MBA USP ESALQ

Agrupamentos Espaciais de Vulnerabilidade Social Através de Métodos de Estatística Multivariada

Renato Godoi da Cruz
Auberth Henrik Venson

Julho/2023

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;
- o Município de Belo Horizonte (em 2010):

- 2.375.151 habitantes;
- território de 331,354 km²;
- 96,2% de domicílios com esgotamento sanitário adequado;
- 82,7% de domicílios urbanos em vias públicas com arborização;
- 44,2% de domicílios urbanos em vias públicas com urbanização adequada;
- Taxa de mortalidade infantil de 11 para 1.000 nascidos vivos.

Descrição do problema

- 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;

- auxiliar na vigilância social de vulnerabilidades e riscos sociais;

- 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.

Objetivo

- O objetivo deste estudo 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;

Materiais 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.

Materiais

- Dados do GeoSES derivado Censo Demográfico de 2010 da cidade de Belo Horizonte (Áreas de Ponderação);
- 10 variáveis utilizadas:
- % de pessoas não escolarizadas ou com ensino fundamental incompleto [P_SEM_INST];
- % de pessoas cujo nível de escolaridade é o ensino superior completo [P_ENSSUP];
- % de pessoas cujo tempo gasto casa/trabalho é de até 5 minutos [P_ATE5];
- % de pessoas cujo tempo gasto casa/trabalho é superior a 2 horas [P_MAISDE2];
- média de densidade de residentes por cômodo [MEDIA_DESMORA];
- % de pessoas na linha de pobreza: [P_POBREZA];
- % de residências com alvenaria sem revestimento [P_ALVESREV];
- % de domicílios com acesso a rede de esgoto, água, coleta de lixo e moradia adequada [P_TUDOADEQ];
- média da renda familiar mensal em julho de 2010, em reais [MED_RENDDOM];
- % de pessoas com 65 anos ou mais com renda mensal igual ou superior a US$ 2.897,72 [P_IDOSO10SM]

Métodos

- Análise de Componentes Principais [PCA] – para identificar as variáveis que melhor expressam a vulnerabilidade social;
- As análises de PCA foram aplicadas sucessivamente até que se atendesse as três regras:

– 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;

– 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;

– e por último, que a hipótese nula de que a matriz de correlação seja uma matriz identidade fosse rejeitada pelo Teste de Esfericidade de Bartlett

Métodos

- Análise de Agrupamento - para capturar comportamentos semelhantes entre observações em relação a determinadas variáveis e criar grupos em que prevaleça a homogeneidade interna;
- Os métodos de agrupamento aplicados foram:

- Ligação Simples;
- Ligação Completa;
- Média das Distâncias;
- K-Médias;
- “Density Based Spatial Clustering of Application with Noise” [DBSCAN];
- “Hierarchical Density Based Spatial Clustering of Application with Noise” [HDBSCAN];

Métodos

- 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 PCA;

Resultados e Discussões

Análise de PCA

- Score fatoriais, cargas fatoriais e comunalidade dos modelos de PCA;

