Universidade de Brasília
Departamento de Estatística

Identificação de clusters de domicílios de Aglomerados Subnormais no Norte do Brasil

Sabrina Lopes França

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Sabrina Lopes França

Identificação de clusters de domicílios de Aglomerados Subnormais no Norte do Brasil

Orientador(a): Prof. Dr. André Luiz Fernandes Cançado

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Dedico o trabalho à Maria do Socorro, minha mãe.

Agradecimentos

Agradeço ao professor André Luiz Fernandes Cançado, por toda orientação e passagem de conhecimento.

Agradeço a minha mãe por todo apoio, educação e cuidado.

Agradeço ao meu cachorro Bento por toda a companhia na realização deste trabalho.

Agradeço aos meus companheiros de curso que se tornaram meus amigos nessa caminhada, Davi, Júlia, Kevyn, Luana, Marcelo e Stephany.

Agradeço aos meus professores do ensino fundamental, Dirlene e Belchior, que sempre me motivaram e foram fundamentais para minha formação.

Agradeço a todos os professores do Departamento de Estatística da UnB, pelo quais tenho muita admiração, por todo aprendizado e disponibilidade.

Resumo

Aglomerados Subnormais são territórios de ocupação irregular com com finalidade habitacional em áreas urbanas. São áreas caracterizadas por um arranjo urbano desorganizado e falta de serviços públicos essenciais, são conhecidas também como favelas, comunidades, grotões etc. Em 2019, o Instituto Brasileiro de Geografia e Estatística informou a existência de 13.151 aglomerados subnormais no país, compostos por mais de 5 milhões de domicílios. Visando identificar clusters desse tipo de domicílio, ou seja, áreas que possuem um número maior que o esperado de domicílios de aglomerados subnormais na região Norte do Brasil, aplicou-se e comparou-se dois métodos da estatística espacial: Scan Poisson e Scan ZIP. O Scan ZIP é uma extensão do primeiro método com o propósito de lidar com dados com excesso de zero. Como resultado, identificou-se clusters nos estados do Amazonas, Amapá e Pará e foi encontrado clusters em comum para os dois métodos.

Palavras-chaves: Estatística Scan Espacial, Detecção de Clusters, Inflacionado de zeros, Favelas, Aglomerados Subnormais.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

O Brasil, marcado pela desigualdade, apresenta diferentes tipos de habitações nas cidades que o compreendem e parte delas são caracterizadas pela precariedade da infraestrutura urbana. Em algumas áreas do país a aquisição de moradia se estabeleceu pela ocupação de terras ociosas e pela construção própria de habitação, o que resultou em assentamentos inadequados por poderem ser de áreas de risco, sem infraestrutura de saneamento ambiental e ainda de difícil acesso a transportes e equipamentos sociais (COSTA; NASCIMENTO, 2005; FILHO, 2015).

Ademais, o conhecimento dessas áreas por meio de dados é escasso. Porém, o Instituto Brasileiro de Geografia e Estatística (IBGE) realiza pesquisas que trazem informações sobre as favelas e semelhantes assentamentos no Brasil (COSTA; NASCIMENTO, 2005). O Instituto utiliza o conceito de aglomerados subnormais para nomear esses tipos de assentamentos precários brasileiros (FILHO, 2015). Define-se pelo IBGE (2020):

> "Os Aglomerados Subnormais (...) são formas de ocupação irregular de terrenos de propriedade alheia (públicos ou privados) para fins de habitação em áreas urbanas e, em geral, caracterizados por um padrão urbanístico irregular, carência de serviços públicos essenciais e localização em áreas que apresentam restrições à ocupação."

Essas áreas também são conhecidas por outros nomes em diferentes partes do país como favelas, comunidades, grotões, vilas, mocambos, entre outros (IBGE, 2010).

Segundo os dados do Censo demográfico do IBGE (2010) estimou-se a existência de 6.329 aglomerados no Brasil. Já em 2019, o Instituto informou a existência de 13.151 aglomerados subnormais no país, compostos por mais de 5 milhões de domicílios, sendo estes distribuídos em 734 municípios brasileiros, abrangendo todas as 27 unidades federativas.

Nota-se, dessa forma, o crescimento dessas habitações no Brasil e isso se mostra como um indicador do cenário urbano no país (COSTA; NASCIMENTO, 2005).

Dessa forma, é proposto com o auxílio da Estatística Espacial, um olhar sobre como esses aglomerados são distribuídos no país e questiona-se a respeito da existência de áreas que possuem uma incidência significativamente elevada de domicílios instalados nesses aglomerados.

1.1 Estatística Espacial

A estatística espacial se define como um conjunto de técnicas para realizar análises geográficas que dependem do arranjo espacial dos eventos de interesse. O objetivo é incorporar o espaço à análise que se quer executar (KREMPI, 2004; CÂMARA et al., 2004).

Os tipos de dados que a estatística espacial envolve são: (a) Dados pontuais: ocorrências identificadas como pontos no espaço, comumente representados por (x_i, y_i); (b) Dados de área ou agregados: dados disponíveis por unidade de análise, como uma unidade administrativa do mapa em estudo; (c) Dados contínuos ou de superfície: dados resultantes de medições em determinadas localizações do mapa, comum para levantamentos de recursos naturais (FERNANDES; REIS, 2011; FERNANDES, 2015).

Quando os dados são pontuais ou de área, podemos estar interessados em identificar um cluster espacial. Ele se caracteriza por ser uma área em que a ocorrência de eventos não é meramente casual, apresentando um número significativamente maior que o esperado de eventos em relação ao restante do mapa (ANDRADE et al., 2007).

A detecção de clusters é bastante útil para a esfera de políticas públicas, já que é possível direcionar serviços e recursos para o controle de criminalidade, trânsito, saúde pública etc., quando se conhece os locais de maior necessidade. Além disso, a identificação de clusters também é comumente utilizada em epidemiologia, arqueologia, botânica, criminologia, demografia, ecologia, economia, engenharia, genética, geografia, história, neurologia, sociologia e zoologia (FERNANDES; REIS, 2011).

Um dos métodos voltados para a detecção de clusters é o Scan Circular de Kulldorff (1997). Trata-se de uma técnica que consiste na varredura do mapa de forma iterativa através de janelas circulares, limitadas pela proporção máxima da população, em busca dos clusters. Com isso, deseja-se localizar um conglomerado que seja caracterizado por conter o número de casos significamente superior ao esperado dentro da janela circular. Esse método é utilizado geralmente com a suposição de que os dados seguem distribuição Poisson (FERNANDES; REIS, 2011).

Uma outra abordagem na identificação de clusters é apresentada por Cançado, Silva e Silva (2014), onde se estende a estatística scan para o caso de excesso de zeros. Entende-se que, nesse contexto, utilizar a aplicação tradicional pode produzir viéses nos resultados e então utiliza-se a distribuição Poisson Inflacionada de Zeros (ZIP) para se desenvolver a técnica de Scan ZIP para a clusterização.

Visando identificar as áreas que possuem um número maior que o esperado de domicílios de aglomerados subnormais, esse estudo se estabelece pela análise de conglomerados espaciais para identificar regiões significativas que permitam avaliar se os domicílios em aglomerados estão distribuídos de modo homogêneo ou não sobre os municípios de parte do Brasil. Portanto, este trabalho se propõe a localizar os clusters significativos utilizando e comparando dois métodos da estatística espacial: o Scan de Poisson e Scan ZIP.

2 Objetivos

2.1 Objetivo Geral

Identificar conglomerados que possuem alta incidência de domicílios de Aglomerados Subnormais através do método Scan de Kulldorf e Scan para dados inflacionados de zeros.

2.2 Objetivos Específicos

- Detectar e identificar os clusters mais verossímeis de domicílios de aglomerados delimitando os 450 municípios da Região Norte do Brasil;
- Comparar os métodos através da aplicação nos dados.

3 Metodologia

3.1 Conjunto de dados

Os dados foram coletados no site do IBGE , da pesquisa de Aglomerados Subnormais do ano de 2019. O conjunto de dados abrange as seguintes variáveis para a Região Norte: Município, Latitude do Município, Longitude do Município, Unidade Federativa, Número de aglomerados do município, Número de domicílios do município e o Número de domicílios do município que são de aglomerados subnormais.

Para atingir o objetivo do trabalho será então aplicado o método de estatística espacial Scan Circular de Kulldorff e Scan para dados com excessos de zeros ao mapa da região Norte. Os métodos serão detalhados a seguir.

3.2 A Estatística Espacial Scan Circular de Kulldorff

A estatística Scan Circular apresentada por Kulldorff (1997) leva em consideração um mapa com m regiões. Sejam n_i e x_i, respectivamente, o tamanho da população e o número de casos na i-ésima região, $i = 1, \ldots, m$. O tamanho da população total do mapa é $N = \sum_i^m n_i$ e o número total de casos é $C = \sum_i^m x_i$. Chama-se zona qualquer subconjunto conexo de regiões do mapa. Assim, para uma zona z particular, tem-se:

- $x_z = \sum_{i \in z} x_i$: número de casos observados na zona z;
- $x_{\bar{z}} = \sum_{i \notin z} x_i$: número de casos observados fora da zona z;
- $n_z = \sum_{i \in z} n_i$: população da zona z;
- $n_{\bar{z}} = \sum_{i \notin z} n_i$: população fora da zona z;
- θ_z: probabilidade de que um indivíduo na zona z seja um caso;
- θ_0: probabilidade de que um indivíduo fora da zona z seja um caso.

Assim, a hipótese nula (H_0) assume que $\theta_z = \theta_0 \ \forall z$, enquanto a hipótese alternativa (H_a) indica $\theta_z > \theta_0$ para pelo menos uma zona z e, nesse último caso, z seria considerado um cluster.

