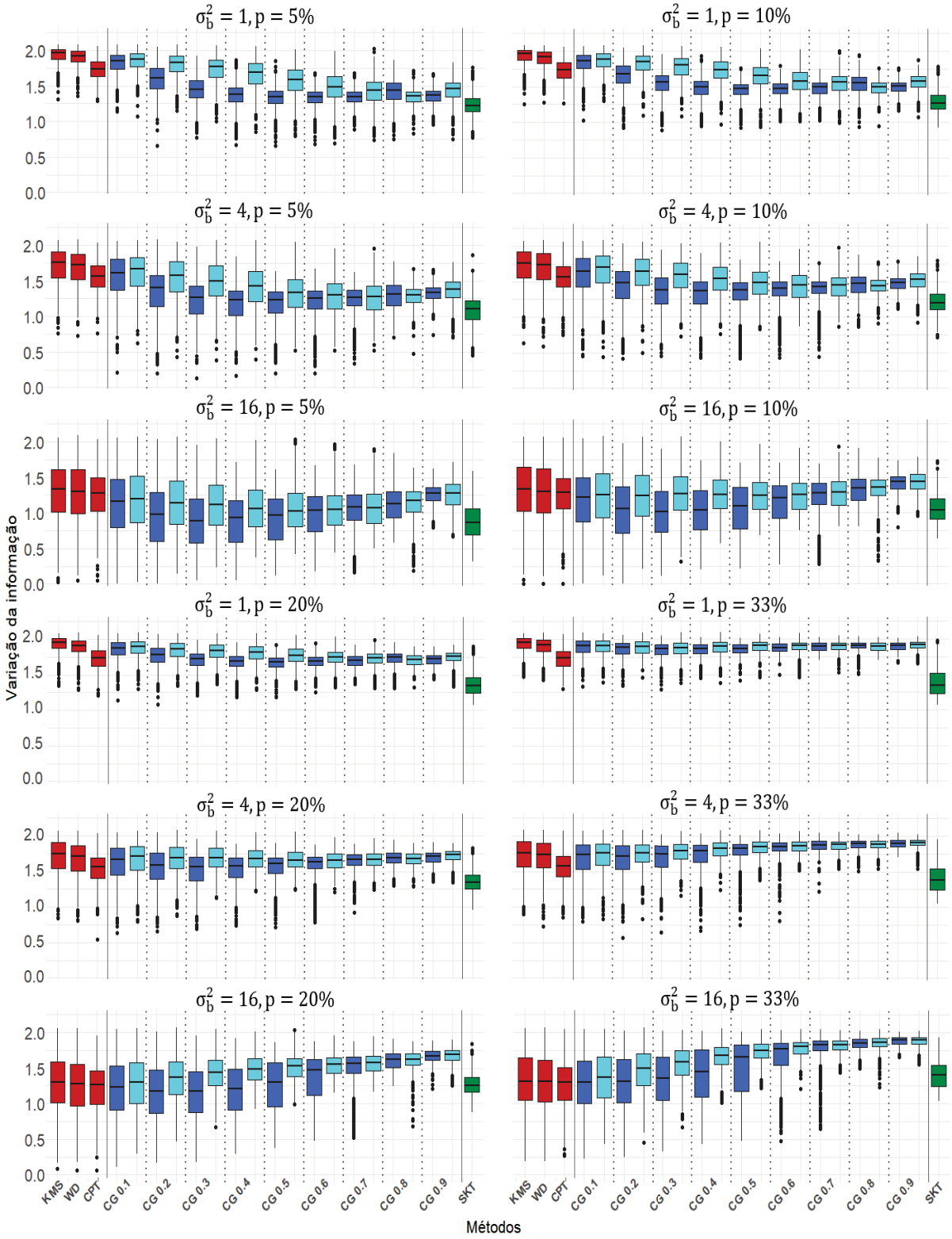
i. Ao fixar os valores desse parâmetro os resultados não apresentam mudanças de reprodutibilidade dos gabaritos em relação aos diferentes níveis de variabilidade intragrupos.

ii. Os resultados do índice de variação da informação não apresentam mudanças causadas pela autocorrelação espacial, em termos de reprodutibilidade do gabarito, para os diferentes valores de ρ.

iii. Na maior parte das vezes, os resultados para o índice de variação da informação com p até 20% não apresentam mudanças significativas em termos da reprodutibilidade do gabarito para os métodos analisados. Quando p = 33% esse parâmetro apresenta resultados em que os métodos não espaciais reproduzem melhor o gabarito do que o método espacial CG.

iv. Para o método CGG o valor de $\alpha = 0,4$ e para CGV o valor de $\alpha = 0,7$ são os que retornam melhores resultados em reproduzir os gabaritos, logo esses valores são os que melhor conferem ponderação, na medida do possível, em produzir grupos homogêneos quanto à variável resposta. Para esses dois métodos os valores α introduzem melhorias, em relação à homogeneidade espacial, e a partir de um certo valor, que é o valor de α considerado o mais adequado, os efeitos introduzidos nos métodos por α estabilizam.

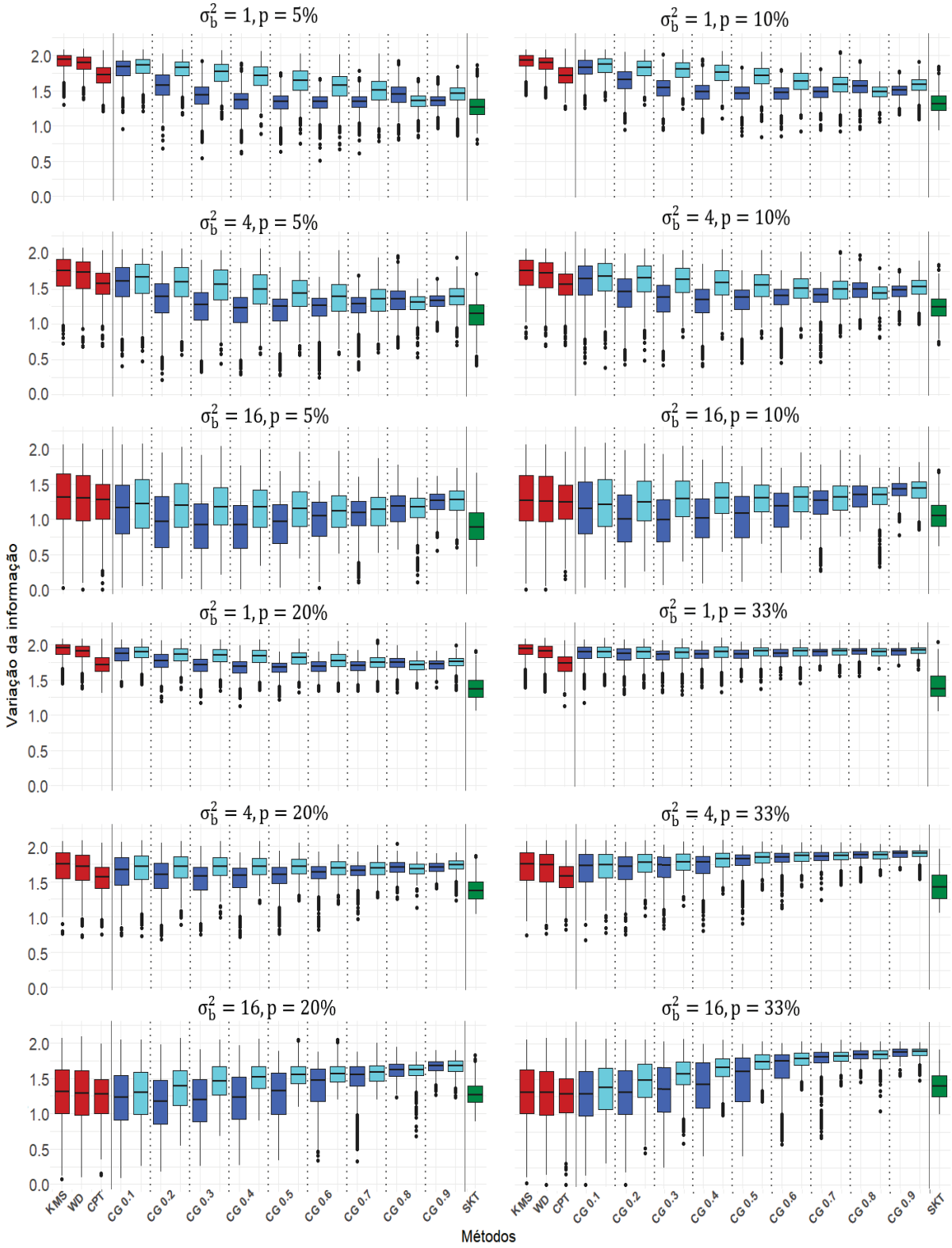
FIGURA 20 – Resultados do estudo por simulação para o índice de variação da informação com k=3 e $\rho = 0{,}5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

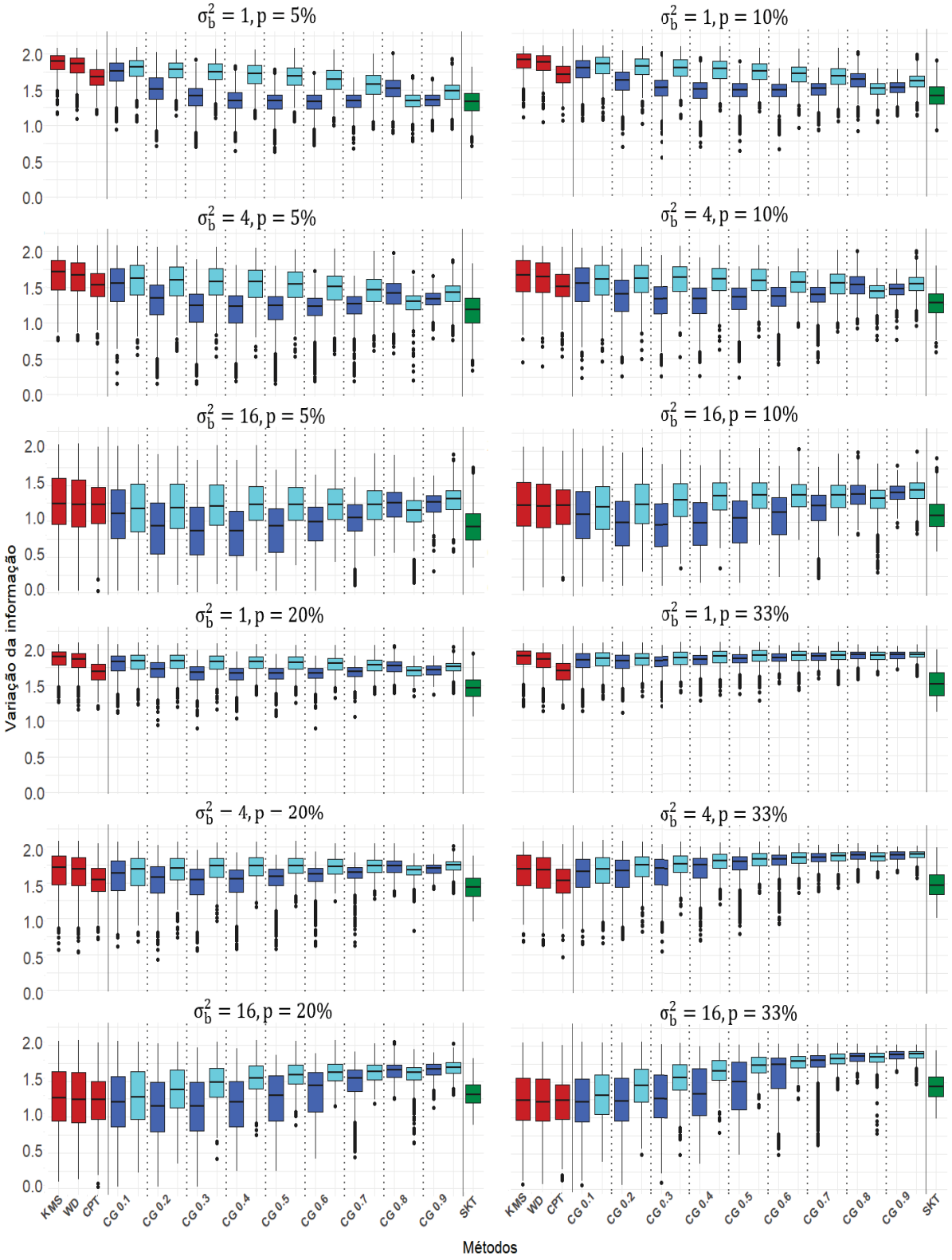
FIGURA 21 – Resultados do estudo por simulação para o índice de variação da informação com k=3 e $\rho = 0{,}7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