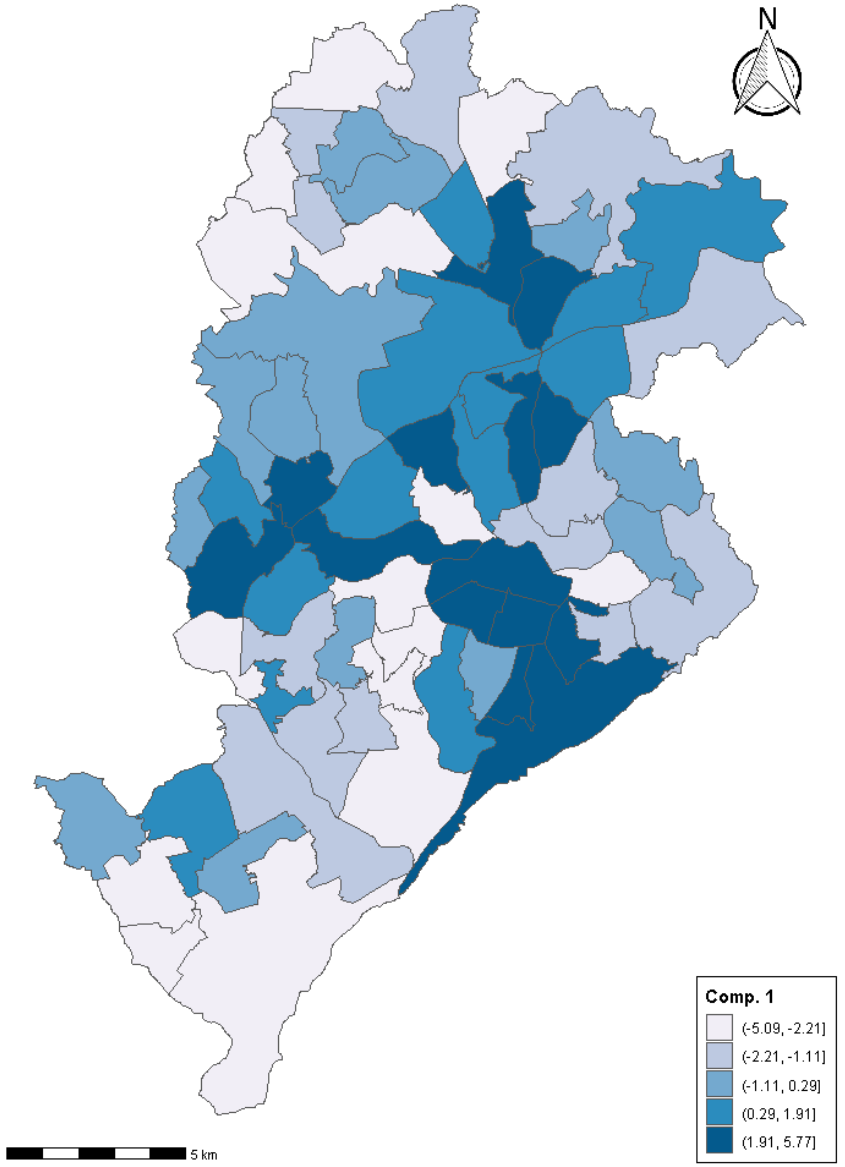
Variável	Ajuste	Fator1	Fator2	X1	X2	Comunalidade
P_SEM_INST	2	-0,14815	0,01208	-0,98302	0,01225	0,96649
P_ENSSUP	2	0,14292	0,26724	0,94829	0,27104	0,97272
M_DENSMORA	2	-0,14806	-0,08898	-0,98243	-0,09025	0,97332
P_POBREZA	2	-0,13950	0,29809	-0,92560	0,30232	0,94815
M_RENDDOM	2	0,13605	0,38100	0,90275	0,38641	0,96427
P_IDOSO10SM	2	0,12513	0,47631	0,83027	0,48307	0,92271
P_ALVSREV	2	-0,13209	0,41149	-0,87645	0,41733	0,94233
P_TUDOADEQ	2	0,12382	-0,52562	0,82155	-0,53308	0,95912

- Teste de Esfericidade de Bartlett = 2 e^{-16} (nível de signif. de 5%);
- Regra de Kaiser definiu número de 2 Fatores;
- As duas componentes principais selecionadas explicam quase 96% da variância total.

Análise de PCA

- Como a primeira 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;

- Ao lado, visualização do quintil da escore fatorial da primeira componente principal da PCA;



Métodos 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.
- Esta etapa contou com os seguintes passos:

- Determinação do número de grupos (dendograma ou método de Elbow);