3.2.1 A estatística Scan Poisson

Como indicado por Kulldorff (1997), pode-se supor que o número de casos em cada região segue distribuição Poisson, $x_i \sim \text{Poisson}(n_i\theta_i)$, e assumindo, sob a hipótese nula, que $\theta_z = \theta_0$, a verossimilhança toma a forma

$$\begin{aligned} L_0(x,\theta_0) = \prod_{i=1}^{m} \frac{e^{-n_i\theta_0}(n_i\theta_0)^{x_i}}{x_i!} = \frac{e^{-\sum_i^m n_i\theta_0} \cdot \prod_i^m [(n_i\theta_0)^{x_i}]}{\prod_i^m (x_i!)} \\ = \frac{e^{-\sum_i^m n_i\theta_0} \cdot \prod_i^m n_i^{x_i} \cdot \theta_0^{\sum_i^m x_i}}{\prod_i^m (x_i!)}. \end{aligned} \tag{3.2.1}$$

Substituindo $N = \sum_i^m n_i$ e $C = \sum_i^m x_i$, tem-se

$$L_0(x,\theta_0) = \frac{e^{-N\theta_0} \cdot \prod_i^m n_i^{x_i} \cdot \theta_0^C}{\prod_i^m (x_i!)}. \tag{3.2.2}$$

Seja $l_0(x,\theta_0) = \log(L_0(x,\theta_0))$. Então a log-verossimilhança é dada por

$$l_0(x,\theta_0) = -N\theta_0 + \sum_i^m x_i \log n_i + C\log\theta_0 - \sum_i^m \log(x_i!). \tag{3.2.3}$$

Para encontrar o ponto que maximiza $l_0(x,\theta_0)$, deriva-se em relação a θ_0 e iguala-se a 0, obtendo

$$\frac{\partial l_0(x,\theta_0)}{\partial\theta_0} = -N + \frac{C}{\theta_0} = 0, \tag{3.2.4}$$

e assim tem-se que

$$\hat{\theta_0} = \frac{C}{N}. \tag{3.2.5}$$

Considerando agora, sob a hipótese alternativa (H_a), que

$$\begin{cases} \theta_i = \theta_z, \text{se } i \in z, \\ \theta_i = \theta_0, \text{se } i \notin z, \end{cases}$$

tem-se que a verossimilhança pode se descrita como

$$\begin{aligned} L_a(z, x, \theta_0, \theta_z) = & \prod_{i \in z} \frac{e^{-n_i\theta_z}(n_i\theta_z)^{x_i}}{x_i!} \cdot \prod_{i \notin z} \frac{e^{-n_i\theta_0}(n_i\theta_0)^{x_i}}{x_i!} \\ = & \frac{e^{-\sum_{i \in z} n_i\theta_z} \cdot \prod_{i \in z}[(n_i\theta_z)^{x_i}]}{\prod_{i \in z}(x_i!)} \cdot \frac{e^{-\sum_{i \notin z} n_i\theta_0} \cdot \prod_{i \notin z}[(n_i\theta_0)^{x_i}]}{\prod_{i \notin z}(x_i!)} \\ = & \frac{e^{-\sum_{i \in z} n_i\theta_z} \cdot \prod_{i \in z} n_i^{x_i} \cdot \theta_z^{\sum_{i \in z} x_i}}{\prod_{i \in z}(x_i!)} \cdot \frac{e^{-\sum_{i \notin z} n_i\theta_0} \cdot \prod_{i \notin z} n_i^{x_i} \cdot \theta_0^{\sum_{i \notin z} x_i}}{\prod_{i \notin z}(x_i!)}. \end{aligned} \tag{3.2.6}$$

A log-verossimilhança $l_a(z, x, \theta_0, \theta_z) = \log(L_a(z, x, \theta_0, \theta_z))$ pode ser escrita da forma

$$\begin{aligned} l_0(z, x, \theta_0, \theta_z) = & -n_z\theta_z + \sum_{i \in z} x_i \log n_i + x_z \log \theta_z - \sum_{i \in z} \log(x_i!) \\ & - n_{\bar{z}}\theta_0 + \sum_{i \notin z} x_i \log n_i + x_{\bar{z}} \log \theta_0 - \sum_{i \notin z} \log(x_i!) \end{aligned} \tag{3.2.7}$$

Maximizando $l_a(z, x, \theta_0, \theta_z)$ em relação a θ_0 e θ_z, tem-se

$$\begin{aligned} \frac{\partial l_a(z, x, \theta_0, \theta_z)}{\partial \theta_0} = & -n_{\bar{z}} + \frac{x_{\bar{z}}}{\theta_0} = 0 \\ \hat{\theta_0} = & \frac{x_{\bar{z}}}{n_{\bar{z}}}, \end{aligned} \tag{3.2.8}$$

$$\begin{aligned} \frac{\partial l_a(z, x, \theta_0, \theta_z)}{\partial \theta_z} = & -n_z + \frac{x_z}{\theta_z} = 0 \\ \hat{\theta_z} = & \frac{x_z}{n_z}. \end{aligned} \tag{3.2.9}$$

Assim a razão de verossimilhança para o Modelo Poisson se define como

$$\lambda = \frac{e^{-\sum_{i\in z} n_i\theta_z} \cdot \prod_{i\in z} n_i^{x_i} \cdot \theta_z^{\sum_{i\in z} x_i}}{\prod_{i\in z}(x_i!)} \cdot \frac{e^{-\sum_{i\notin z} n_i\theta_0} \cdot \prod_{i\notin z} n_i^{x_i} \cdot \theta_0^{\sum_{i\notin z} x_i}}{\prod_{i\notin z}(x_i!)} \cdot \frac{1}{\frac{e^{-\sum_i n_i\theta} \cdot \prod_i n_i^{x_i} \cdot \theta^{\sum_i x_i}}{\prod_i(x_i!)}} = \frac{e^{-(n_z\theta_z + n_{\bar{z}}\theta_0 - n\theta)} \cdot \theta_z^{x_z} \cdot \theta_0^{x_{\bar{z}}}}{\theta^C}. \tag{3.2.10}$$

Assim, tem-se que, para uma zona qualquer z, a razão de verossimilhança será

$$\lambda_z = \frac{\sup_{\theta_z > \theta_0} l_a(z, x, \theta_0, \theta_z)}{\sup_{\theta_z = \theta_0} l_0(z, x, \theta_0, \theta_z)} = \left(\frac{\frac{x_z}{n_z}}{\frac{C}{N}}\right)^{x_z} \cdot \left(\frac{\frac{x_{\bar{z}}}{n_{\bar{z}}}}{\frac{C}{N}}\right)^{x_{\bar{z}}} \cdot I\left(\frac{x_z}{n_z} > \frac{x_{\bar{z}}}{n_{\bar{z}}}\right), \tag{3.2.11}$$

em que I é a função indicadora. A zona z^* que maximiza λ_z é chamada de zona mais verossímil, isto é, $z^* = \arg\sup_z \lambda_z$, e a estatística de teste T é definida como

$$T = \sup_z \lambda_z. \tag{3.2.12}$$

Portanto, a zona com maior valor de λ_z (T) é considerada o cluster mais provável, verossímil. Para encontar esse cluster adotam-se as janelas circulares para diferentes centros e raios que irão permitir encontrá-lo dentre diversas combinações de regiões (FERNANDES, 2015; FERNANDES; REIS, 2011; MELO; MELO; MORAES, 2006). Esse mecanismo será explicado na próxima seção.

3.2.2 Matriz de distâncias

Seja (w_i, y_i) o par de coordenadas geográficas do centroide da i-ésima região $i = 1, \ldots, m$. Através da distância Euclidiana, que mede a distância entre dois pontos, será calculada a distância entre dois centroides. Dessa forma, para regiões i e j quaisquer, sua distância é indicada por

$$d_{i,j} = \sqrt{(w_i - w_j)^2 + (y_i - y_j)^2}. \tag{3.2.13}$$

Será construída então uma matriz $m \times m$ com as distâncias entre as regiões, ou seja, uma matriz quadrada com m linhas e m colunas. Seja $e_{i,j}$ o elemento da i-ésima linha e j-ésima coluna da matriz. Então

$$e_{i,j} = \begin{cases} d_{i,j} & \text{, se } i \neq j \\ 0 & \text{, c.c.} \end{cases} \tag{3.2.14}$$

Assim, a matriz tem a forma

$$\mathrm{D} = \begin{bmatrix} 0 & d_{1,2} & \dots & d_{1,j} & ... & d_{1,m} \\ d_{2,1} & 0 & ... & d_{2,j} & ... & d_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i,1} & d_{i,2} & \dots & 0 & \dots & d_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \dots & d_{m,j} & \dots & 0 \end{bmatrix}.$$

3.2.3 Seleção dos candidatos a cluster

Para encontrar as zonas candidatas a cluster são adotadas as janelas circulares que realizam a varredura pelo mapa captando todas as zonas. Essas zonas são construídas levando em conta as distâncias entre as regiões. Por exemplo, para a região 1, sua distância para com outras regiões é registrada no seguinte vetor coluna:

$$\begin{bmatrix} 0 \\ d_{2,1} \\ d_{3,1} \\ d_{4,1} \\ \vdots \\ d_{m,1} \end{bmatrix}$$

Considere $d_{(j),i}$ como a distância entre a região i e a j-ésima região mais próxima de i, em que $d_{(m),i} > ... > d_{(3),i} > d_{(2),i}$. O vetor coluna da região 1 ordenado de forma crescente pelas distâncias é dado por

$$\begin{bmatrix} 0 \\ d_{(2),1} \\ d_{(3),1} \\ d_{(4),1} \\ \vdots \\ d_{(m),1} \\ . \end{bmatrix}.$$

Dessa forma, a primeira zona é formada pela própria região 1, sendo $z_1 = \{1\}$, que terá o valor de razão de verossimilhança λ_{z_1} calculado. A segunda zona, z_2, será formada pela região 1 mais a região mais próxima de 1, que corresponde à distância $d_{(2),1}$, isto é, $z_2 = \{1, (2)\}$. Essa zona tem então o seu respectivo valor da razão de verossimilhança λ_{z_2} calculado.

Nessa lógica, as zonas são formadas agregando as regiões segundo a distância em diferentes combinações até que se atinja o tamanho máximo de população permitido dentro de uma zona, que corresponde a 50% da população. Nesse ponto o processo é reiniciado a partir da região 2, formando uma zona contendo apenas essa região e em seguida outras zonas, adicionando à zona anterior, a cada passo, as regiões por ordem de distância à região 2. O processo iterativo é repetido então partindo de cada uma das m regiões, formando, ao fim, um conjunto

$$\mathcal{Z}$$

de zonas candidatas. A seguir é exemplificado para a região Norte do Brasil o processo de varredura da janela circular pelo mapa construindo as zonas no processo iterativo.

