Comparação de Agrupamentos Espaciais de Vulnerabilidade Socioeconômica através de Métodos de Estatística Multivariada

Renato Godoi da Cruz[1]*; Auberth Henrik Venson[2]

[1] Universidade de São Paulo. Instituto Pecege. Av. Pádua Dias, 11 – Agronomia; 13418-900. Piracicaba, São Paulo, Brasil.
[2] Orientador do Instituto Pecege. Doutor em Desenvolvimento Econômico. Universidade Estadual de Londrina. Departamento de Economia. Rodovia Celso Garcia Cid, PR 445 Km 380 – Campus Universitário; 86057-970 Londrina, Paraná, Brasil.
*autor correspondente: renatogcruz@hotmail.com

Comparação de Agrupamentos Espaciais da Vulnerabilidade Social através de Métodos de Estatística Multivariada

Resumo

O mapeamento da condição socioeconômica da população brasileira nas cidades pode contribuir nas gestões públicas estaduais e municipais na criação de programas governamentais de cunho social de enfrentamento da vulnerabilidade, direcionando assistência social e econômica orientada para a justiça social e a inclusão social. Assim, o objetivo deste trabalho foi à comparação do desempenho de agrupamentos espaciais da vulnerabilidade social utilizando-se de técnicas de Estatística Multivariada. Os dados utilizados correspondem as Áreas de Ponderação da cidade de Belo Horizonte, capital do estado de Minas Gerais e foram publicados pelo Censo Demográfico de 2010 conduzidas pelo Instituto Brasileiro de Geografia e Estatística [IBGE]. Os resultados mostraram que os métodos de Análises de Agrupamentos de Ligação Completa, K-Médias e “Hierarchical Density Based Spatial Clustering of Application with Noise” se saíram melhores em comparação aos métodos de Ligação Simples, da Média das Distâncias e “Density Based Spatial Clustering of Application with Noise”.

Palavras-chave: Análise de Componente Principal. Análise de Cluster Hierárquico. Análise de Cluster Não Hierárquico. Iniquidade Social.

Introdução

Este trabalho consiste na comparação do desempenho de agrupamentos da vulnerabilidade socioeconômico da população da cidade de Belo Horizonte através de aplicação de métodos de Estatística Multivariada. Segundo Mingoti (2005), a Estatística Multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais variáveis são correlacionadas entre si e são medidas simultaneamente em cada elemento amostral.

Vulnerabilidade socioeconômica pode ser entendida, segundo Semzezem e Alves (2013), como as situações de empobrecimento da classe trabalhadora, relacionadas às dificuldades materiais para a manutenção da sobrevivência, assim como às dificuldades relacionais e culturais, que interferem na forma de viver dos trabalhadores e de suas famílias.

Historicamente, as grandes cidades brasileiras são marcadas pela divisão territorial da pobreza (Santos, 2009). Esta também é a realidade da cidade de Belo Horizonte, capital

do estado de Minas Gerais, e objeto de estudo estatístico deste trabalho. De acordo com o site Panorama Cidades (2022), do Instituto Brasileiro de Geografia e Estatística, o Município de Belo Horizonte, no ano de 2010, apresentava uma população de cerca de 2.375.151 habitantes distribuídos por um território de 331,354 km². Apresentava 96,2% de domicílios com esgotamento sanitário adequado, 82,7% de domicílios urbanos em vias públicas com arborização e 44,2% de domicílios urbanos em vias públicas com urbanização adequada (presença de calçada, pavimentação, meio-fio e bueiro). Comparado com as outras cidades do estado mineiro, ocupava a 225º, 517º e 83º posições num universo de 645 cidades, respectivamente. Já quando comparado aos municípios do Brasil, suas posições eram 317º, 2779º e 419º de 5570, respectivamente.

Quanto à taxa de mortalidade infantil, a média no município era de 11.21 para 1.000 nascidos vivos. As internações devido a diarreias são de 0.3 para cada 1.000 habitantes. Comparado com todas as cidades do estado, ficava nas posições 277º e 332º, respectivamente, do total de 645. Quando comparado a municípios do Brasil, essas posições eram de 2796º e 3907º de 5570, respectivamente.

Além disso, o salário médio mensal, em 2019, era de 4,1 salários mínimos. A proporção de pessoas ocupadas em relação à população total era de 47,1%. Na comparação com as outras cidades do estado, ocupava as posições 4º e 23º de 645, respectivamente. Já na comparação com municípios do país, o município de Belo Horizonte ficava na posição, respectivamente, 17º e 79º de 5570. Tinha 31.6% da população em domicílios com rendimentos mensais de até meio salário mínimo por pessoa, o que o colocava na posição 305º de 645 dentre os municípios do estado e na 4372º posição de 5570 dentre os municípios do Brasil.

Para realizar o trabalho, foram utilizadas Áreas de Ponderação Censitárias do Censo Demográfico de 2010 do Município de Belo Horizonte. Essas áreas serão classificadas e agrupadas de acordo com a maior ou menor presença de características socioeconômicas que contribuem para tornar uma família mais vulnerável socioeconomicamente. Para tanto, será necessário identificar, inicialmente, quais são as variáveis que melhor expressam a vulnerabilidade socioeconômica.

Uma das maneiras de se fazer isso, segundo Mingoti (2005), é aplicando métodos de Estatística Multivariada que consiste em técnicas exploratórias de sintetização da estrutura de variabilidade dos dados e que permitam agrupar as Áreas de Ponderação que possuem características sociais em comuns. Ainda segundo Mingoti (2005), fazem parte desse grupo os métodos de análises de Componentes Principais, de Agrupamentos, Fatorial, de Correlações Canônicas, Discriminante e de Correspondência. Os métodos aplicados neste

trabalho foram, especificamente, Análises de Componentes Principais [PCA] e de Agrupamentos.

Para o Centro de Estudos da Metrópole [CEM] (2004), as construções de mapas de vulnerabilidade social reforçam a relevância da complementação do mapeamento dos grupos de vulnerabilidade a partir da análise de indicadores oriundos de outras fontes como, por exemplo, dados censitários e consistem numa boa metodologia de aproximações as complexas situações observadas na realidade. Já que, para Araújo *et. al.* (2019), os aspectos geográficos, econômicos, sociais e ambientais provocam diferentes níveis de desenvolvimento dentro de um mesmo aglomerado espacial e Barrozo *et. al.* (2019), dizem que compreender esses aspectos socioeconômicos é importante para descrever uma conjuntura dos fenômenos socioeconômicos e orientar na elaboração de politicas públicas intermunicipais.

Assim, fornecer um estudo comparativo de agrupamento espacial de vulnerabilidade socioeconômica nos permite visualizar um cenário mais real das complexas situações da desigualdade social e da iniquidade presente nas cidades brasileiras. Dessa forma, um agrupamento espacial de vulnerabilidade socioeconômica consistente poderia nos auxiliar na vigilância social de vulnerabilidades e riscos sociais e nos orientar na criação de programas sociais de enfrentamento da vulnerabilidade, subsidiando as escolhas de prioridades para politicas pública de assistência social e econômica orientada para a justiça social e a inclusão social.

O objetivo deste estudo, então, foi de avaliar e comparar o desempenho de técnicas de agrupamento espacial no contexto socioeconômico para fins de avaliação, pesquisa e monitoramento das desigualdades de cidades brasileiras utilizando-se de dados do Censo Demográfico a fim de contribuir nas gestões públicas, estaduais e ou municipais.

Material e Métodos

Os procedimentos metodológicos aplicados neste trabalho foram divididos em três partes:

(a) – a primeira parte consistiu de levantamento e escolha da base de dados;

(b) – a segunda, das análises de dados através de métodos de Estatística Multivariada;

(c) – e por fim, discussões e ponderações foram organizadas e descritas.

Para isso, foram utilizados os softwares Excel, R e seu ambiente integrado Rstudio.

1.1 Material

A área de estudo deste trabalho foi o Município de Belo Horizonte. Os dados publicados pelo IBGE estão organizados por Áreas de Ponderação e derivam do Censo Demográfico de 2010, resultados do universo.

Foram escolhidas, inicialmente, dez variáveis: porcentagem de pessoas não escolarizadas ou com ensino fundamental incompleto [P_SEM_INST], percentagem de pessoas cujo nível de escolaridade é o ensino superior completo [P_ENSSUP], porcentagem de pessoas cujo tempo habitual gasto no deslocamento de casa para o trabalho é de até 5 minutos [P_ATE5], porcentagem de pessoas cujo tempo habitual gasto no deslocamento de casa para o trabalho é superior a 2 horas [P_MAISDE2], média de densidade de residentes por cômodo [MEDIA_DESMORA], porcentagem de pessoas na linha de pobreza: cuja renda familiar mensal per capita é menor ou igual a US$ 144,89[1] [P_POBREZA], percentual de residências com alvenaria sem revestimento [P_ALVESREV], porcentagem de domicílios com acesso a rede de esgoto, abastecimento de água, coleta de lixo, energia elétrica e moradia adequada [P_TUDOADEQ], média da renda familiar mensal em julho de 2010, em reais [MED_RENDDOM] e porcentagem de pessoas com 65 anos ou mais com renda mensal igual ou superior a US$ 2.897,72[2] [P_IDOSO10SM].

As variáveis escolhidas representam seis dimensões do contexto socioeconômico que são: Educação, Mobilidade, Pobreza, Renda, Fortuna e Privação material.

Tabela 1: Tabela descritiva das variáveis selecionadas

Variáveis	Média	Variância	Desvio padrão	Coeficiente de variação (%)
P_SEM_INST	42.1950	2.029168e+02	14.2448868	33.75966
P_ENSSUP	16.5126	2.441122e+02	15.6240904	94.61920
P_ATE5	6.3162	4.175200e+00	2.0433306	32.35063
P_MAISDE2	1.1942	1.248500e+00	1.1173630	93.56582
M_DENSMORA	0.5237	1.350000e-02	0.1161895	22.18627
P_POBREZA	15.6235	7.165590e+01	8.4649808	54.18108
M_RENDDOM	4991.8011	1.471408e+07	3835.8939958	76.84389
P_IDOSO10SM	1.0431	2.530200e+00	1.5906602	152.49355
P_ALVSREV	9.5693	4.207610e+01	6.4866093	67.78562
P_TUDOADEQ	86.0409	8.537370e+01	9.2397890	10.73883

Fonte: Dados originais da pesquisa

[1] Equivalente a R$ 255,00 ou meio salário mínimo em 2010.
[2] Equivalente a R$ 5.100,00 ou 10 salários mínimos em 2010.

1.2 Métodos

Para identificar as variáveis que melhor expressam a vulnerabilidade socioeconômica foi utilizada a técnica estatística chamada de Análise de Componentes Principais [PCA], técnica introduzida por Karl Pearson em 1901 e posteriormente fundamentada por Hotelling em 1933 (Mingoti, 2005).