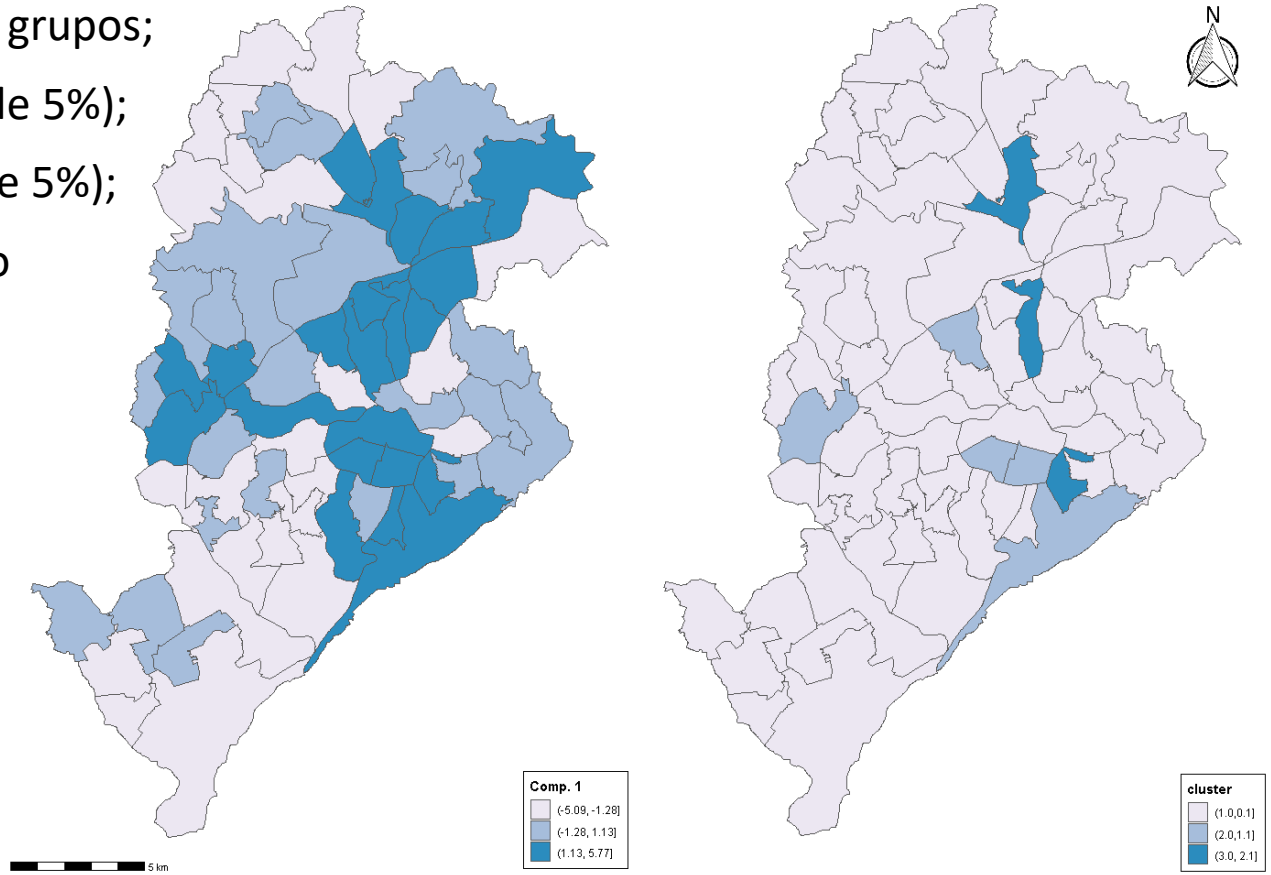
- Verificação se a variabilidade entre os grupos é significativamente superior à variabilidade interna a cada grupo (teste F da análise de variância de um fator/ANOVA);

- Comparação entre os escores da primeira componente principal da PCA com agrupamento do método em questão.

Método da Ligação Simples

- O dendograma sugeriu a quantidade de três grupos;
- Teste F fator 1 = 5.22 e^{-12} (nível de signif. de 5%);
- Teste F fator 2 = 4.67 e^{-12} (nível de signif. de 5%);
- Maior concentração de Áreas de Ponderação