FIGURA 22 – Resultados do estudo por simulação para o índice de variação da informação com k=3 e $\rho = 0{,}9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

4.2.2 Resultados do estudo por simulação para k = 6

As FIGURAS 23, 24 e 25 referem-se aos resultados do estudo por simulação para o índice de Rand ajustado, considerando k = 6 grupos. As análises de reprodutibilidade dos gabaritos pelos métodos de agrupamentos, medidas por esse índice de validação, são apresentadas na sequência:

i. p = 5%, 10% e 20%: Para todas as combinações de ρ e σ_b^2 os resultados são semelhantes, com os métodos espaciais apresentando melhor performance do que os não espaciais, o método CG superando o método Skater e com CGG melhor do que CGV para reproduzir os gabaritos.

ii. p = 33%: Em todos os cenários o método CGG reproduz melhor o gabarito em comparação a todos os outros métodos de agrupamentos. Para todos os valores de ρ com $\sigma_b^2 = 1$ e para $\rho = 0{,}5$ com $\sigma_b^2 = 4$ o método CG apresenta resultados melhores do que o método Skater e CGG apresenta melhores resultados do que o método CGV em termos de reprodutibilidade dos gabaritos. Para os demais cenários, os métodos não espaciais Ward ($\rho = 0{,}7$ com $\sigma_b^2 = 4$) e complete linkage ($\rho = 0{,}7$ e $0{,}9$ com $\sigma_b^2 = 16$) apresentam desempenho superior ao método Skater.

Na sequência são discutidos os efeitos de ρ, σ_b^2 e p, respectivamente, nos resultados da validação externa:

i. Os resultados simulados não apresentam variações significativas para os diferentes valores de ρ.

ii. Diferentemente do que ocorre para os outros valores de σ_b^2, para $\sigma_b^2 = 16$ os resultados apresentam melhor reprodutibilidade do gabarito para o método Skater em comparação ao método CGV. De maneira geral, as variações dos resultados produzidos pelo índice de Rand ajustado para os outros valores de σ_b^2, apresentam mudanças não muito notáveis.

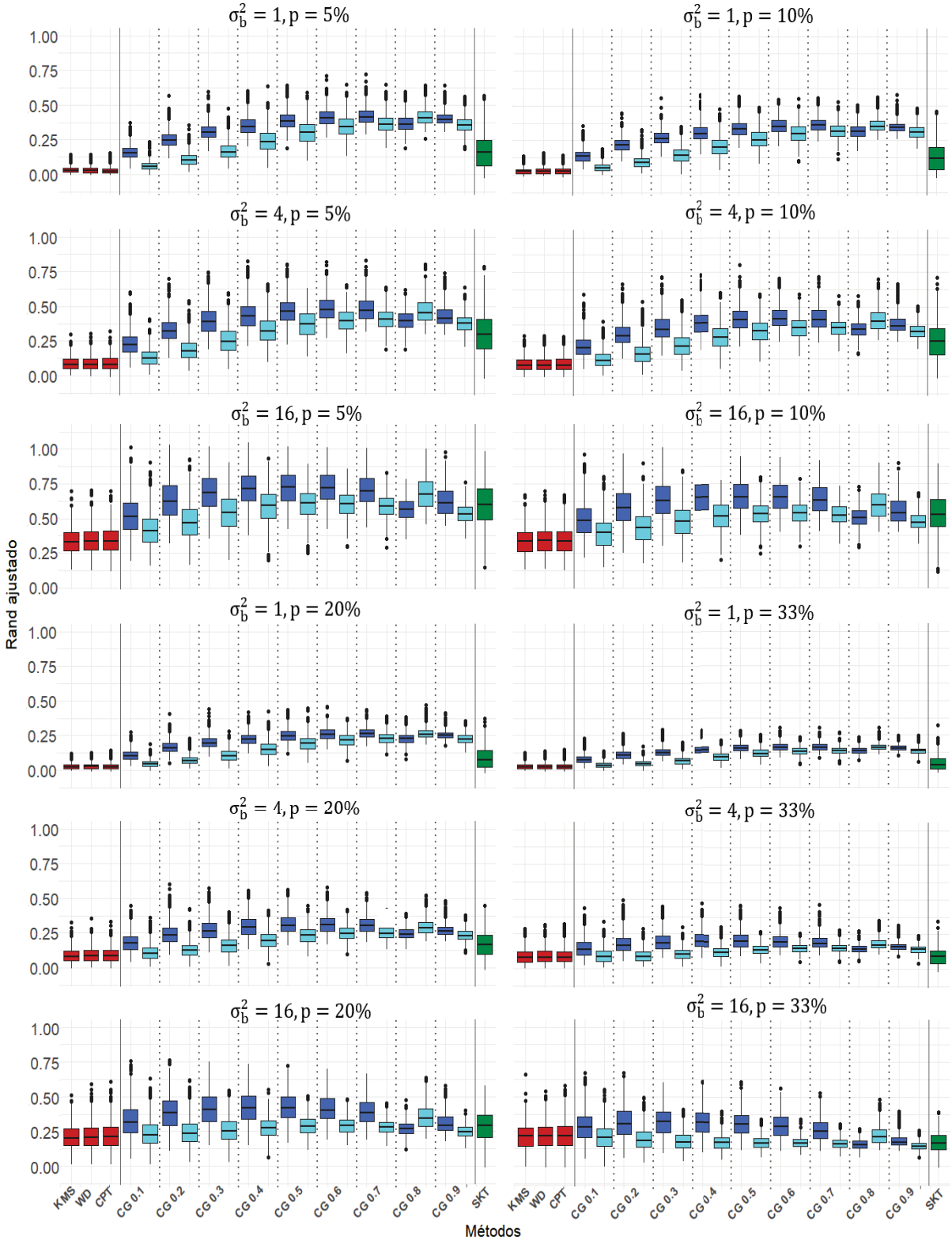
iii. Os resultados simulados não apresentam variações notáveis para os diferentes valores de p.

Dada a similaridade em que são apresentados os resultados para diferentes valores de α para o índice de Rand ajustado apresentados nas FIGURAS 23, 24 e 25 e para o índice de variação da informação apresentados nas FIGURAS 26, 27 e 28, a análise do efeito de α para CGG e CGV será feita conjuntamente na sequência:

i. Para k = 6, de maneira geral, o valor de α adequado está em torno de 0,6 para CGG e 0,8 para CGV. Os valores α produzem resultados muito semelhantes quanto

à validação externa quando p = 5% e 10%. Para p = 20% e 33%, os valores de α são menores do que para as configurações com p = 5% e 10%, mas essas diferenças não são expressivas.

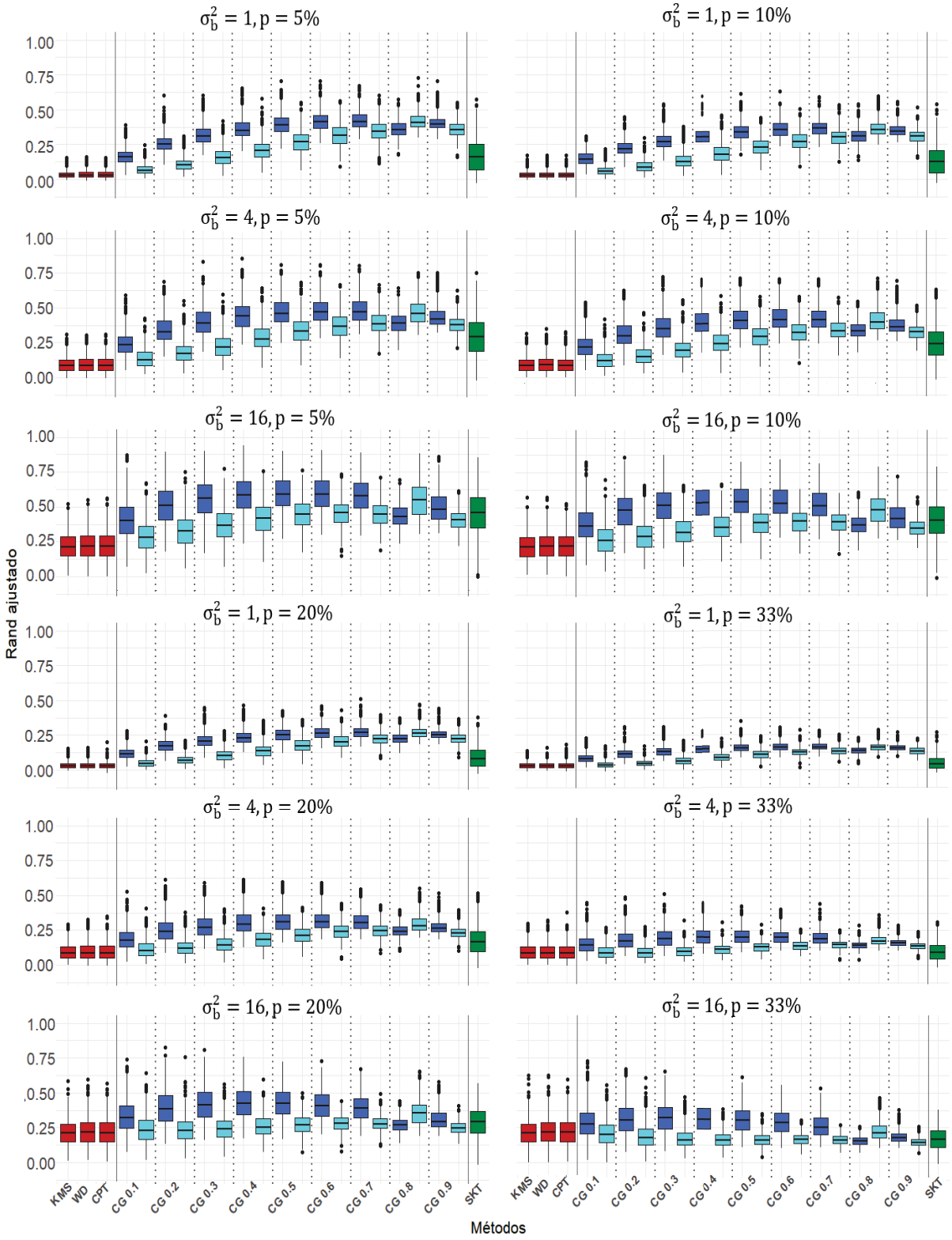
FIGURA 23 – Resultados do estudo por simulação para o índice de Rand ajustado com k=6 e $\rho = 0{,}5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