A intenção em usar essa técnica é procurar identificar uma quantidade relativamente pequena de Fatores que representam o comportamento conjunto de variáveis originais interdependentes. Segundo Fávero e Belfiore (2017), as técnicas de PCA são uteis quando há a intenção de se trabalhar com variáveis que apresentam, entre si, coeficientes de correlação relativamente elevados e, concomitantemente, se deseja estabelecer novas variáveis que captem o comportamento conjunto das variáveis originais.

Como dito anteriormente, essas técnicas são utilizadas em situações nas quais variáveis são correlacionadas entre si.

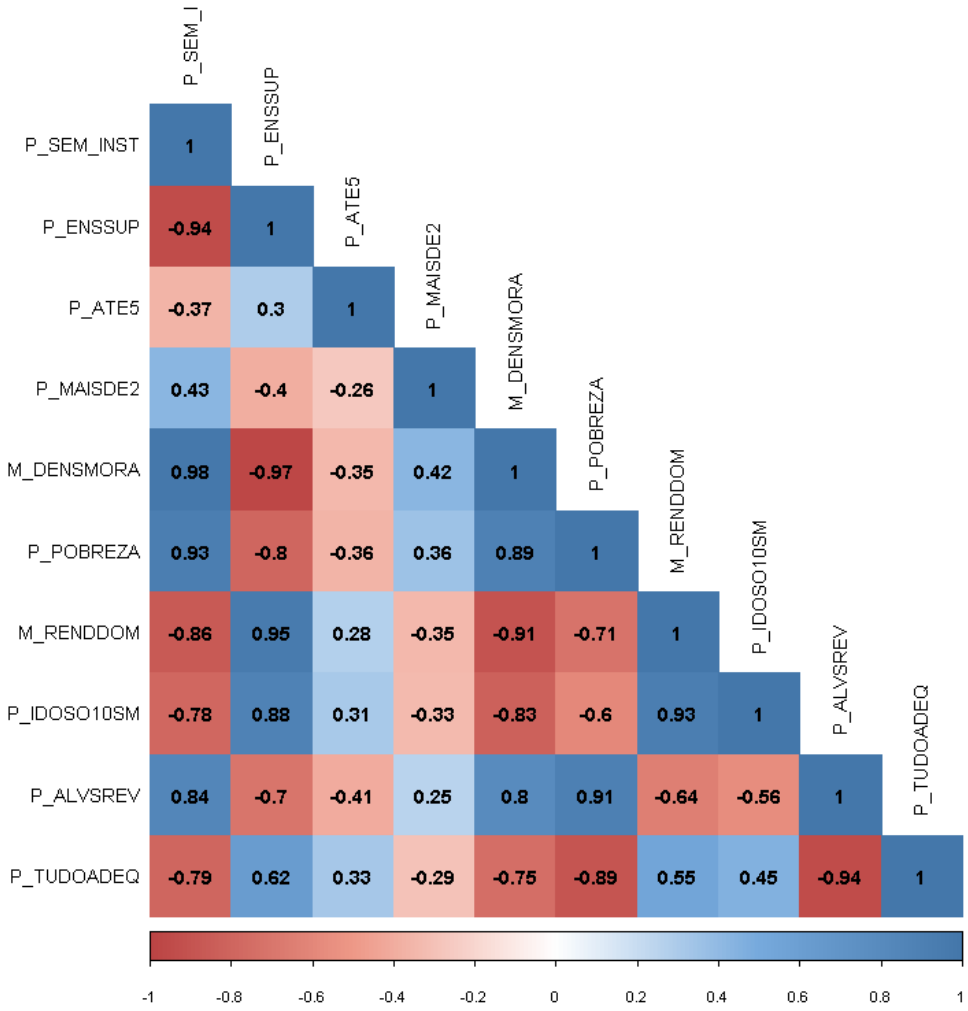

Figura 1: Análise de coeficiente de correlação de Pearson das variáveis escolhidas

Fonte: Dados originais da pesquisa

Na Figura 1, nota-se um alto grau de correlação, negativa ou positiva, entre as variáveis inicialmente escolhidas. A correlação de Pearson é um coeficiente estatístico que pode ser usado para medir o grau da correlação entre variáveis. Este coeficiente assume valores entre -1, que significa uma correlação perfeita negativa entre as duas variáveis e 1, que significa uma correlação perfeita positiva entre as duas variáveis.

A técnica estatística de PCA nos permitiu identificar, ao mesmo tempo, quais são as variáveis que apresentam coeficientes de correlação relativamente elevados entre si e estabelecer novas variáveis, chamado de Fatores, que sejam capazes de captarem o comportamento conjunto das variáveis originais (Fávero e Belfiore, 2017).

Já a técnica de Análise de Agrupamento representa um conjunto de métodos exploratórios muito úteis e que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos semelhantes entre observações em relação a determinadas variáveis e o objetivo de se criarem grupos em que prevaleça a homogeneidade interna.

Nesse sentido, esse conjunto de técnicas tem por objetivo principal a alocação de observações em uma quantidade relativamente pequena de agrupamento das observações (Fávero e Belfiore, 2017).

A ideia é que as Áreas de Ponderação de determinado grupo sejam relativamente semelhantes entre si e consideravelmente diferentes dos setores censitários de outros grupos.

Os métodos de agrupamento aplicados foram:

(a) – de Ligação Simples;
(b) – Ligação Completa;
(c) – e o da Média das Distâncias;
(d) – K-Médias;
(e) – “Density Based Spatial Clustering of Application with Noise” [DBSCAN] e ;
(f) – “Hierarchical Density Based Spatial Clustering of Application with Noise” [HDBSCAN].

Para avaliar o desempenho relativo dos modelos de agrupamento na identificação de vulnerabilidade socioeconômica será conduzida uma análise visual de agrupamentos plotados espacialmente comparando a representatividade dos grupos formados com a divisão, em igual quantidade, da componente principal 1 definida pela técnica anterior.

A intenção é avaliar o desempenho relativo dos modelos de agrupamento na identificação de vulnerabilidade socioeconômica para informar qual das soluções de agrupamento resume melhor a variação na concentração de vulnerabilidade socioeconômica.

Resultados e Discussões

As análises de PCA foram aplicadas sucessivamente até que se atendesse a três regras definidas a seguir:

(a) – a primeira regra, conhecida como critério de Kaiser, tem como princípio básico para o estabelecimento do número de Fatores e sugere reter apenas os aqueles com autovalor maior do que 1;

(b) – a segunda regra foi aplicada considerando que uma variável só deve ficar no modelo se sua Comunalidade– que representam a variância total compartilhada de cada variável em todos os fatores extraídos a partir de autovalores maiores que 1 – fosse maior ou igual a 0,7;

(c) – e por último, que a hipótese nula de que a matriz de correlações é identidade fosse rejeitada pelo Teste de Esfericidade de Bartlett.

A primeira regra tem como base o raciocínio de que autovalores representam a quantidade de variação explicada por um fator e que um autovalor de 1 representa uma quantidade substancial de variação (Matos e Rodrigues, 2019).

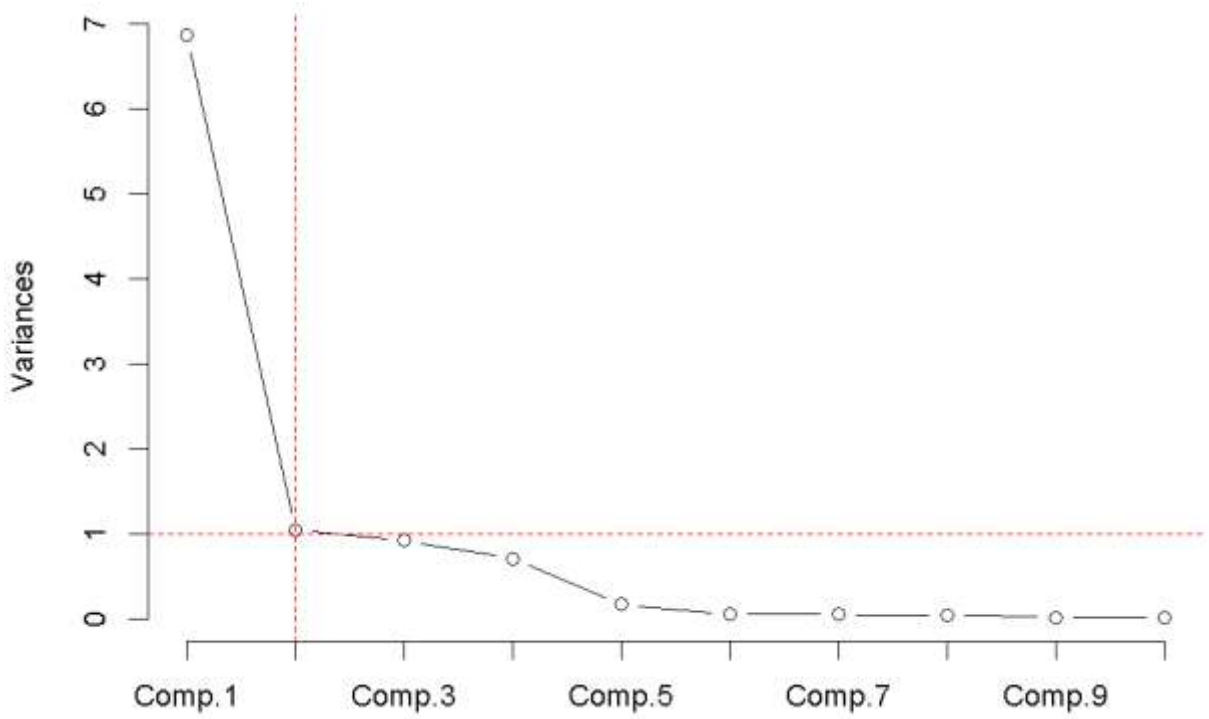

Figura 2: Regra de Kaiser para definir número de Fatores

Fonte: Dados originais da pesquisa

Na Figura 2 podemos ver no gráfico as 10 componentes, correspondentes as 10 variáveis descritas anteriormente e seus respectivos autovalores. Nota-se que, obedecendo à regra definida acima, o número de Fatores da PCA devia ser igual a dois.

Estabelecido que o número de Fatores adotado fosse igual a dois, definiu-se, também, as cargas fatoriais, que são as correlações de Pearson entre as variáveis originais e cada um dos Fatores e suas respectivas Comunalidades para cada uma das variáveis descritas anteriormente.