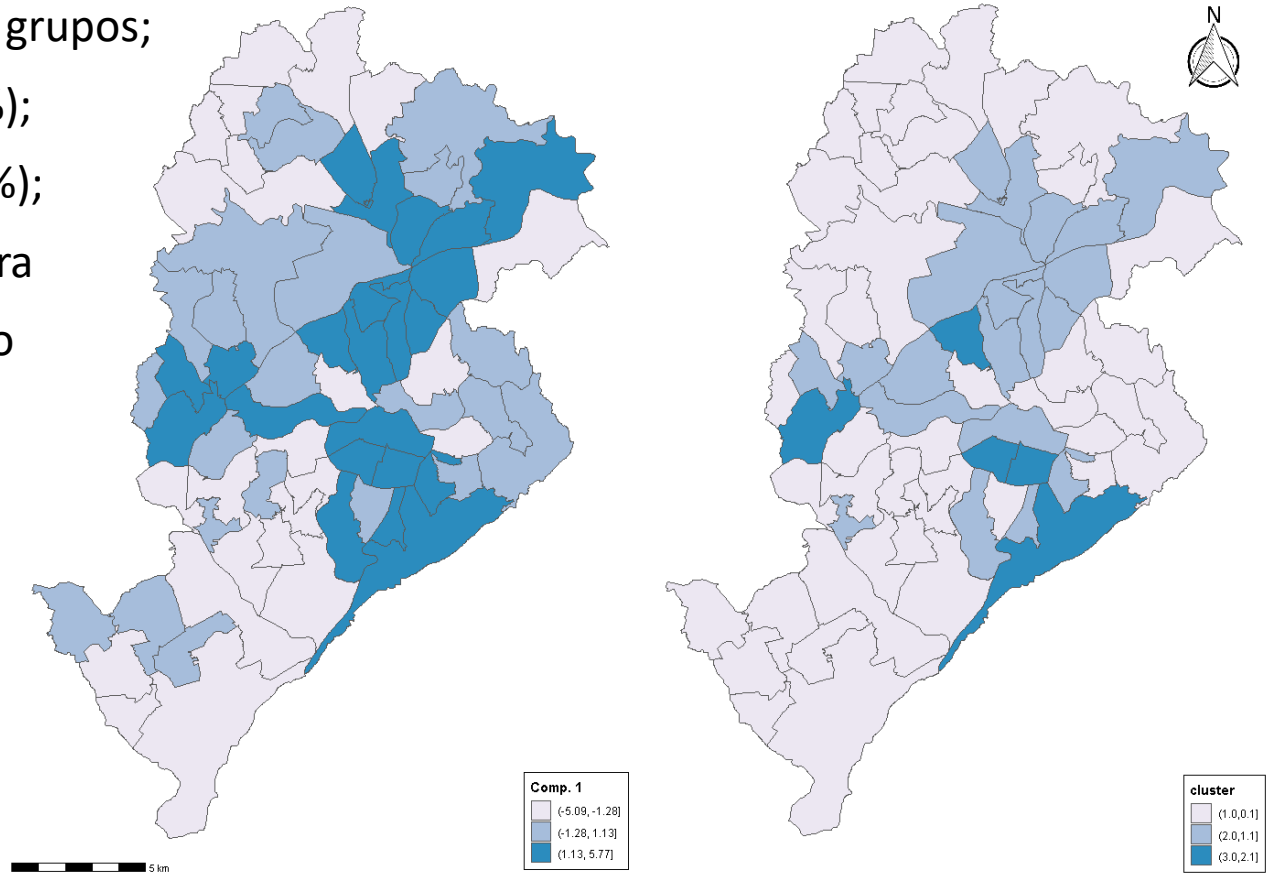
no grupo 1 em comparação com a distribuição do tercil da escore da primeira componente principal da PCA e pouca representatividade dos demais grupos



Método da Ligação Completa

- O dendograma sugeriu a quantidade de três grupos;
- Teste F fator 1 = $2e^{-16}$ (nível de signif. de 5%);
- Teste F fator 2 = $2e^{-16}$ (nível de signif. de 5%);
- Nota-se que neste método ouve uma melhora

na representatividade das Áreas de Ponderação quando comparados aos tercis dos escores originais da componente utilizadas.

