Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Uma Estatística de Varredura Árvore-Espacial

Geiziane Silva de Oliveira

Dissertação apresentada ao Departamento de Estatística do Instituto de Exatas da Universidade de Brasília como parte dos requisitos necessários para o grau de Mestre em Estatística.

Brasília
2019

Geiziane Silva de Oliveira

Uma Estatística de Varredura árvore-Espacial

Dissertação apresentada ao Departamento de Estatística do Instituto de Exatas da Universidade de Brasília como parte dos requisitos necessários para o grau de Mestre em Estatística.

Orientador: Prof. Dr. **André Luiz Fernandes Cançado**

Brasília, 26 de Julho de 2019

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A Deus, agradeço pela força e proteção concebida para a conclusão deste trabalho.

Ao professor Dr. André Luiz Fernandes Cançado, pelos ensinamentos que foram essenciais para o desenvolvimento desse trabalho, por me compreender e se dedicar na orientação desse trabalho.

Ao SAS Institute Brasil, agradeço pelo apoio e disponibilização da licença de uso pessoal como funcionária da equipe de Consultoria e máquinas virtual utilizadas para execução das simulações desse trabalho.

A minha família, meus pais Joaquim e Aurelina e todos os meus irmãos e sobrinhos, por compreenderem minhas ausências e por todo amor e carinho.

A minha irmã Maria José, em especial, pelo apoio e acolhimento em Brasília durante o início da minha vida acadêmica.

Aos professores do Departamento de Estatística que muito contribuíram com minha formação acadêmica. Em especial, aos professores do Programa de Pós Graduação: Dr Bernardo Borba, Dr Raul Matsushita e Drª Cira Etheowalda.

Ao Professor Dr George von Borries, por sempre me incentivar na continuação dos meus estudos, ensinamentos e conversas.

Ao Adolfo, pela amizade, estudos e auxílio, que foi essencial em algumas disciplinas do mestrado.

Aos meus amigos da turma de mestrado, alguns deles da graduação para a vida (Erique Pereira, Márcia Maia, Mateus Carbone), Alisson Silva e Alessandra Moreira, pela amizade, inúmeros dias de estudos e pelos momentos de descontração.

À amiga Márcia Maia, por todo companheirismo durante esse período e pelas contribuições na revisão do texto.

Aos meus amigos de trabalho do SAS Brasília, São Paulo e Rio de Janeiro, por compreenderem meus horários malucos nos projetos que trabalhamos, devido às aulas na UnB, além do carinho e momentos de descontração.

À amiga Carlinha, por todo carinho, muito café e as saladas de fruta que sempre ajudou a segurar a jornada de trabalho e mestrado.

A Luiz Kauffmann e Gabriel Antunes, pela compreensão, incentivo e flexibilização dos meus horários de trabalho.

À minha amiga Jéssica e sua mãe Rute, por todo carinho, apoio, e por serem minha segunda família em Brasília.

Aos meus amigos, Felipe Quintino, Andressa Lima, Claudia Edith, Mariana Fehr, Ágda Galletti e Rodrigo Ferrari, por todo carinho e contribuições acadêmicas.

À Haianne Sampaio, pela disponibilização dos arquivos com a estrutura hierarquia da CID-10, que facilitou muito na construção da base de dados.

As minhas amigas, Denise Rayanne e Cíntia Soares, pela amizade e compreensão.

Aos funcionários do Departamento de Estatística, especialmente a Tathyanna Cordeiro, Edenilson e André.

Aos professores Drs. Antônio Eduardo Gomes e Luiz Henrique Duczmal por aceitarem participar da banca examinadora desse trabalho e pelas contribuições.

Resumo

Nesse trabalho propomos a técnica estatística de varredura árvore-espacial, uma combinação do método *Scan* Circular de Kulldorff, utilizado para detecção de *clusters* espaciais, e a técnica de mineração de dados estatística de varredura baseada em árvore. A ideia principal do método é incluir a informação espacial (geográfica) dos eventos que são naturalmente dispostos em forma hierárquica na técnica de varredura baseada em árvore. A estatística de varredura baseada em árvore dispõe-se a varrer todos os possíveis ramos da árvore a fim de identificar o ramo no qual a probabilidade associada de casos é maior do que o esperado sob a hipótese de homogeneidade dos eventos. Dessa forma, o método de varredura árvore-espacial busca identificar um conjunto de regiões z e um ramo g da árvore para os quais a probabilidade de um indivíduo vir a ser um caso associado ao ramo g é maior dentro desse conjunto de regiões do que para esse mesmo ramo fora dessas regiões. O algoritmo foi avaliado por meio de simulações com cenários hipotéticos considerando a estrutura espacial e hierárquica e apresentou um bom desempenho em relação à capacidade de detecção das estruturas espacial e hierárquica. O método de varredura árvore-espacial foi aplicado a dados de mortalidade infantil segundo a classificação Estatística Internacional de Doença-CID 10 para o estado do Rio de Janeiro no ano de 2016. A aplicação do método permitiu identificar um conjunto de municípios no Rio de Janeiro para os quais uma subcategoria de doenças possui número de óbitos significativamente maior que o esperado sob a hipótese de homogeneidade.

Palavras-chave: Estatística espacial, *tree Scan* , árvore-espacial, *clusters* espacial, CID-10, óbitos infantis, estrutura hierárquica.

Abstract

In this work we propose the statistical technique of spatial-tree scan which is the combination of the statistic scan method used in spatial cluster detection and the data mining technique, tree-based scan statistic. The main idea of the method is to include the spatial (geographical) information of the events that are naturally arranged hierarchically in the tree-based scan statistic. The tree-based scan statistic is arranged to scan all possible branches in order to identify the most likely tree branch, that is, the branch in which the associated probability of cases is greater than expected under the hypothesis of homogeneity of events. Thus, the spatial-tree scan method seeks to identify a set of regions z and a branch g of the tree in which the probability of an individual happens to be a case associated with the g branch is larger within this set of regions than for that same branch outside these regions. The algorithm was evaluated by means of simulations with hypothetical scenarios considering the spatial and hierarchical structures and presented a good performance in terms of the capacity of detection of the spatial and hierarchical structures. The statistical method of spatial-tree scan was applied to infant mortality data according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) for the state of Rio de Janeiro in the year 2016. The application of the method allowed to identify a set of municipalities in Rio de Janeiro for which a subcategory of diseases has a significantly higher number of deaths than expected under the homogeneity hypothesis.

Key Words: Spatial statistics, tree Scan , Spatial-tree, Spatial clusters, ICD-10, infant deaths, Hierarchical structure.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

Este trabalho tem como objetivo apresentar uma adaptação da estatística de varredura baseada em árvore (*tree scan*) (Kulldorff et al., 2003a) para a abordagem espacial. A estatística de varredura baseada em árvore consiste em um método de mineração de dados utilizado para a identificação de conglomerado (*cluster*) em dados hierárquicos apresentados em formato de árvore.

Usualmente, compreender a distribuição espacial de eventos tornou-se um aspecto importante nas áreas epidemiológicas, financeiras, agronômicas, socioeconômicas, entre outras. Para as organizações de vigilância à saúde é comum estudar a incidência de casos de doenças segundo a região geográfica a fim de identificar um conjunto de regiões ou sub-área em que a incidência de casos é maior, se comparada às demais e, com base nisso, adotar políticas especificas para essas áreas (Câmara et al., 2000).

A estatística espacial compreende um conjunto de técnicas que tem como objeto identificar modelos inferenciais para estudar as características de eventos levando em consideração a localização espacial do evento de interesse de maneira direta. A fase inicial se dedica a análise exploratória dos fenômenos no espaço e visualização através de mapas. (Câmara et al., 2000).

Os métodos em análise espacial consideram três tipos de dados:

- Dados pontuais: os eventos são representados por suas ocorrências no espaço (pontos no espaço). Esses dados são geralmente associados aos pares de coordenadas (x_i, y_i). Alguns exemplos são localização de crimes, casos de doenças, localização de espécies de plantas.

- Dados de área ou agregados: os eventos são normalmente obtidos por meio de levantamento censitário e não se conhece a localização exata da ocorrência de cada evento, apenas o número total de ocorrências em cada região. Por exemplo, dados de saúde normalmente são agregados por unidade de análise, como municípios.

- Dados contínuos ou de superfície: correspondem a conjuntos de amostras observadas em determinadas localizações do mapa. Os dados podem ser representados por meio das coordenadas geográficas (x_i, y_i) e da medição z_i realizada naquela localização,

formando triplas (x_i, y_i, z_i). O objetivo mais recorrente é reconstruir a superfície de onde as amostras foram retiradas, por exemplo, perfis de solo ou temperatura em determinadas localizações.

Na análise de dados pontuais, o evento de interesse é a própria localização espacial dos eventos, portanto, o objetivo principal da análise é testar hipótese sobre o padrão observado, ou seja, se os dados são distribuídos no espaço de forma aleatória ou se há algum padrão de distribuição espacial. Dessa forma, um *cluster* ou conglomerado espacial é dado por uma área ou sub-região com maior probabilidade de ocorrência, ou um número de casos significativamente maior que o esperado.

Inicialmente, muitas técnicas foram desenvolvidas com objetivo de entender a distribuição espacial de eventos e determinar a presença de algum padrão espacial (conglomerado). E muitos desses estudos tiveram motivação na área da saúde, como o *Geographical Analysis Machine* (GAM) de Openshaw et al. (1988), *Mapas Baseados em Probabilidade* de (Choynowski, 1959) e o *Scan* Circular de Kulldorff (1997). Porém, a maior parte desses métodos eram exploratórios, isto é, apenas permitia identificar um *cluster* mas não testava sua significância estatística. O método *Scan* de Kulldorff foi o primeiro método que permitiu fazer inferência e, portanto, tornou-se uma técnica muito utilizada para detecção de *clusters* espaciais, por sua fácil implementação computacional. A ideia do algoritmo é varrer regiões geográficas, em janelas circulares, determinando as zonas candidatas (conjuntos de regiões) em que o número de casos é maior que o esperado. Para a modelagem do número de casos, são utilizadas usualmente as distribuições Binomial e Poisson.

Posteriormente, propôs-se a estatística de varredura baseada em árvore (*Tree Scan*) Kulldorff et al. (2003a), similar ao método *Scan* Circular, porém não se considera a informação espacial. Esse método consiste em identificar o ramo da árvore para o qual o número de casos é significativamente maior que o esperado sob a hipótese de uniformidade na distribuição dos casos.

A proposta deste trabalho é incluir a informação espacial dos eventos na técnica de varredura baseada em árvore; um método de varredura árvore-espacial. Portanto, se pensarmos em dados estruturados em forma de árvore, relacionados a doenças por exemplo, tem-se o sistema de Classificação de Estatística Internacional de Doenças e Problemas Relacionados a Saúde-CID 10, sob o qual as doenças estão divididas em capítulos, os capítulos são formados por agrupamentos de doenças, seguidos pelas categorias, essas por sua vez em sub-categorias. As sub-categorias finalmente armazena as informações das doenças específicas. A ideia é detectar uma anomalia em uma região geográfica na qual um ramo dessa árvore seja mais incidente.

Os métodos serão implementados por meio do *Software SAS* e avaliados por simulação e dados reais.

Objetivos

Objetivo Geral

Implementar um abordagem espacial para a estatística de varredura baseada em árvore. Isto é, fazer a junção dos algoritmos *Scan* Circular e *Scan* baseado em árvore.