Tabela 2. Score fatoriais, cargas fatoriais e comunalidade dos modelos de PCA

Variável	Ajuste	Fator1	Fator2	X1	X2	Comunalidade
P_SEM_INST	1	-0,14087	0,00720	-0,98204	0,00761	0,96447
	2	-0,14815	0,01208	-0,98302	0,01225	0,96649
P_ENSSUP	1	0,13527	-0,26708	0,94299	-0,28225	0,96890
	2	0,14292	0,26724	0,94829	0,27104	0,97272
P_ATE5	1	0,06161	0,32239	0,42953	0,34070	0,30057
	2	-	-	-	-	-
P_MAISDE2	1	-0,06470	0,14265	-0,45102	0,15075	0,22615
	2	-	-	-	-	-
M_DENSMORA	1	-0,14059	0,10053	-0,98006	0,10624	0,97182
	2	-0,14806	-0,08898	-0,98243	-0,09025	0,97332
P_POBREZA	1	-0,13243	-0,25605	-0,92321	-0,27060	0,92554
	2	-0,13950	0,29809	-0,92560	0,30232	0,94815
M_RENDDOM	1	0,12844	-0,35981	0,89536	-0,38026	0,94626
	2	0,13605	0,38100	0,90275	0,38641	0,96427
P_IDOSO10SM	1	0,11873	-0,42290	0,82772	-0,44692	0,88485
	2	0,12513	0,47631	0,83027	0,48307	0,92271
P_ALVSREV	1	-0,12513	-0,39037	-0,87232	-0,41255	0,93114
	2	-0,13209	0,41149	-0,87645	0,41733	0,94233
P_TUDOADEQ	1	0,11734	0,46284	0,81796	0,48913	0,90832
	2	0,12382	-0,52562	0,82155	-0,53308	0,95912

Fonte: Dados originais da pesquisa

Como se podem notar na Tabela 2, no primeiro ajuste do método PCA, as variáveis P_ATE5 e P_MAISDE2 obtiveram Comunalidades abaixo de 0,70 e, Aplicando-se a segunda regra, essas duas variáveis foram retirada para que o segundo ajuste fosse realizado.

Atendida as duas primeiras regras, partiu-se para o Teste de Esfericidade de Bartlett, terceira e última regra, para efeitos de decisão sobre a adequação global da PCA a fim de validar o segundo ajuste como modelo final da PCA.

Tabela 3. Resultados dos Testes de Esfericidade de Bartlett dos modelos de PCA

Ajuste	X^2	df	Valor-p
2	7668.6	66	< 2.2e-16

Fonte: Dados originais da pesquisa

A partir da tabela 3, podemos verificar que o Testes de Esfericidade de Bartlett para o ajuste 2 foi menor que 0,05, ou seja, a regra numero 3, ao nível de significância de 5%, foi atendida.

Tabela 4: Importância das componentes principais

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Sd	2.5758	1.0071	0.4253	0.2396	0.2242	0.2023	0.1118	0.0942
PV	0.8294	0.1268	0.0226	0.0071	0.0062	0.0054	0.0051	0.0011
CP	0.8294	0.9561	0.9787	0.9859	0.9922	0.9894	0.9973	1.0000

Fonte: Dados originais da pesquisa

Nota: Desvio padrão [SD], Proporção de Variância [PV] e Proporção Cumulativa [CP]

Na Tabela 4, podemos observar algumas informações estatísticas importante do ajuste final da PCA, como por exemplo, a PV da primeira componente principal onde se nota que cerca de 80% da variância total é explicada por ela. Isso significa que quase dois terços dos dados no conjunto de variáveis podem ser representados apenas pelo primeiro componente principal, justificando o uso dessa componente na comparação com os agrupamentos gerados posteriormente. Já a segunda componente principal explica 13% da variância total.

Na CP, as duas componentes principais selecionadas explicam quase 96% da variância total. Isso implica que os dois primeiros componentes principais podem representar com precisão os dados e não causaria prejuízos em sua utilização como “inputs” para os modelos de Análises de Agrupamentos dos passos seguintes.

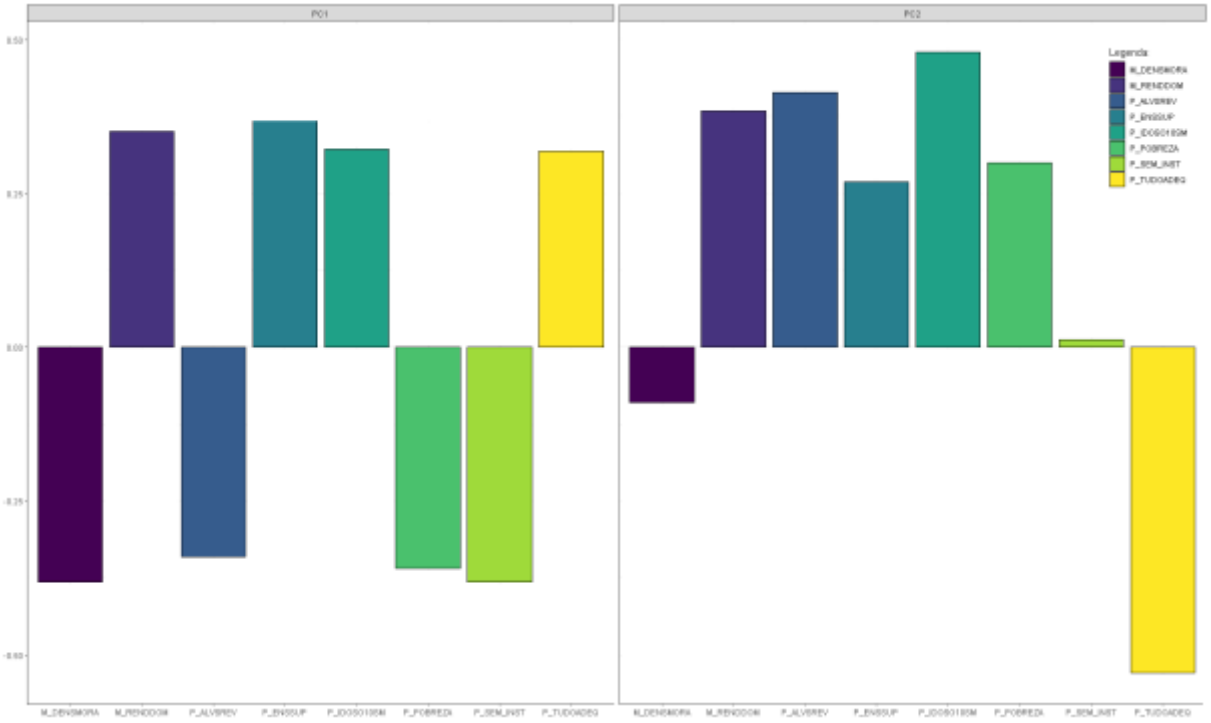

Figura 3: Pesos que cada variável tem em cada componente principal
Fonte: Dados originais da pesquisa

Na Figura 3 podemos ver a contribuição dos pesos que cada variável exerceu em cada componente principal. Com as componentes principais definidas, pode-se aplicar um algoritmo de visualização espacial. Como a componente principal, sozinha, contém a maior porcentagem da variância total do modelo, ela foi escolhida para ser a referência de representatividade das informações das variáveis trabalhadas.

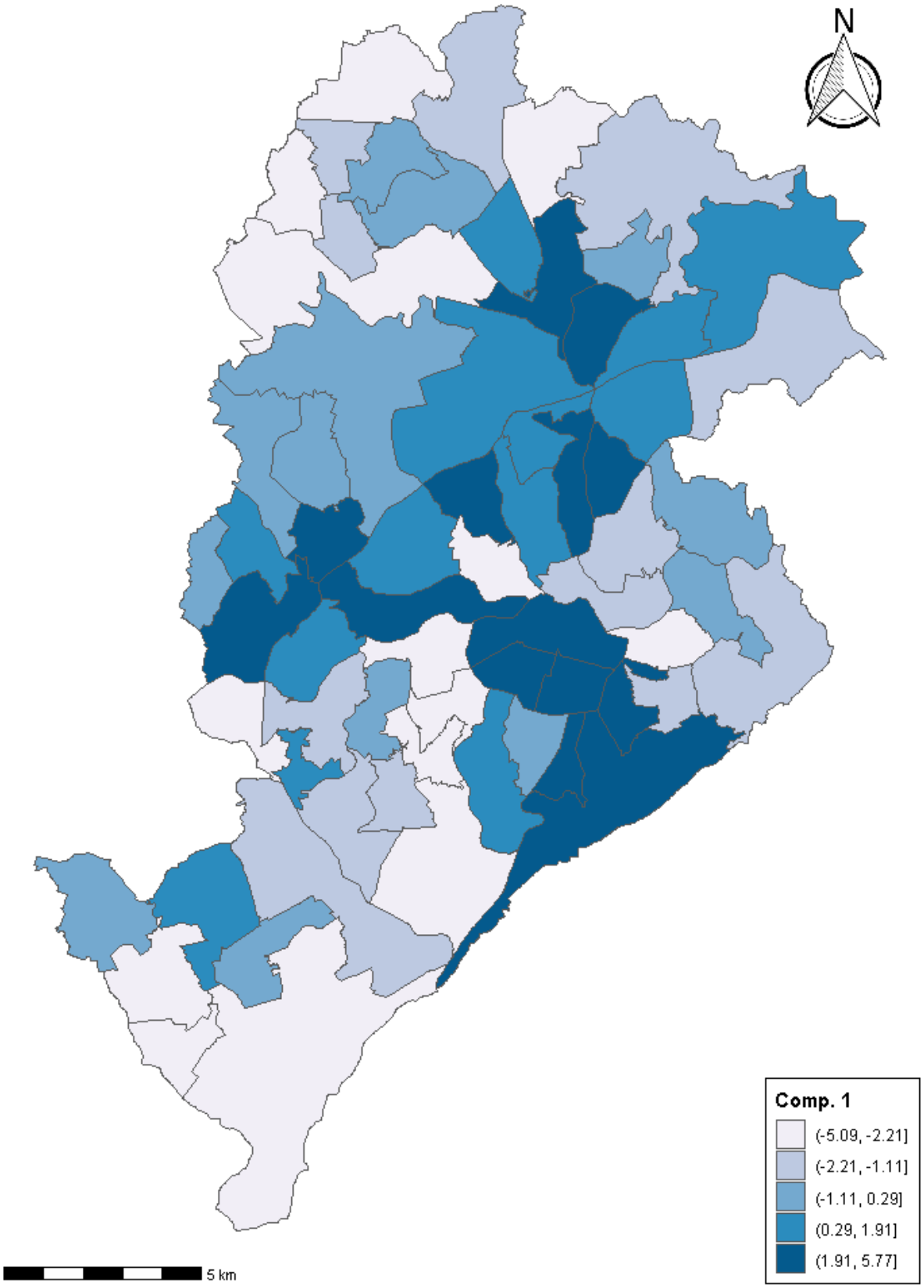

Figura 4: Quintil do escore da primeira componente principal da PCA

Fonte: Dados originais da pesquisa

Na Figura 4, pode-se visualizar os valores da componente principal 1 da análise de PCA distribuídos em quintis. Na estatística descritiva, um quintil é um dos valores de uma variável que divide o seu conjunto ordenado em cinco partes iguais.

Definido os Fatores das variáveis na PCA, podemos utiliza-los agora os comparando através de medidas de distâncias aplicando os modelos de Análises de Agrupamentos. Nessa etapa foram ajustados os modelos de agrupamentos de Ligação Simples, da Média das Distâncias, de Ligação Completa, K-Médias, DBSCAN e HDBSCAN.