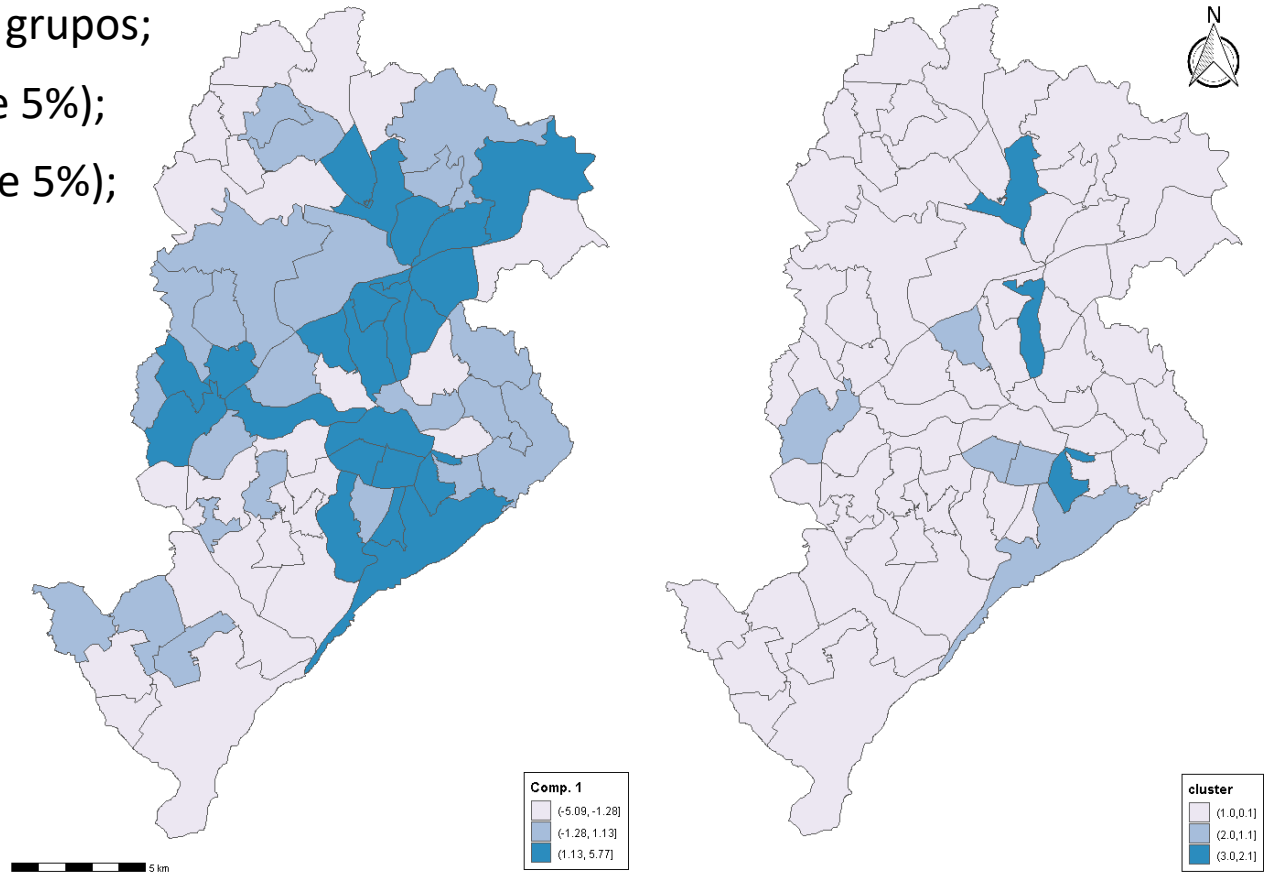


Método da Média das Distâncias

- O dendograma sugeriu a quantidade de três grupos;
- Teste F fator 1 = 5.22 e^{-16} (nível de signif. de 5%);
- Teste F fator 2 = 4.67 e^{-16} (nível de signif. de 5%);
- Nota-se que neste método representou,

