**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Renato Godoi da Cruz | | **Data início curso**: 11/05/2021 |
| **Orientador**: Auberth Henrik Venson | | **Defesa em:**  12/2022 |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 202 |

1. **Título do projeto**

**Comparação de algoritmos de agrupamento espacial: um estudo de caso sobre a pobreza de bairros da cidade de São Paulo**

1. **Introdução**

*(Contextualizar e apresentar a problemática do tema geral, ou seja, a importância do tema proposto e sua relevância. O texto deverá ser escrito de forma impessoal e toda informação utilizada deverá ser embasada por meio de trabalhos de fontes confiáveis com as devidas citações dos autores)*

Nosso estudo se propõe a analisar trajetórias de pobreza de bairros em um período mais longo de 20 anos. Estatisticamente, isso será alcançado aplicando e comparando ~~duas~~ técnicas de agrupamento para agrupar bairros que seguiram trajetórias semelhantes de mudança nos níveis de pobreza.

Veremos mais adiante que essas abordagens permitem maximizar a variação entre trajetórias e minimizar a variação dentro das trajetórias. Esta etapa é crucial se o objetivo é identificar os determinantes da mudança da vizinhança e medir sua importância relativa. Este artigo baseia-se em trabalhos anteriores definindo trajetórias de pobreza usando cinco pontos de tempo, permitindo que a magnitude e a mudança de direção da pobreza variem em cada ponto de tempo.

Essa precisão, no entanto, vem com desafios metodológicos, incluindo a construção de um banco de dados longitudinal na escala intrametropolitana (ou seja, no nível do setor censitário) com dados socioeconômicos e fronteiras geográficas comparáveis e harmonizados ao longo de vários anos censitários.

Outro desafio metodológico é encontrar a abordagem mais precisa para grupos de bairros caracterizados por uma evolução semelhante de sua população pobre ao longo do tempo, com cada grupo (ou seja, trajetória) sendo mais diferente entre si. Na próxima seção, discutiremos duas técnicas possíveis.

(APPARICIO, RIVA e SÉGUIN, 2015) (FÁVERO e BELFIORE, 2017)

Setor Censitário do Município de São Paulo <http://dados.prefeitura.sp.gov.br/es/dataset/setor-censitario-do-municipio-de-sao-paulo>

1. **Objetivo**

*(Qual o objetivo principal do trabalho, ou seja, qual pergunta deve ser respondida ao final da sua pesquisa)*

Objetivos do estudo

Com o objetivo de caracterizar melhor as trajetórias de mudança da pobreza nos bairros, o objetivo deste estudo é aplicar e comparar duas técnicas de agrupamento a 20 anos de dados do censo (cinco momentos) para identificar grupos de bairros que seguiram trajetórias semelhantes de pobreza entre 1986 e 2006. Aplicamos as técnicas k-means e LCGM para avaliar qual método tem melhor desempenho na identificação de trajetórias de pobreza. A seleção da classificação mais precisa representa um passo crucial antes de desenvolver modelos explicativos das mudanças socioeconômicas que operam nas áreas metropolitanas.

Vários estudos demonstraram que a precisão de agrupamento do k-means é superior à do HCA, especialmente quando calculado em grandes conjuntos de dados (ver, por exemplo, Abbas, 2008). Além disso, os resultados do HCA variam de acordo com a métrica de distância (distância euclidiana, distância euclidiana quadrada, distância de Mahalanobis, etc.) Para evitar a comparação dos resultados dos modelos LCGM com várias variantes do HCA, optou-se pelo agrupamento k-means.

1. **Material e Métodos**

*(Descrever o(s) método(s) de coleta de dados e a(s) ferramenta(s) de análise a ser(em) utilizada(s) no trabalho de conclusão de curso, ou seja, como será a condução da pesquisa e a forma de obtenção dos resultados, por exemplo, fontes de dados, técnicas, procedimentos, índices, entre outros)*