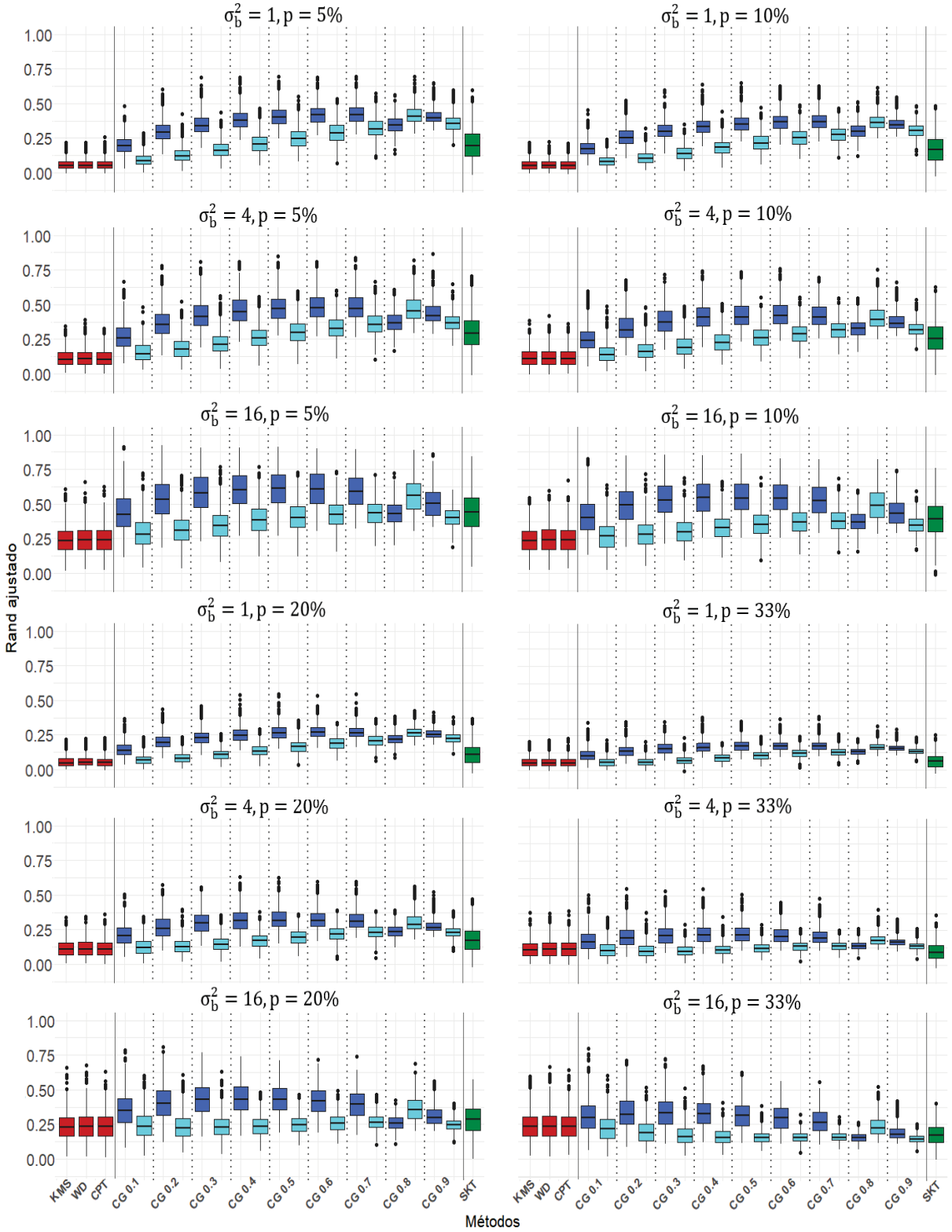
FIGURA 24 – Resultados do estudo por simulação para o índice de Rand ajustado com k=6 e $\rho = 0{,}7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

FIGURA 25 – Resultados do estudo por simulação para o índice de Rand ajustado com k=6 e $\rho = 0{,}9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

As FIGURAS 26, 27 e 28 referem-se aos resultados do estudo por simulação para o índice de variação da informação, considerando k = 6 grupos. As análises de reprodutibilidade dos gabaritos pelos métodos de agrupamentos, medidas por esse índice de validação, são apresentadas na sequência:

i. p = 5%: Para todos as combinações de valores de ρ e σ_b^2 os resultados são sempre os métodos espaciais melhores do que os métodos não espaciais, ou seja, os valores dos métodos espacias para o índice de variação da informação são mais próximos de zero que os métodos não espaciais. O método CG melhor do que o método Skater e CGG melhor que CGV para reproduzir os gabaritos.

ii. p = 10%: Para todos os valores de ρ com $\sigma_b^2 = 1$ e 4 os resultados sempre indicam os métodos espaciais melhores do que os não espaciais, o método CG com resultados superiores que o método Skater e CGG superior do que CGV em termos de reprodutibilidade dos gabaritos. Com $\sigma_b^2 = 16$ os resultados são similares aos relatados para os demais valores de σ_b^2. No entanto, o método Skater tem desempenho superior do que CGV e inferior que CGG em termos de reprodutibilidade dos gabaritos.

iii. p = 20%: Para $\rho = 0{,}5$ com todos os valores de σ_b^2, para $\rho = 0{,}7$ com $\sigma_b^2 = 1$ e 4 e para $\rho = 0{,}9$ com $\sigma_b^2 = 1$, os resultados indicam que os métodos espaciais tem desempenho superior do que os métodos não espaciais. O método Skater têm resultados mais próximos de zero que CG. Por fim, CGG tem desempenho elevado em termos de reprodutibilidade dos gabaritos do que CGV. Para $\rho = 0{,}7$ com $\sigma_b^2 = 16$ e para $\rho = 0{,}9$ com $\sigma_b^2 = 4$ e 16, ocorre o mesmo em relação ao relatado anteriormente. No entanto, o método CGG apresenta resultados superiores ao Skater, mas o mesmo não ocorre para o método CGV, que tem resultados inferiores ao método Skater.

iv. p = 33%: Para todos os valores de ρ com $\sigma_b^2 = 1$ e 4, os resultados do método complete linkage apontam maior reprodutibilidade do gabarito do que o método CGV. Considerando os demais métodos, o método Skater possui resultados mais satisfatórios nesses cenários. Com $\sigma_b^2 = 16$ os métodos não espaciais, em particular o método complete linkage tem melhor desempenho do que os demais métodos em reproduzir o gabarito e, entre os métodos espaciais, CGG tem melhor performance do que o método Skater.

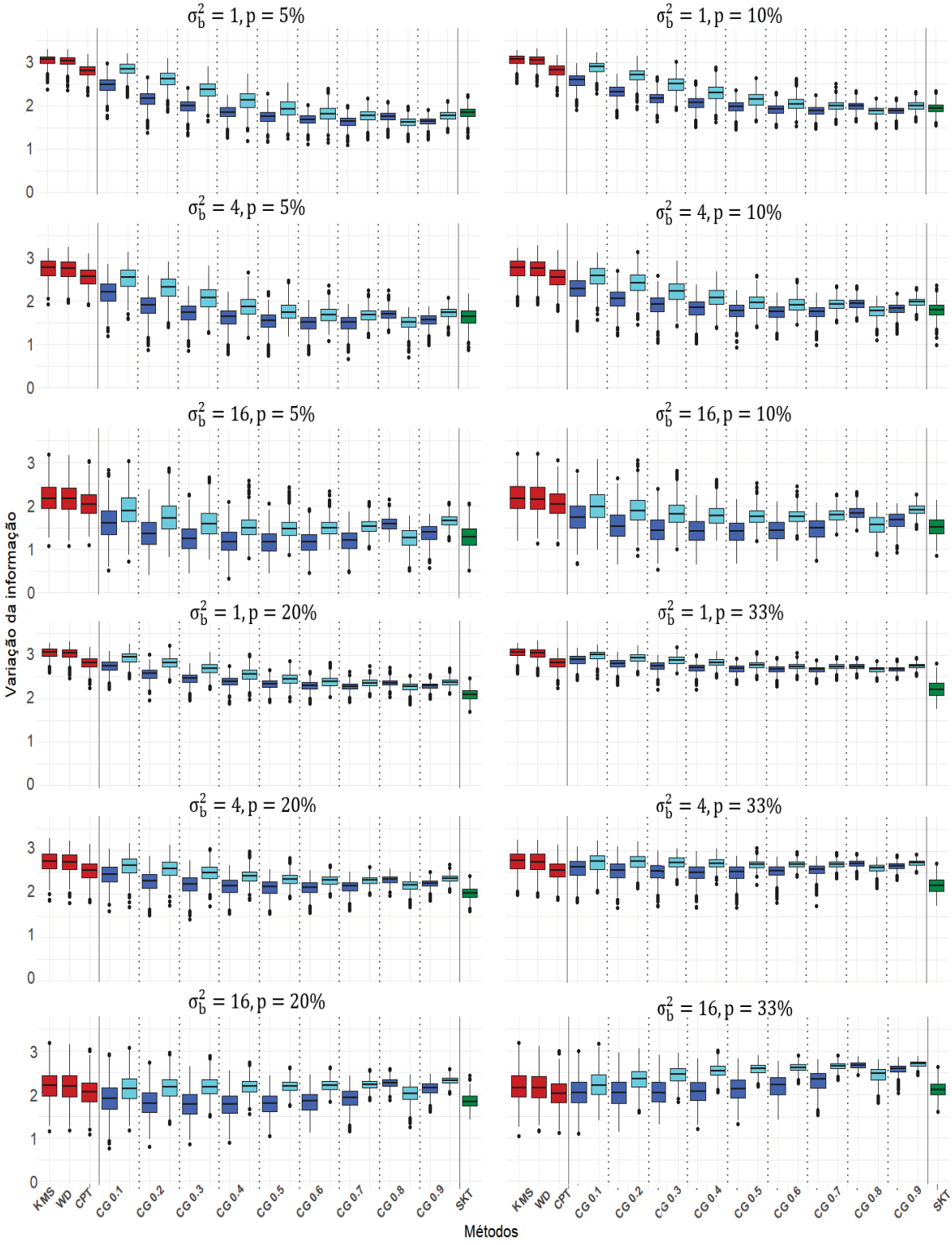
Na sequência são discutidos os efeitos de ρ, σ_b^2 e p, respectivamente, nos resultados da validação externa:

i. Para $\rho = 0{,}9$, os resultados apresentam o método Skater com desempenho superior ao ser comparado com o método CG, não ocorrendo o mesmo para os valores de $\rho =$ 0,5 e 0,7.

ii. Os métodos CGG e Skater se alternam como sendo os melhores em reproduzir os agrupamentos do gabarito para $\sigma_b^2 = 1$ e 4. Para $\sigma_b^2 = 16$, quando o p = 33% indicou o método não espacial complete linkage com resultados mais próximos de zero, indicando maior reprodutibilidade do gabarito, do que os métodos espaciais.

iii. Para p = 5% e 10% os resultados do índice de variação da informação não apresentam variações para os diferentes métodos de agrupamentos. Para p = 20%, o método Skater, e 33%, o método complete linkage, apresentam desempenhos mais satisfatórios do que o método CG.