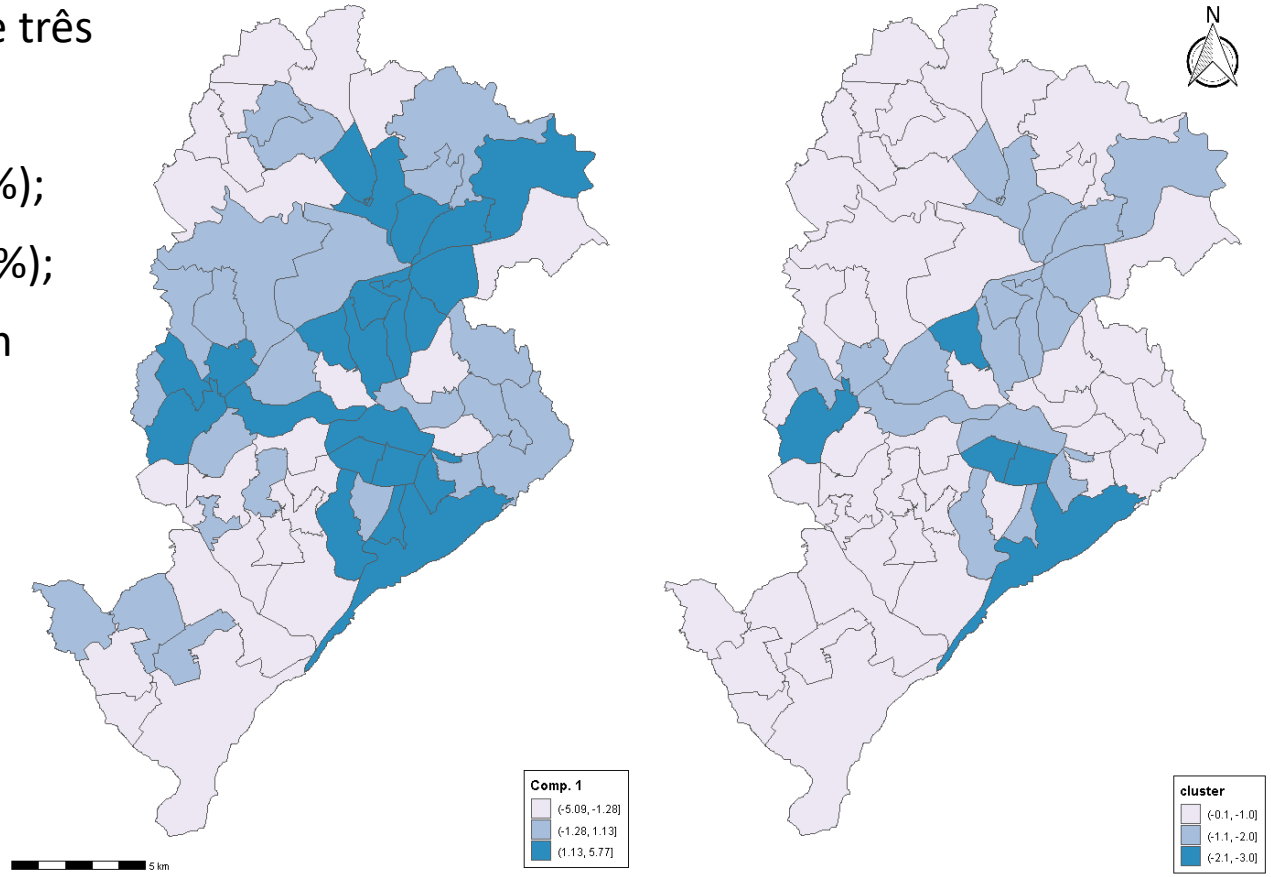
assim como no primeiro caso, pouco o

tercil da componente principal 1 da PCA.