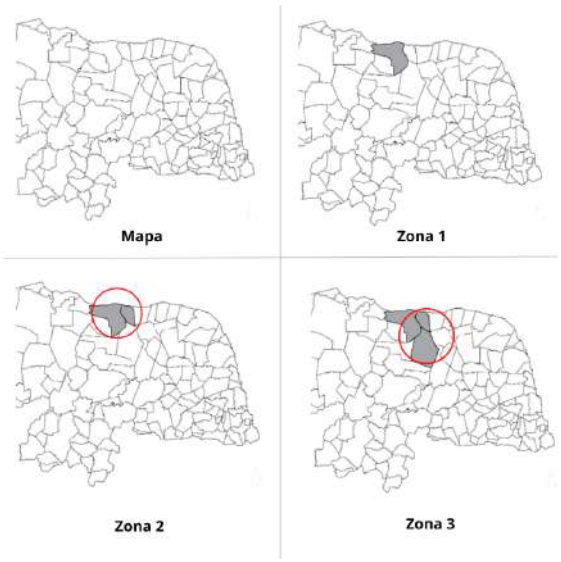

Figura 1: Exemplo: Construção das zonas

O cluster mais verossímil será a zona que obtiver o maior λ_z, correspondendo ao valor da estatística de teste T.

3.2.4 Significância do Cluster

Encontrado o cluster mais verossímil, deve-se testar sua significância. Esse processo é realizado com simulações de Monte Carlo, já que a distribuição de T é desconhecida.

As simulações baseiam-se na construção de V réplicas do mapa original em que o total de casos C é distribuído entre as regiões da área de estudo de forma aleatória, sob H_0.

Assim, para cada réplica tem-se um valor da estatística T. Considerando a aplicação de V réplicas e armazendando os valores de T, determina-se uma distribuição empírica da estatística de teste sob H_0. Essa distribuição é então comparada com o valor de T dos casos observados que chamaremos de T_o. Então, com um nível de significância de 5% se rejeita a hipótese H_0 de ausência de clusters se $T_o > P_{95}$, onde P_{95} representa o 95° percentil da distribuição de T sob a hipótese nula. Dessa forma, obtém-se uma estimativa do p-valor do cluster mais provável. (FERNANDES, 2015; MELO; MELO; MORAES, 2006; FIGUEIREDO, 2010).

3.3 Scan para dados com excesso de zeros

Sabe-se que a distribuição Poisson requer que a média seja igual à variância. Porém, essa suposição pode ser descaracterizada possibilitando a geração de viéses nos cálculos da identificação de clusters. Uma das causas desse desequilíbrio entre média e variância é a presença de excesso de zeros nos dados (ARAÚJO, 2012).

O valor nulo presente dos dados pode ser reconhecido de duas formas:

- **Zero amostral**: definido com a ausência de valor/característica no período do estudo.
- **Zero estrutural**: se dá pela inexistência de fato da característica na populção. Segundo Agresti (2002), zero estrutural é uma observação impossível de ser obtida, independente do tamanho da amostra ou seu período amostrado.

Uma maneira de lidar com o excesso de zeros nos dados é através do modelo de Poisson Inflacionado de Zeros (ZIP), pois este leva em conta a existência de zeros estruturais nos dados.

3.3.1 Modelo Poisson Inflacionado de Zeros - ZIP

Como indicado por Araújo (2012, p. 46), "o modelo ZIP pode ser visualizado como uma mistura entre uma distribuição de Poisson e uma distribuição degenerada no ponto zero, com probabilidade de mistura igual a p.". É proposto com esse modelo observar o valor zero com probabilidade p e observar uma distribuição Poisson (λ) com probabilidade $(1-p)$. A distribuição ZIP é definida como

$$\begin{aligned} P(Y=0|p,\lambda) &= p + (1-p)e^{\lambda}, \\ P(Y=y|p,\lambda) &= (1-p)\frac{e^{-\lambda}\lambda^{y}}{y!}, y > 0. \end{aligned} \tag{3.3.1}$$

Essa distribuição pode ser entendida no sentindo em que uma proporção p é retirada da distribuição Poisson (λ) e conferida ao evento $\{y=0\}$. É justamente essa adição da probabilidade de observar o valor zero que tem intenção de reproduzir a presença dos zeros estruturais na distribuição (ARAÚJO, 2012).

A esperança da distribuição ZIP é dada por

$$\begin{aligned} E(Y|p,\lambda) &= 0 \cdot P(Y=0|p,\lambda) + \sum_{y>0} y \cdot P(Y=y|p,\lambda) \\ &= \sum_{y>0} y(1-p)\frac{e^{-\lambda}\lambda^y}{y!} \\ &= (1-p)e^{-\lambda}\sum_{y>0} y\frac{\lambda^y}{y!} = (1-p)e^{-\lambda}\lambda e^{\lambda} \\ &= (1-p)\lambda \end{aligned} \tag{3.3.2}$$

Calcula-se também $Var(Y|p,\lambda) = (1-p)\lambda(1+p\lambda)$.

3.3.2 A estatística Scan ZIP

Uma maneira de lidar com dados com excesso de zeros na identificação de clusters espaciais é substituir a distribuição Poisson pela distribuição ZIP na estatística Scan de Kulldorf (ARAÚJO, 2012).

Dessa forma, como proposto no trabalho de Cançado, Silva e Silva (2014) e considerando a notação dada na seção 3.2 mas assumido que o número de casos X_i da i-ésima região segue distribuição ZIP, $X_i \sim ZIP(p; n_i, \theta_i)$, essa distribuição é caracterizada por

$$\begin{aligned} P(X_i = 0|p, n_i\theta_i) &= p + (1-p)e^{-n_i\theta_i}, \\ P(X_i = x_i|p, n_i\theta_i) &= (1-p)\frac{e^{-n_i\theta_i}(n_i\theta_i)^{x_i}}{x_i!}, x_i > 0, \end{aligned} \tag{3.3.3}$$

onde p representa a probabilidade da ocorrência de zero estrutural, e para a i-ésima região, n_i é a população e θ_i é a probabilidade de que um indivíduo seja um caso.

O modelo ZIP, dessa forma, é capaz de fornecer a probabilidade de ocorrência de um zero estrutural, p, e a probabilidade de ocorrência do zero amostral, $(1-p)e^{-n_i\theta_i}$ (ARAÚJO, 2012).

São consideradas as hipóteses, onde $z \in \mathcal{Z}$,

$$\begin{cases} H_0 : \theta_0 \text{ é o mesmo para todo } z \\ H_a : \theta_z \neq \theta_0 \text{ para algum } z \end{cases}$$

Com o objetivo de identificar as regiões com zero estrutural, considere a variável binária δ_i que indica para cada região i a ausência ou presença de zero estrutural. Se

$\delta_i = 1$, há zero estrutural na região i e caso contrário $\delta_i = 0$, sendo $\delta_i \sim Bernoulli(p)$. Então, seja $\delta = (\delta_1, ..., \delta_m)$, tem-se $P(\delta_i = 1) = p$. A distribuição conjunta de X_i e δ_i é dada por

$$P(X_i = x_i, \delta_i = d_i | p, n_i\theta_i) = \begin{cases} (1-p)e^{-n_i\theta_i} & \text{, se } x_i = 0 \text{ e } d_i = 0 \\ \text{p} & \text{, se } x_i = 0 \text{ e } d_i = 1 \\ \frac{(1-p)(n_i\theta_i)^{x_i}e^{-n_i\theta_i}}{x_i!} & \text{, se } x_i > 0 \text{ e } d_i = 0 \\ 0 & \text{, c.c.} \end{cases} \tag{3.3.4}$$

Sucintamente, pode-se entender que

$$P(X_i = x_i, \delta_i = d_i | p, n_i\theta_i) = \begin{cases} p^{d_i}\left[(1-p)\dfrac{e^{-n_i\theta_i}(n_i\theta_i)^{x_i}}{x_i!}\right]^{(1-d_i)} & \text{, se } x_i = 0 \text{ ou } d_i = 0 \\ 0 & \text{, c.c.} \end{cases} \tag{3.3.5}$$

Assumindo que a natureza de δ_i é conhecida, ou seja, os zeros estruturais e amostrais são informados, a função de verossimilhança tendo referÊncia a zona $z \in \mathcal{Z}$, sob H_a é definida como

$$\begin{aligned} L(p, \theta_0, \theta_z) &= L(p, \theta_0, \theta_z | z, X, \delta) \\ &= \left[\prod_{i \in z} P(X_i = x_i, \delta_i = d_i)\right] \times \left[\prod_{j \notin z} P(X_j = x_j, \delta_j = d_j)\right] \\ &= \left\{\prod_{i \in z} p^{d_i}\left[(1-p)\frac{e^{-n_i\theta_z}(n_i\theta_z)^{x_i}}{x_i!}\right]^{(1-d_i)}\right\} \\ &\times \left\{\prod_{j \notin z} p^{d_j}\left[(1-p)\frac{e^{-n_j\theta_0}(n_j\theta_0)^{x_j}}{x_j!}\right]^{(1-d_j)}\right\}. \end{aligned} \tag{3.3.6}$$

A log-verossimilhança, $L(p, \theta_0, \theta_z) = \log(l(p, \theta_0, \theta_z))$, sob H_a,é dada por:

$$\begin{aligned} l(p, \theta, z | X, \delta) = p^{\sum_{i=1}^m d_i} \times (1-p)^{m - \sum_{i=1}^m d_i} \times e^{-\theta_z \sum_{i \in z} n_i(1-d_i)} \times \theta_z^{\sum_{i \in z} x_i(1-d_i)} \\ \times e^{-\theta_0 \sum_{j \notin z} n_j(1-d_j)} \times \theta_0^{\sum_{j \notin z} x_j(1-d_J)}. \end{aligned} \tag{3.3.7}$$

E sob H_0, tem-se:

$$l_0(p,\theta,z|X,\delta) = p^{\sum_{i=1}^m d_i} \times (1-p)^{m-\sum_{i=1}^m d_i} \times e^{-\theta_0 \sum_{i=1}^m n_i(1-d_i)} \theta_0^{\sum_{i=1}^m x_i(1-d_i)}. \tag{3.3.8}$$

Como indicado por Cançado, Silva e Silva (2014), maximizando a log-verossimilhança sob cada hipótese, encontram-se os estimadores:

Sob H_0:

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^m x_i(1-d_i)}{\sum_{i=1}^m n_i(1-d_i)}, \\ \hat{p} &= \frac{\sum_{i=1}^m d_i}{m}. \end{aligned} \tag{3.3.9}$$

E sob H_a:

$$\begin{aligned} \hat{\theta_z} &= \frac{\sum_{i\in z} x_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)}, \\ \hat{\theta_0} &= \frac{\sum_{j\notin z} x_j(1-d_j)}{\sum_{j\notin z} n_j(1-d_j)}, \\ \hat{p} &= \frac{\sum_{i=1}^m d_i}{m}. \end{aligned} \tag{3.3.10}$$

Portanto, a razão de verossimilhança irá ser definida como:

$$\lambda_{ZIP} = sup_{z\in\mathcal{Z}} \left(\frac{sup_{\theta_z>\theta_0} l(p,\theta,z;X,\delta)}{sup_{\theta_z=\theta_0} l_0(p,\theta,z;X,\delta)} \right). \tag{3.3.11}$$

Temos que

$$sup_{\theta_z>\theta_0} l_0(p,\theta,z;X,\delta) = e^{-\sum_{i=1}^m x_i(1-d_i)} \left(\frac{\sum_{i=1}^m x_i(1-d_i)}{\sum_{i=1}^m n_i(1-d_i)} \right)^{\sum_{i=1}^m x_i(1-d_i)} \tag{3.3.12}$$

e,

$$
\begin{aligned}
sup_{\theta_z=\theta_0} l_0(p,\theta,z;X,\delta) = e^{-\sum_{i\in z} x_i(1-d_i)-\sum_{j\notin z} x_j(1-d_j)} \\
\times \left(\frac{\sum_{i\in z} x_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)}\right)^{\sum_{i\in z} x_i(1-d_i)} \\
\times \left(\frac{\sum_{j\notin z} x_i(1-d_j)}{\sum_{j\notin z} n_i(1-d_j)}\right)^{\sum_{j\notin z} x_j(1-d_j)}.
\end{aligned}
\tag{3.3.13}
$$

Dessa forma, a razão de verossimilhança é

$$
\begin{aligned}
\lambda_{ZIP} = sup_{z\in\mathcal{Z}} \frac{\left(\frac{\sum_{i\in z} x_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)}\right)^{\sum_{i\in z} x_i(1-d_i)} \left(\frac{\sum_{j\notin z} x_i(1-d_j)}{\sum_{j\notin z} n_i(1-d_j)}\right)^{\sum_{j\notin z} x_j(1-d_j)}}{\left(\frac{\sum_{i=1}^{m} x_i(1-d_i)}{\sum_{i=1}^{m} n_i(1-d_i)}\right)^{\sum_{i=1}^{m} x_i(1-d_i)}} \\
\times I\left(\frac{\sum_{i\in z} x_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)} > \frac{\sum_{j\notin z} x_i(1-d_j)}{\sum_{j\notin z} n_i(1-d_j)}\right)
\end{aligned}
\tag{3.3.14}
$$

para pelo menos uma zona z em que $\left(\frac{\sum_{i\in z} x_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)} > \frac{\sum_{j\notin z} x_i(1-d_j)}{\sum_{j\notin z} n_i(1-d_j)}\right)$, e $\lambda_{ZIP} = 1$ caso contrário.

A diferença entre a razão de verossimilhança da estatística Scan Poisson para a da estatística utilizando a distribuição ZIP está no fato de que a última leva em consideração a ocorrência de zeros estruturais.

Entende-se que uma região que representa um zero estrutural é aquela que não possui condição favorável para possuir casos. Dessa forma, no contexto da estatística Scan Poisson, uma região que é caracterizada por um zero estrutural é naturalmente incrementada à população, e esta por sua vez passa a ser maior e o número esperado de casos proporcionalmente também aumenta. Porém, como se trata de um zero estrutural essa expectativa para o aumento no valor esperado resulta na redução do valor da razão de verossimilhança, o que distorce as possíveis descobertas de clusters.

Em contrapartida, a estatística Scan ZIP, pondera cada região por $(1-d_i)$, sendo que d_i atua como uma variável indicadora pois seu valor é definido como 1 se a região apresenta zero estrutural ou 0 caso contrário. Por isso, ao incrementar uma região com zero estrutural, a estatística Scan ZIP permite que isso não aumente o valor esperado de casos para a região, adequando-se à situação do zero estrutural (ARAÚJO, 2012).

No entanto, saber se um zero é amostral ou estrutural não é usual e então d_i é desconhecido. Para esse caso, através do algoritmo EM, é possível estimar d_i dado o valor

observado de x_i e assim torna-se possível a ponderação da população pela probabilidade de cada região ser um zero estrutural.

O algoritmo EM, responsável por possibilitar a estimação do vetor δ, será descrito na próxima seção.

3.3.3 Algoritmo EM para δ_i desconhecido

O algoritmo EM para encontrar o máximo da função de máxima verossimilhança para dados faltantes foi apresentado por Dempster, Laird e Rubin (1977).

No algoritmo EM as etapas de expectância e maximização são executadas alternadamente, em cada iteração, até que se atinja a convergência pré estabelecida. Assim, os passos são descritos a seguir: .

- **Passo E**: como explicado em Cançado, Silva e Silva (2014), δ_i é estimado com base em sua esperança condicional dado X_i. Seja $(\delta_i|X_i) \sim Bernoulli(\zeta_i)$ e utilizando o Teorema de Bayes, então:

$$\begin{aligned}\zeta_i = E(\delta_i|X_i) &= P(\delta_i = 1|X_i) \\ &= \frac{P(X_i = x_i|\delta_i = 1)P(\delta_i = 1)}{P(X_i = x_i|\delta_i = 1)P(\delta_i = 1) + P(X_i = x_i|\delta_i = 0)P(\delta_i = 0)} \cdot I(x_i = 0) \\ &= \frac{p}{p(1-p)e^{-n_i\theta_i}} \cdot I(x_i = 0).\end{aligned} \tag{3.3.15}$$

Dessa forma, tem-se

$$\zeta_i = \begin{cases} \frac{p}{p+(1-p)e^{-n_i\theta_i}} \text{ , se } x_i = 0 \\ 0 \text{ , se } x_i = 1, 2, ... \end{cases} . \tag{3.3.16}$$

Logo, na m-ésima iteração do algoritimo EM, tem-se

$$\hat{\delta}_i^{(m)} = \frac{\hat{p}^{(m)}}{\hat{p}^{(m)} + (1-\hat{p}^{(m)})e^{-n_i\hat{\theta_i}^{(m)}}} I(x_i = 0), \quad i = 1, 2, ..., k. \tag{3.3.17}$$

- **Passo M**: Na $(m+1)$-ésima iteração, considerando o vetor $\hat{\delta}^{(m)} = (\hat{\delta}_1^{(m)}, ..., \hat{\delta}_m^{(m)})$,

os estimadores de p, θ_z, θ_0 são calculados conforme visto em (3.3.14),substituindo d_i pelos $\hat{\delta}^{(m)}$ estimados anteriormente . Desse modo, sob H_a, define-se:

$$\hat{\delta}_i^{(m+1)} = \begin{cases} \dfrac{\hat{p}^{(m+1)}}{\hat{p}^{(m+1)} + (1-\hat{p}^{(m+1)})e^{-n_i\hat{\theta_Z}^{(m+1)}}} \text{ , se } x_i = 0 \text{ e } i \in Z \\ \dfrac{\hat{p}^{(m+1)}}{\hat{p}^{(m+1)} + (1-\hat{p}^{(m+1)})e^{-n_i\hat{\theta_0}^{(m+1)}}} \text{ , se } x_i = 0 \text{ e } i \notin Z \\ 0 \text{ , se } x_i = 1, 2, ... \end{cases} . \quad (3.3.18)$$

Isto posto, repete-se o passo E e o passo M para cada candidato a cluster, até que se alcançe a convergência definida. Na inicialização do algoritmo, uma possibilidade é definir $\delta_i^{(0)} = 0.5$ se $x_i = 0$ e $\delta_i^{(0)} = 0$ se $x_i > 0$, como utilizado nesse trabalho. Os valores de δ_i são portanto estimados para $i \in Z$ e para $i \notin Z$.

3.3.4 Significância do Cluster

Como a estatística Scan ZIP não possui distribuição conhecida, serão utilizadas simulações de Monte Carlo na verificação da significância dos clusters. Para cada caso a seguir se é definido o processo.

- **Se δ_i é conhecido**: As regiões com ($\delta_i = 1$), representando os zeros estruturais, são removidas e pode-se executar o mesmo processo desenvolvido por Kulldorf descrito em 3.2.4, utilizando a distribuição Poisson vista anteriormente.

- **Se δ_i é desconhecido**: Para esse caso, as simulações de Monte Carlo serão realizadas por meio de um *bootstrap* com estrutura paramétrica como apresentado no trabalho de Cançado, Silva e Silva (2014). Dado $\hat{p} = \frac{\sum_i \hat{\delta}_i}{m}$, representando a estimação de p para os dados observados, onde $\hat{\delta}_i$ é estimado pelo algoritmo EM para o cluster mais verossímil, se realiza o seguinte procedimento na simulação (FERNANDES, 2015):

  1. Atribuir zero estrutural com probabilidade p para as regiões do mapa.
  2. Para as regiões que não foram registradas como zero estrutural anteriormente distribuir aleatoriamente, sob H_0, o número de casos C entre elas.

3. Encontrar o cluster mais verossímil por meio da estatística Scan ZIP.
4. Construir uma distribuição empírica de λ_{ZIP}, repetindo as etapas 1 a 3, V vezes.
5. Rejeitar a hipótese H_0, com nível de significância de 5%, se $\lambda_{ZIP} > P_{95}$, em que P_{95} representa o percentil da distribuição empiríca de λ_{ZIP}.

.

4 Resultados

4.1 Análise descritiva

É apresentado a seguir a Tabela 1 contendo dados resumos sobre os aglomerados subnormais pelas regiões do país e Brasil, além do gráfico da Figura 2 que representa a distribuição dos aglomerados subnormais no mapa brasileiro.

Como exibido na Tabela 1, em 2019 o Brasil possuía 13.151 aglomerados subnormais que abrigavam pouco mais de 5 milhões de domicílios. É possível visualizar como esses aglomerados eram distribuídos no país através do mapa apresentado na Figura 2.