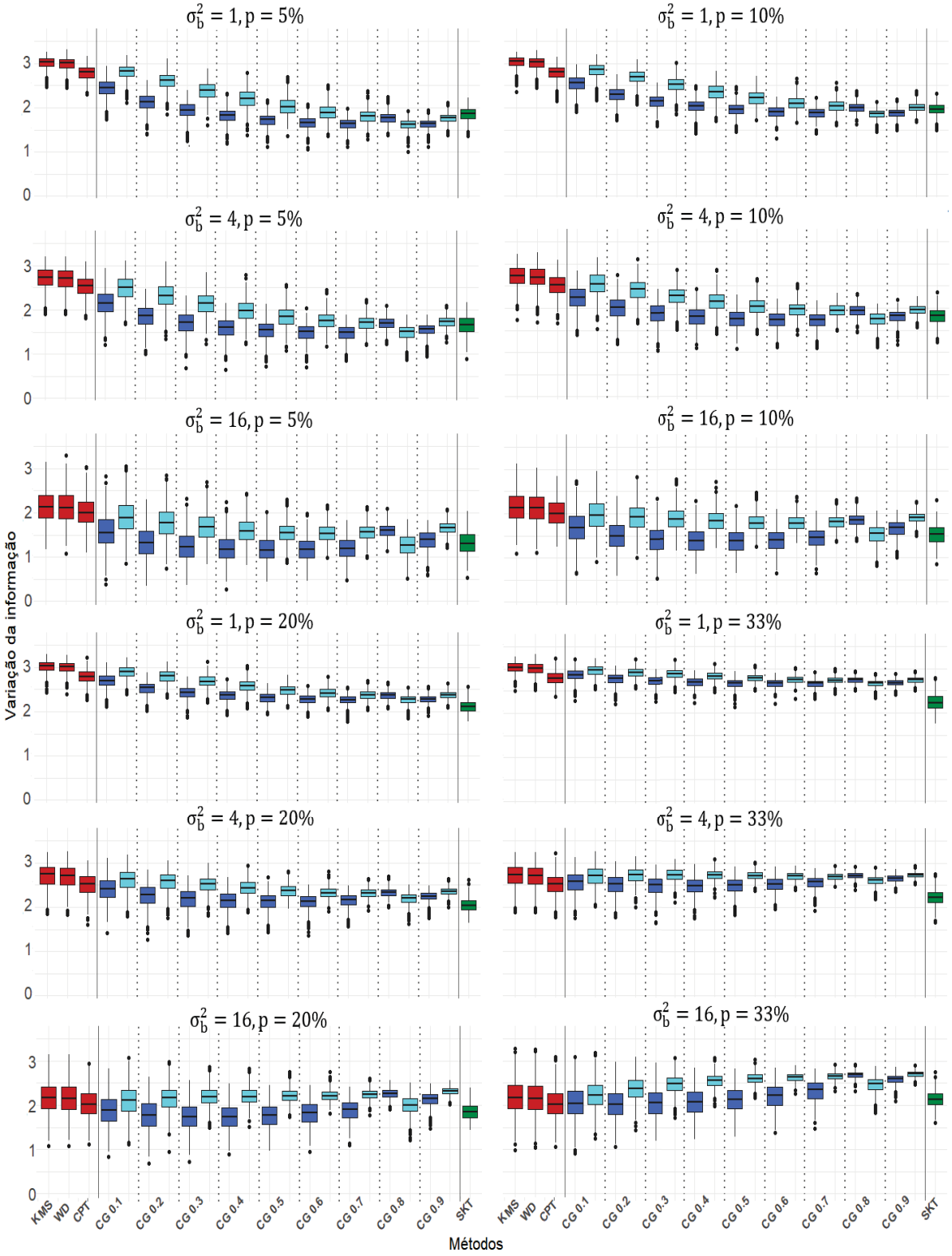
FIGURA 26 – Resultados do estudo por simulação para o índice de variação da informação com k=6 e $\rho = 0{,}5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

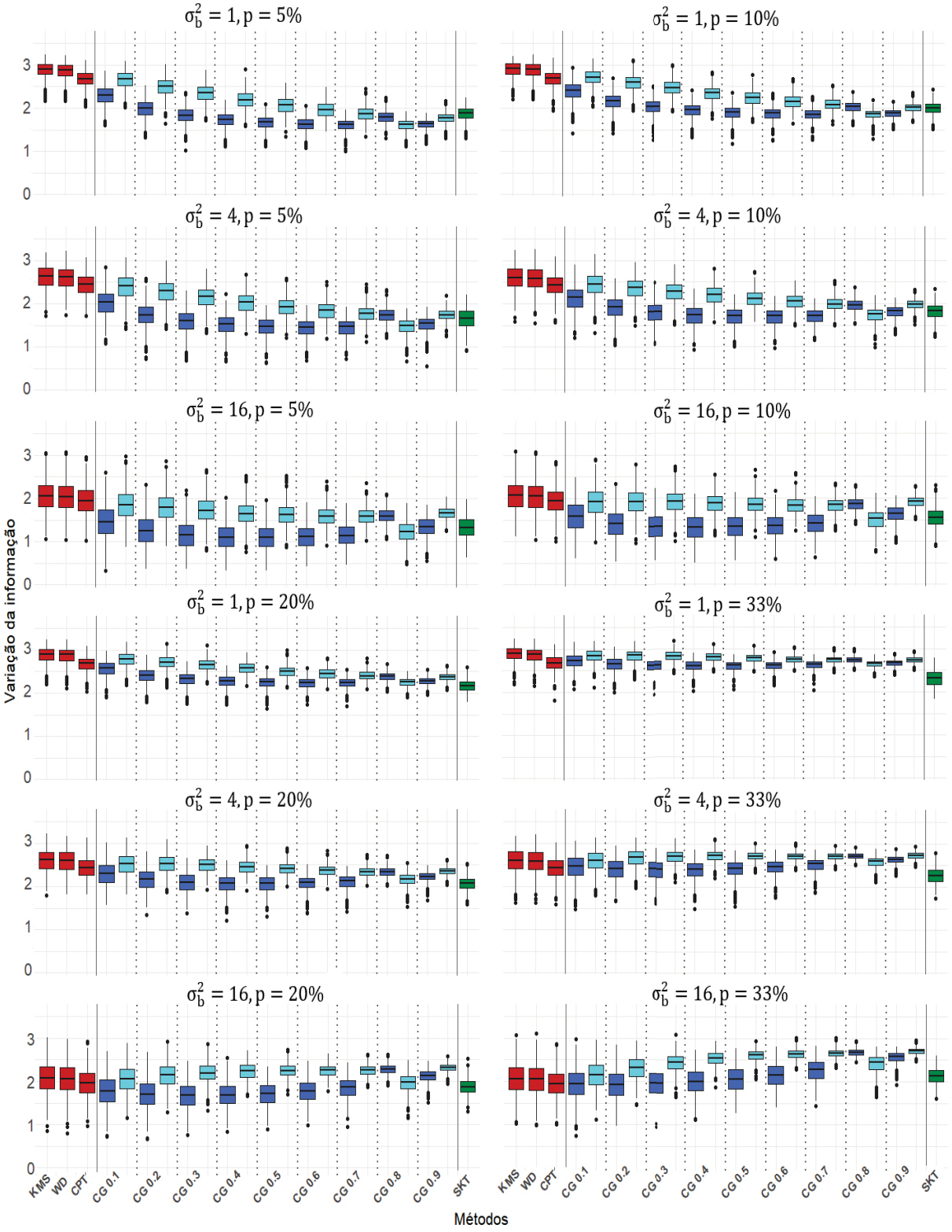
FIGURA 27 – Resultados do estudo por simulação para o índice de variação da informação com k=6 e $\rho = 0{,}7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

FIGURA 28 – Resultados do estudo por simulação para o índice de variação da informação com k=6 e $\rho = 0{,}9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

4.2.3 Resultados do estudo por simulação para k = 10

As FIGURAS 29, 30 e 31 referem-se aos resultados do estudo por simulação para o índice de Rand ajustado e as FIGURAS 32, 33 e 34 referem-se aos resultados do estudo por simulação para o índice de variação da informação, considerando k = 10 grupos. As análises de reprodutibilidade dos gabaritos pelos métodos de agrupamentos, medidos por esses dois índices de validação, são apresentados em conjunto na sequência:

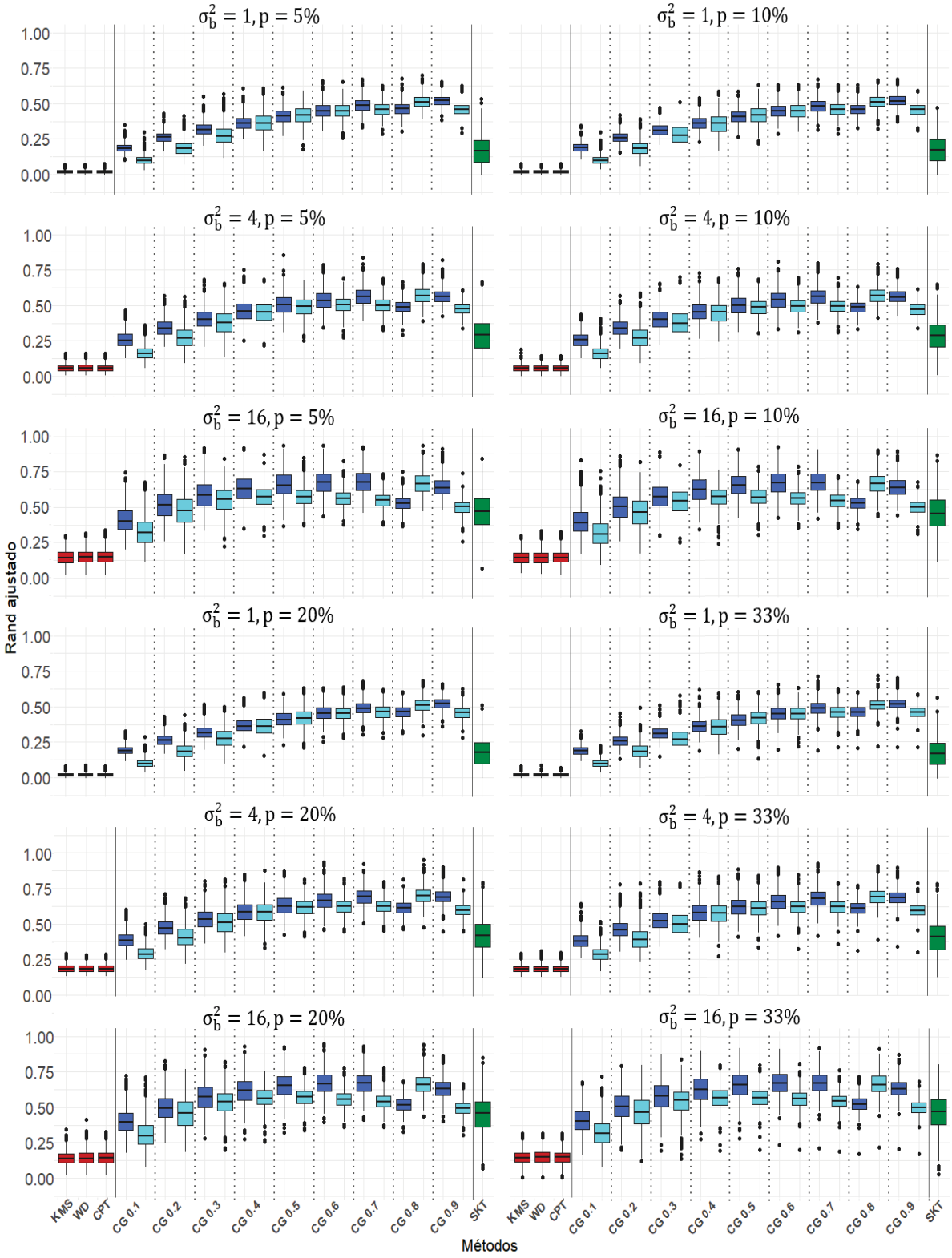
i. Os métodos espaciais sempre têm desempenho superior aos métodos não espaciais em termos de reprodutibilidade dos gabaritos pelos métodos de agrupamentos. O método CG sempre tem resultados superiores ao método Skater, tanto para o índice de Rand ajustado quanto para o índice de variação da informação, para todos os cenários simulados.

ii. O método CGG, quanto aos resultados para o índice de Rand ajustado com σ_b^2 = 1 e 16, tem melhores resultados se comparados ao método CGV. Para $\sigma_b^2 = 4$ ocorre o contrário, ou seja, o método CGV apresenta resultados superiores que o método CGG em termos da reprodutibilidade do gabarito. Para o índice de variação da informação, os resultados do método CGG são melhores que CGV em todos os cenários simulados.

iii. Analisando os efeitos dos parâmetros simulados para a reprodutibilidade do gabarito pelos diferentes métodos de agrupamentos, os parâmetros ρ e p não apresentam, na maioria das vezes, variações expressivas nos resultados. Para o índice de Rand ajustado σ_b^2 também não apresenta variações nos resultados para os diferentes valores desse parâmetro. Para o índice de variação da informação o método Skater apresenta resultados superiores com $\sigma_b^2 = 16$ do que o método CGV, no entanto isso não ocorre para os outros valores de σ_b^2.

iv. Para CGV e CGV, analisando o efeito de α, tanto para o índice de Rand ajustado quanto para o índice de variação da informação, para todos os valores de ρ com σ_b^2 = 1 e 4 e todos os valores p os resultados não apresentam alterações nesses cenários, em termos de reprodutibilidade dos gabaritos. Quando $\sigma_b^2 = 16$, os cenários apresentam melhores resultados para valores de α menores em relação aos cenários anteriores. No entanto, na maior parte da vezes, $\alpha = 0{,}8$ para as duas variações do método CG, é a escolha mais adequada se comparado aos outros valores desse parâmetro. Logo esse valor é aquele que, em geral, confere melhor ponderação quanto à produção de grupos homogêneos para a variável resposta.

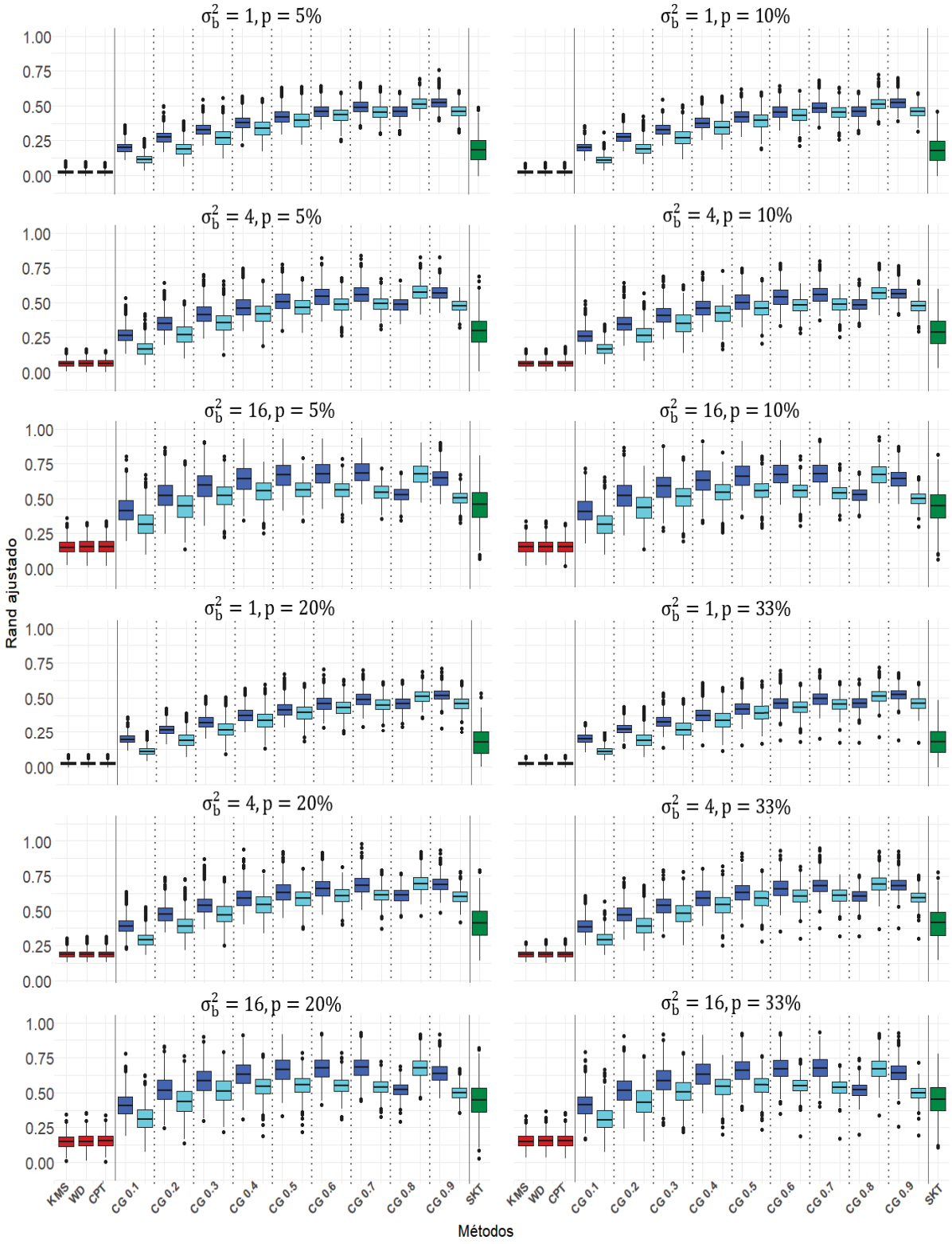
FIGURA 29 – Resultados do estudo por simulação para o índice de Rand ajustado com k=10 e $\rho = 0{,}5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

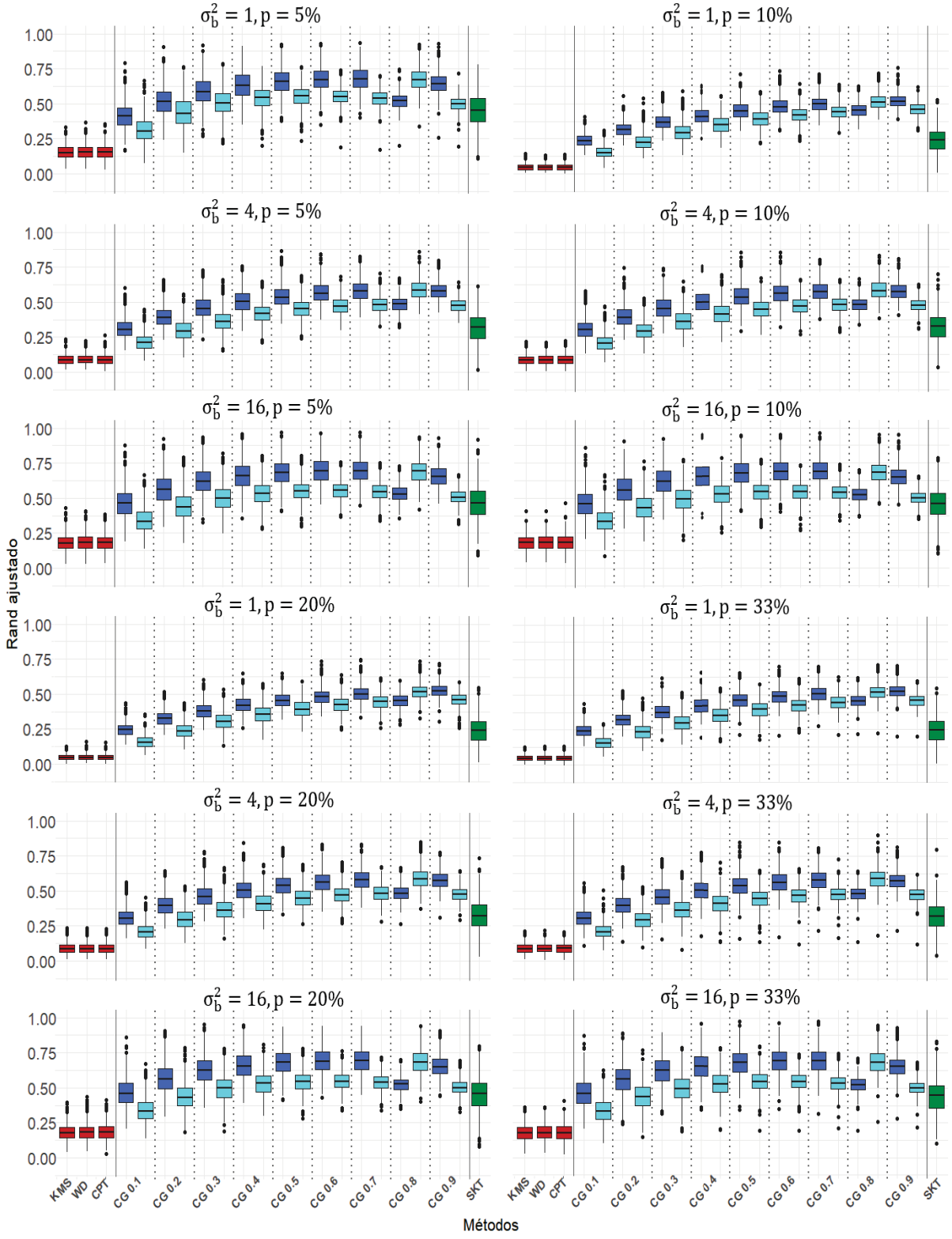
FIGURA 30 – Resultados do estudo por simulação para o índice de Rand ajustado com k=10 e $\rho = 0{,}7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