Após os ajustes de todos os modelos, foram verificados se a variabilidade entre os grupos é significativamente superior à variabilidade interna a cada grupo produzido para cada um dos modelos. Essa verificação foi necessária para validar, estatisticamente, os métodos de agrupamento realizado a fim de prosseguir para a etapa de avaliar o desempenho relativo dos modelos. Para isso, aplicou-se o teste F da análise de variância de um fator [ANOVA] para todos os modelos.

Tabela 5. Resultados da ANOVA por Fatores dos modelos de agrupamento

Métodos	Fator	Sum Sq	Mean Sq	F value	Pr(>F)	Grupos
Ligação Simples	1	9603824	9603824	70.92	5.22e-12	3
	2	73625057	73625057	71.38	4.67e-12	
Média das Distâncias	1	9603824	9603824	70.92	5.22e-12	3
	2	73625057	73625057	71.38	4.67e-12	
Ligação Completa	1	15886859	15886859	409.9	<2e-16	3
	2	121194312	121194312	404.5	<2e-16	
K-Médias	1	16152619	16152619	465.9	<2e-16	3
	2	123366401	123366401	463.4	<2e-16	
DBSCAN	1	9603824	9603824	70.92	5.22e-12	3
	2	73625057	73625057	71.38	4.67e-12	
HDBSCAN	1	11424787	11424787	106.4	2.57e-15	3
	2	87027689	87027689	105.5	3.07e-15	

Fonte: Dados originais da pesquisa

A partir da Tabela 5, podemos verificar que o teste F para os Fatores 1 e 2 de todos os modelos foi menor que 0,05, ou seja, existe pelo menos um grupo que apresenta média estatisticamente diferente dos demais ao nível de significância de 5%.

Os primeiros modelos de agrupamentos aplicados são conhecidos como hierárquicos por possuírem a característica de privilegiar uma estrutura hierárquica para a formação dos grupos (Fávero e Belfiore, 2017). Desses modelos, foram aplicados os métodos da Ligação Simples, da Ligação Completa e da Média das Distâncias.

Enquanto o método da Ligação Simples favorece as menores distâncias para que sejam formados novos grupos a cada passo do processo de agrupamento, o método da Ligação Completa privilegia as maiores distância entre as observações para que sejam formados novos grupos.

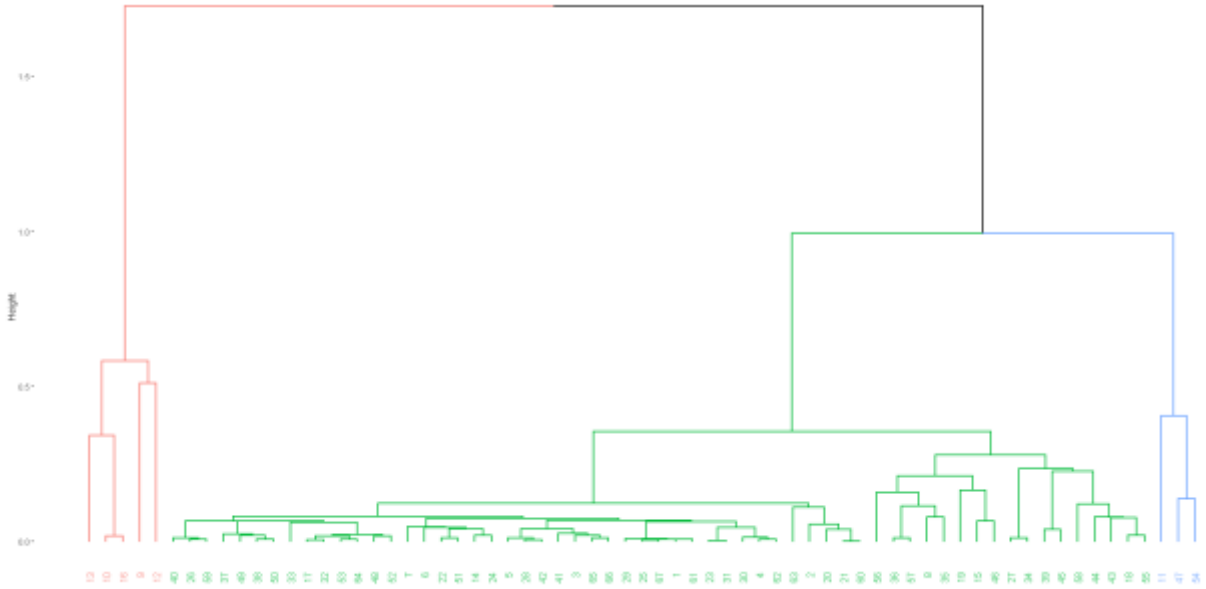

Figura 5: Dendograma do método da Ligação Simples
Fonte: Dados originais da pesquisa

A Figura 5 apresenta o dendograma do método da Ligação Simples, que nada mais é que um diagrama de arvore que exibe os grupos formados em cada passo e em seus níveis de distâncias, medido ao longo do eixo vertical.

Pelo dendograma podemos visualizar como os grupos são formados em cada etapa e consequentemente avaliar os níveis de distância dos agrupamentos que são formados. O passo onde os valores de distâncias mudam abruptamente de uma etapa para outra nos ajudam a escolher o número de grupos final. Neste caso, o dendograma acima sugere a quantidade de três grupos.

Neste método, o nível de distância entre grupos é definido pelas duas observações mais semelhantes entre si (Sneath, 1957). Como exemplo, Mingoti (2005) sugere considerar que, num determinado momento do algoritmo, tenhamos dois grupos, um composto dos elementos amostrais 1, 3 e 7 e outro dos elementos dos elementos 2 e 6, isto é:

$$C1 = \{X_1, X_3, X_7\}\ e\ C1 = \{X_2, X_6\} \quad (1)$$

Então, a distância entre esses dois grupos seria definida por:

$$d(C_1, C_{2)} = \min\{\, d(X_l, X_k, l \neq k, l = 1{,}3{,}7 \; e \; k = 2{,}6\} \quad (2)$$

Isto é, a cada fase do processo, os dois grupos mais similares com relação à distância em (2) são combinados em um único grupo.

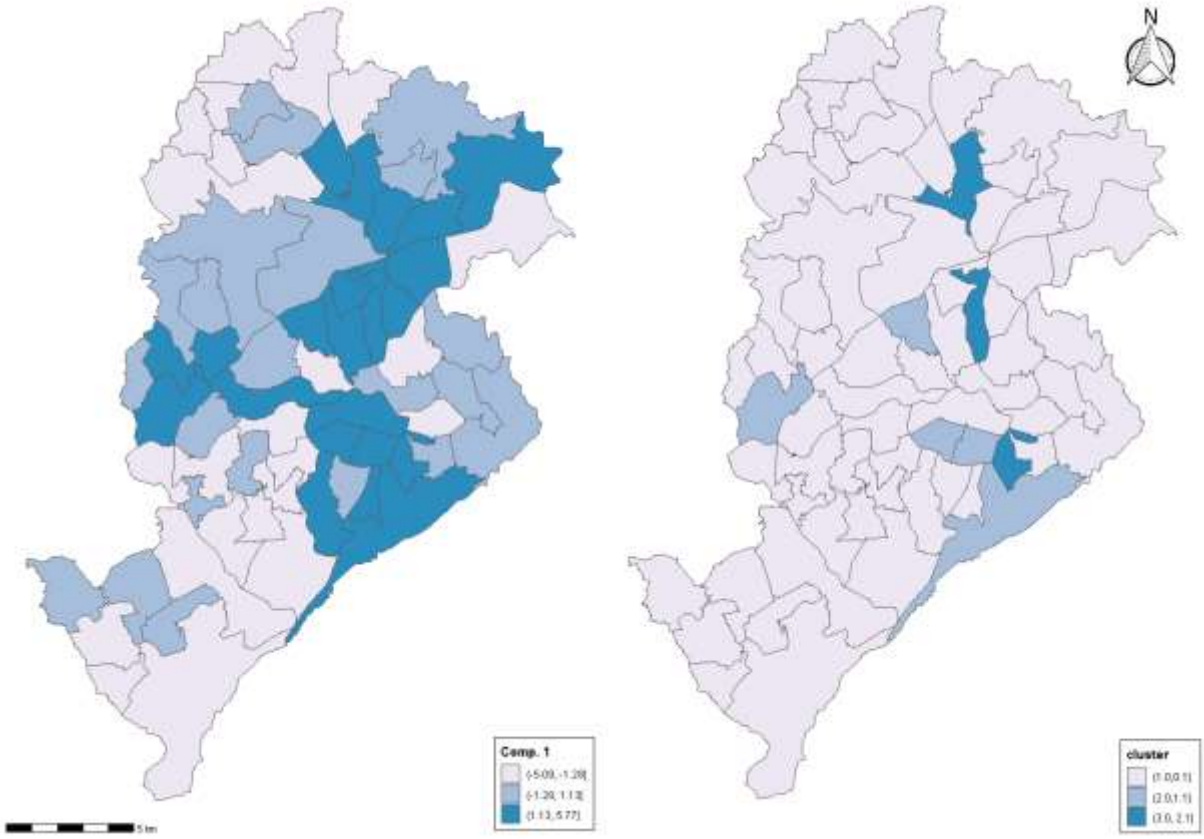

Figura 6: Tercil do escore da primeira componente principal da PCA x agrupamento do método da Ligação Simples
Fonte: Dados originais da pesquisa

A Figura 6 mostra os grupos das Áreas de Ponderação da cidade de Belo Horizonte definidos pelo método da Ligação Simples plotados espacialmente, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que neste método ouve uma maior concentração de Áreas de Ponderação no grupo 1 em comparação com a distribuição do tercil da score da primeira componente principal da PCA e pouca representatividade dos demais grupos.