Tabela 1: Dados sobre os aglomerados subnormais pelas regiões e Brasil

Região	População estimada	N° de aglomerados subnormais	N° de domicílios em aglomerados subnormais	% de domicílios de aglomerados subnormais
Norte	18.430.980	1.237	918.498	18,9%
Nordeste	57.071.654	3.203	16.943.328	8,6 %
Sudeste	88.371.433	6.573	2.321.963	8,1%
Sul	29.975.984	1.749	300.625	3,1%
Centro-Oeste	16.297.074	389	127.175	2,4%
Brasil	210.147.125	13.151	5.127.747	7,8%

Figura 2: Mapa da distribuição dos aglomerados subnormais no Brasil (2019)

Observa-se na Tabela 1 que a região Nordeste é o lugar que possui o maior número de domicílios instalados em seus aglomerados subnormais, correspondendo a mais de 16 milhões. Já a região Sudeste apresenta o maior número de aglomerados subnormais (6.573) em comparação com as demais regiões.

No próximo gráfico, Figura 3, observa-se os dez maiores aglomerados subnormais do Brasil, adotando como critério o número de domicílios desses locais. Dessa forma percebe-se que o aglomerado com mais domicílios no Brasil é a Rocinha, localizada no Rio de Janeiro, região Sudeste, que possui 25.742 domicílios. Em seguida, o aglomerado Sol Nascente, com 25.441 domicílios, localizado no Distrito Federal, região Centro-Oeste.

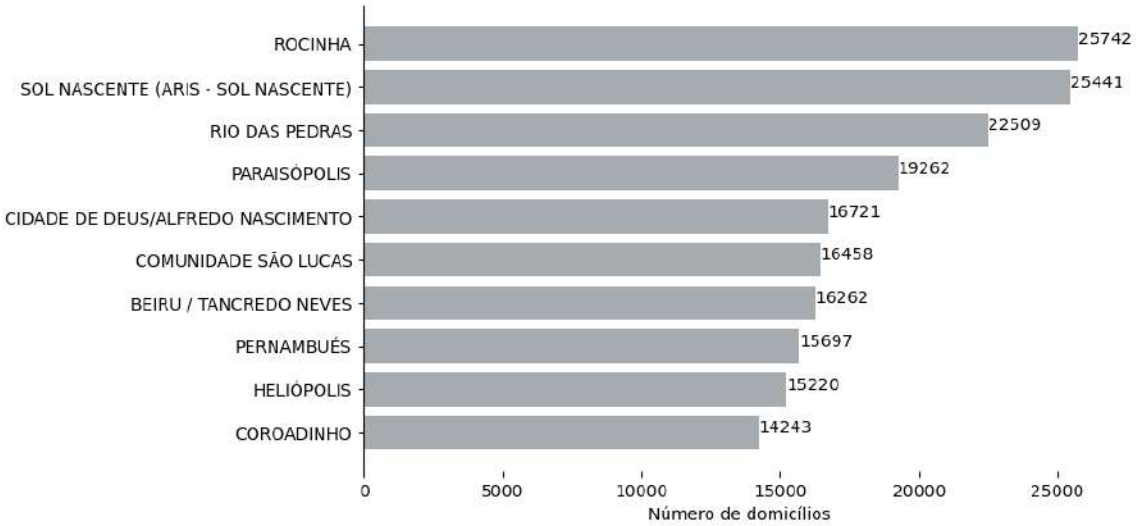

Figura 3: Gráfico dos dez aglomerados subnormais com mais domicílios no Brasil em 2019

Na Figura 4 verifica-se o gráfico da porcentagem de domicílios de aglomerados subnomais sobre o total de domicílios das regiões do Brasil. Através dessa Figura e da Tabela 1 observa-se que apesar de não ser a região que mais possui aglomerados subnormais no Brasil, o Norte é a localidade que apresenta a maior proporção de domicílios instalados em aglomerados subnormais (18,9%).

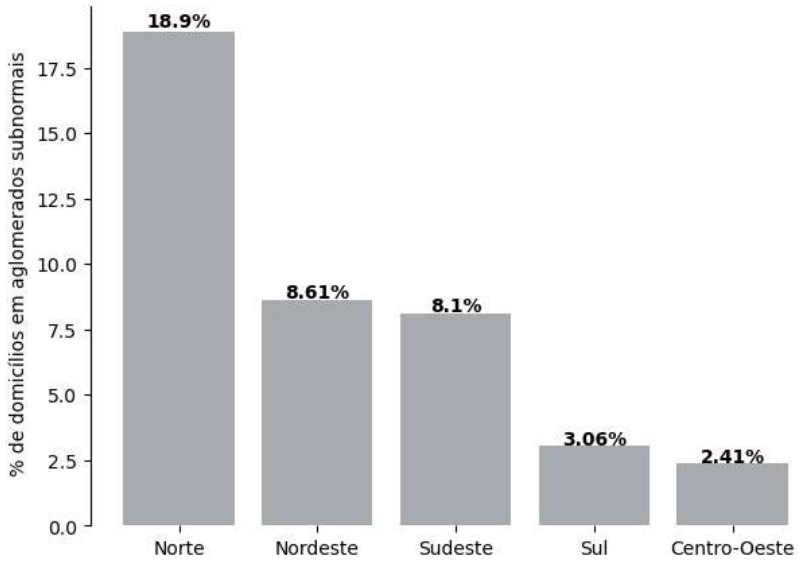

Figura 4: Gráfico da porcentagem de domicílios de aglomerados subnomais em 2019

Possibilitando uma análise geral, o Gráfico 5 apresenta como as porcentagens de domicílios de aglomerados por município das regiões do Brasil se distribuem.

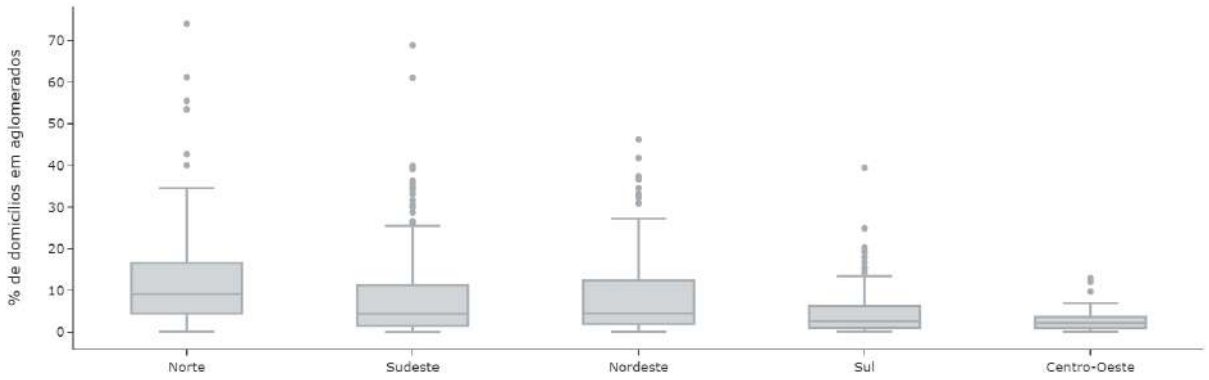

Figura 5: Box plot da distribuição da porcentagem de domicílios em aglomerados subnormais por município das regiões do Brasil

Observando a Figura 5, comparada as outras regiões, a região Norte é a que possui a maior amplitude de porcentagens de domicílios em aglomerados subnormais por município, e também é a região que abriga o município com a maior porcentagem desses domicílios no Brasil.

Além disso, a mediana da porcentagem de domicílios em aglomerados subnormais da região Norte é superior às demais. Percebe-se que essa mediana é menor que 10% e o limite superior do box plot fica pouco abaixo de 40%. Portanto tem-se outliers que caracterizam os municípios com porcentagens atípicas de domicílios em aglomerados subnormais.

4.1.1 Região Norte

Em busca de entender melhor como a distribuição de aglomerados subnormais na região Norte se estabelece, apresenta-se o mapa da Figura 6, segmentado pelos 450 municípios.

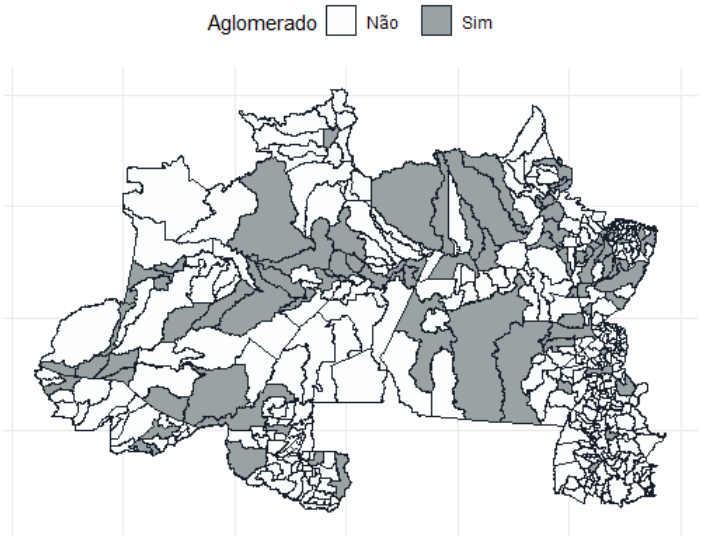

Figura 6: Mapa dos municípios da região Norte com a indicação da presença ou não de aglomerados subnormais em 2019

Com o Gráfico 6, é possível identificar a eminente presença de zeros nas regiões, ou seja, municípios que não possuem aglomerados subnormais e portanto não apresentam domicílios pertecentes a essas localizações. Precisamente, em 2019, a região Norte apresentava 92 municípios com registros de aglomerados subnomais e os outros 358 constavam não ter aglomerados, representando 79,56% dos municípios.

O Gráfico 7 disponibiliza a informação das proporções de domicílios em aglomerados subnormais naqueles municípios que os abrigam. A Tabela 2 resume as medidas descritivas para essas proporções.

