**A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal**

Uma comparação de dois métodos de classificação de trajetórias: um estudo de caso sobre a pobreza de bairros em nível intrametropolitano em Montreal

**Resumo**:

Nos últimos anos, vários estudos examinaram as mudanças na distribuição da pobreza nas cidades norte-americanas, com a maioria dos trabalhos empíricos avaliando a mudança de bairro entre dois momentos. Este artigo visa dar uma contribuição metodológica ao estudo da mudança de bairro, comparando dois métodos de classificação, um clássico (*k-means clustering*) e outro mais inovador (*Latent Class Growth Modeling; LCGM*) para identificar grupos de setores censitários que seguiram trajetórias de pobreza na área metropolitana de Montreal, Canadá. Aqui as trajetórias da pobreza são medidas ao longo de um período de vinte anos, usando cinco pontos de tempo. O desempenho relativo do LCGM versus o agrupamento k-means foi avaliado usando uma série de regressões logísticas multinomiais examinando como diferentes variáveis socioeconômicas foram associadas às trajetórias da pobreza. Os resultados mostraram que k-médias e LCGM identificaram grupos semelhantes de setores censitários caracterizados por níveis de pobreza ascendente, descendente ou estável ao longo do período, com LGCM superando apenas marginalmente o agrupamento de k-médias.

**Fundo**

**Estudando a mudança da pobreza no bairro**

Vários estudos preocupados com as mudanças nos bairros de alta pobreza, pelo menos nos EUA, focaram no crescimento desses bairros em áreas suburbanas, e especialmente em subúrbios internos. Alguns estudos visam identificar a evolução na distribuição das zonas de pobreza nas áreas metropolitanas, notadamente opondo o centro da cidade aos subúrbios (Cooke et al., 2006; Lee, 2011; Lee et al., 2007). Outras pesquisas como o estudo de McConville e Ong (2003), se interessam pelas trajetórias dos bairros pobres, ou seja, se permaneceram pobres, pioraram ou melhoraram ao longo do tempo, em relação à mudança nas condições de outros bairros, por exemplo: etnia, imigração, educação, emprego e perfis familiares.

No Canadá, poucos estudos examinaram as transformações da vizinhança da pobreza. Em 2000, Ley e Smith (2000) notaram a mudança na natureza de alguns setores censitários em Toronto, Montreal e Vancouver ao longo de um período de vinte anos, observando que bairros carentes em 1971 não eram necessariamente carentes em 1991 e inversamente, bairros não carentes em 1971 poderiam ter se tornados carentes em 1991. Suas observações basearam-se em vários indicadores cumulativos associados à privação (medidos usando limiares) medidos primeiro em 1971 e depois em 1991.

Usando dados do censo, Heisz e McLeod (2006) mostraram que tanto a proporção de bairros de baixa renda quanto sua distribuição espacial nas diferentes áreas metropolitanas canadenses variaram entre 1981 e 2001. Por exemplo, eles observaram que, em comparação com 1981, os bairros de baixa renda bairros em Montreal e Toronto eram menos propensos a estar localizados em bairros do centro da cidade em 2001 e mais em áreas suburbanas do anel interno. Embora este estudo de Heisz e McLeod tenha identificado tendências amplas para cada área metropolitana, os autores não identificaram as trajetórias dos bairros per se. Um estudo de Kitchen e Williams (2009) realizado em Saskatoon, área metropolitana canadense de tamanho moderado, analisou a mudança de bairro entre 1991 e 2001, considerando dois “subperíodos”, 1991-1996 e 1996-2001. Suas análises consideraram o perfil socioeconômico de 58 bairros no início do período (1991), classificando-os como bairros de baixo, médio ou alto nível socioeconômico (SES). Os bairros foram então caracterizados como seguindo três trajetórias possíveis de mudança no SES entre 1991 e 2001, ou seja, declínio, melhoria ou estabilidade. O estudo de Kitchen e Williams (2009) é interessante por considerar as condições socioeconômicas no início do período de estudo e suas mudanças ao longo de um período de 10 anos. No entanto, dez anos de observação podem ser muito curtos para identificar mudanças importantes nos processos urbanos, como filtragem para baixo (migração de domicílios para novos e mais elaborados residências e imigração de famílias de menor riqueza e menor status social), gentrificação e suburbanização da pobreza.