Objetivos Específicos

- Apresentar uma revisão bibliográfica dos métodos *Scan* Circular e *Scan* baseado em árvore.
- Propor a estatística de varredura árvore-espacial.
- Implementar os algoritmos no *Software SAS* por meio do módulo de linguagem matricial iterativa-*SAS/IML*, *SAS/STAT* e *SAS/GRAPHICS*;
- Testar os métodos por meio de dados simulados e reais.

Capítulo 1

Revisão Bibliográfica

1.1 Visão geral: Detecção de *Clusters*

A análise de *clusters*, ou análise de conglomerados, tem como objetivo identificar padrões, ou seja, dividir elementos ou indivíduos em grupos, de forma que elementos com característica similares sejam alocados dentro de um mesmo grupo e os elementos com características diferentes em grupos heterogêneos (Mingoti, 2005). Na análise de *clusters*, não se conhece *a priori* o número ou a estrutura de grupos. Um grupo é identificado por meio de medidas de similaridade ou dissimilaridade (distâncias). Essas medidas fornecem uma ideia de quão próximas ou distantes estão as observações (Johnson e Wichern, 1988).

Entre as principais medidas de dissimilaridade (distância) e similaridade estão, entre outras

- Dissimilaridade: distância euclidiana, distância euclidiana média, distância de Minkowsky, distância generalizada ou ponderada;
- Similaridade: coeficiente de concordância de Jaccard, coeficiente de concordância simples, coeficiente de concordância positiva.

Para as medidas de dissimilaridade, tem-se que quanto menor forem os valores, mais similares serão os elementos comparados. E, quanto às medidas de similaridade, quanto maior for o valor, maior é a similaridade.

As metodologias para análise de *clusters* são divididas em técnicas hierárquicas e não-hierárquicas. As técnicas de *cluster* hierárquicas são compostas por uma série de uniões sucessivas ou por uma série de divisões, sendo, portanto, subdivididas em métodos aglomerativos e divisivos. Os métodos aglomerativos iniciam com n objetos individuais, em que cada elemento é um *cluster*. O algoritmo classifica e agrupa os elementos, de acordo com a relação de similaridade (ou com base em uma métrica de dissimilaridade), de modo que, em cada passo, sejam formados novos grupos. O processo continua, formando subgrupos cada vez maiores, até que todos os elementos sejam alocados em um único *cluster* ou algum critério de parada estabelecido seja satisfeito. Quando não se atinge um

único *cluster*, pode-se observar uma similaridade pequena entre os subgrupos formados. Já os métodos divisivos executam o processo oposto. Um grupo inicial formado por todos os objetos é dividido em dois subgrupos de forma que os elementos de um subgrupo estejam distantes dos elementos do outro grupo. Esse algoritmo é repetido até que cada elemento forme um grupo.

Os métodos hierárquicos são técnicas exploratórias comumente usadas para identificar os possíveis *clusters* e a quantidade de grupos. O gráfico dendrograma é utilizado para identificar o número de grupos por nível da medida de similaridade. Algumas técnicas hierárquicas são: Método de Ligação Simples (*Single Linkage*), Método da Média das Distâncias (*Average Linkage*), Método do Centroide, Método da Ligação Completa, e Método de Ward.

As técnicas não hierárquicas, ao contrário dos métodos hierárquicos, necessitam que a quantidade de grupos seja previamente estabelecida pelo pesquisador. A ideia é encontrar uma partição inicial em k grupos para a qual, internamente, os grupos sejam semelhantes e possíveis de serem isolados (separação dos *clusters* selecionados). Esse processo é inciado selecionando aleatoriamente pontos inciais entre os itens ou dividindo itens em grupos inciais. Uma matriz de distâncias não precisa ser determinada e a base de dados não precisa ser armazenada em cada passo do algoritmo, logo, pode-se aplicar em grandes conjuntos de dados. O método *K-means* é um dos procedimentos mais populares.

Para mais detalhes sobre esses métodos, o leitor pode recorrer aos trabalhos de Johnson e Wichern (1988) e Mingoti (2005).

Ressalta-se que o processo de detecção de *cluster* espacial será o foco desse trabalho, no entanto as técnicas de identificação de *cluster* tradicional foram apresentadas resumidamente, apenas no intuito de mostrar que apesar dos objetivos serem os mesmos, o processo de identificação de um *cluster* espacial difere das técnicas tradicionais.

1.2 Visão Geral: Detecção de *Clusters* Espaciais

A estatística espacial é uma área da estatística que estuda a ocorrência de eventos no espaço. A ideia principal é obter algum significado, ou seja, encontrar padrões de dados distribuídos no espaço a partir da localização e contexto de fenômenos do mundo real, como por exemplo ocorrência de crimes, doenças, acidentes de trânsito, crescimento da vegetação, preço médio de residência em diferentes regiões, entre outros.

Um *cluster* espacial é representado por uma área ou região que possui alta incidência ou predominância de casos. Em outras palavras, uma região que contém um número de casos (evento em estudo, ex: o número de casos de dengue) significativamente maior que o esperado se denomina um *cluster*. Em muitos estudos, principalmente na área médica e epidemiológica, tem-se o interesse em saber se o número de casos apresenta algum padrão em determinadas regiões ou pode-se atribuir simplesmente ao acaso (Araújo, 2012).

Os primeiros estudos baseavam-se em calcular as frequências absolutas ou relativas de algum fator de risco, como por exemplo, apurar a quantidade de pessoas com determinada infecção em algumas regiões e verificar se essa doença tinha presença mais forte em uma determinada região em relação às demais.

Para responder essas questões, muitos estudos foram desenvolvidos na área de estatística espacial. Uma revisão detalhada das primeiras técnicas de análise de *cluster* espacial foi apresenta por Marshall (1991).

Os itens abaixo compreendem uma ideia dos primeiros métodos apontados para identificação de *clusters* espaciais:

- **Mapas Baseados em Probabilidades de Choynowski**:
  Este foi o primeiro método proposto para detecção de *clusters* espaciais baseado na distribuição de probabilidade do evento de interesse. O procedimento foi ilustrado por Choynowski (1959), mostrando-se contrário aos estudos anteriores, que simplesmente mostravam a distribuição de algum fenômeno sobre uma área geográfica em termos de frequências absolutas ou porcentagens. No estudo de Choynowski, o objetivo era modelar a distribuição do número de casos em algumas regiões, parte da Polônia, divididas em municípios. Portanto, modelou-se o número de casos de tumores cerebrais em cada região por meio da distribuição Poisson e obteve-se a probabilidade do número de casos observados ser superior ou inferior ao número de casos esperados em cada área. Dessa forma, para cada uma das áreas do mapa em estudo tem-se uma probabilidade, mas não existe uma medida de significância global em relação à presença do *cluster*. A verificação da significância do *cluster* é realizada em cada área separadamente, ou seja, avalia se uma região possui um número de casos significativamente alto, considerando um nível de significância α. Portanto, o método é restrito à ocorrência de *clusters* apenas dentro de uma mesma área, não sendo possível detectar a formação de *cluster* na vizinhança de outras áreas ou municípios. Além disso, para averiguar a existência ou não de *clusters*, baseado na informação para área total, haveria o problema de múltiplos testes. Uma alternativa é a correção de Bonferroni.

- **Distribuição do Tamanho do Máximo de *Clusters* em uma linha - Naus (1965a)**
  Seguindo a mesma linha de Choynowski (modelagem baseada em probabilidade), o trabalho de Naus (1965a) busca determinar a probabilidade de que, de uma amostra de N pontos de uma distribuição uniforme (0,1), exista um subintervalo de (0,1) de tamanho p que possua pelo menos n dos N pontos observados, com $n > 2$.

  No contexto espacial, como o cálculo das probabilidades era obtido unidimensionalmente, a aplicabilidade da técnica era limitada. Logo, seria necessário somar o número de casos segundo coordenadas geográficas (latitude ou longitude), e assim, supor que o número de casos (unidimensional) se distribuem uniformemente e obter

as probabilidades de ocorrência. Em seguida, com base nisto, seria preciso averiguar se os dados seguem uma distribuição uniforme.

- **Agrupamento de pontos aleatórios em duas dimensões - Naus (1965b)** Considerando a limitação do método proposto por Naus (1965a), no mesmo ano foi proposta uma correção para o problema de trabalhar apenas em uma dimensão (Naus, 1965b). A ideia principal é obter um limite inferior e superior da probabilidade de ocorrência de um evento em uma determinada área. Assim, em uma das suas aplicações, um valor é fixado para uma área, e busca-se determinar o formato da janela de varredura que fornece a maior probabilidade de se obter o maior *cluster* e se esse formato está associado com o número esperado de casos.

- **Máquina de Análise Geográfica - Openshaw et al. (1988)**

  A máquina de análise geográfica, denotada pela sigla GAM (do inglês *Geographical Analysis Machine*), foi proposta por Openshaw et al. (1988) e tem inspiração no trabalho de Choynowski. A análise foi construída desenhando em cima de uma região geográfica (mapa) vários círculos sobrepostos de raio R com tamanhos variados, espaçados regularmente e abrangendo a área total em estudo. O tamanho do raio (tamanho do *cluster*) deve ser informado *a priori*. Como não existe uma medida exata, em geral, nos estudos, a análise era realizada com diferentes tamanhos de raios. Esse processo garante que todos os locais possíveis sejam examinados e que as regiões em estudo tenham formas diferentes das delimitações geográficas naturais, por exemplo, dos municípios da área em estudo. O valor do raio R de cada círculo varia conforme os valores predefinidos. A análise é repetida para cada tamanho de raio e a significância do *cluster* (círculos selecionados) é determinada com base em simulações de Monte Carlo, comparando-se o número de casos observados no interior do círculo com o percentil $P_{99,8}$ da distribuição empírica sob hipótese nula (para maiores detalhes ver trabalho de (Openshaw et al., 1988)). Esse processo, resulta em um mapa com vários círculos de diferentes tamanhos, onde cada círculo é individualmente significativo. Apesar de ser de fácil compreensão, não foi apresentada uma forma geral para verificar a presença ou não de um *cluster* espacial. Esse procedimento leva à execução de múltiplos testes simultaneamente, à medida que se aumenta o número de raios. Portanto, novamente faz-se necessário um ajuste para múltiplos testes. Com isso, o método de Bonferroni produz um teste bem conservador, visto que aumenta o nível de significância dos testes.

Para fins de contornar esses problemas e produzir um método capaz de verificar a significância do *cluster* espacial, foi proposta a Estatística *Scan* Circular.

1.3 A Estatística *Scan* Circular de Kulldorff

A Estatística *Scan* Circular é uma metodologia para análise de *clusters* espaciais com dados pontuais ou agregados que permite a identificação, ou seja, detecta a localização de um *cluster* e testa sua significância estatística.

Esta metodologia foi apresenta em uma forma concreta por Kulldorff (1997), mas já havia sido desenvolvida por Naus (1965a) para o caso unidimensional e no artigo de Kulldorff e Nagarwalla (1995) como uma maneira de detectar a localização e testar a significância de um *cluster* espacial, considerando o modelo Poisson. Já no trabalho de Kulldorff (1997) foram apresentados os modelos Poisson e Binomial e as formas do teste para verificação da significância de um *cluster* espacial.

As distribuições Binomial e Poisson são frequentemente utilizadas para modelar dados de contagem, por exemplo o número de casos de uma determinada doença. Assim, a escolha entre essas duas distribuições depende muito do tipo de aplicação e suas características, mas se o número de casos em estudo é pequeno em relação à população total em risco, não existem diferenças significativas, pois os modelos Poisson e Binomial se aproximam. Para mais detalhes ver Kulldorff (1997),pgs 1484-1485.