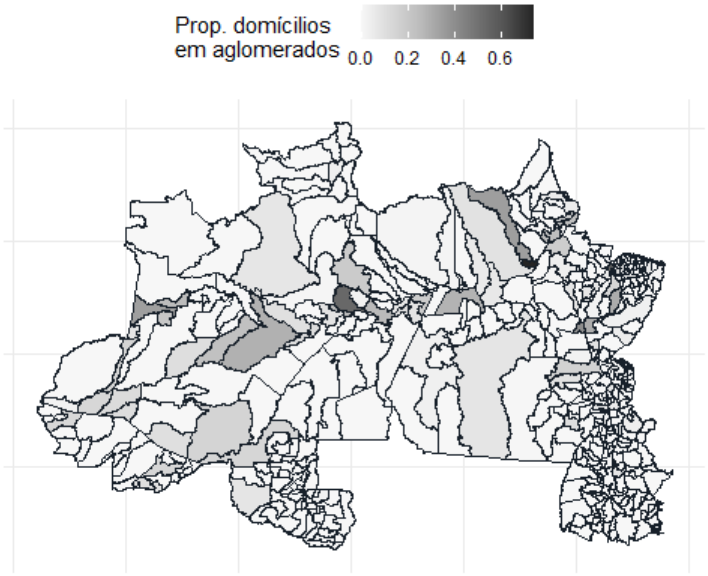

Figura 7: Mapa da proporção de domicílios em aglomerados subnomais por município na região Norte em 2019

Tabela 2: Medidas descritivas da proporção de domicílios em aglomerados por municípios da região Norte

Medida descritiva	
Média	0,03
Desvio padrão	0,09
1° Quartil	0,00
Mediana	0,00
3° Quartil	0,00
Máximo	0,74
Mínimo	0,00

A Tabela 2 reflete também a grande quantidade de zeros presente no mapa. É observado para a região Norte uma média de 0,03 na proporção de domicílios isntalados em aglomerados subnormais por município da região. Além disso, até o 3° quartil essa proporção é igual a zero.

No mapa da Figura 7, destaca-se a região em cor preta, que tem como valor de proporção o máximo apresentado na Tabela 2: 0,74. Observa-se, nesse gráfico que para os locais que possuem aglomerados, a distribuição da proporção não parece ser regular, isto é, se percebe concentrações de locais que se destacam pela proporção de domicílios em aglomerados em relação às demais.

A Figura 8 apresenta um gráfico que mostra os dez municípios da região Norte

que possuem as maiores proporções de aglomerados subnormais.

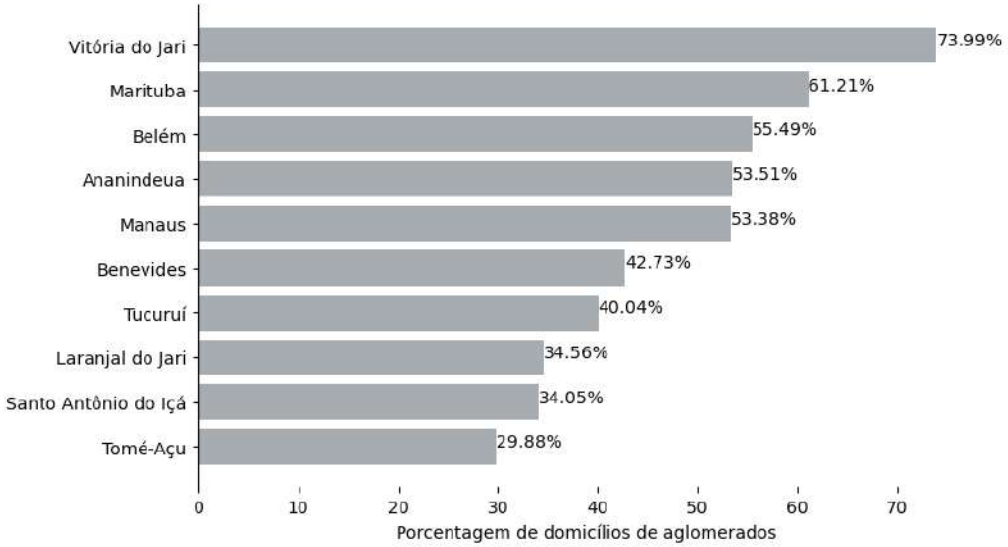

Figura 8: Gráfico dos dez municípios do Norte com as maiores proporções de domicílios de aglomerados subnormais em 2019

Dessa forma, visualiza-se que é o município Vitória do Jari que apresenta a porcentagem máxima (73,99%) de domicílios em aglomerados. Além disso, Marituba, Belém, Ananindeua e Manaus são municípios que possuem mais da metade de seus domicílios compostos por domicílios de aglomerados subnormais.

4.2 Aplicação Scan

Apresentam-se nessa seção os resultados obtidos para a aplicação da teoria desenvolvida na seção 3.2, que chamaremos de Scan Poisson, além da sua extensão apresentada na seção 3.3, que chamaremos de Scan ZIP-EM. Escolheu-se Scan ZIP-EM pois foi empregado o uso do algoritmo EM para δ_i desconhecido, já que não se conhece as regiões com zeros estruturais. Os dois algoritmos foram usados com as janelas circulares na construção do conjunto $\mathcal{Z}$ de candidatos a clusters. A significância dos clusters foi calculada a partir de 1000 simulações de Monte Carlo para ambos também.

Além disso, tanto para o método de clusterização utilizando Scan Circular quanto para o método utilizando Scan ZIP-EM foi elencado que a população $N = \sum_i^m n_i = 4.860.554$ representa o número de domicílios dos municípios da região Norte e o número total de casos $C = \sum_i^m x_i = 918.498$ é caracterizado pelo número de domicílios de aglomerados subnormais.

O algoritmo do Scan Poisson foi implementado no software R utilizando o pacote

scanstatistics (ALLÉVIUS, 2018). Já o Scan ZIP, foi aplicado em linguagem C++, referente ao algoritmo usado no trabalho de Cançado, Silva e Silva (2014), devido ao seu tempo de processamento, que se demonstrou mais rápido para esse caso que o R.

4.2.1 Scan Poisson

Os resultados da aplicação da técnica Scan Poisson para identificar os clusters de domicílios de aglomerados subnormais são apresentados a seguir pela Figura 9, Figura 10 e Tabela 3.

A Figura 9 revela o cluster 1 obtido com a técnica Scan Poisson. Esse cluster apresentou a maior razão de log-verossimilhança LLR (271.194,31) com significância estatística, como visto na Tabela 3.

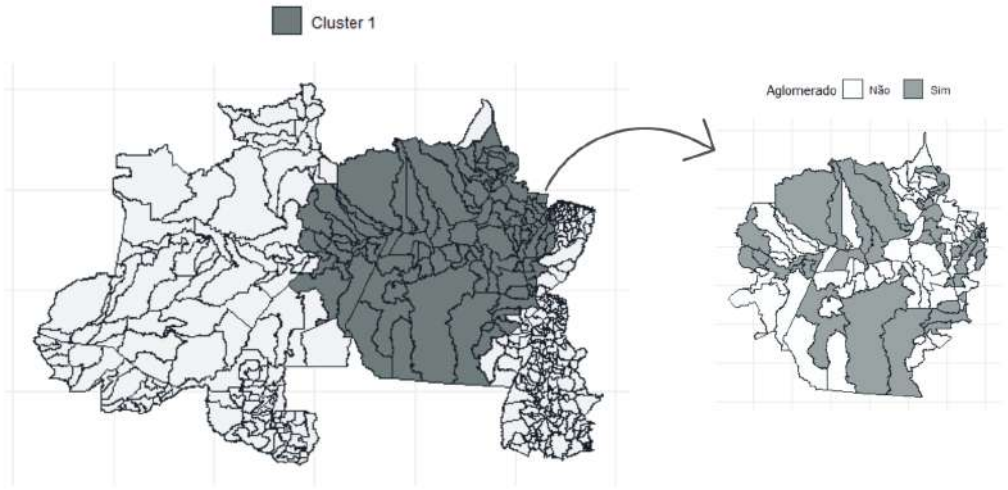

Figura 9: Mapa do Cluster 1 obtido com Scan de Poisson.
No mapa à esquerda é apresentado o cluster 1 na região Norte e à direita tem-se a indicação de municípios com aglomerados dentro do cluster.

O cluster 1 abrage 113 municípios, sendo eles pertencentes aos estados do Pará, Amazonas e Amapá. O gráfico da Figura 9 mostra também, dentro do cluster, quais regiões possuem aglomerados. É interessante perceber a quantidade de regiões que não apresentam essa característica, são 48 municípios que não possuem aglomerados, representando 42,48% das regiões do conglomerado. Assim, entende-se que esse cluster é formado por regiões com alta incidência de domicílios com aglomerados subnormais e por outras regiões onde isso não ocorre, indicando uma heterogeneidade na distribuição dos aglome-

rados dentro do cluster identificado.

Outra questão a ser observada é a de que o cluster 1 inclui em suas regiões os municípios que o Gráfico 8 apresenta, exceto Santo Antônio do Içá, que são os municípios com as maiores proporções de domicílios em aglomerados da região Norte. Percebe-se, que, devido à restrição de formato circular das zonas candidatas, o cluster que engloba esses municípios de alta incidência tem que, necessariamente, incluir diversos outros municípios com ausência de casos. O resultado é a identificação de um cluster tão grande que, do ponto de vista geográfico e de políticas públicas, não tem utilidade, pois não indica as localidades específicas onde as autoridades competentes devem atuar e onde devem ser aplicados recursos com prioridade.