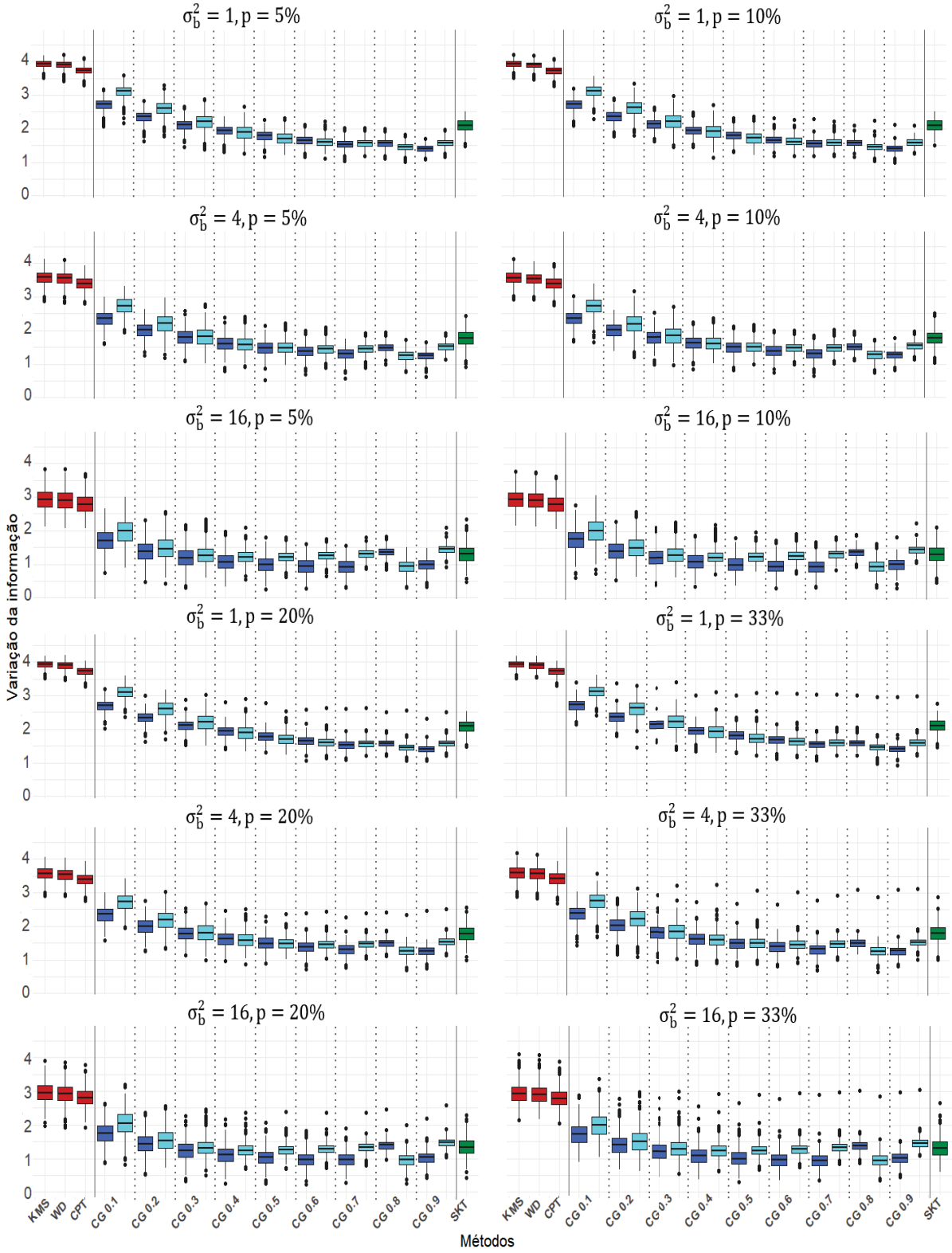
FIGURA 31 – Resultados do estudo por simulação para o índice de Rand ajustado com k=10 e $\rho = 0{,}9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

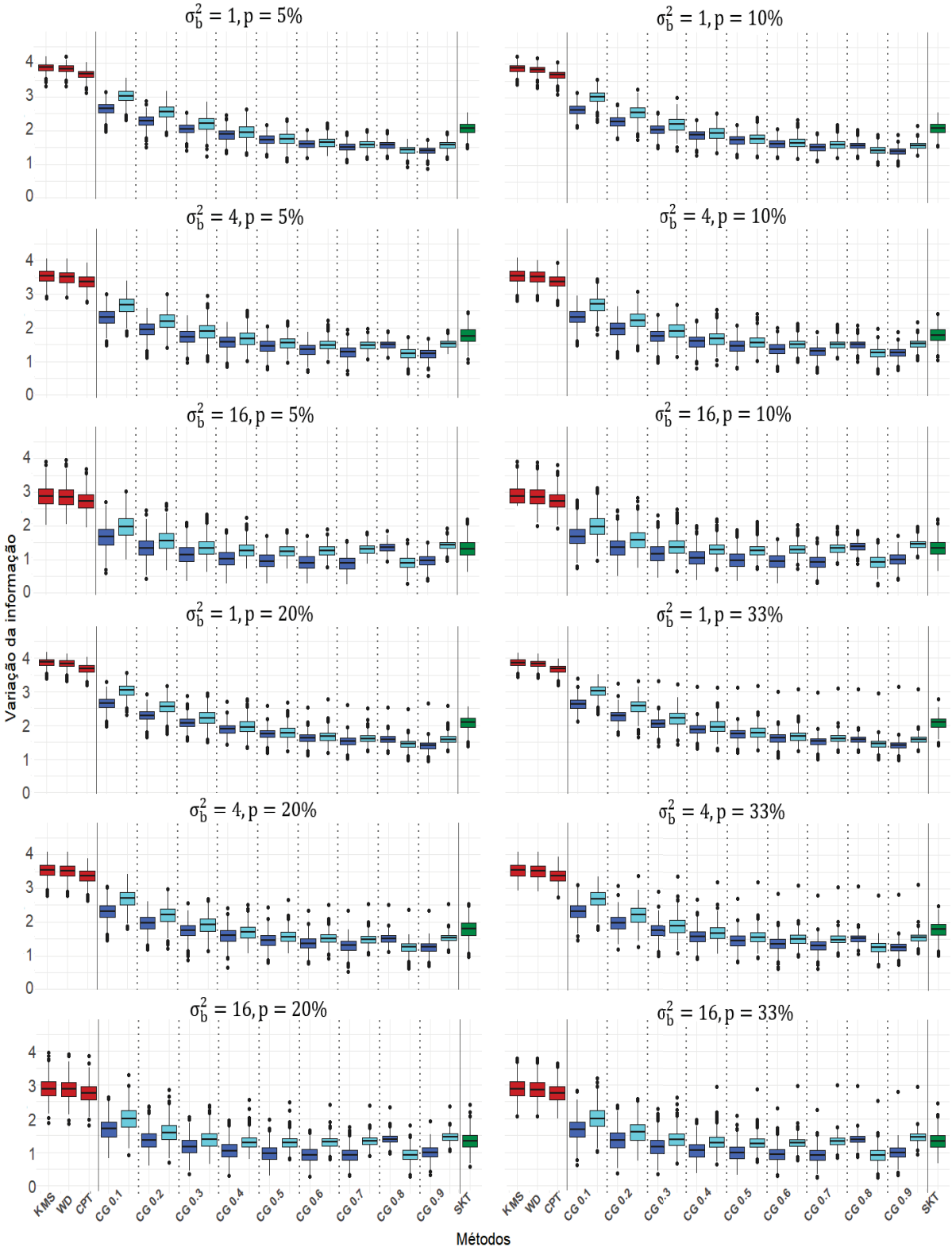
FIGURA 32 – Resultados do estudo por simulação para o índice de variação da informação com k=10 e $\rho = 0{,}5$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