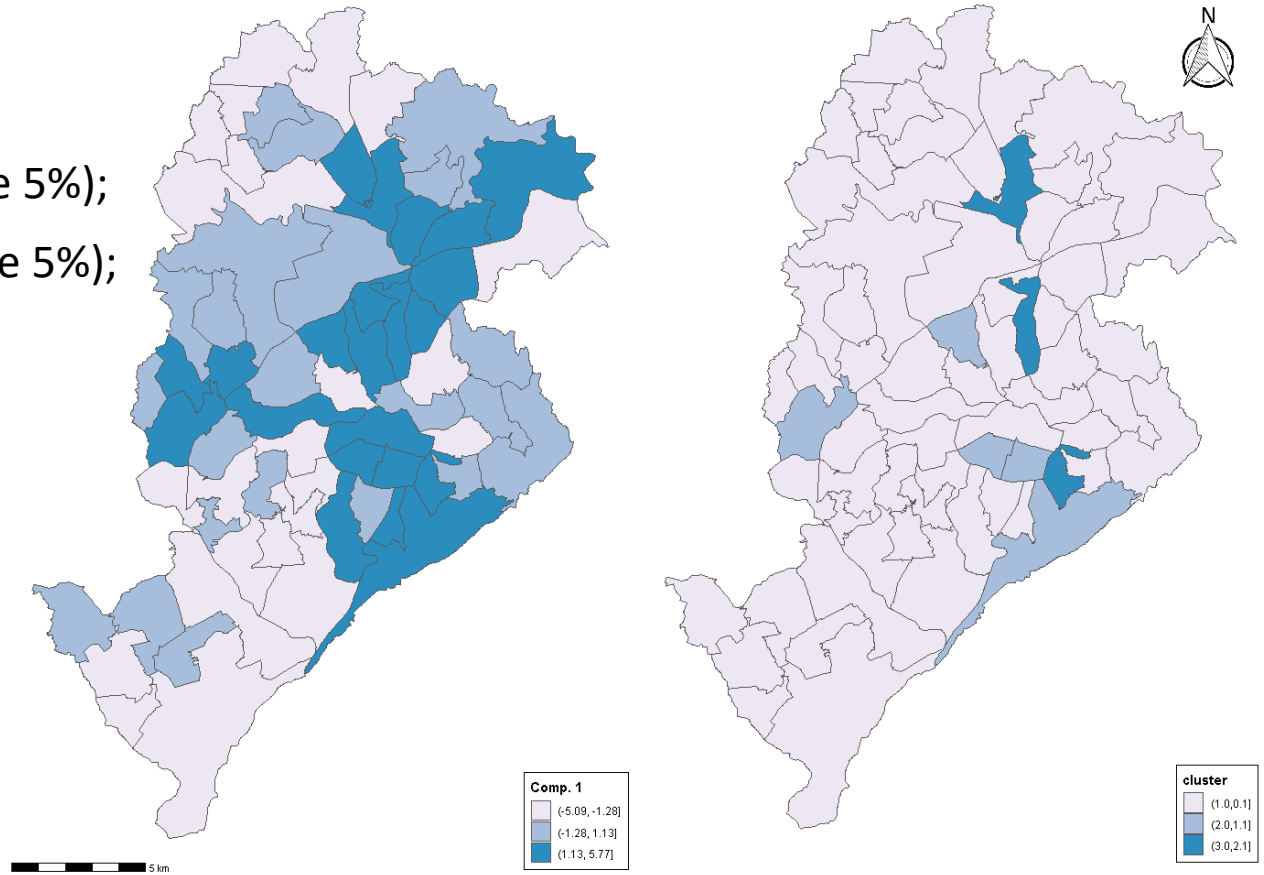
Método K-Médias

- O Método de Elbow sugeriu a quantidade de três grupos;
- Teste F fator 1 = 2 e^{-16} (nível de signif. de 5%);
- Teste F fator 2 = 2 e^{-16} (nível de signif. de 5%);
- Nota-se que este método representou, assim como no segundo caso, razoavelmente bem os tercil da componente principal 1 da PCA.