1.3.1 Algoritmo Scan Circular Kulldorff

O método *Scan* circular consiste em um processo iterativo que faz uma varredura em uma área geográfica com janelas circulares com tamanho variável. O processo é replicado até atingir o tamanho da janela, como por exemplo, uma proporção máxima da população (Naus, 1965a). A proporção de pontos ou eventos são mensurados e comparados com uma distribuição de probabilidade que modela o número de pontos.

Considere M como uma área geográfica subdivida em k regiões, onde em cada região $i = 1, 2, \ldots, k$ existe uma população em risco n_i, que é dada pelo total de expostos/suscetíveis a algum evento de interesse. Sejam x_i o número de casos na i-ésima região, $C = \sum_{i=1}^{k} x_i$ é o número total de casos e $N = \sum_{i=1}^{k} n_i$ a população total em risco.

Considere $z \subset M$ um subconjunto de regiões do mapa M, denominado de zona, $\bar{z} = M - z$, e Z o conjunto de zonas. Um *cluster* espacial é uma zona z para a qual a probabilidade p_z de que um indivíduo de sua população de risco venha a ser um caso é maior que em $\bar{z}$.

Com base na descrição acima, considere as seguintes definições:

- $x_z = \sum_{i \in z} x_i$ é o número de casos na zona z;
- $x_{\bar{z}} = \sum_{i \notin z} x_i = C - x_z$ é o número de casos fora da zona z ;
- $n_z = \sum_{i \in z} n_i$ é a população em risco na zona z;

- $n_{\bar{z}} = \sum_{i \notin z} n_i = N - n_z$ é população em risco fora da zona z.

A estatística *Scan* de Kulldorff é construída com base em um teste de razão de verossimilhança, com as seguintes hipóteses:

$$\begin{cases} H_0 : p_z = p_{\bar{z}} = p_0, \forall z \in Z \\ H_a : \exists\ z \text{ tal que } p_z > p_{\bar{z}} \end{cases}$$

Sendo p_z a probabilidade de ocorrer um caso na zona z e $p_{\bar{z}}$ a probabilidade de ocorrer casos fora da zona z.

Logo, sob a hipótese nula H_0, a probabilidade p que um indivíduo venha a ser um caso é a mesma em qualquer região do mapa M. E a hipótese alternativa é de que existe pelo menos uma zona z, tal que a probabilidade p de que um indivíduo venha a ser um caso em z é maior do que a probabilidade $p_{\bar{z}}$ de que um indivíduo venha a ser um caso em qualquer região fora da zona z. Portanto, se a hipótese nula do teste for rejeitada a zona z é considerada um *cluster*.

1.3.2 Modelo Poisson

A distribuição Poisson é frequentemente utilizada para modelar dados de contagem. Sendo x_i o número de casos do evento de interesse, na i-ésima região do mapa M, tem-se que:

$$x_i \sim Poisson\left(n_i p_i\right),$$

em que, $\mu_i = n_i p_i$ é o número esperado de casos na i-ésima região.

Dessa forma, sob a hipótese nula da estatística *Scan*, tem-se que $p_i = p_0$. Fazendo $\mu_i = n_i p_i$ como número esperado de casos, a função de verossimilhança é dada por:

$$\mathcal{L}_0(p_0, \mathbf{x}) = \prod_{i=1}^{k} \frac{e^{-\mu_i} \mu_i^{x_i}}{x_i!} = \frac{e^{-\sum_i \mu_i} \prod_{i=1}^{k} \mu_i^{x_i}}{\prod_{i=1}^{k} (x_i!)}$$

$$= \frac{e^{-p_0 \sum_i n_i} p_0^{\sum_i x_i} \prod_{i=1}^{k} (n_i)^{x_i}}{\prod_{i=1}^{k} (x_i!)} = \frac{e^{-p_0 N} p_0^{C} \prod_{i=1}^{k} (n_i)^{x_i}}{\prod_{i=1}^{k} (x_i!)}. \tag{1.1}$$

Aplicando logaritmo na Equação (1.1) tem-se a log-verossimilhança:

$$\ell_0(p_0, \mathbf{x}) = \log\left(\mathcal{L}_0\right) = -p_0 N + C \log p + \sum_{i=1}^{k} x_i \log n_i - \sum_{i=1}^{k} \log\left(x_i!\right). \tag{1.2}$$

Derivando a Equação 1.2 com relação a p_0 e igualando a zero, obtém-se o estimador de máxima verossimilhança para p_0.

$$\frac{\partial \ell_0}{\partial p_0} = -N + \frac{C}{p_0} = 0 \rightarrow \hat{p}_0 = \frac{C}{N}, \tag{1.3}$$

De forma análoga, sob a hipótese alternativa H_a, para alguma zona $z \in Z$ temos que;

$$\begin{cases} p_i = p_z & se \quad i \in z \\ p_i = p_{\bar{z}} < p_z & se \quad i \notin z \end{cases}$$

A função de verossimilhança sob H_a é dada por:

$$\begin{aligned} \mathcal{L}(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x}) &= \prod_{i \in z}^{k} \frac{e^{-p_z n_i} \left(p_z n_i\right)^{x_i}}{x_i!} \times \prod_{i \notin z}^{k} \frac{e^{-p_{\bar{z}} n_i} \left(p_{\bar{z}} n_i\right)^{x_i}}{x_i!} \\ &= \frac{e^{-\sum_{i \in z} n_i p_z} \prod_{i \in z} n_i^{x_i} p_z^{\sum_{i \in z} x_i}}{\prod_{i \in z}^{k} x_i!} \times \frac{e^{-\sum_{i \notin z} n_i p_{\bar{z}}} \prod_{i \notin z} n_i^{x_i} p_{\bar{z}}^{\sum_{i \notin z} x_i}}{\prod_{i \notin z}^{k} x_i!} \end{aligned} \tag{1.4}$$

Aplicando o logaritmo na Equação 1.4, obtém-se a função log-verossimilhança:

$$\begin{aligned} \ell(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x}) = \log\left(\mathcal{L}\right) = -p_z n_z + \sum_{i \in z}^{k} x_i \log n_i + x_z \log p_z - \sum_{i \in z}^{k} \log\left(x_i!\right) \\ -p_{\bar{z}} n_{\bar{z}} + \sum_{i \notin z}^{k} x_i \log n_i + x_{\bar{z}} \log p_{\bar{z}} - \sum_{i \notin z}^{k} \log\left(x_i!\right). \end{aligned} \tag{1.5}$$

Derivando a equação 1.5 e igualando a zero são encontrados os valores de p_z e $p_{\bar{z}}$ que maximizam a função log-verossimilhança.

$$\frac{\partial \ell}{\partial p_z} = -n_z + \frac{x_z}{p_z} = 0 \rightarrow \hat{p}_z = \frac{x_z}{n_z}, \tag{1.6}$$

$$\frac{\partial \ell_0}{\partial p_{\bar{z}}} = -n_{\bar{z}} + \frac{x_{\bar{z}}}{p_{\bar{z}}} = 0 \rightarrow \hat{p}_{\bar{z}} = \frac{x_{\bar{z}}}{n_{\bar{z}}}. \tag{1.7}$$

Razão verossimilhança - Modelo Poisson

A razão de verossimilhança para o modelo Poisson será, portanto, dada por:

$$\lambda = \frac{\mathcal{L}}{\mathcal{L}_0} = \frac{e^{-\sum_{i\in z} n_i p_z} \prod_{i\in z} n_i^{x_i} p_z^{\sum_{i\in z} x_i}}{\prod_{i\in z}^{k} (x_i!)} \times \frac{e^{-\sum_{i\notin z} n_i p_{\bar{z}}} \prod_{i\notin z} n_i^{x_i} p_{\bar{z}}^{\sum_{i\notin z} x_i}}{\prod_{i\notin z}^{k} (x_i!)}$$

$$\times \left(\frac{e^{-p_0 \sum_i n_i} p_0^{\sum_i x_i} \prod_{i=1}^{k} (n_i)^{x_i}}{\prod_{i=1}^{k} (x_i!)} \right)^{-1} = \frac{e^{-(n_z p_z + n_{\bar{z}} p_{\bar{z}} - p_0 N)} p_z^{x_z} p_{\bar{z}}^{x_{\bar{z}}}}{p_0^C} \tag{1.8}$$

1.3.3 Modelo Binomial

A distribuição Binomial também é utilizada para modelar dados de contagem, principalmente quando o processo já é naturalmente binário (estudos do tipo caso-controle). Portanto, o número de casos x_i pode ser modelado assumindo uma distribuição Binomial:

$$x_i \sim Bin\left(n_i, p_i\right).$$

Sob a hipótese nula da estatística *scan*, tem-se que $p_i = p_0$, logo a função de verossimilhança é dada por :

$$\mathcal{L}_0(p_0, \mathbf{x}) = \left[\prod_{i=1}^{k} \binom{n_i}{x_i} p_0^{x_i} (1-p_0)^{n_i - x_i} \right] = \left[\prod_{i=1}^{k} \binom{n_i}{x_i} \right] p_0^C (1-p_0)^{N-C}. \tag{1.9}$$

Aplicando logaritmo na Equação 1.9 tem-se a log-verossimilhança:

$$\ell_0(p_0, \mathbf{x}) = \log\left(\mathcal{L}_0\right) = \sum_{i=1}^{k} \log \binom{n_i}{x_i} \log n_i + C \log p_0 + (N-C) \log\left(1-p_0\right) \tag{1.10}$$

Derivando a Equação 1.10 em relação a p_0 e igualando a zero, obtêm-se o estimador de máxima verossimilhança para p_0.

$$\frac{\partial \ell_0}{\partial p_0} = \frac{C}{p_0} - \frac{(N-C)}{(1-p_0)} = 0 \rightarrow \hat{p}_0 = \frac{C}{N}, \tag{1.11}$$

Note que o estimador para p_0 na Equação 1.11 é o mesmo obtido para o modelo Poisson.