Nosso estudo se propõe a analisar trajetórias de pobreza de bairro em um período mais longo de 20 anos. Conforme Kitchen e Williams (2009), consideramos os níveis de pobreza no início do período, ou seja, 1986, mas também os níveis de pobreza do bairro em 1991, 1996, 2001 e 2006. Assim, ao longo do tempo, um bairro pode ser caracterizado por uma trajetória ascendente , depois uma trajetória descendente, seguida por um nível estável de pobreza, para uma inclinação decrescente no final do período de estudo. Em cada momento, a pobreza é medida como uma variável contínua. Isso nos permitirá identificar trajetórias com mais precisão e identificar um número ótimo de trajetórias.

Estatisticamente, isso será alcançado aplicando e comparando duas técnicas de agrupamento para agrupar bairros que seguiram trajetórias semelhantes de mudança nos níveis de pobreza. Portanto, o objetivo deste artigo é essencialmente metodológico, ou seja, identificar um método de agrupamento mais adequado para medir a mudança de vizinhança. 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.

**Identificando trajetórias de mudança de bairro no nível intrametropolitano: as possibilidades oferecidas pelo agrupamento k-means e modelagem de crescimento de classes latentes**

A modelação da mudança social tem sido abordada principalmente a partir de séries temporais e perspectivas econométricas, nomeadamente para estudar os ciclos económicos e as mudanças nos mercados de trabalho a níveis geográficos alargados. Também é possível visualizar trajetórias de mudança em termos de “grupos” de áreas que seguiram um padrão semelhante de mudança ao longo de uma variável (ou variáveis) de interesse ao longo do tempo. Até o momento, técnicas de agrupamento não supervisionadas, como análise hierárquica de agrupamento (HCA) e agrupamento k-means, ou extensões de k-means (por exemplo, fuzzy k-means (Friedman et al., 1998), particionamento em torno do medoid (Kaufman et al. . no tempo (ver, por exemplo, Mikelbank, 2004; Vicino et al., 2011) e menos frequentemente em dois momentos (Reibel et al., 2007, 2011; Vicino, 2008).

No entanto, é possível aplicar essas técnicas de agrupamento “clássicas” a um conjunto de dados longitudinal para identificar trajetórias de vizinhança. Por exemplo, Mikelbank (2011) aplicou um HCA para agrupar setores censitários ao longo de várias variáveis demográficas, habitacionais e socioeconômicas extraídas de quatro censos (1970, 1980, 1990, 2000). Essas variáveis foram padronizadas (zscores) em cada momento e, em seguida, foram anexadas em uma tabela, na qual um HCA foi calculado. Isso permitiu identificar cinco tipos de bairros ao longo do período: lutando, lutando afro-americano, estabilidade, novos começos e subúrbio. Assim, um setor censitário pode pertencer ao mesmo tipo de bairro durante os quatro anos censitários, ou pode mudar de tipo uma ou mais vezes. Finalmente, Mikelbank (2011) construiu várias tabelas de transição para identificar a estabilidade ou mudança nas trajetórias dos bairros ao longo de dois anos ou todos os períodos de tempo.

Considerando que é possível aplicar técnicas clássicas de agrupamento a conjuntos de dados espaciais longitudinais, novos procedimentos analíticos semi-paramétricos foram recentemente desenvolvidos para agrupar “objetos” que seguiram tendências semelhantes ao longo do tempo, ou seja, Modelos de Crescimento de Classes Latentes (LCGM) (Nagin, 2005). Até o momento, o LCGM tem sido aplicado principalmente em psicologia (Nagin, 2005) e epidemiologia, por exemplo, para agrupar indivíduos com trajetórias semelhantes de mudança em comportamentos relacionados à saúde (Barnett et al., 2008; Brookmeyer et al., 2009). Até onde sabemos, ele foi aplicado a dados espaciais para examinar trajetórias de mudança dentro de uma área metropolitana ou de um país em apenas três estudos (Apparicio et al., 2012; Pearson et al., 2013; Riva et al., 2012).