Área de estudo e dados

|  |
| --- |
| Os dados publicados pelo IBGE para o Município de São Paulo (MSP) estão organizados por Distritos  Administrativos (DA) ou Setores Censitários. Para o Censo 2010 a cidade foi dividida em 18.952 setores censitários. A organização dos dados por Distrito e por Subprefeitura é feita pela Secretaria de Desenvolvimento Urbano e os dados podem ser consultados em <https://www.prefeitura.sp.gov.br/cidade/secretarias/licenciamento/desenvolvimento_urbano/dados_estatisticos/info_cidade/index.php/>. Os dados relativos às CRS e STS foram trabalhados pela Gerência de Geoprocessamento e Informações Socioambientais (GISA) da Coordenação de Epidemiologia e Informação (CEInfo). |

Este estudo está situado no Canadá, na Área Metropolitana do Censo de Montreal (CMA) compreendendo uma população de cerca de 3,64 milhões de habitantes distribuídos por um território de 4.259 km2 em 2006 (Statistics Canada, 2007). As áreas intrametropolitanas são definidas usando os limites dos setores censitários.

Os limites geográficos administrativos e censitários na CMA de Montreal mudaram consideravelmente entre 1986 e 2006: o número de setores censitários aumentou de 698 para 825 nesse período. A harmonização dos limites geográficos dos setores censitários era, portanto, necessária. Começando com os limites geográficos dos setores censitários de 1986 (o ponto mais antigo), isso foi alcançado pela agregação de setores censitários contíguos para obter fronteiras em todos os anos censitários. Obteve-se um total de 611 setores censitários.

A pobreza relativa foi medida a cada cinco anos entre 1986 e 2006 usando dados do Censo Canadense usando a variável “cut-offs de baixa renda” calculada pelo Statistics Canada. Essa variável corresponde ao nível de renda em que uma família gasta 20% ou mais de sua renda (antes de impostos) com o básico (ou seja, alimentação, abrigo e roupas) do que a média de uma família de tamanho similar (Statistics Canada, 2011).

Essa medida é a única do censo canadense que permite identificar pessoas ou famílias de baixa renda em pequena escala geográfica, por exemplo. setores censitários (Apparicio et al., 2007; Séguin et al., 2012). Como a comparação das taxas de níveis de pobreza "brutos" entre setores censitários e ao longo do tempo pode ser influenciada pela economia em mudança (ou seja, períodos de recessão ou prosperidade econômica), a pobreza foi modelada como um "quociente de localização" para que, em cada ponto do tempo, a pobreza a taxa de cada setor censitário foi dividida pela taxa observada para o CMA como um todo; estamos, portanto, usando uma medida de concentração de pobreza relativa. A proporção da população de baixa renda na CMA em cada momento é mostrada na Tabela 2. O quociente de localização fornece uma visão geral de como, a qualquer momento, os níveis de pobreza locais se comparam à média da CMA. Essa medida de concentração é amplamente utilizada em estudos urbanos e regionais (Mikelbank, 2006; Shearmur et al., 2008; Shearmur et al., 2009; Vicino et al., 2011; Walks et al., 2008), e é calculada como Segue:

Um quociente de localização maior que 1 indica uma concentração de pobreza (ou seja, uma porcentagem da população de baixa renda maior que a da CMA), enquanto um valor abaixo de 1 indica uma sub-representação da pobreza (ou seja, uma porcentagem da população de baixa renda menor que a de o CMA)

Métodos de agrupamento LCGM e k-means para identificar trajetórias de pobreza relativa

A geração de trajetórias de pobreza de bairros foi realizada pela primeira vez usando LCGM, pois esta técnica fornece estatísticas diagnósticas sobre a solução ótima de cluster. Estabelecemos como critério inicial que cada conglomerado/trajetória necessitasse de um mínimo de 5% dos 611 setores censitários, ou seja, um mínimo de 30 setores censitários por trajetória. Isso foi definido para garantir um mínimo de observações por trajetória no estágio de validação posterior (por exemplo, e de acordo com o requisito mínimo de observação para análise de regressão). Como não tínhamos a priori o número ideal de aulas, o LCGM foi realizado de 5 a 20 aulas; um mínimo de 5 aulas foi definido para ter um mínimo de diferenciação entre grupos de setores censitários. A solução ótima de agrupamento é identificada por: 1) mínimo de 5% de setores censitários por trajetória; e 2) o menor valor de BIC. As análises foram realizadas no software LatentGOLD (Statistical Innovations).