Sob a hipótese alternativa H_a temos:

$$\begin{cases} p_i = p_z & se \quad i \in z \\ p_i = p_{\bar{z}} < p_z & se \quad i \notin z \end{cases}$$

A função de verossimilhança sob H_a é dada por:

$$\begin{aligned} \mathcal{L}(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x}) = & \left[\prod_{i \in z}^{k} \binom{n_i}{x_i}\right] p_z^{\sum_{i \in z} x_i} (1-p_z)^{\sum_{i \in z} n_i - \sum_{i \in z} x_i} \\ & \times \left[\prod_{i \notin z}^{k} \binom{n_i}{x_i}\right] p_{\bar{z}}^{\sum_{i \notin z} x_i} (1-p_{\bar{z}})^{\sum_{i \notin z} n_i - \sum_{i \notin z} x_i}. \end{aligned} \tag{1.12}$$

Aplicando o logaritmo na Equação 1.12, obtém-se a função log-verossimilhança:

$$\begin{aligned} \ell(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x}) = \log(\mathcal{L}) = & \sum_{i \in z} \log \binom{n_i}{x_i} + x_z \log p_z + (n_z - x_z) \log(1-p_z) \\ & + \sum_{i \notin z} \log \binom{n_i}{x_i} + x_{\bar{z}} \log p_{\bar{z}} + (n_{\bar{z}} - x_{\bar{z}}) \log(1-p_{\bar{z}}). \end{aligned} \tag{1.13}$$

Derivando a Equação 1.13 e igualando a zero são obtidos os valores de $p_{\bar{z}}$ e p_z que maximizam a função log-verossimilhança:

$$\frac{\partial \ell}{\partial p_z} = \frac{x_z}{p_z} - \frac{(n_z - x_z)}{(1-p_z)} = 0 \rightarrow \hat{p}_z = \frac{x_z}{n_z}, \tag{1.14}$$

$$\frac{\partial \ell}{\partial p_{\bar{z}}} = \frac{x_{\bar{z}}}{p_{\bar{z}}} - \frac{(n_{\bar{z}} - x_{\bar{z}})}{(1-p_{\bar{z}})} = 0 \rightarrow \hat{p}_{\bar{z}} = \frac{x_{\bar{z}}}{n_{\bar{z}}}. \tag{1.15}$$

Razão de verossimilhança para o modelo Binomial

Semelhante ao modelo Poisson, a razão de verossimilhança para o modelo binomial é definida na Equação 1.16:

$$\lambda = \frac{\mathcal{L}}{\mathcal{L}_0} = \left[\prod_{i\in z}^{k}\binom{n_i}{x_i}\right] p_z^{\sum_{i\in z} x_i}(1-p_z)^{\sum_{i\in z} n_i - \sum_{i\in z} x_i} \times \left[\prod_{i\notin z}^{k}\binom{n_i}{x_i}\right] p_{\bar{z}}^{\sum_{i\notin z} x_i}(1-p_{\bar{z}})^{\sum_{i\notin z} n_i - \sum_{i\notin z} x_i} \times \frac{1}{\prod_{i=1}^{k}\binom{n_i}{x_i} p_0^{\sum_i x_i}(1-p_0)^{\sum_i n_i - \sum_i x_i}} = \frac{p_z^{x_z}(1-p_z)^{n_z-x_z}\, p_{\bar{z}}^{x_{\bar{z}}}(1-p_{\bar{z}})^{n_{\bar{z}}-x_{\bar{z}}}}{p_0^C(1-p_0)^{N-C}}. \tag{1.16}$$

1.3.4 Teste da Razão de Verossimilhança: *Cluster* mais verossímil

Após a escolha da distribuição utilizada para a modelagem do número de casos x_i, é realizado o processo de identificar *clusters* candidatos. O *cluster* mais verossímil é a zona z a que corresponde o maior valor da função de verossimilhança entre as zonas candidatas. Portanto, é realizado o teste da razão de verossimilhança para os modelos Poisson e Binomial para cada zona z, conforme Equação (1.17).

$$\lambda_z = \frac{\sup_{p_z>p_{\bar{z}}}\mathcal{L}(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x})}{\sup_{p_z=p_0}\mathcal{L}(p_0, p_z, \mathbf{z}, \mathbf{x})} = \frac{\mathcal{L}(\hat{p}_{\bar{z}}, \hat{p}_z, \mathbf{z}, \mathbf{x})}{\mathcal{L}(\hat{p}_0, \mathbf{x})} = \frac{e^{-x_z}\left(\frac{x_z}{n_z}\right)^{x_z} e^{-x_{\bar{z}}}\left(\frac{x_{\bar{z}}}{n_{\bar{z}}}\right)^{x_{\bar{z}}}}{e^{-C}\left(\frac{C}{N}\right)^{C}} = \left(\frac{x_z/n_z}{C/N}\right)^{x_z} \times \left(\frac{x_{\bar{z}}/n_{\bar{z}}}{C/N}\right)^{x_{\bar{z}}} \times I(x_z/n_z > x_{\bar{z}}/n_{\bar{z}}). \tag{1.17}$$

em que *I(.)* é a função indicadora. Portanto, a estatística *Scan* é definida como:

$$T = \sup_z \lambda_z = \frac{\sup_{p_z>p_{\bar{z}}}\mathcal{L}(p_{\bar{z}}, p_z, \mathbf{z}, \mathbf{x})}{\sup_{p_z=p_0}\mathcal{L}(p_0, p_z, \mathbf{z}, \mathbf{x})}. \tag{1.18}$$

Observa-se na Equação (1.18) que esta razão pode crescer rapidamente. Portanto, pode-se utilizar a função logarítmica, que é estritamente crescente. Logo, o valor que maximiza λ_z também maximiza $\log\lambda_z$, (Kulldorff, 1997)[p. 1485:1487].

O *cluster* mais verossímil é dado por um conjunto de regiões - zonas z - com o maior valor de λ_z em relação a todas as zonas candidatas. Logo, para encontrar o *cluster* mais verossímil define-se um conjunto de zonas candidatas Z, de forma que para cada elemento $z \in Z$ é obtido um valor de λ_z.

Uma maneira de construir um conjunto de zonas candidatas Z é por meio de janelas circulares, que podem conter diferentes configurações de áreas vizinhas. Uma área

é considerada dentro da janela circular se o seu centroide está contido dentro dessa janela. Na Seção 1.3.5 é apresentada uma forma de se obter as janelas e suas respectivas zonas (Fernandes, 2015).

1.3.5 Identificando zonas Candidatas: Matriz de distâncias

Considere um mapa M dividido em k regiões e que para cada uma dessas regiões têm-se as coordenadas (x_i, y_i), $i = 1, 2, \ldots k$, que representam os centroides dessas regiões. O primeiro passo do algoritmo iterativo é encontrar as zonas candidatas a *clusters*. A distância entre duas regiões é dada pela distâncias entre seus centroides. Logo, a distância Euclidiana para duas regiões i e j quaisquer é dada por:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{1.19}$$

As distâncias entre os centroides de cada região resulta em uma matriz **D** quadrada, simétrica com k linhas e k colunas $(k \times k)$ dada por

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1j} & \cdots & d_{1k} \\ d_{21} & 0 & \cdots & d_{2j} & \cdots & d_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \cdots & 0 & \cdots & d_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & \cdots & d_{kj} & \cdots & 0 \end{bmatrix}. \tag{1.20}$$

A construção das zonas formadas pelas janelas circulares é feita através da matriz de distâncias D.

1.3.6 Construindo os Candidatos a *Cluster*

Consideremos a primeira coluna da matriz D, dada por

$$D_{1,k} = \begin{bmatrix} 0 \\ d_{21} \\ \vdots \\ d_{i1} \\ \vdots \\ d_{k1} \end{bmatrix}.$$

Agora considere que $d_{(j),1}$ seja a distância da j-ésima região mais próxima da região 1, isto é, $d_{(k),1} > \ldots > d_{(3),1} > d_{(2),1}$. Dessa forma, tem-se o vetor coluna anterior ($D_{1,k}$) ordenado de forma crescente de distâncias (Fernandes, 2015).

$$D_{1,k} = \begin{bmatrix} 0 \\ d_{(2),1} \\ d_{(3),1} \\ \vdots \\ d_{(k),1} \end{bmatrix}.$$

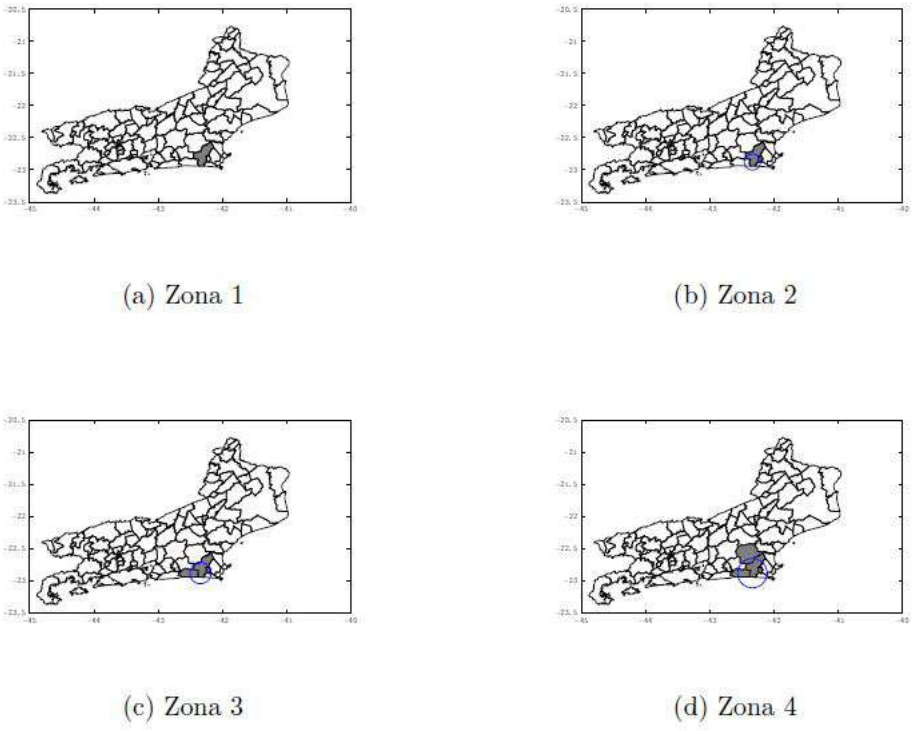
A primeira zona z_1 selecionada é formada apenas pela região 1, logo, $z_1 = \{1\}$. Então, calcula-se o valor da estatística de razão de verosimilhança λ_{z_1} conforme Equação 1.18. A segunda zona z_2 é formada pela região $\{1\}$ e a região mais próxima de 1, ou seja, a região associada à distância $d_{(2),1}$. Portanto, a segunda zona é denotada por $z_2 = \{1, (2)\}$, com os dados dessa zona é calculado a estatística λ_{z_2}. A estatística T é dada pelo maior valor observado de λ_z, logo, se λ_{z_2} é maior que λ_{z_1} esta zona é considerada temporariamente um candidato a *cluster* espacial. Contudo, apenas poderá ser considerada um *cluster* detectado depois da verificação da significância estatística.

O processo iterativo é repetido, acrescentando em cada passo uma região vizinha segundo a distância, e calculando uma estatística λ_{z_i}, até atingir o tamanho máximo da janela. Uma regra natural indicada na literatura é de que o processo seja repetido até atingir uma porcentagem máxima da população total, por exemplo, 50% do tamanho máximo da população (Naus, 1965b).

Nos primeiros métodos para detecção de *clusters* espaciais, como o *GAM* de Openshaw et al. (1988), o raio (tamanho das zonas candidatas) era fixado e varria áreas de interesse. A estatística *scan* de Kulldorff possibilita que o tamanho das zonas candidatas variem, mas não é interessante que o raio seja muito grande, incluindo quase toda população. Isso porque, caso seja detectado um *cluster* que abranja mais do que 50% da população, considera-se que, na verdade, o seu complementar é que é um *cluster*, mas de baixa incidência.

O *cluster* mais verossímil, será formado pelo conjunto de regiões (a zona z^*) que apresentar o maior valor de λ_z.

Na Figura 1.1 é ilustrado o processo de construção das zonas candidatas.



(a) Zona 1 (b) Zona 2

(c) Zona 3 (d) Zona 4

Figura 1.1: Ilustração do processo de construção de zonas Circulares
Fonte: Adaptado de (Fernandes, 2015) (Fernandes, 2015)

1.3.7 Verificação da significância do *cluster*

Conforme descrito na Seção 1.3.6, a primeira parte do algoritmo é executada para a identificação do *cluster* mais verossímil com base nos dados observados. A zona $z*$, correspondente ao valor máximo da razão de verossimilhança no conjunto de todas as janelas circulares com tamanho variando até o percentual máximo definido a priori centrado nos centroides de cada região, será considerada uma zona candidata a *cluster*, porém, apenas poderá ser dado como um *cluster* após verificar sua significância estatística.