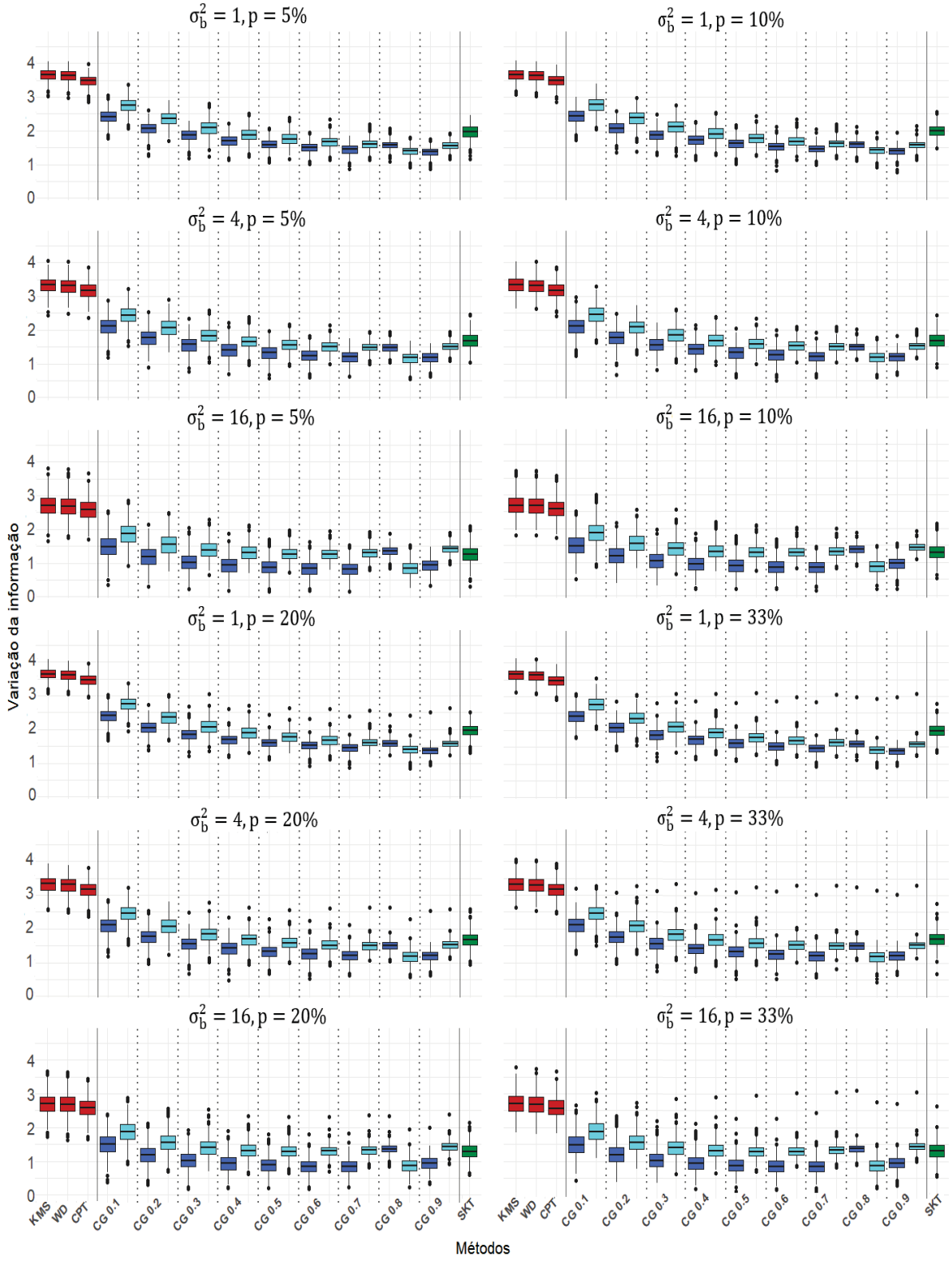
FIGURA 33 – Resultados do estudo por simulação para o índice de variação da informação com k=10 e $\rho = 0{,}7$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

FIGURA 34 – Resultados do estudo por simulação para o índice de variação da informação para k=10 e $\rho = 0{,}9$**



FONTE: A autora (2020)

** CG 0.1 corresponde ao método CG para $\alpha = 0.1$ e o mesmo para os demais valores de α. Os box plots em azul escuro correspondem ao método CGG e os em azul claro ao método CGV.

4.2.4 Comparação dos resultados para os diferentes números de grupos (k)

Para todos os diferentes valores de k a comparação de métodos espaciais *versus* métodos não espaciais resulta em ao menos um dos métodos espaciais obtendo resultados superiores aos métodos não espaciais. Portanto, ao menos um dos métodos espaciais reproduz de maneira mais eficaz os gabaritos gerados para os 108 diferentes cenários simulados.

Pode-se perceber também que à medida que o número de grupos aumenta os índices de validação passam a apresentar resultados de reprodutibilidade mais semelhantes. Essa semelhança ocorre, como explicado pelos autores Hoef e Warrens (2019), pelo fato de que os índices de validação apresentam dificuldade quando aplicados a partições que resultam em grupos pequenos e de tamanhos heterogêneos.

Em todos os cenários, o método CGV apresenta valores adequados de α maiores que os valores adequados de α para o método CGG, com exceção de k = 10, para o qual os valores de α para as duas variações do método CG são semelhantes. O método CGG também tem os resultados superiores, em termos de reprodutibilidade dos gabaritos, para todos os valores de k, do que o método CGV.

5 CONCLUSÃO

Considerando que a análise de agrupamentos é uma técnica bastante utilizada para analisar dados e com vasto campo de aplicação. Por conta disso, o existem diversos métodos de agrupamentos, logo a escolha adequada do método a ser utilizado é uma das questões principais para uma análise de agrupamentos mais coerente com os dados. Por este motivo e também pela maior disponibilidade de dados espaciais, os métodos com restrições espaciais ganham maior destaque e interesse do que os métodos não espaciais, hierárquicos e não hierárquicos, para agrupar esse tipo de dados.

Com esse trabalho fomentamos o entendimento dos métodos de agrupamento por meio de dois estudos. O estudo de caso realizado ajuda a entender como os métodos de agrupamentos se comportam com esse tipo de dados, já que é um campo ainda com poucos estudos. Já o estudo por simulação fomenta a comparação de métodos de agrupamentos com maior, média e pouca restrição espacial reforçando o entendimento dos mesmos em diferentes cenários.

Os dados educacionais do Estado do Paraná consistiram de seis indicadores educacionais, coletados de repositórios públicos disponíveis livremente na internet. Considerando que a educação é um dos principais temas de interesse para um bom desenvolvimento da sociedade, ao determinar agrupamentos segundo indicadores educacionais, permite-se identificar similaridades e particularidades entre municípios, bem como identificar padrões regionais. Esses resultados podem auxiliar no planejamento de políticas públicas e na produção de maior conhecimento sobre a situação educacional no Estado, tanto quanto pesquisas futuras envolvendo esse tipo de dados, já que é um campo com poucos estudos realizados tendo esse objetivo.

A partir de análises preliminares, foi verificada autocorrelação na distribuição espacial dos indicadores. Tendo-se detectado o número adequado de grupos igual a três, a análise dos resultados produzidos pelos métodos de agrupamentos aplicados aos dados educacionais seguiu as seguintes etapas: visualização dos grupos resultantes por meio de mapas, avaliação do número de municípios por grupo, análise das distribuições dos indicadores em cada grupo usando box plots e comparação e validação dos resultados com base em três índices de validação interna.

Com base na análise dos resultados, concluímos que as diferenças entre os métodos espaciais e não espaciais foram muito sutis e não foi possível afirmar que um método é superior aos demais. No entanto, podemos concluir que os métodos espaciais apresentam melhores indicativos de reproduzir com mais eficácia as informações contidas nos indicadores educacionais, baseados nos resultados dos índices de validação interna e, também, pelo

o fato dos métodos espaciais produzirem soluções mais regionalizadas com semelhanças apontadas nos mapas e nos box plots. Obter resultados mais regionalizados podem facilitar a aplicação de iniciativas públicas. Logo, dadas as análises feitas nesse trabalho, como os resultados em conjunto não apontam métodos que se sobressaem fortemente aos outros, reforçamos que para uma análise mais concreta também devem ser considerados os contextos dos dados e objetivos da aplicação desses resultados.

O estudo por simulação, teve a motivação de comparar os desempenhos dos métodos de agrupamentos na reprodutibilidade de uma configuração predefinida de grupos (gabarito). Neste estudo, os dados espaciais foram simulados com diferentes parâmetros de autocorrelação (ρ), variância (σ_b^2) e porcentagem de permutação dos municípios (p). Para poder comparar os métodos de agrupamentos, dois índices de validação externa foram usados, já que índices de validação externa são os mais adequados para comparar soluções produzidas por métodos de agrupamentos com soluções predeterminadas (neste trabalho, as soluções gabarito).

Os dois índices de validação externa considerados (índice de Rand ajustado e variação da informação) apresentam resultados mais semelhantes conforme o número k de grupos aumenta. Na maioria dos cenário simulados, tanto para o índice de Rand ajustado quanto para o índice de variação da informação, o método CGG foi detentor dos melhores resultados em termos de reprodutibilidade dos gabaritos. Pode-se perceber também que em alguns cenários com maiores valores de σ_b^2 e com maior ruído (maiores percentuais de trocas de municípios entre grupos), os métodos não espaciais começam a indicar resultados superiores do que pelo menos um dos métodos espaciais em alguns cenários com essas característica. Isso ocorre justamente devido aos elevados valores desses dois parâmetros, que acabam introduzindo maior ruído aos dados simulados, fazendo assim os dados perderem espacialização.

O método CGG apresentou melhor desempenho do que o método CGV e os efeitos de α nos resultados foram diferentes, para CGG e CGV, para k = 3 e 6, mas semelhantes para k = 10. Vale mencionar que como a construção da matriz D_1, para as duas variações do método CG, são diferentes, usar o mesmo valor de α para as ambas variações nem sempre é adequado, como apontado para os diferentes números de grupos estudados. O método Skater que é um método de agrupamentos espacial que apresenta restrições mais impositivas quanto à regionalização dos grupos produziu, em geral, grupos com tamanhos mais discrepantes. Logo, por produzir grupos com muito mais municípios que outros, o método Skater pode se tornar menos eficaz do que os outros métodos de agrupamentos em cenários com maior variação não espacial.

Como sugestão de pesquisas futuras, para os estudo de caso pode-se utilizar mais ou outros indicadores educacionais. Algumas cidades do Brasil possuem indicadores municípais, que podem ajudar em análises especificas, como também, utilizar separadamente os

indicadores por etapa de ensino (ens. fund. 1 ou ens. fund.2 ou ensino médio). Para o estudo por simulação, pode-se aplicar a outros métodos de agrupamentos espaciais e não espacias, também analisar a construção dos grupos por meio dos índices de validação interna. Como mencionado, os índices de validação externa apresentam baixa sensibilidade para números pequenos de grupos, por conta disso, sugere-se estudar esses métodos com outros números de grupos também.

REFERÊNCIAS

AGARWAL, Jyoti; NAGPAL, Renuka; SEHGAL, Rajni. Crime analysis using k-means clustering. **International Journal of Computer Applications**, v. 83, n. 4, 2013.

ASSUNÇÃO, R.M. et al. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. **International Journal of Geographical Information Science**, v. 20, n. 7, p. 797–811, 2006. DOI: 10.1080/13658810600665111.

AZUR, J. M. et al. Multiple imputation by chained equations: what is it and how does it work? **Int. J. Methods Psychiatr**, v. 20, p. 40–49, 2011. DOI: 10.1002/mpr.329.

BDEWEB. **IPARDES – Instituto Paranaense de Desenvolvimento Econômico e Social. Indicadores Sociais.** [S.l.: s.n.]. http://www.ipardes.pr.gov.br/Pagina/Indicadores-Sociais. Acessado em 20/03/2020.

BIVAND, Roger S.; PEBESMA, Edzer; RUBIO, Virgilio Gómez. **Applied Spatial Data Analysis with R**. New York, NY: Springe, 2008. 405 p.

CHARRAD, Malika et al. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. **Journal of Statistical Software**, v. 61, p. 1–36, 2014. DOI: 10.18637/jss.v061.i06.