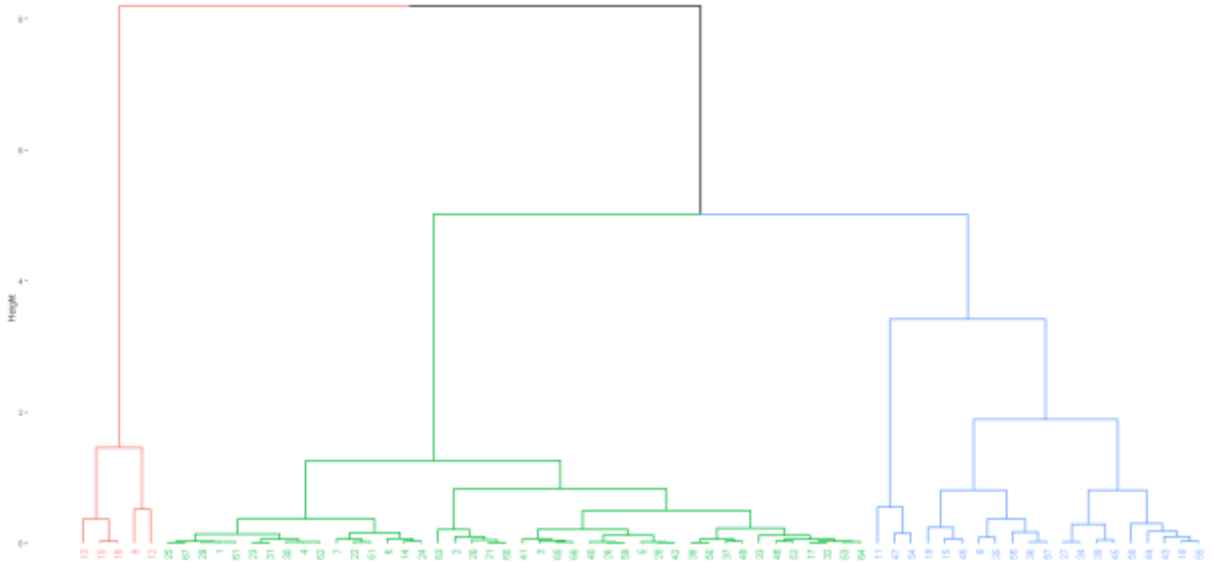

Figura 7: Dendograma do método da Ligação Completa
Fonte: Dados originais da pesquisa

Na Figura 7, assim como anteriormente, o dendograma do método da Ligação Completa sugere a quantidade de três grupos.

Neste método, o nível de distância entre dois grupos é definido, ainda segundo Sneath (1957) pelas observações que são mais distintos entre si. Como exemplo, Mingoti (2005) ainda nos sugere considerar os grupos C1 e C2 em (1).

Então, a distância entre eles seria definida por:

$$d(C_1, C_{2)} = \max\{\, d(X_l, X_k, l \neq k, l = 1{,}3{,}7 \; e \; k = 2{,}6\} \qquad (3)$$

Em cada etapa deste processo de formação de grupos, a medida em (3) é calculada para todos os pares de grupos, sendo, então, unificados aqueles que apresentarem o menor valor da distância, isto é, o menor valor de máximo (Mingoti, 2005).

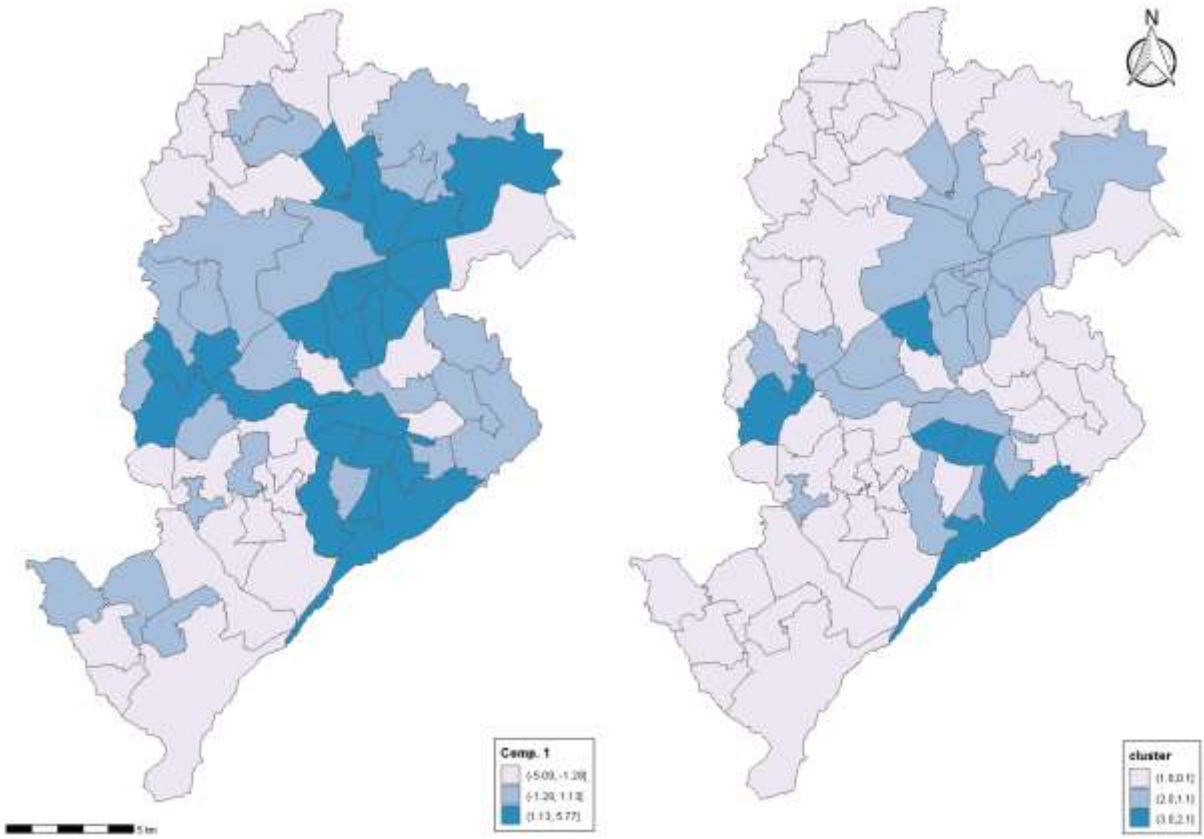

Figura 8: Tercil do escore da primeira componente principal da PCA versus agrupamento do método da Ligação Completa
Fonte: Dados originais da pesquisa

A Figura 8 mostra os grupos das Áreas de Ponderação da cidade de Belo Horizonte definidos pelo método da Ligação Completa plotados como dados espaciais, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que neste método ouve uma melhora na representatividade das Áreas de Ponderação quando comparados aos tercis dos scores originais da componente utilizada.

Por último, o método da Média das Distâncias faz dois grupos sofrerem fusão com base na distância média entre todos os pares de observações pertencentes a esses grupos (Fávero e Belfiore, 2017). Este método trata a distância entre dois grupos como a média das distâncias entre todos seus pares. Portanto, se o grupo C1 tem n1 observações e o grupo C2 tem n2 observações, a distância entre eles será definida por:

$$d(C_1, C_2) = \frac{\sum}{l \in C_1} \frac{\sum}{k \in C_2} \left(\frac{1}{n_1 n_2}\right) d\,(X_l, X_k) \quad (5)$$

Assim, a distância entre os grupos $C1$ e $C2$ é dado por:

$$d(C_1, C_2) = \frac{1}{6}\,[d(X_1, X_2) + d(X_1, X_6) + d(X_3, X_2) + d(X_3, X_6) + d(X_7, X_2) + d(X_7, X_6)] \quad (6)$$

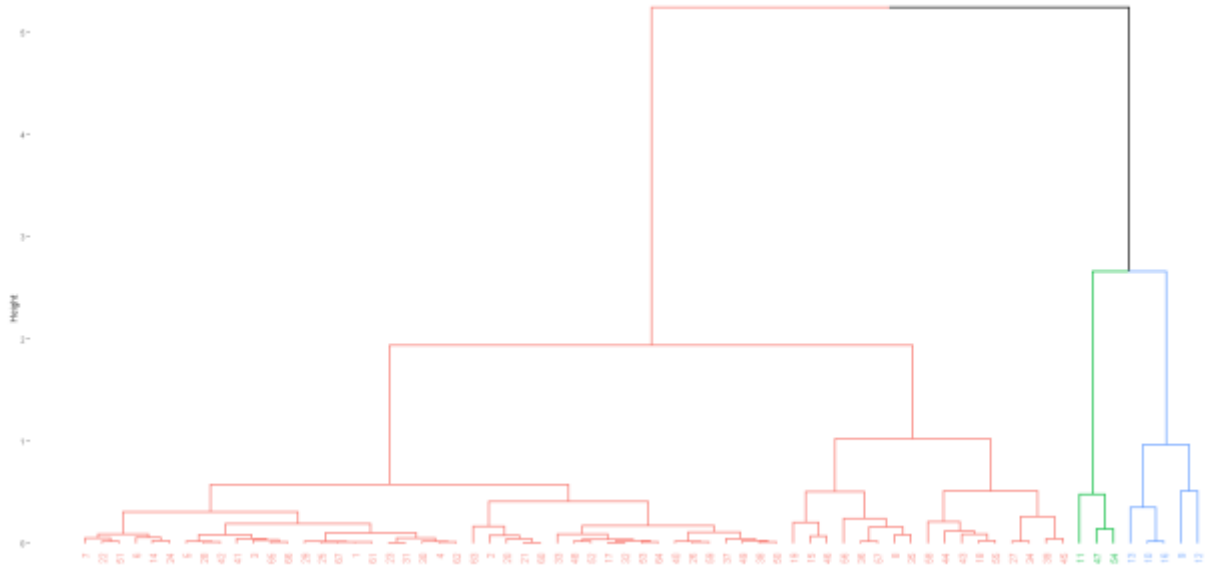

Figura 9: Dendograma do método da Média das Distâncias
Fonte: Dados originais da pesquisa

Na Figura 9, o dendograma do método da Média das Distâncias sugere, assim como nos dois últimos casos, a quantidade de três grupos.

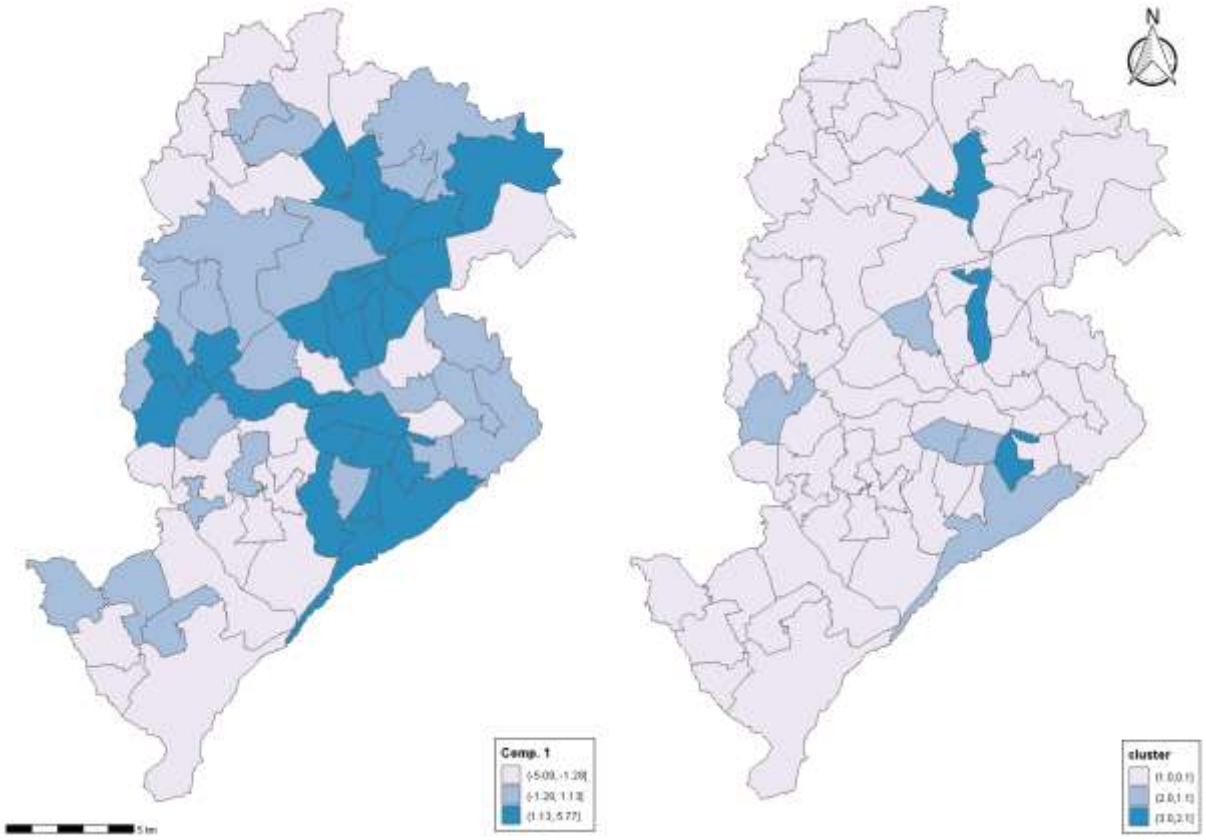

Figura 10: Tercil do escore da primeira componente principal da PCA versus agrupamento do método da Média das Distâncias

Fonte: Dados originais da pesquisa

A Figura 10 mostra os grupos das Áreas de Ponderação deste método plotados como dados espaciais, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que neste método representou, assim como no primeiro caso, pouco o quintil da componente principal 1 da PCA.