A distribuição da estatística *Scan* de Kulldorff não possui forma analítica fechada, ou seja, a distribuição exata da estatística T não é conhecida. Uma solução apresentada por Kulldorff e Nagarwalla (1995) recorre ao processo de simulação de Monte Carlo para obter a distribuição empírica da estatística T, permitindo o cálculo do *p-valor*. Esta solução foi apontada por Dwass (1957).

O processo de simulação de Monte Carlo consiste em gerar J réplicas do mapa M original sob a hipótese nula. Assim, o número total de casos C é fixado e o número de casos em cada região é obtido aleatoriamente dada a hipótese nula, ou seja, a probabilidade de um indivíduo ser um caso é a mesma em qualquer região. O processo é repetido J vezes e em cada passo é obtido o valor da estatística T^*. Portanto, ao fim do processo tem-se uma distribuição empírica formada por uma amostra de tamanho J da estatística

T^* sob H_0. Em seguida compara-se o valor de T observado com a distribuição empírica, para fins de obter o p-valor e seu valor crítico.

1.3.8 Resumo do Algoritmo *Scan* Circular

Em resumo, algoritmo *Scan* de Kulldorff pode ser apresentado por:

1. Obter a matriz de distâncias entre os centroides das regiões em estudo, conforme 1.3.5.

2. Ordenar a i-ésima coluna da matriz de distâncias em ordem crescente, obtendo a lista de regiões em ordem crescente de distância em relação à região i.

3. Encontrar as zonas candidatas e obter o valor da estatística λ_z conforme 1.3.6. Registrar a zona com o valor máximo de λ_z até o momento, limitando-se à proporção máxima da população prefinida.

4. Repetir os passos 2 a 3, para cada região $i = 1, \ldots, n$ do mapa em estudo, e armazenar o valor de T.

5. Testar a significância do *cluster* identificado com base em simulação de Monte Carlo:

   - Criar J réplicas do mapa original, onde em cada réplica o número de casos C seja igual ao do conjunto de dados original. Dessa forma, pode-se distribuir aleatoriamente os C casos com base em uma distribuição multinomial com parâmetros C (número total de casos) e n_i/N (proporção de casos em cada região).

   - Para cada réplica do mapa original obter o valor da estatística T, conforme passos 3 e 4.

6. Rejeitar a hipótese nula, isto é, de ausência de *cluster* espacial ao nível $\alpha = 5\%$ se $T > P_{95}$, onde T é o valor observado da estatística de teste obtido por meio dos dados reais (ver seção 1.3.4), e P_{95} é o percentil 95% da distribuição empírica de T sob H_0.

   A Figura 1.2 ilustra o processo de verificação da significância de um *cluster* espacial com base em simulação de Monte Carlo.

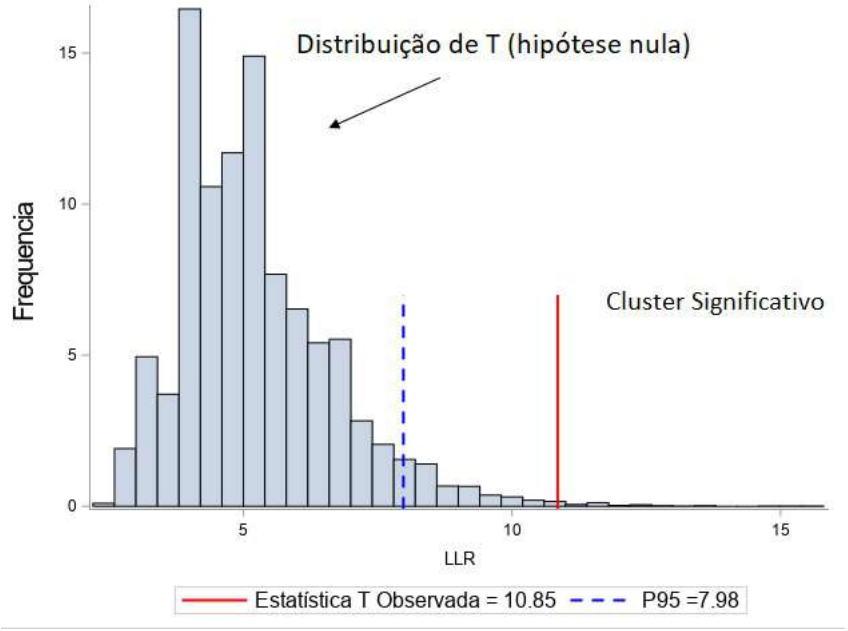


Figura 1.2: Distribuição empírica da estatística T obtida via Simulação de Monte de Carlo

Com base na Figura 1.2, note que dado um nível de significância $\alpha = 5\%$, se a estatística do teste T observada (conjunto de dados real) estiver entre os 5% maiores valores da distribuição da estatística T empírica, isto é, valores baseados no conjunto dados gerados aleatoriamente, deve-se rejeitar a hipótese nula.

Conforme descrito por Dwass (1957), o p-valor pode ser obtido com base no posto ocupado pela estatística T observada em relação à distribuição empírica sob H_0. Portanto, considerando J réplicas do mapa original, se para $(j-1)$ dessas replicações a estatística de teste for maior ou igual à estatística de teste observada (dos dados reais), então o p-valor para a estatística T será dado por $j/(J+1)$.

1.4 Estatística de Varredura baseada em árvore

Um conjunto de dados hierárquico é definido por uma sequência natural de atributos aninhados, onde as variáveis que representam diversas características podem estar ligadas a outras variáveis dentro do mesmo nível hierárquico e em outros níveis hierárquicos.

Os dados hierárquicos podem ser representados no formato de árvore, como por exemplo, na área da saúde, uma doença específica pode ter vários subtipos associados, e esses subtipos podem ser divididos em especificações mais gerais, representando os diversos níveis da árvore. Na área da educação, tem-se escolas, turmas, alunos constituindo uma sequência de agrupamento natural. Na área financeira, os estabelecimentos comerciais podem ser divididos por área geográfica. Dentro das instituições, as profissões são naturalmente organizadas segundo área de trabalho. Essas áreas de trabalho podem ser classificadas em subtipos (área de trabalho mais gerais) aninhados, que compõem os ramos ou todos os possíveis níveis da árvore (Prates, 2008).

Kulldorff et al. (2003a) abordaram a técnica de varredura baseada em árvore como uma forma de identificar *clusters* em árvores hierárquicas e aplicaram-na em um estudo de vigilância em saúde com objetivo de identificar sub-conjuntos de profissões que possuíam evidências incomuns de taxas de mortes por doenças relacionadas à profissão. Porém, não havia conhecimento *a priori* sobre qual grupo de profissões ou profissão específica poderia estar relacionada com um maior risco de morte. O interesse era em identificar o *cluster* em particular, no qual o número de casos era muito frequente. Em estudos com dados hierárquicos, o número de combinações possíveis entre todas as hierarquias pode ser elevado, o que torna a análise mais complicada devido ao problema de múltiplos testes. E nem todas as combinações são de interesse. Logo, ao invés de procurarem entre todas as possíveis combinações, considerara-se uma classe menor de *clusters* possíveis.

A construção de uma variável em forma de uma árvore hierárquica se inicia pela definição das folhas. As folhas armazenam a informação para um evento específico. Portanto, as folhas armazenam toda a informação sobre o conjunto de dados. Por exemplo, o número total de casos e demais características (ex: número de mortes, gênero, idade, profissão). Cada uma das folhas pertence a um ramo e no final de cada ramo têm-se os nós. As folhas relacionadas a um ramo são conectadas por meio de um mesmo nó. Dessa forma, cada nó nasce de um ramo acima de uma determinada folha e se conecta com um nó de nível superior. O processo é repetido até chegar ao nó de nível mais alto, que é denominado raiz (R).

A Figura abaixo ilustra um exemplo hipotético de árvore hierárquica com três níveis.

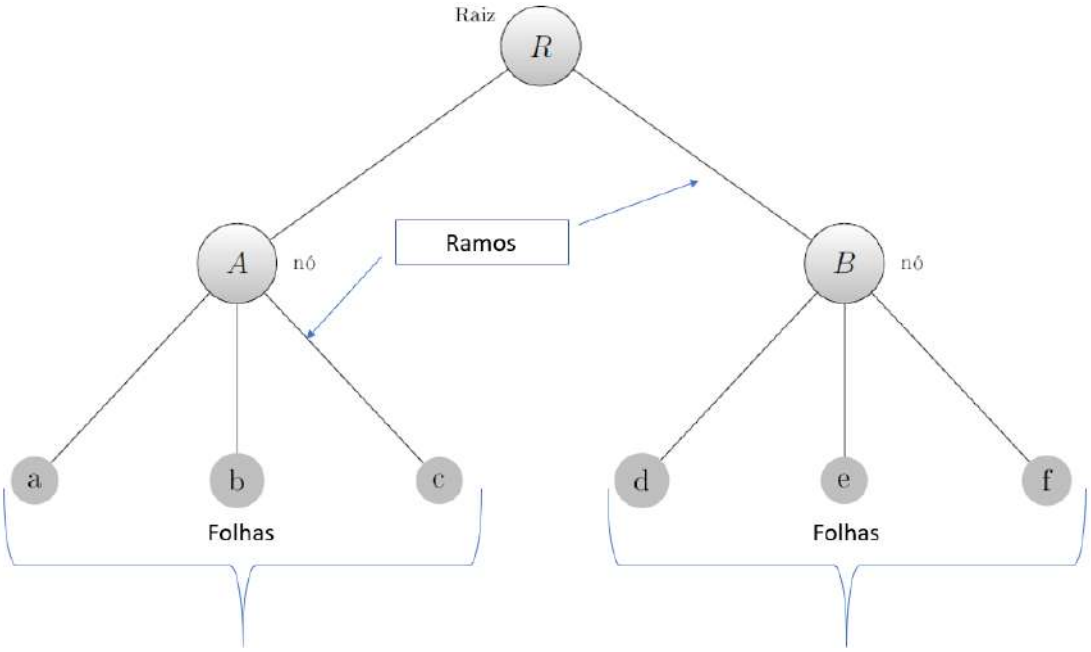


Figura 1.3: Exemplo de uma árvore hierárquica (3 níveis), com definição de ramos, folhas e nós.

No contexto da ciência da computação, a figura 1.3 é dada como uma representação gráfica de um banco de dados em formato hierárquico. A nomenclatura desse método é derivada da teoria dos grafos e da ciência da computação (Kulldorff (2018), pgs 8-9).

- **Árvore**. Uma variável hierárquica. Na terminologia matemática, pode ser representada por um grafo simples que consiste em nós (ou vértices) e arestas (ou arcos).
- **Nó**. Corresponde a cada vértice na árvore.
- **Aresta**. Uma linha que conecta pelo menos dois nós em uma árvore.
- **Pai**. Um nó pai é o nó que está posicionado imediatamente acima de outro na árvore, conectado por uma aresta.
- **Filho**. Um nó filho é um nó que está posicionado imediatamente abaixo de outro na árvore, conectado por uma aresta. Dessa forma, um nó pode ser pai de um nó enquanto é filho de outro nó.
- **Folha**. Nó que não possui nenhum filho.
- **Raiz**. Nó que não possui pai, isto é, representa o maior nível da árvore ou primeiro nível.
- **Corte**. Poda que pode ser feita em qualquer uma das arestas, imediatamente acima de um nó. O corte é identificado pelo nome do nó que está imediatamente abaixo do corte

- **Ramo**. O conjunto de nós definidos por um corte, ou seja, representa os nós que foram separados da árvore. Inclui o nó imediatamente abaixo do corte (nó que identifica o corte) e todos os seus descendentes.