O agrupamento K-means foi realizado no SAS 9.2 (SAS Institute Inc), especificando novamente 5 a 20 agrupamentos. O valor da distância média ao centroide do cluster para cada solução de cluster foi plotado para identificar uma "quebra natural" na distribuição, indicando a solução de cluster ideal. No final, a escolha do número ideal de soluções de cluster foi informada pela solução LCGM fornecendo o melhor ajuste aos dados.

**análise estatística**

Para avaliar o desempenho relativo do LCGM *versus* o agrupamento k-means na identificação de trajetórias de pobreza relativa, dois conjuntos de análises foram conduzidos. Primeiro, em uma regressão logística multinomial (MLR), as variáveis usadas para a classificação (ou seja, os quocientes de localização de 1986 a 2006) foram modeladas como preditoras das trajetórias (as trajetórias são modeladas como uma variável dependente categórica). Esta abordagem é uma forma de análise discriminante, usada para testar o desempenho de diferentes métodos de classificação (ver, por exemplo, Magidson et al., 2002) ou diferentes números de soluções de cluster. A ideia aqui é usar o R-Square e as estatísticas de ajuste do modelo desta análise para informar qual das soluções de cluster k-means e LCGM resume melhor a variação na concentração da pobreza.

Uma segunda série de MLR foi então conduzida para examinar empiricamente como um conjunto de preditores teoricamente associados à pobreza explica cada trajetória: taxa de desemprego, famílias monoparentais (%), famílias unipessoais (%), idosos (≥ 65 anos) ( %), imigrantes recentes (%), população com baixa escolaridade (%), educação universitária (%) e inquilinos (%) (ver Tabela 2 para uma descrição dos valores tomados pelos preditores entre 1986 e 2006 para a região de estudo).

Essas variáveis foram mantidas porque estudos recentes demonstraram que elas estão fortemente associadas à distribuição espacial da pobreza na CMA de Montreal no setor censitário nível (Apparicio et al., 2007; Séguin et al., 2012). Em modelos de MLR separados, esses preditores foram modelados no início do período, ou seja, 1986, no final, ou seja, em 2006, e como variação entre 1986 e 2006 (por exemplo, taxa de desemprego 2006 - taxa de desemprego 1986). Um modelo final incluindo preditores de linha de base e variação entre 1986 e 2006 foi executado. Para cada modelo, o foco está na força (R-Quadrado) e no ajuste (Critério de Informação de Akaike [AIC] e Critério de Informação Bayesiano [BIC]; valor mais baixo do AIC e BIC são indicativos de melhor ajuste do modelo) do modelo.

1. **Resultados Esperados**

*(Descrever os resultados que são esperados após a realização da coleta e análise dos dados, ou seja, quais resultados são esperados ao final da pesquisa)*

1. **Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Pesquisa bibliográfica |  |  | **x** | **x** |  |  |  |  |  |  |
| Identificação de oportunidades no  mercado “plant-based”, utilizando-se a  matriz SWOT |  |  | **x** | **x** |  |  |  |  |  |  |
| Contato com indústrias de alimentos e  profissionais que atuem no mercado  “plant-based”, com a finalidade de  obter informações específicas e  enriquecer o trabalho |  |  | **x** | **x** |  |  |  |  |  |  |
| Aplicação inicial do BMC |  |  |  | **x** | **x** | **x** |  |  |  |  |
| Avaliação individual de cada bloco que  compõe do BMC |  |  |  |  |  |  |  |  |  |  |
| Confrontamento entre BMC e matriz  SWOT |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Redação da monografia |  |  | **x** | **x** | **x** | **x** | **x** | **x** | **x** |  |
| Elaboração de manuscrito para publicação |  |  |  |  |  |  |  |  |  |  |
| Elaboração da apresentação da defesa |  |  |  |  |  |  |  |  | **x** |  |
| Defesa |  |  |  |  |  |  |  |  |  | **x** |
| Ajustes da monografia |  |  |  |  |  |  |  |  |  | **x** |
| Elaboração de artigo para publicação em revista especializada |  |  |  |  |  |  |  |  |  | **x** |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa.

1. **Referências Bibliográficas**

APPARICIO, P.; RIVA, M.; SÉGUIN, A.-M. 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**, 2015.

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