Enquanto os esquemas hierárquicos permitem a possibilidades para que a quantidade de agrupamentos formados seja avaliada e decidida no decorrer do ajuste, nos esquemas não hierárquicos, parte-se de uma quantidade conhecida de grupos e, a partir disso, as observações são alocadas nos grupos (Fávero e Belfiore, 2017).

O algoritmo k-Médias constrói o agrupamento baseada em centros. Seu algoritmo inicia escolhendo n centroides iniciais, em que n é o número de grupos definido inicialmente. Cada observação é então atribuída ao centroide mais próximo, e cada coleção de observação atribuída ao centroide forma um grupo. Em seguida, recalculam-se os valores dos centroides de cada grupo para cada novo grupo formado. Os processos anteriores são refeitos ate que os centroides se estabilizem nas mesmas posições (Mingoti, 2005).

Para determinar o valor n foi utilizado o método de Elbow. O gráfico Elbow mostra o incremento do valor de n até a soma das distâncias quadráticas das observações, medido ao longo do eixo vertical. O ponto de corte para o valor n deve ser onde começa a suavizar a queda soma das distâncias quadráticas.

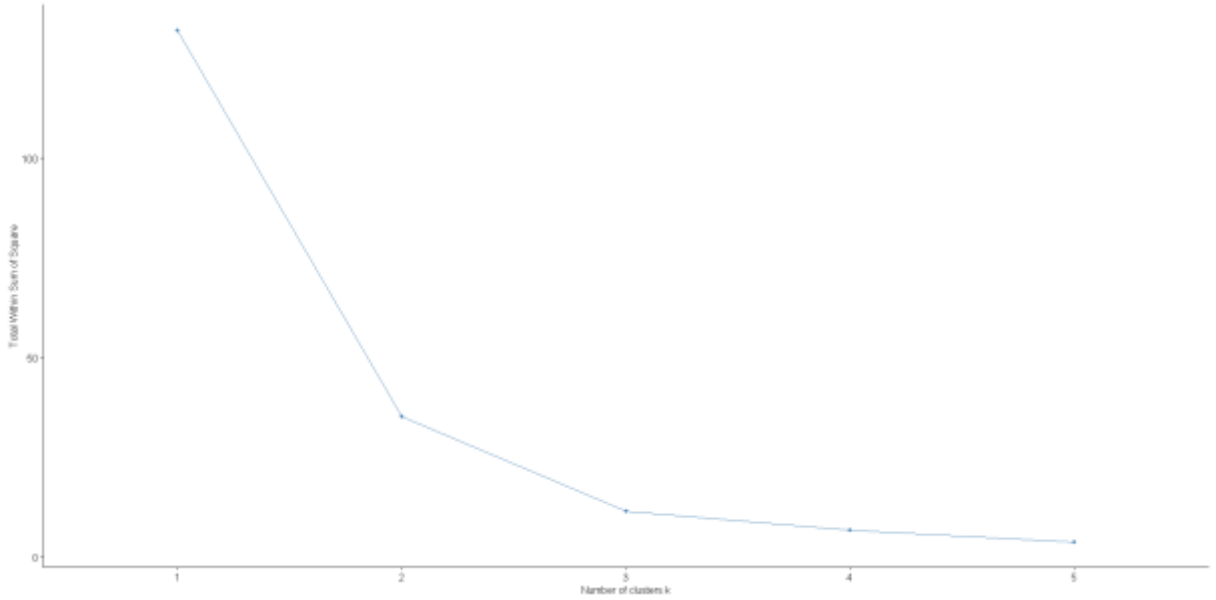

Figura 11: Método de Elbow para identificação do número ótimo de clusters
Fonte: Dados originais da pesquisa

Assim, Figura 11 nos sugere a adoção de três grupos, o que vai ao encontro com a média de grupos sugeridos nos métodos hierárquicos.

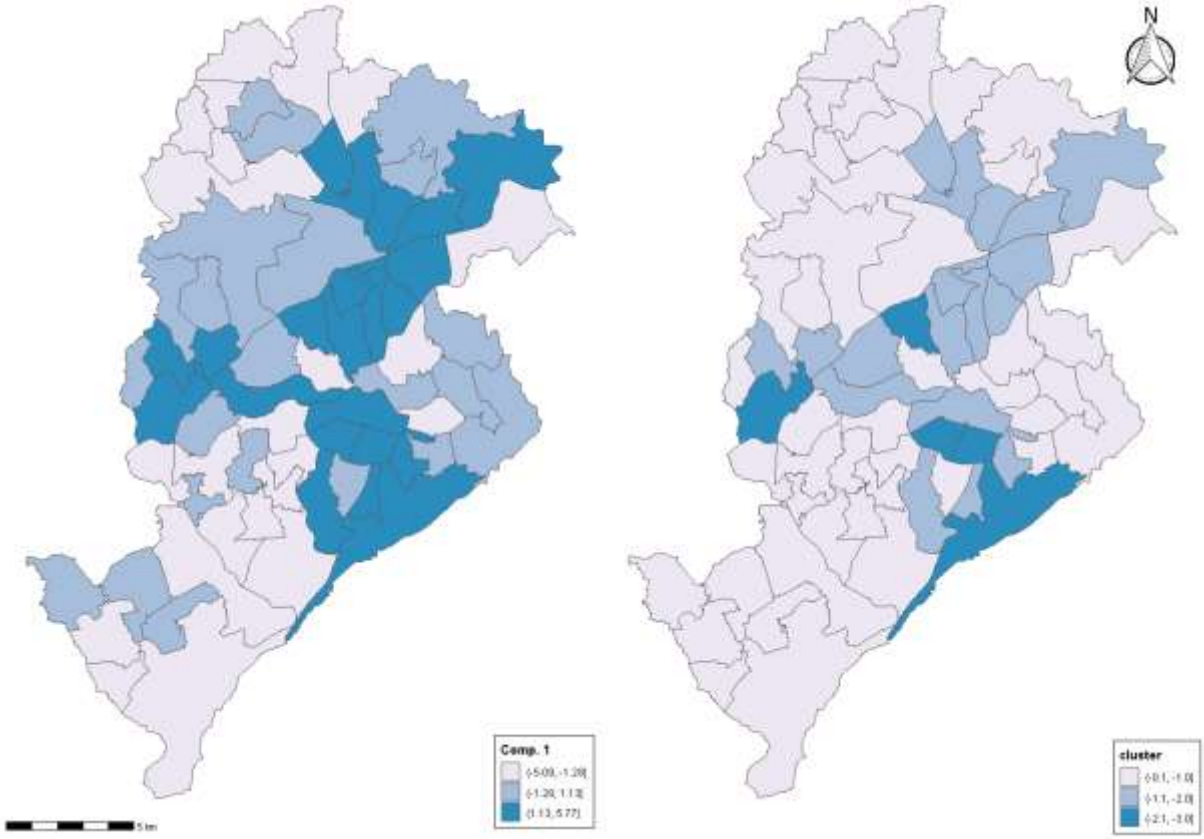

Figura 12: Tercil do escore da primeira componente principal da PCA versus agrupamento do método K-Médias
Fonte: Dados originais da pesquisa

A Figura 12 mostra os grupos das Áreas de Ponderação deste método plotados como dados espaciais, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que este método representou, assim como no segundo caso, razoavelmente bem os quintis da componente principal 1 da PCA.

O método DBSCAN é baseado em número de pontos (densidade) dentro de um raio específico (ESP) que buscam identificar regiões de alta densidade que estejam separadas entre si por região de baixa densidade. Nesse esquema, dois parâmetros básicos necessitam ser definidos, sendo eles o ESP que determina o raio de vizinhança para cada observação e o $MinPts$ que especifica o número mínimo de observações, no dado raio ESP, que uma observação precisa possuir para ser considerado ponto central e

consequentemente, conforme as definições de grupo baseado em densidade inicia a formação de um grupo.

A implementação do DBSCAN foi descrito por Ester *et al*. (1996) e executa as seguintes etapas: estima-se a densidade em torno de cada observação contando o número de observações no raio de vizinhança ESP respeitando o número mínimo de observações $MinPts$ e identificando os pontos como centrais, de fronteira ou de ruído; os pontos centrais formam densidade de grupos e, por fim, os pontos de fronteira são atribuídos a grupos observações no raio de vizinhança ESP respeitando o número mínimo de observações $MinPts$.