- **Cluster**. Um ramo, coleção de nós ou folhas, em que o número observado de casos é significativamente maior que o esperado.

Para a estatística de varredura baseada em árvore, todos os dados de um evento que carrega uma informação específica estão nas folhas. Cada folha corresponde às informações sobre o número total de indivíduos com essa característica específica (Ex: número de mortes, gênero, idade).

A janela de varredura para a estatística de varredura baseada em árvore (*tree scan*) é definida pelos cortes que definem os ramos da árvore. Todos os possíveis cortes são considerados, e são denominados de cortes simples. Na estatística de varredura espacial circular, define-se um cluster como um conjunto de regiões dentro das quais a probabilidade de um individuo vir a ser um caso é maior se comparada com o esperado sob a hipótese de aleatoriedade. O mesmo ocorre na estatística de varredura baseada em árvore, mas neste caso, as zonas são os ramos da árvore. Portanto o método se baseia em identificar um ramo da árvore cujas folhas podem conter uma quantidade de eventos maior que o esperado (Kulldorff et al., 2003a).

Analogamente à estatística de varredura espacial, em que o número de casos de um determinado evento é avaliado para cada zona e o *cluster* mais verosímil é dado pela zona cuja probabilidade de um evento vir a ser um caso é maior se comparado às demais, na estatística de varredura baseada em árvore o procedimento é similar. Porém, as zonas candidatas são os ramos, logo o número de casos de um dado evento é avaliado dentro e fora do ramo. Esse procedimento de estimar o risco dentro e fora do ramo também é baseado na razão de verossimilhança.

Mais especificamente, considere um conjunto de dados classificados em uma estrutura de árvore hierárquica, conforme a Figura 1.3. Para cada folha i tem-se o número observado de casos x_i. O número de casos x_i pode ser modelado por uma distribuição Poisson com média $n_i p_i$, em que n_i representa a população na folha e p_i é a probabilidade de cada indivíduo na folha i ser um caso. No contexto de uma árvore hierárquica o interesse é em identificar em qual ramo da árvore a incidência de casos é maior em relação aos demais. No estudo de vigilância em saúde apresentado por (Kulldorff et al., 2003a), não se tem o interesse no número total de mortes em cada folha e sim na sua distribuição relativa entre as diferentes folhas. Portanto, a análise está condicionada ao número de casos observados, o que leva a uma distribuição multinomial.

Dessa forma estamos interessados no seguinte teste de hipóteses:

$$\begin{cases} H_0 : p_g = p_{\bar{g}} = p_0, \forall g \in G \\ H_a : \text{Existe um ramo } g \text{ tal que } p_g > p_{\bar{g}} \end{cases}$$

sendo p_g a probabilidade de caso no ramo g e $p_{\bar{g}}$ a probabilidade de caso fora do ramo g.

Em outras palavras a hipótese nula é de que, para todas as folhas da árvore, a probabilidade de um indivíduo ser um caso é igual. Já a hipótese alternativa é de que existe um conjunto de nós, denominado ramo g, para o qual $p_i = p_g$ se $i \in g$ e $p_i = p_{\bar{g}} < p_g$ se $i \notin g$. Por exemplo, no contexto sobre dados de mortalidade, a hipótese nula equivale a afirmar que a probabilidade de morte é a mesma em qualquer ramo da árvore, enquanto a hipótese alternativa é de que existe algum ramo da árvore para o qual a probabilidade de morte é maior em relação aos demais.

Considere x_i e n_i o número de casos e a população na i-ésima folha, respectivamente. Sejam ainda $C = \sum_{i=1}^{m} x_i$ o número total de casos e $N = \sum_{i=1}^{m} n_i$ a população total da árvore.

Seja g um ramo da árvore e G o conjunto de todos os ramos. Um *cluster* em uma árvore hierárquica é formado por qualquer ramo g em que a probabilidade p_g de um evento ocorrer é maior que nos demais.

O processo de modelagem do número de casos em cada zona com base na distribuição Poisson visto na Seção 1.3.2 e definido por (Kulldorff, 1997) pode ser replicado no contexto de árvore. Porém, ao invés de um conjunto de regiões, consideramos um grupo g de nós. O *cluster* mais verossímil é dado pelo ramo g da árvore em que a incidência de casos é significativamente maior que nos demais. Dessa forma, o cálculo da razão de verosimilhança λ_g com base no modelo Poisson é realizado para cada ramo da árvore.

A razão de verossimilhança é definida na Equação 1.21, em que $x_g = \sum_{i \in g} x_i$ é o número total de casos no ramo g, $x_{\bar{g}} = \sum_{i \notin g} x_i$ é o número de casos fora do ramo g, $n_g = \sum_{i \in g} n_i$ é população do ramo g e $n_{\bar{g}} = \sum_{i \notin g} n_i$ é população fora do ramo g.

$$\lambda_g = \frac{\sup_{p_g > p_{\bar{g}}} \mathcal{L}\left(p_{\bar{g}}, p_g, \mathbf{g}, \mathbf{x}\right)}{\sup_{p_g = p_0} \mathcal{L}\left(p_0, p_g, \mathbf{g}, \mathbf{x}\right)} = \frac{\mathcal{L}\left(\hat{p}_{\bar{g}}, \hat{p}_g, \mathbf{g}, \mathbf{x}\right)}{\mathcal{L}\left(\hat{p}_0, \mathbf{x}\right)} = \frac{e^{-x_g}\left(\frac{x_g}{n_g}\right)^{x_g} e^{-x_{\bar{g}}}\left(\frac{x_{\bar{g}}}{n_{\bar{g}}}\right)^{x_{\bar{g}}}}{e^{-C}\left(\frac{C}{N}\right)^{C}} \times I(x_g/\mu_g > x_{\bar{g}}/\mu_{\bar{g}}). \tag{1.21}$$

Portanto, a estatística de varredura baseada em árvore é definida por:

$$T = \sup_g \lambda_g = \frac{\sup_{p_g > p_{\bar{g}}} \mathcal{L}\left(p_{\bar{g}}, p_g, \mathbf{g}, \mathbf{x}\right)}{\sup_{p_g = p_0} \mathcal{L}\left(p_0, p_g, \mathbf{g}, \mathbf{x}\right)}. \tag{1.22}$$

As definições descritas acima podem ser resumidas da seguinte forma;

- $x_g = \sum_{i \in g} x_i$ é o número total de casos no ramo g;

- $x_{\bar{g}} = \sum_{i \notin g} x_i = C - x_g$ é o número de casos fora do ramo g ;
- $n_g = \sum_{i \in g} n_i$ é a população em risco do ramo g;
- $n_{\bar{g}} = \sum_{i \notin g} n_i = N - n_g$ é população em risco fora do ramo g.

1.4.1 Identificando candidatos a *cluster* em uma estatística de varredura baseada em árvore

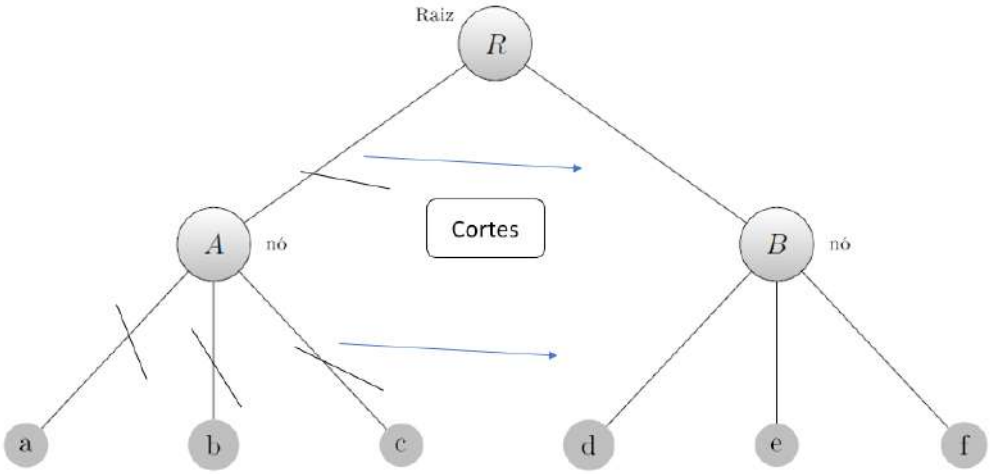


Figura 1.4: Identificando cortes na árvore

A Figura 1.4 ilustra uma árvore hipotética. Todos os cortes possíveis aqui apresentados são denominados cortes simples e o processo iterativo em busca do ramo mais verossímil pode-se iniciar realizando um corte em qualquer parte da árvore, pois não se tem conhecimento a *priori* sobre os grupos.

Suponha que o primeiro corte seja em A. O ramo correspondente será dado, então, por $g_1 = [A, a, b, c]$. Portanto, tem-se que $\bar{g}_1 = [R, B, d, e, f]$. Em seguida, pode-se realizar um corte no ramo a, então tem-se $g_2 = [a]$ e todo o resto da árvore forma o conjunto fora do corte g_2, ou seja, $\bar{g}_2 = [R, A, b, c, B, d, e, f]$. Esse processo é repetido sucessivamente até que todos os possíveis cortes sejam avaliados, escolhendo-se ao fim o corte com o maior valor de λ_g. Esse será considerado o *cluster* detectado.

Nesse trabalho, apenas os cortes simples serão considerados. Por exemplo, o corte no nó A gera o ramo que contém, necessariamente, o próprio A e todos os seus nós descendentes (a, b e c). Porém, a complexidade dos cortes pode ser maior, visto que também é possível analisar os cortes combinatórios e ordinais (Kulldorff et al., 2003a).

A estatística de varredura baseada em árvore também pode ser utilizada considerando uma extensão para cortes complexos em uma árvore que possui um nó com pelos menos dois nós filhos. Um corte simples é dado por uma poda ou qualquer corte

que é realizado imediatamente acima de um nó (será considerado o nó com todos os seus descendentes), ou por um corte, em qualquer um dos ramos descendentes do nó, isto é, uma folha apenas.

Suponha que exista um nó identificado por A, que possui quatros nós folhas: a, b, c, d. Dessa forma, ao invés de considerar-se apenas os quatros cortes simples $g_1 = [a]$, $g_2 = [b]$, $g_3 = [c]$ e $g_4 = [d]$, ou seja, todas as folhas separadamente, pode-se considerar todas as possíveis combinações entre esses nós. Assim, teríamos uma sequência de cortes adicionais, dada pelos seguintes ramos: $g_5 = [a, b]$, $g_6 = [a, c]$, $g_7 = [a, d]$, $g_8 = [b, c]$, $g_9 = [b, d]$, $g_{10} = [c, d]$, $g_{11} = [a, b, c]$, $g_{12} = [a, b, d]$, $g_{13} = [a, c, d]$ e $g_{14} = [b, c, d]$. Esses tipos de cortes denominam-se os cortes combinatórios (Kulldorff et al., 2003a).

A ideia desses cortes é possibilitar avaliar se as possíveis combinações do nó A podem estar relacionadas uns aos outros, mas com aspectos diferentes. Por exemplo, se o nó A representa doenças do sistema circulatório, onde as três primeiras folhas (três cortes simples) são: Hipertensão essencial (primária), Doença cardíaca hipertensiva e Doença renal hipertensiva respectivamente. Dessa forma, utilizando os cortes combinatórios é analisado a relação entre os dados de: (Hipertensão essencial (primária) + Doença cardíaca hipertensiva), (Hipertensão essencial (primária) + Doença renal hipertensiva) e (Doença cardíaca hipertensiva + Doença renal hipertensiva).