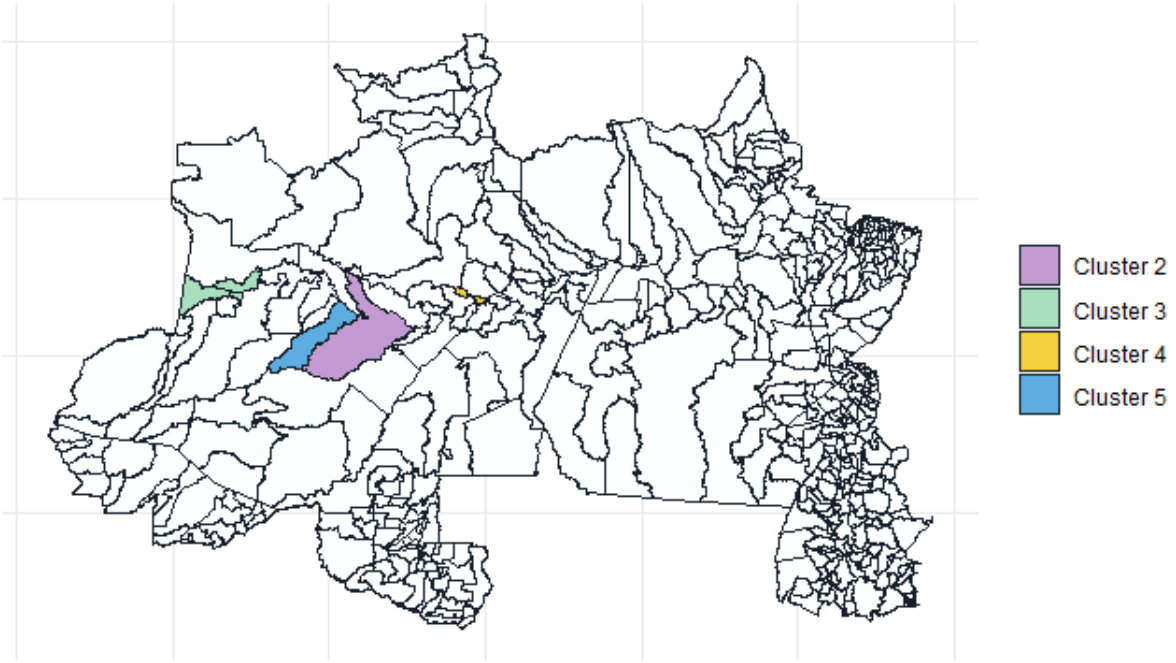

Figura 10: Mapa dos outros clusters obtidos com Scan de Poisson

Os outros clusters detectados com significância são apresentados na Figura 10. É notável que os outros clusters possuem tamanhos bem menores que o primeiro, sendo que o cluster 3 é composto por dois municípios e os clusters 2, 4 e 5 possuem apenas um município contabilizado. Além disso, excluindo o cluster 1, todos os outros clusters foram identificados no estado do Amazonas.

A Tabela 3 apresenta um resumo dos resultados obtidos à partir da clusterização com o Scan Poisson e exibe as características dos clusters encontrados.

Tabela 3: Resultados para a clusterização utilizando a técnica Scan Poisson

Cluster	Municípios	Pop.	N° esperado de casos	N° obs. de casos	LLR	p-valor
1	Grupo de municípios ∗	2.616.913	494.517,6	820.140	271.194,31	$< 0{,}001$
2	Coari (AM)	19.836	3.748,41	5.617	405,21	$< 0{,}001$
3	Tonantins (AM), Santo Antônio do Içá (AM)	9.213	1.740,98	2.817	280,22	$< 0{,}001$
4	Iranduba (AM)	21.093	3.985,94	5.352	212.18	$< 0{,}001$
5	Tefé (AM)	15.192	2.870,82	3.489	62,44	$< 0{,}001$

∗ Os municípios do cluster 1 se encontram no anexo do documento.

Pela Tabela 3, verifica-se que para todos o clusters detectados, o número de casos observados ultrapassa o número de casos esperados para os conglomerados. Além disso, nota-se que o LLR para o segundo cluster já é bastante inferior àquele calculado para o primeiro cluster.

4.2.2 Scan ZIP-EM

Como já citado anteriormente, na inicialização do algoritmo EM para estimação dos parâmetros do modelo, foi definido:

$$\delta_i^{(0)} = \begin{cases} 0,5 \text{ , se } x_i = 0 \\ 0, \text{ se } x_i > 0 \end{cases} \tag{4.2.1}$$

e assume-se que a convergência é atingida quando $|\delta_i^{(m+1)} - \delta_i^m| < 0,01$.

Os resultados da aplicação da técnica Scan ZIP-EM para identificar os clusters de domicílios de aglomerados subnormais são apresentados a seguir pela Figura 11, Figura 12 e Tabela 4.

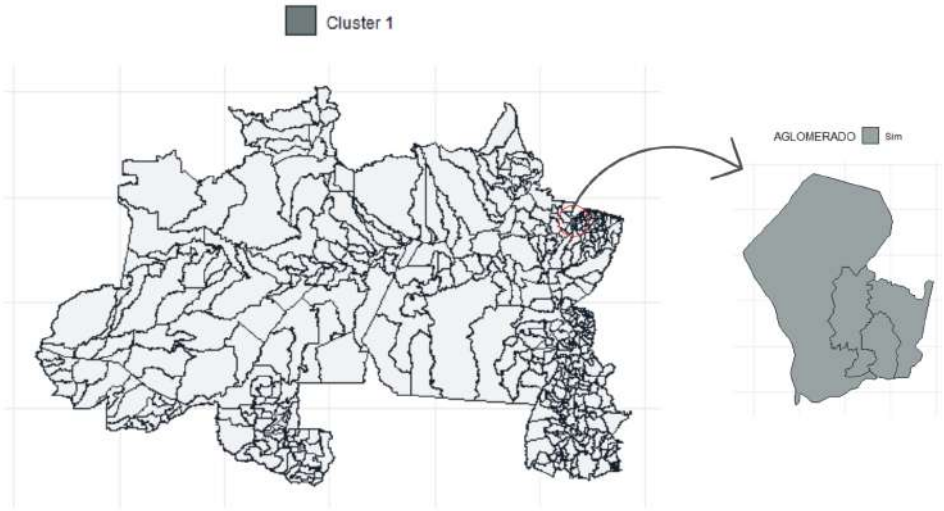

Figura 11: Mapa do Cluster 1 obtido com Scan ZIP-EM
No mapa a esquerda é apresentado o cluster 1 na região Norte e a direita tem-se a indicação de municípios com aglomerados dentro do cluster.

O cluster 1 obtido com Scan ZIP-EM é apresentado na Figura 11, e observa-se se tratar de uma região pequena em relação ao mapa e em relação ao cluster 1 obtido com o Scan Poisson (Figura 9). O cluster 1 resultante abrange 4 municípios, sendo todos eles do estado do Pará. Além disso, todos esses municípios possuem aglomerados subnormais e todos eles estão presentes no gráfico da Figura 8, onde são exibidos os 10 municípios da região Norte com maiores proporções de domicílios de aglomerados subnormais. O cluster 1 do Scan Kulldorff inclui o cluster 1 Scan ZIP-EM.

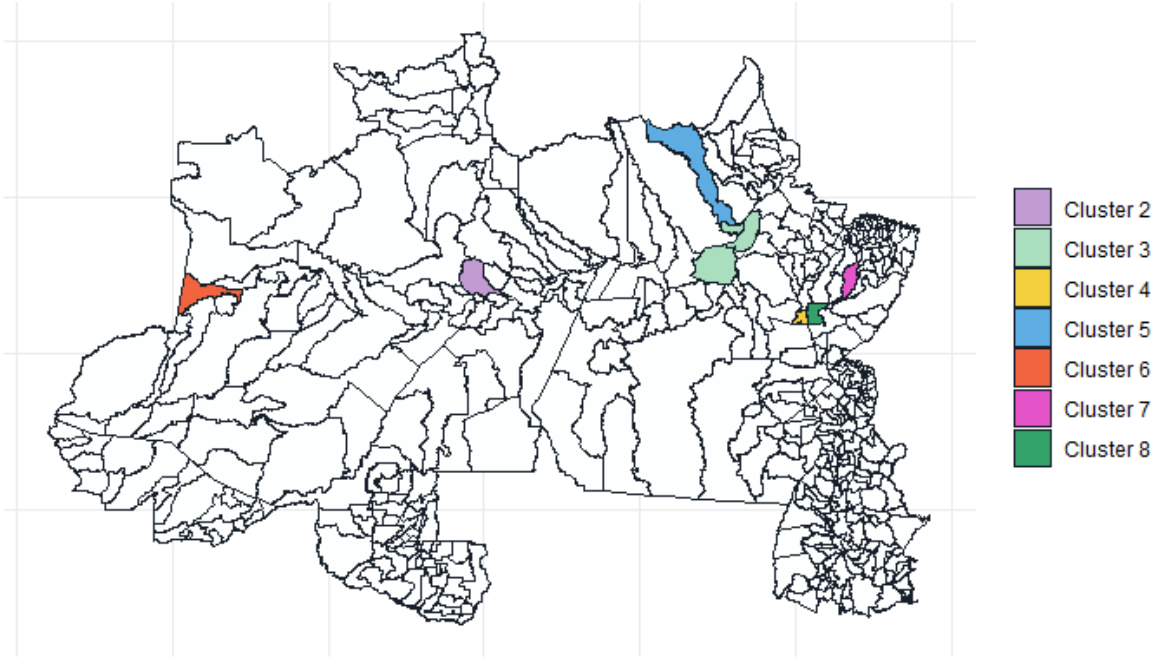

Figura 12: Mapa dos outros clusters obtido com Scan ZIP-EM

Na Figura 12 são apresentados os outros 7 cluster significativos obtidos. Seus números identificadores estão elencados em ordem decresente em relação ao valor da máxima razão de verossimilhança respectivamente.

Os clusters obtidos estão espalhados nos estados do Pará, Amapá e Amazonas. Nota-se que que esses clusters também são pequenos, sendo compostos por no máximo 3 municípios.

Dos clusters apresentados, somente o cluster 6 e o cluster 8 não estão incluídos no cluster 1 do Scan Poisson. Ou seja, um cluster obtido pelo Scan Poisson inclui vários clusters obtidos com Scan ZIP-EM nesse contexto. A vantagem agora é que, ao contrário do cluster imenso identificado anteriormente, um analista teria um direcionamento melhor, identificando os municípios que deveriam ter prioridade na aplicação de recursos e de políticas públicas adequadas.

É importante mencionar que apenas o cluster 3 possui municípios que não incluem aglomeados subnormais. De seus três municípios englobados dois não possuem aglomerados, Gurupá e Porto de Moz. Em contrapartida Vitória do Jari, o terceiro município que o cluster 3 possui, é o município que conta com a maior proporção de domicílios em aglomerados da região.