CHAVENT, Marie et al. ClustGeo: an R package for hierarchical clustering with spatial constraints. **Computational Statistics**, v. 33, p. 1799–1822, 2018. DOI: 10.1007/s00180-018-0791-1.

DISTEFANO, Veronica; MAMELI, Valentina; POLI, Irene. Identifying spatial patterns with the Bootstrap ClustGeo technique. **Spatial Statistics**, v. 38, 2020. DOI: https://doi.org/10.1016/j.spasta.2020.100441.

DUTT, Ashish et al. Clustering Algorithms Applied in Educational Data Mining. **International Journal of Information and Electronics Engineering (IJIEE)**, v. 5, 2015. DOI: 10.7763/IJIEE.2015.V5.513.

GEARY, R. C. The Contiguity Ratio and Statistical Mapping. **The Incorporated Statistician**, v. 5, n. 3, p. 115–146, 1954. DOI: https://doi.org/10.2307/2986645.

GRUBESIC, Tony H.; WEI, Ran; MURRAY, Alan T. Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense. **Annals of the Association of American Geographers**, v. 104, n. 6, p. 1134–1156, 2014. DOI: 10.1080/00045608.2014.958389.

HAIR, J. F. et al. **Multivariate Data Analysis**. 7. ed. [S.l.]: Prentice Hall, 2010. (Always learning).

HASLBECK, Jonas M. B.; WULFF, Dirk U. Estimating the number of clusters via a corrected clustering instability. **Computational Statistics**, v. 35, p. 1879–1894, 2020. DOI: 10.1007/s00180-020-00981-5.

HASSANI, Marwan; SEIDL, Thomas. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. **Vietnam Journal of Computer Science**, v. 4, p. 171–183, 2017. DOI: 10.1007/s40595-016-0086-9.

HOEF, Hanneke van der; WARRENS, Matthijs J. Understanding information theoretic measures for comparing clusterings. **Behaviormetrika**, v. 49, p. 353–370, 2019. DOI: 10.1007/s41237-018-0075-7.

IPEA -INSTITUTO DE PESQUISA ECONÔMICA E APLICADA E FUNCAÇÃO JOÃO PINHEIRO, PNUD – Programa das nações unidas para o desenvolvimento e. **Atlas de desenvolvimento humano do Brasil de 2013**. [S.l.: s.n.], 2013. http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/. Acessado em 20/03/2020.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters"**, v. 31, n. 8, p. 651–666, 2010. DOI: 10.1016/j.patrec.2009.09.011.

JOHNSON, Richard Arnold; WICHERN, Dean W. **Applied multivariate statistical analysis**. 5. ed. Upper Saddle River, NJ: Prentice Hall, 2002. xviii, 767 p.

MEC. **INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira;Indicadores Educacionais**. [S.l.: s.n.]. http://portal.inep.gov.br/web/guest/indicadores-educacionais. Acessado em 20/03/2020.

MEILĂ, Mariana. Comparing Clusterings by the Variation of Information. In: **Learning Theory and Kernel Machines**. Edição: Bernhard Schölkopf e Manfred K. Warmuth. [S.l.]: Springer Berlin Heidelberg, 2003. P. 173–187. ISBN 978-3-540-45167-9.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. 1. ed. [S.l.]: Editora UFMG, 2005. 295 p.

PFITZNER, Darius; LEIBBRANDT, Richard; POWERS, David. Characterization and evaluation of similarity measures for pairs of clusterings. **Knowl. Inf. Syst.**, v. 19, p. 361–394, 2009. DOI: 10.1007/s10115-008-0150-6.

PUNJ, Girish; STEWART, David W. Cluster Analysis in Marketing Research: Review and Suggestions for Application. **Journal of Marketing Research**, v. 20, n. 2, p. 134–148, 1983. DOI: 10.1177/002224378302000204.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

SHIRKHORSHIDI, Ali Seyed; AGHABOZORGI, Saeed; WAH, Teh Ying. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. **PLoS One**, v. 10, 2015. DOI: 10.1371/journal.pone.0144059.

SIDDIQUI, Mohammad Khubeb et al. Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis. **Journal of Pure and Applied Microbiology**, v. 14, 2020. DOI: 10.22207/JPAM.14.SPL1.40.

TIWARI, Mamta; MISRA, Bharat. Application of Cluster Analysis in Agriculture-A Review Article. **International Journal of Computer Applications**, v. 36, n. 4, p. 43–47, 2011.

VAN BUUREN, Stef; GROOTHUIS-OUDSHOORN, Karin. mice: Multivariate Imputation by Chained Equations in R. **Journal of Statistical Software**, v. 45, n. 3, p. 1–67, 2011. Disponível em: <https://www.jstatsoft.org/v45/i03/>.

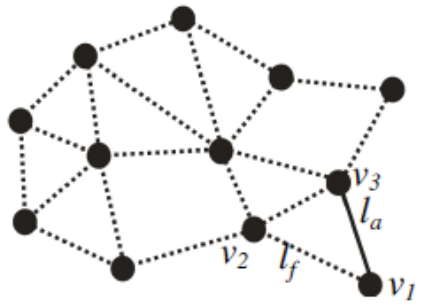
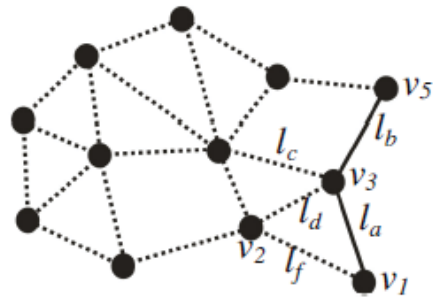
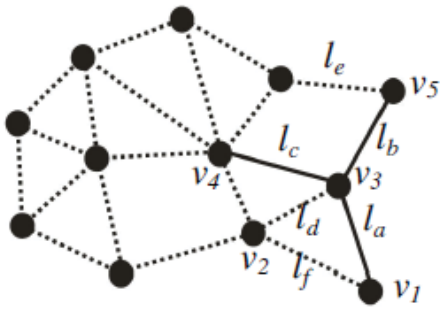
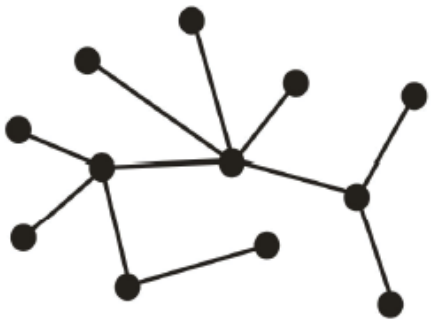
VERICA, Weverton Rodrigo; VILLWOCK, Rosangela; JOHANN, Jerry Adriani. Uso de técnicas de mineração de dados para agrupamento E espacialização de dados educacionais no Estado do Paraná. In: XVII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO - SBSR, XVII., 2015, João Pessoa-PB. ANAIS. Local da publicacao: IMPE, 2015. P. 1753–1760.

WEBBER, Carine G; ZAT, Daline; PRADO, Maria de Fátima Webber do et al. Utilização de algoritmos de agrupamento na mineração de dados educacionais. **RENOTE-Revista Novas Tecnologias na Educação**, v. 11, n. 1, 2013.

ZHU, Xuwen. Probability of misclassification in model-based clustering. **Computational Statistics**, v. 34, p. 1427–1442, 2019. DOI: 10.1007/s00180-019-00868-0.

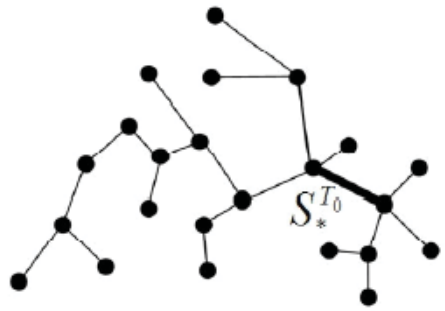
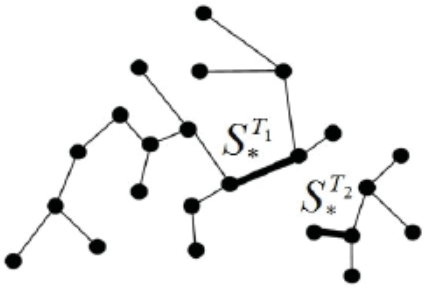
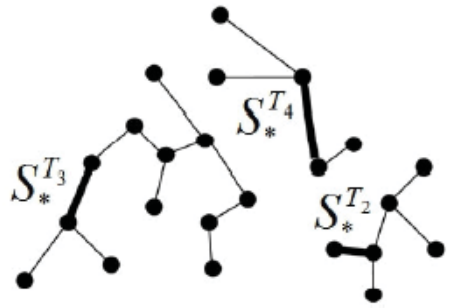
ANEXOS

FIGURA 35 – Primeira etapa do algoritmo Skater: Construção da árvore geradora mínima

<table>
<tr>
<td> 
</td>
<td>Primeira iteração:
Seja $T_1 = (V_1, L_1)$, onde $V_1 = \{v_1\}$ e $L_1 = \phi$.
Encontre a aresta de me menor custo $(l_a \langle l_f)$.
Passo 3: $T_2 \Rightarrow V_2 = \{v_1, v_3\}$ e $L_2 = \{l_a\}$.
Passo 4: repita o passo 2</td>
</tr>
<tr>
<td> 
</td>
<td>Segunda iteração:
Encontre a aresta de menor custo $(l_b \langle l_c \langle l_d \langle l_f)$.
Seja $T_3 \Rightarrow V_3 = \{v_1, v_3, v_5\}$ e $L3 = \{l_a, l_b\}$.</td>
</tr>
<tr>
<td> 
</td>
<td>Terceira iteração
Encontre a aresta de menor custo $(l_c \langle l_d \langle l_e \langle l_f)$.
Seja $T_4 \Rightarrow V_4 = \{v_1, v_3, v_4, v_5\}$ e $L3 = \{l_a, l_b, l_c\}$.</td>
</tr>
<tr>
<td>  </td>
<td>Iteração final:
$V_n = V$.</td>
</tr>
</table>

FONTE: Reproduzido de Assunção et al. (2006)

FIGURA 36 – Segunda etapa do algoritmo Skater: Particionamento da árvore geradora mínima

	Iteração 0: G*=MST. Selecionamos a aresta que tem a maior função objetivo. Cortemos essa aresta, formando 2 árvores (T_1 e T_2).
	Iteração 1: G* = (T_1, T_2). Comparamos a função objetivo para T_1 e T_2. Cortamos a árvore T_1 uma vez que $f_1(S_*^{T_2}) \leq f_1(S_*^{T_1})$
	Iretação 2: G* = (T_2, T_3, T_4). Comparamos qual das árvores têm a maior função objetivo para T_2, T_3 e T_4. Cortamos a árvore T_3 uma vez que $f_1(S_*^{T_2}) \leq f_1(S_*^{T_4}) \leq f_1(S_*^{T_3})$

FONTE: Reproduzido de Assunção et al. (2006)