O terceiro corte é o ordinal. Nesse caso, as combinações consideram a ordem da sequencia hierárquica para montar os ramos relacionados. Por exemplo, considere uma lista de profissões naturalmente definida com base na idade de alunos: professores do jardim de infância, ensino fundamental, ensino médio e professores universitários. Nesse cenário, não faz muito sentido considerar uma ramo formado por professores do jardim de infância e professores universitários. Portanto, além dos cortes simples a, b, c e d, tem-se os cortes $[a, b]$, $[b, c]$, $[c, d]$, $[a, b, c]$, $[b, c, d]$, mas o corte $[a, d]$ não é considerado Kulldorff et al. (2003a).

Os cortes simples são subconjuntos dos cortes combinatórios e ordinais. Dessa forma, em uma árvore na qual um nó pai só possui apenas dois nós filhos, todos os tipos de cortes serão iguais (Kulldorff et al., 2003a).

Todos os possíveis cortes na árvore formam um determinado grupo de ramos G e para cada um dos cortes, por sua vez, é obtido o valor de λ_g, sendo que a estatística do teste será dada pelo maior valor de λ_g observado. A significância estatística do *cluster* é avaliada usando simulações de Monte Carlo sob a hipótese nula, pois a estatística do teste não tem expressão analítica fechada. Esse processo é equivalente ao descrito na Seção 1.3.7 para a estatística de varredura espacial. Um grande número de réplicas da árvore original será gerado, sendo que o número de casos em cada folha é gerado com base em uma distribuição multinomial. Cada uma das árvore geradas são analisadas da mesma maneira que para os dados reais. A estatística do teste observada é comparada com a distribuição empírica da estatística T. O *p-valor* da estatística T é dado por $r/(j+1)$. Onde r é o *rank*, o valor correspondente à posição da estatística de teste observada em

relação à distribuição empírica da estatística T, e J é o número de simulações de Monte Carlo.

1.4.2 Resumo do algoritmo *Tree Scan*

Em resumo, o algoritmo *tree Scan* de Kulldorff, pode ser apresentado pela sequência de passos abaixo:

1. Varrer a árvore considerando todos os possíveis cortes.
2. Para cada corte, calcular a razão de verossimilhança λ_g.
3. O valor da estatística de teste T é dado pelo maior valor de λ_g.
4. Testar a significância do *cluster* identificado (ramo da árvore correspondente ao maior valor de λ_g), com base em simulação de Monte Carlo sob H_0, condicionado ao número total de casos.
    - Criar J réplicas da árvore original (conjunto de nós, ramos e folhas), sendo que em cada réplica o número de casos C corresponde à mesma quantidade de casos dos dados originais. Logo, pode-se distribuir aleatoriamente C casos com base em uma distribuição multinomial com parâmetros C (número total de casos) e n_i/N (proporção de população em cada folha).
    - Para cada réplica da árvore, obter o valor da estatística T, conforme passos 1 a 3.
5. Compare o valor de T obtido para o conjunto de dados real com os valores obtidos para as árvores geradas com base em simulação. Não rejeitar a hipótese nula significa que não existe quaisquer ramos na árvore em que o número de casos é mais frequente se comparado ao esperado (ausência de *cluster*). Ao nível $\alpha = 5\%$ se $T > P_{95}$ rejeita-se a hipótese nula, onde T é o valor observado da estatística do teste obtido por meio dos dados reais e P_{95} é o percentil da distribuição empírica de T sob H_0.

1.5 Estatística de Varredura Árvore-Espacial

Propomos neste trabalho uma estatística de varredura árvore-espacial, incorporando uma estrutura espacial à estatística de varredura baseada em árvore. Como motivação, consideremos o sistema de Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde-Décima Revisão, conhecido pela sigla **CID-10**. A CID-10 é fruto de um esforço internacional para padronizar, catalogar e classificar doenças conforme nomenclatura internacional de doenças estabelecida pela Organização Mundial de

Saúde (trecho disponível em*http://datasus.saude.gov.br/sistemas-e-aplicativos/cadastros-nacionais/cid-10*). A listagem de doenças é dividida por capítulos, os capítulos vinculados a códigos divididos por categorias, as categorias em subcategorias, e estas contém os códigos das doenças específicas. A CID-10 constitui um banco de dados em forma hierárquica. Portanto, a estatística de varredura baseada em árvore pode ser utilizada como uma forma de identificar algum ramo nessa árvore em que a incidência de casos é significativamente maior que nos demais. Mas poderíamos estar interessados também em incorporar a informação espacial (local da ocorrência do agravo ou contaminação por uma determinada doença). Incorporando a informação espacial, estaríamos interessados em detectar a anomalia em algum ramo dessa árvore em algum lugar do espaço.

Considere uma mapa M ou uma área geográfica subdividida em k regiões, em que $n_i = n(i)$, com $i \in \{1, 2, 3, \cdots k\}$ é a população exposta na i-ésima região e $N = \sum_{i=1} n(i)$ é a população total em risco. Seja $x_g(i)$ o número de casos de algum evento de risco na *i-ésima* região no ramo g. Logo, $C_g = \sum_{i=1}^{k} x_g(i)$ corresponde ao número total de casos no ramo g em todo o mapa. Para uma determinada zona espacial z, consideramos $c_g(z) = \sum_{i \in z} x_g(i)$ o número de casos no ramo g na zona z, $c_g(\bar{z}) = \sum_{i \notin z} x_g(i)$ o número de casos no ramo g fora da zona z, e $n_z = \sum_{i \in z} n_i$ e $n_{\bar{z}} = \sum_{i \notin z} n_i$ são a população em risco dentro e fora da zona z, respectivamente .

As definições apresentadas acima, podem ser resumidas da seguinte forma:

- $c_g(z) = \sum_{i \in z} x_g(i)$ é o número de casos no ramo g e na zona z;
- $c_g(\bar{z}) = \sum_{i \notin z} x_g(i) = C_g - C_g(z)$ é o número de casos do ramo g fora da zona z ;
- $n_z = \sum_{i \in z} n_i$ é a população em risco na zona z;
- $n_{\bar{z}} = \sum_{i \notin z} n_i = N - n_z$ é população em risco fora da zona z.

A estatística de varredura baseada em árvore será construída sob as seguintes hipóteses;

$$\begin{cases} H_0 : p_g(z) = p_g(\bar{z}) = p_0 \ \forall g \text{ e } \forall z \\ H_a : \text{Existem } z \text{ e } g \text{ tais que } p_g(z) > p_g(\bar{z}), \end{cases}$$

em que $p_g(z)$ é a probabilidade de que um indivíduo na zona z venha a ser um caso relacionado ao ramo g.

As hipóteses apresentadas acima podem ser reescritas da seguinte forma:

- H_0: A probabilidade de algum indivíduo vir a ser um caso em um ramo g é a mesma em qualquer lugar do mapa.

- H_a: Existem pelo menos um ramo g e pelo menos uma zona z tais que, se um indivíduo está em z, então a probabilidade de que ele venha a ser um caso relacionado ao ramo g é maior do que se estivesse fora de z.

Assim, caso H_0 seja rejeitada para um certo par formado pela zona z e por um ramo g, então (z, g) é um *cluster* árvore-espacial.

Dessa forma, um *cluster* árvore espacial pode ser definido por um par formado por um conjunto de regiões(zonas) do mapa e um conjunto de ramos da árvores (nós que foram separados da árvore):

$$\{Z = z_1, \cdots z_2, z_k, G\},$$

Onde Z representa um conjunto de zonas, e G é um conjunto de ramos da árvore.

O número de casos $x_g(i)$ do ramo g na *i-ésima* região pode ser modelado como uma variável com distribuição Poisson com média $n_i p_g(i)$ que representa por exemplo a proporção de agravos por doenças do ramo g na i-ésima região em relação à população da região i.

O algoritmo de varredura árvore-espacial será construído por meio da interação entre o algoritmo de varredura espacial circular e o baseado em árvore. A primeira etapa do método de varredura árvore-espacial consiste em varrer áreas geográficas em janelas circulares limitando-se à proporção máxima da população dentro da zona z, em que, para cada conjunto de zona candidatas são identificados todos os possíveis cortes da árvore e para cada corte calcula-se a estatística de razão de verossimilhança. Após a identificação do *cluster* mais verossímil, o processo de verificação da significância é realizado por meio de simulações de Monte Carlo, num procedimento análogo ao descrito anteriormente (Dwass, 1957) e (Kulldorff, 1997).

1.5.1 Modelo Poisson para Estatística de varredura árvore espacial

Considere que $x_g(i)$, o número de casos do evento no ramo g na i-ésima região, siga distribuição Poisson, isto é

$$x_g(i) \sim Poisson\left(n_i p_g(i)\right).$$

Equivalentemente ao apresentado na Seção 1.3.2 sob a hipótese nula da estatística de varredura árvore-espacial, tem-se que $p_g(i) = p_0$, fazendo $\mu_g(i) = n_i p_g(i)$ como número esperado de casos do ramo g na zona z, a função de verossimilhança é dada por:

$$\mathcal{L}_0(p_0, \mathbf{x}) = \prod_{i=1}^{k} \frac{e^{-\mu_g(i)} \mu_g(i)^{x_g(i)}}{x_g(i)!} = \frac{e^{-\sum_i \mu_g(i)} \prod_{i=1}^{k} \mu_g(i)^{x_g(i)}}{\prod_{i=1}^{k} (x_g(i)!)}$$

$$= \frac{e^{-p_0 \sum_i n_i} p_0^{\sum_i x_g(i)} \prod_{i=1}^{k} (n_i)^{x_g(i)}}{\prod_{i=1}^{k} (x_g(i)!)} = \frac{e^{-p_0 N} p_0^{C_g} \prod_{i=1}^{k} (n_i)^{x_g(i)}}{\prod_{i=1}^{k} (x_g(i)!)}. \tag{1.23}$$

Aplicando logaritmo na Equação (1.23) tem-se a log-verossimilhança

$$\ell_0(p_0, \mathbf{x}) = \log(\mathcal{L}_0) = -p_0 N + C_g \log p_0 + \sum_{i=1}^{k} x_g(i) \log n_i - \sum_{i=1}^{k} \log(x_g(i)!). \tag{1.24}$$

Derivando a Equação 1.24 com relação a p_0 e igualando a zero, obtém-se o estimador de máxima verossimilhança para p_0.

$$\frac{\partial \ell_0}{\partial p_0} = -N + \frac{C_g}{p_0} = 0 \rightarrow \hat{p}_0 = \frac{C_g}{N}, \tag{1.25}$$

Observa-se na Equação 1.25 que C_g representa o número total de casos no ramo g, somados sobre todas as regiões.