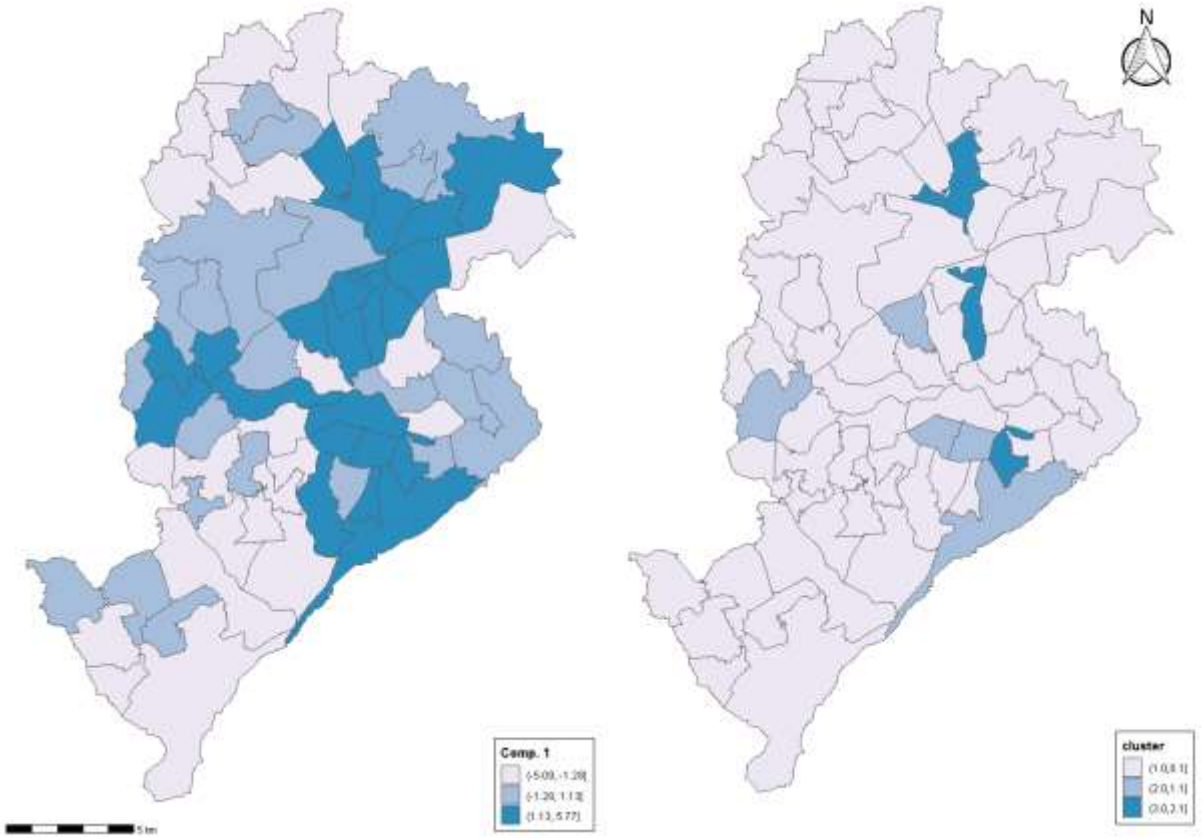

Figura 13: Tercil do escore da primeira componente principal da PCA versus agrupamento do método DBSCAN
Fonte: Dados originais da pesquisa

A Figura 13 mostra os grupos das Áreas de Ponderação deste método plotados como dados espaciais, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que este método representou, novamente, pouco a componente principal 1 da PCA.

Por fim, o modelo HDBSCAN é um algoritmo desenvolvido sobre o DBSCAN que, ao contrário de seu precursor, é capaz de identificar grupos de densidade variável. No HDBSCAN, de acordo com Campello *et al*. (2013), calculam-se os grupos por estimativas de densidade baseada em estabilidade. Essencialmente, ainda segundo os autores, o HDBSCAN calcula a hierarquia de todos os grupos utilizando o método DBSCAN e, em seguida, usa um método de extração baseado em estabilidade para encontrar cortes, produzindo assim uma solução estável executando os seguintes passos: calculam-se as distâncias entre os pontos; usam-se as distâncias mútuas calculadas, do passo anterior anteriormente, como medida para construir uma árvore genitora; corta-se a árvore baseado em estabilidade, por fim, extraiam-se os grupos.

Uma boa maneira de visualizar a árvore genitora é, assim como nos métodos hierárquicos, utilizando-se do dendograma.

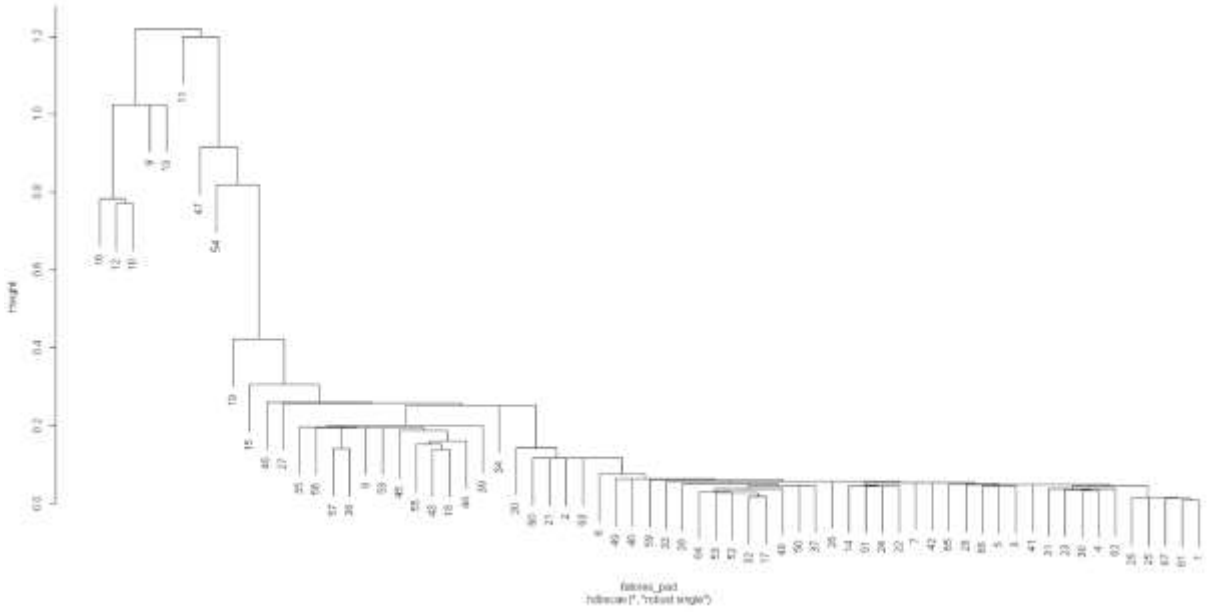

Figura 14: Dendograma do método HDBSCAN
Fonte: Dados originais da pesquisa

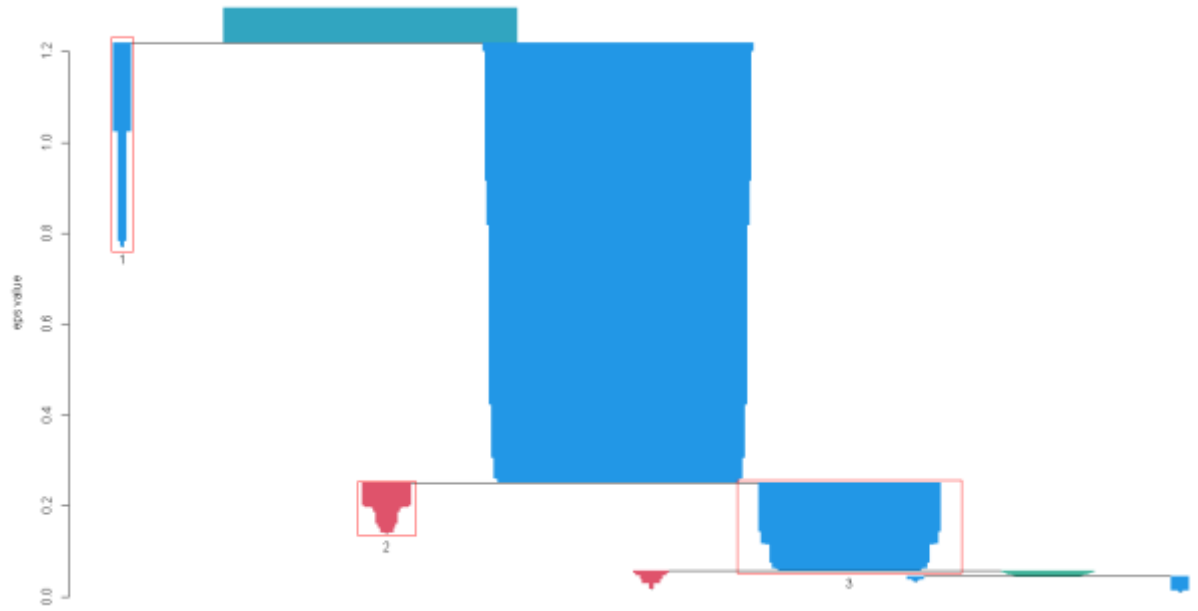

Figura 15: Dendograma Simplificada do método HDBSCAN

Fonte: Dados originais da pesquisa

Sendo assim, tanto a Figura 14 quanto a Figura 15 nos sugerem, assim como todos os métodos anteriores, a quantidade de três grupos.

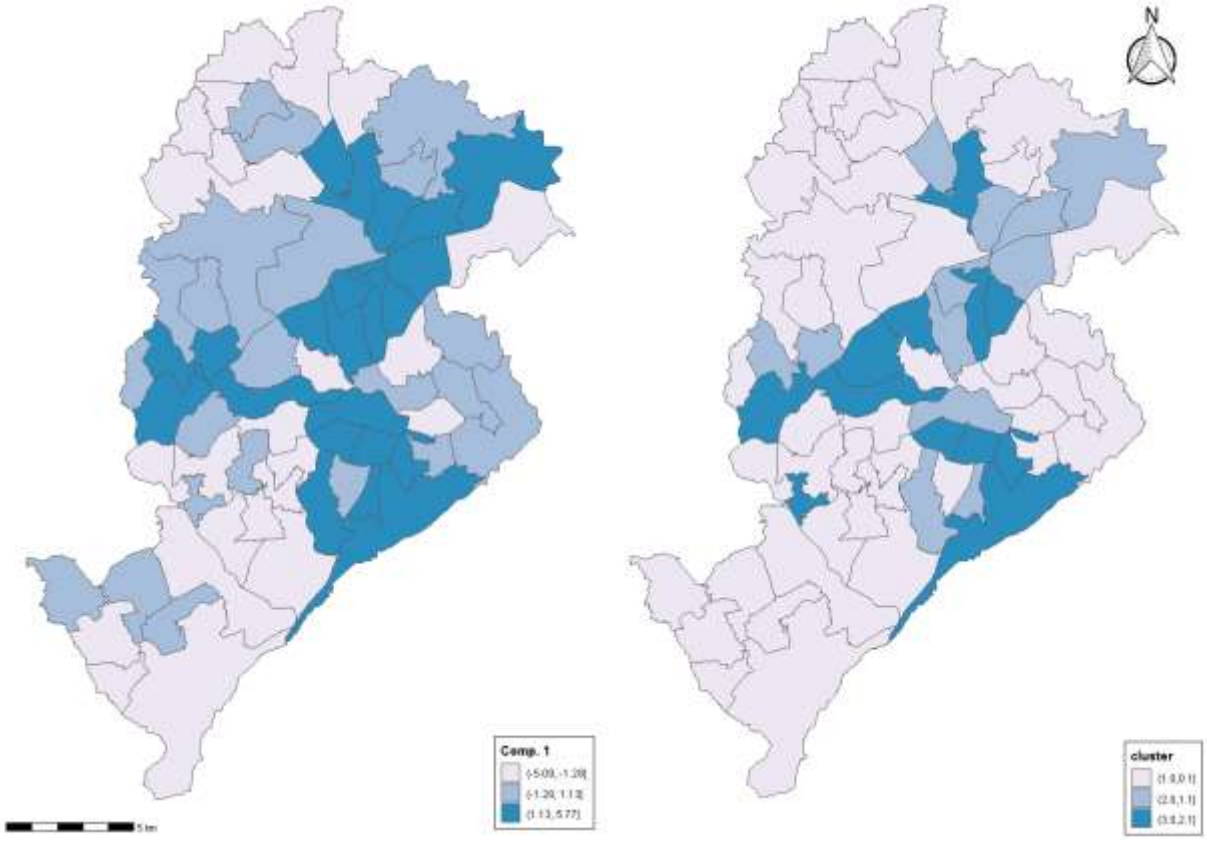

Figura 16: Tercil do escore da primeira componente principal da PCA versus agrupamento do método HDBSCAN

Fonte: Dados originais da pesquisa

A Figura 16 mostra os grupos das Áreas de Ponderação deste método plotados como dados espaciais, à direita, em comparação com o tercil do escore da primeira componente principal da PCA, à esquerda. Nota-se que este método representou, assim como os métodos método da Ligação Completa e K-Médias, razoavelmente bem os quintis da componente principal 1 da PCA.