Tabela 4: Resultados para a clusterização utilizando a técnica Scan ZIP-EM

Cluster	Municípios	Pop.	N° esperado de casos	N° obs. de casos	LLR	p-valor
1	Ananindeua (PA), Marituba (PA), Benevides (PA), Belém (PA)	607.576	114.813,5	333.957	81.794,14	$< 0{,}001$
2	Manaus (AM)	653.218	123.438,5	348.684	80.460,18	$< 0{,}001$
3	Gurupá (PA), Porto de Moz (PA) Vitória do Jari (AP)	14.915	2.818,4	2.114	747,56	$< 0{,}001$
4	Tucuruí (PA)	26.021	4.917,18	10.418	613,97	$< 0{,}001$
5	Laranjal do Jari (AP)	8.428	1.592,64	2.913	63,15	$< 0{,}001$
6	Santo Antônio do Içá (AM)	5.706	1.078,26	1.943	36,63	$< 0{,}001$
7	Tomé-Açu(PA)	17.007	3.213,81	5.081	12,11	$< 0{,}001$
8	Breu Branco (PA)	13.251	2.504,04	3.954	9,09	0,009

Em adendo, na Tabela 4, observa-se que o cluster 3 é o único dos clusters apresentados em que seu número de casos observados não ultrpassa o número esperado de casos.

5 Conclusão

Esse trabalho teve por objetivo identificar clusters de domicílios de aglomerados subnormais na região Norte do Brasil, analisando essa região dividida por municípios, com dados de 2019.

Dessa forma, foram aplicados os algoritmos do Scan Poisson e Scan ZIP. Esse último é usado para dados com excessos de zeros, e como os dados de aglomerados subnormais de 2019 apresentavam 79,56% de valores nulos, se julgou uma maneira interessante de analisar o processo de clusterização também pelo Scan ZIP.

O Scan ZIP caracteriza-se por considerar a existência de zeros estruturais. Estes se definem como regiões onde não é possível se observar casos, independentemente da amostra ou período de estudo. Em contrapartida, o Scan Poisson leva em conta apenas os zeros amostrais, ou seja, as regiões com contagem nula de casos.

Os resultados evidenciaram que para os dois métodos foram identificados clusters nos estados do Amazonas, Amapá e Pará. Além disso, nos dois procedimentos, os municípios que mais possuíam casos, ou seja, aqueles com maiores proporções de domicílios de aglomerados foram incluídos nos clusters detectados.

O cluster com a maior evidência estatística detectado com o método Scan Poisson abrange 113 municípios, dos quais 42,48% não possuíam aglomerados subnormais, o que representa a heterogeneidade na distribuição dos aglomerados dentro desse conglomerado. Notou-se ainda que esse cluster inclui 6 dos 8 clusters apresentados na detecção pelo Scan ZIP, sugerindo uma convergência entre os resultados obtidos pelos dois métodos. No entanto, compreende-se que, por não considerar a ocorrência de zeros estruturais na clusterização, o Scan Poisson resulta na inclusão dos zeros, considerados amostrais, no processo da incorporação das regiões com alta incidência. A restrição ao formato circular das zonas candidatas contribui nesse processo, pois ao englobar as áreas com elevado número de casos necessariamente inclui áreas com ausência de casos. Esse cluster, então, se torna tão grande, que do ponto de vista geográfico e de políticas públicas, não é últil e viável.

Já o cluster mais verossímil identificado com o método Scan ZIP apresenta 4 municípios localizados no estado do Pará. São eles: Ananindeua, Marituba, Benevides e Belém. Todos possuíam aglomerados subnormais e com proporção de domicílios em aglomerados igual ou maior a 42,73%. Em comparação ao cluster anterior, é uma detecção que permite priorizar áreas na aplicação de recursos e de políticas públicas assertivas.

Destaca-se ainda o tempo de processamento dos algoritmos: para o Scan Poisson foi de 46,18 segundos e para o Scan ZIP, com tempo superior, foi de 4325 segundos.

5.1 Trabalhos Futuros

Como trabalhos futuros, imagina-se estender o estudo para as outras regiões do país, visando enriquecer os estudos sobre habitações no Brasil e até mesmo identificar possíveis padrões nos clusters, sendo possível serem utilizados como apoio para desenvolvimento de políticas públicas.

Além disso, seria interessante aplicar o estudo para dados mais atualizados do Censo de 2022.

Considera-se também realizar aplicações que não utilizem a varredura em formato circular, como por exemplo o Scan flexível.

Referências

AGRESTI, A. *Categorical Data Analysis . 2nd Edition.* [S.l.]: New York: John Wiley Sons, Inc., p. 320-332., 2002.

ALLÉVIUS, B. scanstatistics : Space-time anomaly detection using scan statistics. *Journal of Open Source Software.*, 2018.

ANDRADE, A. L. et al. *Introdução à Estatística Espacial para a Saúde Pública.* [S.l.]: Ministério da Saúde, 2007.

ARAÚJO, T. C. Extensão da Estatística Scan para detecção de conglomerados espaço-temporais em dados com excesso de zeros. *Brasília, UnB*, 2012.

CÂMARA, G. et al. *Análise espacial e geoprocessamento.* [S.l.]: EMBRAPA, 2004. (Capítulo 1).

CANÇADO, A. L. F.; SILVA, C. Q. da; SILVA, M. F. A spatial scan statistic for zero-inflated poisson process. *Environ Ecol Stat, v. 21, p. 627–650*, 2014.

COSTA, V. G.; NASCIMENTO, J. A. S. O conceito de favelas e assemelhados sob o olhar do ibge, das prefeituras do brasil e da onu. *Anais do X Encontro de Geógrafos da América Latina: USP*, 2005.

DEMPSTER, A. P.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. , pp. 1-38.*, 1977.

FERNANDES, L. B. Uma estatística scan espacial bayesiana para dados com excesso de zeros. *Brasília, UnB*, 2015.

FERNANDES, L. B.; REIS, S. D. S. Detecção, identificação e inferência de conglomerados espaciais de fraudes bancárias em uma instituição financeira no centro-oeste do brasil. *Brasília, UnB*, 2011.

FIGUEIREDO, R. L. Detecção de clusters usando a estatística scan espacial circular em conjuntos seletivos e um fator de penalização: a ocupaçõa circular. *Belo Horizonte, Universidade Federal de Minas Gerais*, 2010.

FILHO, A. P. Q. As definições de assentamentos precários e favelas e suas implicações nos dados populacionais: abordagem da análise de conteúdo. *Revista Brasileira de Gestão Urbana (Brazilian Journal of Urban Management): USP*, 2015.

IBGE. Instituto brasileiro de geografia e estatística (ibge) - aglomerados subnormais: Informações terrioriais. censo demográfico 2010. *Rio de Janeiro*, 2010.

IBGE. Instituto brasileiro de geografia e estatística - aglomerados subnormais 2019: Classificação preliminar e informações de saúde para o enfrentamento à covid-19. notas técnicas. *Rio de Janeiro*, 2020.

KREMPI, A. P. Explorando recursos de estatística espacial para análise da acessibilidade da cidade de bauru. *São Carlos, USP*, 2004.

KULLDORFF, M. *A spatial scan statistic. Communications in Statistics: Theory and Methods*. [S.l.]: Wiley New York, 1997.

MELO, J. C. S.; MELO, A. C. O.; MORAES, R. M. Comparação dos métodos scan circular e flexível na detecção de aglomerados espaciais de dengue. *Anais da 1ª Escola de Informática Teórica e Métodos Formais : Universidade Federal da Paraíba*, 2006.

Anexo

A Municípios do cluster 1 do modelo Scan de Kulldorff

Autazes (AM) , Barreirinha (AM) , Boa Vista Do Ramos (AM) , Borba (AM) , Careiro Da Várzea (AM) , Itacoatiara (AM) , Itapiranga (AM) , Manaus (AM) , Maués (AM) , Nhamundá (AM) , Nova Olinda Do Norte (AM) , Parintins (AM) , Presidente Figueiredo (AM) , Rio Preto Da Eva (AM) , São Sebastião Do Uatumã (AM) , Silves (AM) , Urucará (AM) , Urucurituba (AM) , Abaetetuba (PA) , Acará (PA) , Afuá (PA) , Água Azul Do Norte (PA) , Alenquer (PA) , Almeirim (PA) , Altamira (PA) , Anajás (PA) , Ananindeua (PA) , Anapu (PA) , Aveiro (PA) , Bagre (PA) , Baião (PA) , Barcarena (PA) , Belém (PA) , Belterra (PA) , Benevides (PA) , Brasil Novo (PA) , Breu Branco (PA) , Breves (PA) , Bujaru (PA) , Cachoeira Do Arari (PA) , Cametá (PA) , Chaves (PA) , Colares (PA) , Curralinho (PA) , Curuá (PA) , Faro (PA) , Goianésia Do Pará (PA) , Gurupá (PA) , Igarapé-Miri (PA) , Itaituba (PA) , Itupiranga (PA) , Jacareacanga (PA) , Jacundá (PA) , Juruti (PA) , Limoeiro Do Ajuru (PA) , Marabá (PA) , Marituba (PA) , Medicilândia (PA) , Melgaço (PA) , Mocajuba (PA) , Moju (PA) , Mojuí Dos Campos (PA) , Monte Alegre (PA) , Muaná (PA) , Nova Ipixuna (PA) , Novo Progresso (PA) , Novo Repartimento (PA) , Óbidos (PA) , Oeiras Do Pará (PA) , Oriximiná (PA) , Ourilândia Do Norte (PA) , Pacajá (PA) , Parauapebas (PA) , Placas (PA) , Ponta De Pedras (PA) , Portel (PA) , Porto De Moz (PA) , Prainha (PA) , Rurópolis (PA) , Salvaterra (PA) , Santa Bárbara Do Pará (PA) , Santa Cruz Do Arari (PA) , Santa Izabel Do Pará (PA) , Santarém (PA) , Santo Antônio Do Tauá (PA) , São Félix Do Xingu (PA) , São Sebastião Da Boa Vista (PA) , Senador José Porfírio (PA) , Soure (PA) , Tailândia (PA) , Terra Santa (PA) , Tomé-Açu (PA) , Trairão (PA) , Tucumã (PA) , Tucuruí (PA) , Uruará (PA) , Vigia (PA) , Vitória Do Xingu (PA) , Serra Do Navio (AP) , Amapá (AP) , Pedra Branca Do Amapari (AP) , Calçoene (AP) , Cutias (AP) , Ferreira Gomes (AP) , Itaubal (AP) , Laranjal Do Jari (AP) , Macapá (AP) , Mazagão (AP) , Porto Grande (AP) , Pracuúba (AP) , Santana (AP) , Tartarugalzinho (AP) , Vitória Do Jari (AP).