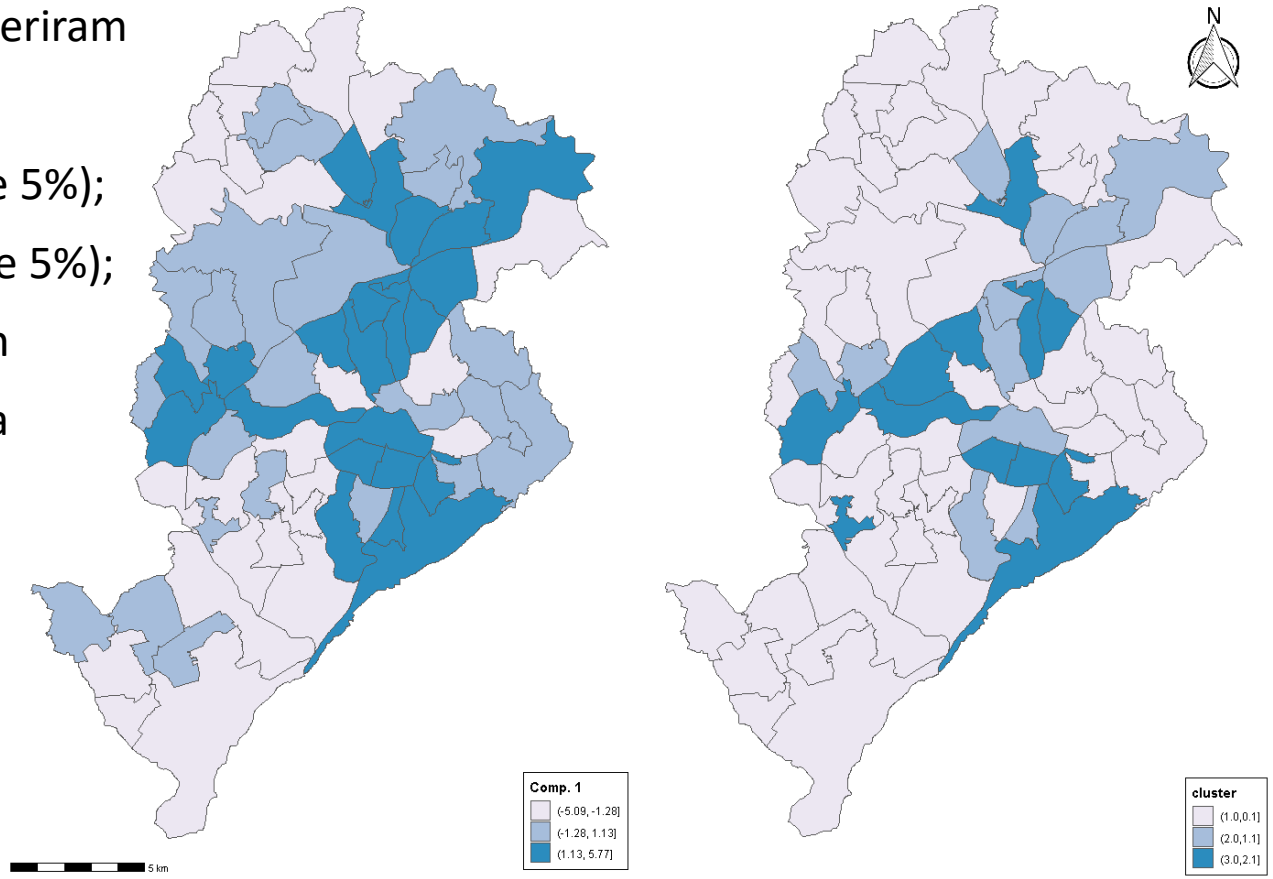
Método DBSCAN

- O Método de Elbow sugeriu a quantidade de três grupos;
- Teste F fator 1 = $5.22\ e^{-12}$ (nível de signif. de 5%);
- Teste F fator 2 = $4.67\ e^{-12}$ (nível de signif. de 5%);
- Nota-se que este método representou, novamente, pouco a componente principal 1 da PCA.



Método HDBSCAN

- O dendograma, completo e simplificado, sugeriram a quantidade de três grupos;
- Teste F fator 1 = 2.57 e^{-15} (nível de signif. de 5%);
- Teste F fator 2 = 3.07 e^{-15} (nível de signif. de 5%);
- Nota-se que este método representou, assim como os métodos método da Ligação Completa e K-Médias, razoavelmente bem os tercil da componente principal 1 da PCA.



Considerações finais

- Os resultados mostraram que os métodos de Análises de Agrupamentos de Ligação Completa, K-Médias e HDBSCAN se saíram melhores em comparação aos métodos de Ligação Simples, da Média das Distâncias e DBSCAN para o conjunto de dados utilizado;
- Em relação aos resultados insatisfatórios, supõe-se que, além das características matemáticas inerentes aos respectivos métodos, o comportamento linear da distribuição dos dados ao utilizar dois Fatores pode ter contribuído para um agrupamento menos eficiente;
- Adicionalmente, concluiu-se 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 com base nos escores e comparar seus R^2 e curvas ROC poderia proporcionar uma classificação de desempenho mais precisa

Sugestões para trabalhos futuros

- O desdobramento metodológico desse trabalho através da exploração de outras variáveis dos dados do censo;
- 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

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
- 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

MBA USP ESALQ

Referências

- 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, Brasi

MBA USP ESALQ

Referências

- 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

MBA
USP
ESALQ

OBRIGADO