K-médias e LCGM são técnicas de agrupamento estatístico que podem ser aplicadas para classificar objetos (ou seja, setores censitários) em k número de grupos (ou seja, trajetórias) com mudança semelhante em uma variável (ou seja, pobreza) ao longo do tempo. K-means é uma técnica exploratória de agrupamento estatístico que usa um algoritmo de alocação/re-alocação para reatribuir de forma otimizada setores censitários ao centróide de agrupamento mais próximo (Everitt et al., 2001). O objetivo é maximizar as variações entre os clusters e minimizar as variações dentro do cluster e, assim, agrupar em k tipos de áreas locais que seguiram trajetórias semelhantes de mudança na pobreza durante o período de estudo. Alguns softwares estatísticos, por ex. SAS, otimiza a escolha de centros de cluster iniciais; assim, a seleção aleatória de centros de cluster, potencialmente levando a soluções diferentes quando o modelo é executado novamente, não é mais um “problema”.

Comparado aos algoritmos HCA, o número de k deve ser escolhido a priori no agrupamento k-means e existem vários métodos para identificar a solução ótima de agrupamento (Milligan et al., 1985) incluindo, entre outros: a estatística Pseudo-F (Calińskia et al., 2007); o Critério de Agrupamento Cúbico (Sarle, 1983); ou um método mais discursivo (Tibshirani et al., 2001) que consiste em traçar a distância média ao centroide do cluster para cada solução de cluster e identificar visualmente a solução de cluster ótima onde há um nivelamento natural na distribuição (o 'critério do cotovelo') .

LCGM é uma abordagem semi-paramétrica para classificação (Andruff et al., 2009; Collins et al., 2009; Duncan et al., 2009). Embora cada setor censitário siga uma trajetória única de mudança nos níveis de pobreza, a heterogeneidade na distribuição dos setores censitários é resumida por um conjunto finito de funções polinomiais, cada uma correspondendo a uma classe ou trajetória discreta (Andruff et al., 2009; Nagin, 2005). Como a magnitude e a direção da mudança podem variar livremente entre as trajetórias, um conjunto de parâmetros do modelo, ou seja, interceptação e inclinação, é estimado para cada trajetória (Andruff et al., 2009; Nagin, 2005). Para cada trajetória, a inclinação e o intercepto são tratados como fixos (iguais) entre os setores censitários. No LCGM, o número ideal de classes é informado por uma abordagem de modelagem construída em que a modelagem começa com um modelo de uma classe e as classes são posteriormente adicionadas para avaliar a melhoria no ajuste do modelo. O modelo que fornece o melhor ajuste aos dados é identificado pela interpretação e comparação de várias ferramentas de diagnóstico, incluindo o modelo com o Critério de Informação Bayesiano (BIC) mais baixo e probabilidades posteriores de associação ao grupo (uma regra de atribuição de probabilidade máxima é usada para atribuir a cada indivíduo à trajetória para a qual detém a maior probabilidade de pertinência posterior) (Andruff et al., 2009).

O LCGM agora é relativamente fácil de aplicar em softwares como SAS (ProcTRAJ) (Jones et al., 2007), Mplus (Muthén & Muthén) e LatentGOLD (Statistical Innovations). As principais diferenças entre k-means e LCGM estão resumidas na Tabela 1 (adaptado do site LatentGold: <http://www.statisticalinnovations.com/articles/kmeans2a.htm>).

Os princípios subjacentes de k-médias e LCGM são, portanto, diferentes: um é uma técnica descritiva/exploratória, enquanto o outro adota uma abordagem semi-paramétrica para a classificação. Também vale a pena notar que K-Means requer variáveis ​​contínuas ou dicotômicas enquanto LCGM pode ser aplicado a qualquer tipo de distribuição (contínua, ordinal, nominal, contagem e binomial). Neste estudo, a variável utilizada para a classificação –i.e. o quociente de localização – é contínuo. Ainda não

estudos compararam como os métodos se saem na geração de grupos de unidades espaciais que seguiram trajetórias de mudança semelhantes.