Considerações Finais

Este trabalho teve como objetivo comparar o desempenho de 6 métodos de Análises de Agrupamentos distintos com o apoio de uma Análise de PCA.

Os resultados mostraram que os métodos de Análises de Agrupamentos de Ligação Completa, K-Médias e “Hierarchical Density Based Spatial Clustering of Application with Noise” se saíram melhores em comparação aos métodos de Ligação Simples, da Média das Distâncias e “Density Based Spatial Clustering of Application with Noise” para o conjunto de dados utilizado. Dos resultados que não foram satisfatórios supõe-se que, além das características matemáticas inerentes aos respectivos métodos, o comportamento linear da distribuição dos dados quando se optou pela utilização de dois Fatores possa ter contribuído no resultado de um agrupamento pouco eficiente. Assim, concluiu-se também que, para as variáveis trabalhadas neste estudo, aplicar uma avaliação utilizando um algoritmo de regressão multinomial para indicar as diferenças entre as probabilidades de pertencer a cada cluster em função dos escores e comparar seus R^2 e curvas ROC poderia ser mais assertivo na classificação de desempenho proposto.

Fica como outras sugestões os seguintes caminhos: o desdobramento metodológico desse trabalho através da exploração de outras variáveis; um maior aprofundamento nas questões sobre a parametrização dos modelos de agrupamento e a utilização de dados de entradas originais, ao invés do uso de Fatores fazendo, em seguida, uma comparação de resultados; a utilização de Setores Censitários, ao invés das Áreas de Ponderação, aumentando assim o número de observações trabalhadas o que poderia resultar em melhores agrupamentos nos métodos propostos.

Agradecimentos

A todos aqueles que contribuíram de alguma forma para a realização deste trabalho, em especial: Ana Caroline F. Nonato, Caroline de Oliveira Faria, Guilherme L. de Oliveira, Gustavo Miranda, José Paschoal, Leandro Araújo, Lucas Scanavini e Ludmilla Conti.

Referências

APPARICIO, P.; RIVA, M.; SÉGUIN, A.-M. 2015. A comparison of two methods for classifying trajectories: a case study on neighbourhood poverty at the intrametropolitan level in Montreal. Cybergeo: European Journal of Geography.

ARAÚJO, K. F; GOMES, R. L.; GOMES, A. S. 2019. Análise da distribuição espacial da pobreza multidimensional em um município do nordeste brasileiro. Revista Contribuciones a las Ciencias Sociales.

BARROZO, L.V; FORNACIALI, M. ANDRE, C. D. S; MORAIS, G. A. Z; MANSUR, G.; CABRAL-MIRANDA, W.; et al. 2020. GeoSES: A socioeconomic index for health and social research in Brazil. PLoS ONE 15(4):e0232074. https://doi.org/10.1371/journalpone.0232074.

CAMPELLO, R. J. G. B., MOULAVI D., SANDER J. 2013. Density-Based Clustering Basead on Hierarchical Density Estimates. Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery in Databases, PAKDD 2013, Lecture Notes in Computer Science 7819, p. 160

CENTRO DE ESTUDOS DA METRÓLE [CEM]. 2004. O Mapa da Vulnerabilidade Social da População da Cidade de São Paulo. Centro Brasileiro de Análise e Planejamento, Serviço Social do Comércio e Secretária Municipal de Assistência Social de São Paulo. São Paulo, p. 01-115.

COUTO, B. R. et al. O sistema único de assistência social no Brasil: Uma realidade em movimento. 1ª edição. ed. São Paulo: Cortez Editora, 2010.

ESTER, M.; KRIEGEL, H. P.; SANDER, J.; XU, X. 1996. A Density-Based Algorithm for Discovering Cluster in Large Spatial Databases with Noise. Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231

FÁVERO, L. P.; BELFIORE, P. 2017. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel, SPSS e Stata. 1ª edição. Elsevier Editora Ltda, Rio de Janeiro, Rio de Janeiro, Brasil.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA [IBGE]. 2011. Base de informações do Censo Demográfico 2010: Resultados do Universo por setor censitário. Centro de Documentação e Disseminação de Informações do Instituto Brasileiro de Geográfia e Estatística. Rio de Janeiro, p. 125.

INSTITUTO BRASILEIRO DE GEOGRÁFIA E ESTATÍSTICA [IBGE]. 2022. Panorama Cidades. Cidades IBGE. Disponivel em: <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama>. Acesso em: 17 abril 2022.

MATOS, D. A. S.; RODRIGUES, E. C. 2019. Análise fatorial. 1ª edição. Enap, Brasília, Distrito Federal, Brasil.

MINGOTI, S. A. 2005. Análise de dados atráves de métodos de Estatística Multivaridas: Uma abordagem aplicada. 1ª edição. Editora UFMG, Belo Horizonte, Minas Gerais, Brasil.

SANTOS, M. 2009. Pobreza Urbana. 1ª edição. Udusp, São Paulo, São Paulo, Brasil.

SEMZEZEM, P.; ALVES, J. D. M. 2013. Vulnerabilidade social, abordagem territorial e proteção na política de assistência social. Serv. Soc. Rev. v. 16, p. p. 143-166

SNEATH, P. H. A. 1957. The application of computer to taxonomy. Journal of General Microbiology, 17, p. 201-226.

Anexo 1 - Variáveis de entrada para criar GeoSES (Barrozo *et al.*, 2020)

VARIABLE	MEANING
	"Education" Dimension
P_GRAD	Percentage of people for whose kind of the highest completed degree was higher education
P_MEST	Percentage of people for whose kind of the highest completed degree was master
P_DOUTOR	Percentage of people whose kind of the highest completed degree was doctorate
P_SEM_INST	Percentage of people whose level of education is unschooled or incomplete Primary school
P_FUND	Percentage of people whose level of education is complete primary school and incomplete high school
P_ENSMED	Percentage of people whose level of education is complete high school and incomplete higher education
P_ENSSUP	Percentage of people whose level of education is complete higher education
	"Mobility" Dimension
P_OUTROMUNC	Percentage of people working in another municipality
P_CASADIA	Percentage of people returning home from work daily
P_ATE5	Percentage of people whose usual time spent commuting from home to work is up to 5 minutes
P_6A30	percentage of people whose usual time spent commuting from home to work is up to 6 to 30 minutes
P_1A2	percentage of people whose usual time spent commuting from home to work is 1-2 hours
P_MAISDE2	percentage of people whose usual time spent commuting from home to work is more than 2 hours
	"Poverty" Dimension
MEDIA_DENSMORA	Resident density per room
P_POBREZA	% of people in poverty line: whose per capita household income per month is less than or equal to R$ 255.00 or US$144.89 (half minimum wage in 2010)
P_PPI_POBREZA	% of people in the poverty line and race, black, brown or indigenous
P_BOLSA_FAM	percentage of people who in July 2010 had a regular monthly income from the *Bolsa Família* Social Program or the Child Labor Eradication Program (PETI)
P_OUTROSPROG	percentage of people who in July 2010 had regular monthly income from other social programs or transfers
	"Material deprivation" Dimension
P_ALVSREV	Percentage of homes with uncoated masonry
P_REDE_ESG	Percentage of households with general sewerage
P_REDE_AGUA	Percentage of households with general water distribution network
P_LIXO	Percentage of households with garbage collected directly by cleaning service
P_ENERGIA	Percentage of households with electricity from electricity distribution company
P_TV	Percentage of households with TV
P_MAQLAV	Percentage of households with washing machine
P_GELADEIRA	Percentage of households with refrigerator

P_MAQTVGEL	Percentage of households with washing machine, TV and refrigerator
P_CELULAR	Percentage of households with cell phones
P_COMP_INT	Percentage of households with computer with internet access
P_CELCOMPINT	Percentage of households with mobile phone and internet computer
P_MOTO	Percentage of households with motorcycle for private use
P_CARRO	Percentage of households with private car
P_ADEQ	Percentage of households with adequate housing
P_TUDOADEQ	Percentage of households with access to sewerage, water supply, garbage collection, electricity and adequate housing
P_NEM_MOTO_CARRO	Percentage of households without motorcycles or cars ownerships for private use
P_SO_MOTO	Percentage of households with only motorcycles ownership for private use
P_SO_CARRO	Percentage of households with only private car ownership
	"Income" Dimension
MED_RENDDOM	monthly household income in July 2010, in Brazilian *Reais*
	"Wealth" Dimension
P_ALUG1000	percentage of rented households with rental value of R$1,000.00 (US$ 568.20) or more
P_BANH4OUMAIS	Percentage of households with 4 or more bathrooms
P_IDOSO10SM	% of people aged 65 years and over with a monthly income equal to or above R$ 5,100.00 (US$ 2,897.72 or 10 Brazilian minimum wages)
	"Segregation" Dimension
ICE_RENDA	(number of people with income above R$ 5,400.00 - number of people with income below R$ 1,000.00) / number of respondents [figures were calculated based on the 20 and 80 percentiles of income V6529 in the PERSON spreadsheet 2010 Census microdata]
ICE_EDU	(Number of persons with completed higher education - Number of persons without education and incomplete elementary school)/Total respondents [V6400]
ICE_RENDA_PRETO	(number of whites with income over R$ 5,400.00 - number of blacks with income equal to or less than R$ 1,000.00) / total number of people who answered both questions [V6529 and V0606]
ICE_RENDA_PPI	(number of whites with income over R$ 5,400.00 - number of black + brown + indigenous with income equal to or less than R$ 1,000.00) / total number of people who answered both questions [V6529 and V0606]
ICE_BRANCO_RENDA	(number of whites with income over R$ 5,400.00 - number of whites with income equal to or less than R$ 1,000.00) / total number of people who answered both questions [V6529 and V0606]