**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.

**material e métodos**

**Área de estudo e dados**

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 .

**Resultado**

O objetivo deste artigo foi comparar duas técnicas de agrupamento, agrupamento k-means e modelos de crescimento de classes latentes, a fim de melhor caracterizar trajetórias de pobreza de bairros ao longo de um período de vinte anos. Assim, a principal contribuição deste artigo para o estudo da mudança de vizinhança é metodológica, contribuindo com novas aplicações de técnicas de agrupamento estatístico para dados longitudinais em nível de área. Os resultados mostraram que as técnicas de k-médias e LCGM foram bem-sucedidas no agrupamento de setores censitários caracterizados por trajetórias semelhantes de mudança nos níveis de pobreza entre 1986 e 2006. Conforme observado em outros estudos, diferentes trajetórias de mudança de bairro na CMA de Montreal foram capturadas, incluindo trajetórias de alta concentração de população de baixa renda em áreas urbanas; áreas suburbanas caracterizadas por uma concentração crescente da pobreza ao longo do período de vinte anos.

Esses resultados corroboram outros estudos que relatam uma suburbanização da pobreza em algumas áreas metropolitanas da América do Norte (Cooke et al., 2006; Madden, 2003); a gentrificação de certas áreas centrais (Heisz et al., 2006; Ley, 1986, 1988; Walks et al., 2008) juntamente com o aumento da concentração da pobreza em outras partes da cidade central durante os últimos vinte anos (Bourne, 1993; Madden, 2003). No entanto, a maioria dos setores censitários são caracterizados por níveis de pobreza ao longo do período, quer em termos de níveis estáveis mas elevados de concentração da pobreza, quer em termos de representação estável mas baixa da pobreza. No entanto, a maioria dos estudos de trajetórias de bairros estão principalmente preocupados com a mudança, ou seja, aumento ou diminuição dos níveis de pobreza. No entanto, nosso estudo mostra que algumas áreas têm experimentado persistentemente altos níveis de pobreza entre 1986 e 2006, enquanto em outras, os níveis de pobreza são continuamente mais baixos em comparação com a CMA como um todo.

No geral, os resultados das duas soluções de cluster foram relativamente semelhantes. No entanto, encontramos algumas discrepâncias entre os mapas de trajetórias de privação relativa obtidos pelos métodos k-means e LCGM. Para os setores censitários localizados na periferia, a solução LCGM identificou três trajetórias estáveis com baixos níveis de pobreza contra duas para o k-médias. Nas áreas centrais, a solução LCGM agrupou em uma trajetória 45 setores censitários caracterizados por um declínio da concentração da pobreza como resultado de um processo de gentrificação. No entanto, esses 45 setores censitários estão divididos em duas trajetórias de diminuição da concentração da pobreza na solução k-means. Em outras palavras, a solução de cluster LCGM forneceu mais detalhes nos subúrbios, enquanto o k-means forneceu mais detalhes nas áreas centrais.

Finalmente, os resultados das regressões logísticas multinomiais mostraram que o LCGM superou marginalmente o agrupamento k-means. Uma vez que hoje em dia o método LCGM é mais complexo de implementar do que o agrupamento k-means, o k-means ainda parece ser um método robusto e relevante.

abordagem para estimar trajetórias de mudança de vizinhança onde a variável de agrupamento é medida continuamente (por exemplo, quocientes de localização). Deve-se lembrar, no entanto, que o agrupamento k-means só pode ser aplicado em variáveis contínuas ou binárias, enquanto o LCGM permite a identificação de trajetórias usando variáveis com uma gama de distribuições diferentes (contínua, ordinal, nominal, contagem e binomial). Além disso, com o LCGM, também é possível maximizar trajetórias em função de variáveis externas (como os preditores examinados aqui por sua associação com as trajetórias). Consequentemente, a abordagem LCGM é promissora dada a sua aplicação mais ampla. Por exemplo, LCGM pode ser usado para identificar trajetórias de mudança de vizinhança com base em uma variável de contagem ou uma variável qualitativa. Estudos futuros são necessários para testar aplicações de LCGM para modelar a mudança de vizinhança.