De forma análoga, sob a hipótese alternativa H_a:

$$\begin{cases} p_g(i) = p_g(z) & se \quad i \in z \\ p_g(i) = p_g(\bar{z}) & se \quad i \notin z \end{cases}$$

A função de máxima verossimilhança sob H_a é dada por:

$$\mathcal{L}(p_g(\bar{z}), p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}) = \prod_{i \in z}^{k} \frac{e^{-p_g(z) n_i} (p_g(z) n_i)^{x_g(i)}}{x_g(i)!} \times \prod_{i \notin z}^{k} \frac{e^{-p_g(\bar{z}) n_i} (p_g(\bar{z}) n_i)^{x_g}}{x_g(i)!}$$

$$= \frac{e^{-\sum_{i \in z} n_i p_g(z)} \prod_{i \in z} n_i^{x_g(i)} p_g(z)^{\sum_{i \in z} x_g(i)}}{\prod_{i \in z}^{k} x_g(i)!} \times \frac{e^{-\sum_{i \notin z} n_i p_g(\bar{z})} \prod_{i \notin z} n_i^{x_g(i)} p_g(\bar{z})^{\sum_{i \notin z} x_g(i)}}{\prod_{i \notin z}^{k} x_g(i)!} \tag{1.26}$$

Aplicando o logaritmo na Equação 1.26, obtêm-se a função log-verossimilhança:

$$\ell_a(p_g(\bar{z}), p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}) = \log(\mathcal{L}) = -p_g(z)n_z + \left[\sum_{i\in z}^{k} x_g(i)\log n_i + x_g(z)\log p_g(z) - \sum_{i\in z}^{k}\log(x_g(i)!)\right] -p_g(\bar{z})n_{\bar{z}} + \left[\sum_{i\notin z}^{k} x_g(i)\log n_i + x_g(\bar{z})\log p_g(\bar{z}) - \sum_{i\notin z}^{k}\log(x_g(i)!)\right]. \quad (1.27)$$

Derivando a equação 1.27 e igualando a zero são encontrados os valores de $p_g(\bar{z})$ e $p_g(z)$ que maximizam a função log-verossimilhança:

$$\frac{\partial \ell}{\partial p_g(z)} = -n_z + \frac{x_g(z)}{p_g(z)} = 0 \rightarrow \hat{p}_g(z) = \frac{x_g(z)}{n_z}, \quad (1.28)$$

$$\frac{\partial \ell}{\partial p_g(\bar{z})} = -n_{\bar{z}} + \frac{x_g(\bar{z})}{p_g(\bar{z})} = 0 \rightarrow \hat{p}_g(\bar{z}) = \frac{x_g(\bar{z})}{n_{\bar{z}}}. \quad (1.29)$$

Razão verossimilhança - Modelo Poisson

A razão de verossimilhança para o modelo Poisson é definida por:

$$\lambda = \frac{\mathcal{L}}{\mathcal{L}_0} = \frac{e^{-\sum_{i\in z} n_i p_g(z)} \prod_{i\in z} n_i^{x_g(i)} p_g(z)^{\sum_{i\in z} x_g(i)}}{\prod_{i\in z}^{k}(x_g(i)!)} \times \frac{e^{-\sum_{i\notin z} n_i p_g(\bar{z})} \prod_{i\notin z} n_i^{x_g(i)} p_g(\bar{z})^{\sum_{i\notin z} x_g(i)}}{\prod_{i\notin z}^{k}(x_g(i)!)} \times \left(\frac{e^{-p_0\sum_i n_i} p_0^{\sum_i x_g(i)} \prod_{i=1}^{k}(n_i)^{x_g(i)}}{\prod_{i=1}^{k}(x_g(i)!)}\right)^{-1} = \frac{e^{-n_z p_g(z) + n_{\bar{z}} p_g(\bar{z}) - p_0 N} p_g(z)^{x_g(z)} p_g(\bar{z})^{x_g(\bar{z})}}{p_0^{C_g}}. \quad (1.30)$$

1.5.2 Encontrando o *cluster* mais verossímil

A estatística do teste da razão de verossimilhança para o modelo Poisson é dada por:

$$\lambda_g(z) = \frac{\sup_{p_g(z)>p_g(\bar{z})} \mathcal{L}\left(p_g(\bar{z}), p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}\right)}{\sup_{p_g(z)=p_0} \mathcal{L}\left(p_0, p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}\right)} = \frac{\mathcal{L}\left(\hat{p}_g(\bar{z}), \hat{p}_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}\right)}{\mathcal{L}\left(\hat{p}_0, \hat{p}_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x},\right)} =$$

$$= \frac{e^{-x_g(z)}\left(\frac{x_g(z)}{n_z}\right)^{x_g(z)} e^{-x_{g(\bar{z})}}\left(\frac{x_{g(\bar{z})}}{n_{\bar{z}}}\right)^{x_{g(\bar{z})}}}{e^{-C_g}\left(\frac{C_g}{N}\right)^{C_g}}$$

$$= \left(\frac{x_g(z)/n_z}{C_g/N}\right)^{x_g(z)} \times \left(\frac{x_{g(\bar{z})}/n_{\bar{z}}}{C_g/N}\right)^{x_{g(\bar{z})}} \times I(x_g(z)/n_z > x_g(\bar{z})/n_{\bar{z}}), \tag{1.31}$$

em que $I(.)$ é a função indicadora. Portanto, a estatística de varredura árvore-espacial é definida como:

$$T = \sup_{g(z)} \lambda_{g(z)} = \frac{\sup_{p_g(z)>p_g(\bar{z})} \mathcal{L}\left(p_g(\bar{z}), p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}\right)}{\sup_{p_g(z)=p_0} \mathcal{L}\left(p_0, p_g(z), \mathbf{z}, \mathbf{g}, \mathbf{x}\right)}. \tag{1.32}$$

O *cluster* mais verossímil é dado por um ramo g, em um conjunto de regiões ou zonas z com o maior valor de $\lambda_g(z)$ em relação a todas as zonas candidatas. Isto é, um candidato a *cluster* é dado pela zona z, na qual a probabilidade de um excesso de casos ter ocorrido no ramo g ao acaso é menor.

De modo geral, o algoritmo da estatística de varredura árvore-espacial pode ser dividido em duas etapas, a primeira se refere a encontrar todos os cortes possíveis da árvore (conjunto de ramos) e pode ser feito por meio da primeira etapa do algoritmo de varredura baseada em árvore. A segunda etapa compreende em utilizar o algoritmo de varredura espacial para encontrar as zonas z, ou seja, conjunto de regiões candidatas e para cada par (z, g) calcular o valor de $\lambda_g(z)$.

O *cluster* árvore-espacial mais verossímil é dado pelo conjunto de regiões que contêm um ramo g com maior valor de $\lambda_g(z)$ em relação a todas as zonas candidatas.

O Processo de identificação e inferência de um *cluster* árvore-espacial é idêntico ao método do algoritmo *Scan* Circular de Kulldorff (Seção 1.3.5) porém é adicionado ao mapa as informações de cortes possíveis da árvore.

1.5.3 Resumo algoritmo de varredura árvore-espacial

As etapas propostas para implementação do algoritmo de varredura árvore-espacial são apresentadas abaixo:

1. Varrer a árvore considerando todos os possíveis ramos.
2. Distribuir todos os possíveis ramos da árvore em todo mapa, ou seja, em todas regiões geográficas em estudo.

3. Encontrar as zonas candidatas, ou seja, varrer o mapa em janelas circulares, conforme Seção 1.3.5. Esse processo pode ser resumido nos seguintes itens:

   - Para cada zona z, conjunto de regiões vizinhas, é calculado um valor de $\lambda_g(z)$, isto é, para cada ramo g (conjunto de cortes) dentro dessa região.
   - Registrar a zona que contém um ramo g com o valor máximo de $\lambda_g(z)$ até o momento, limitando-se à proporção máxima da população predefinida.

4. Armazenar o valor de $T = \max_{z,g} \lambda_g(z)$.

5. Testar a significância do *cluster* árvore-espacial identificado, com base em simulação de Monte Carlo;

   - Criar J réplicas do mapa original, considerando a geração do conjunto de ramos da árvore. Portanto, para cada réplica o número de caso C_g é igual ao conjunto de dados original. Logo, pode-se distribuir aleatoriamente C_g casos com base em uma distribuição multinomial com parâmetros C_g (número total de casos dos ramos) e $\mu_g(z)$ (proporção de casos em cada ramo).
   - Para cada réplica do mapa, obter o valor da estatística T conforme passos 1 a 4.

6. Compare a zona z que possui o ramo g mais provável do conjunto de dados real, com a zona que possui os cortes mais prováveis das árvores geradas com base em simulação. Logo, rejeitar a hipótese nula significa que existe uma zona z que possui um ramo g da árvore, para os quais o número de casos é mais frequente se comparado às demais regiões. Ao nível $\alpha = 5\%$, se $T > P_{95}$, rejeita-se H_0, onde T é o valor observado da estatística de teste obtido por meio dos dados reais e P_{95} é o percentil da distribuição empírica de T sob H_0.

Todos os métodos abordados foram implementados no *Software SAS 9.04.01M5*, por meio dos módulos de linguagem iterativa matricial -*SAS/IML*, *SAS/STAT* e *SAS/-GRAPHICS*.

Para avaliar a implementação do método de varredura árvore-espacial foi realizada uma série de simulações. Posteriormente, apresentamos uma aplicação em dados reais, utilizando dados sobre mortalidade infantil do Sistema de Informações sobre Mortalidade (SIM). Esses dados são naturalmente organizado de forma hierárquica (categorias de doenças) e contêm o código do município da ocorrência da morte, possibilitando a incorporação de estrutura espacial.

Capítulo 2

Resultados e Discussões

2.1 Estudo de Simulações

O algoritmo para a Estatística de Varredura árvore-espacial foi avaliado por meio de simulações. Para testar a metodologia construiu-se quatro cenários (A, B, C, D), cada um representa um *cluster* artificial diferente. Esses *clusters* artificiais são considerados como verdadeiros. Os cenários foram montados com base em um mapa hipotético com 203 regiões em forma de hexágonos (Figura 2.2) e uma árvore com 4 níveis e 15 nós, sendo que um desses nós é a raiz . Um *cluster* árvore-espacial é dado por uma par (z, g), formado pela zona z e por um ramo g. Portanto, se H_0: ausência de *cluster* é rejeitada para um certo par (z, g), esse é um *cluster* árvore-espacial (ver 1.5). Cada um dos cenários possuem quatro *clusters* espacialmente diferentes, mas sempre no mesmo nó da árvore.

Na Tabela 2.1, são apresentadas as informações da árvore hipotética utilizada na construção dos cenários.

Tabela 2.1: Informações sobre a árvore utilizada nos cenários (A, B, C, D).

Informação da árvore	**Quantidade**
Ramos/Nós	15
Nós-raízes	1
Nós com filhos	4
Nós-folhas	8
Níveis na árvore	4

A Figuras 2.1 e 2.2 representam respectivamente a árvore hipotética e o mapa com 203 regiões em forma de hexágonos. As característica dos eventos são armazenadas nas folhas da árvore, ou seja, no menor nível.

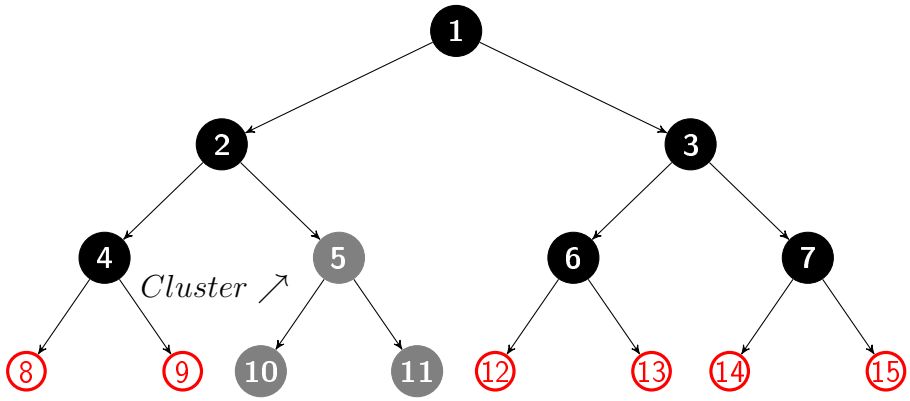


Figura 2.1: Árvore Hipotética

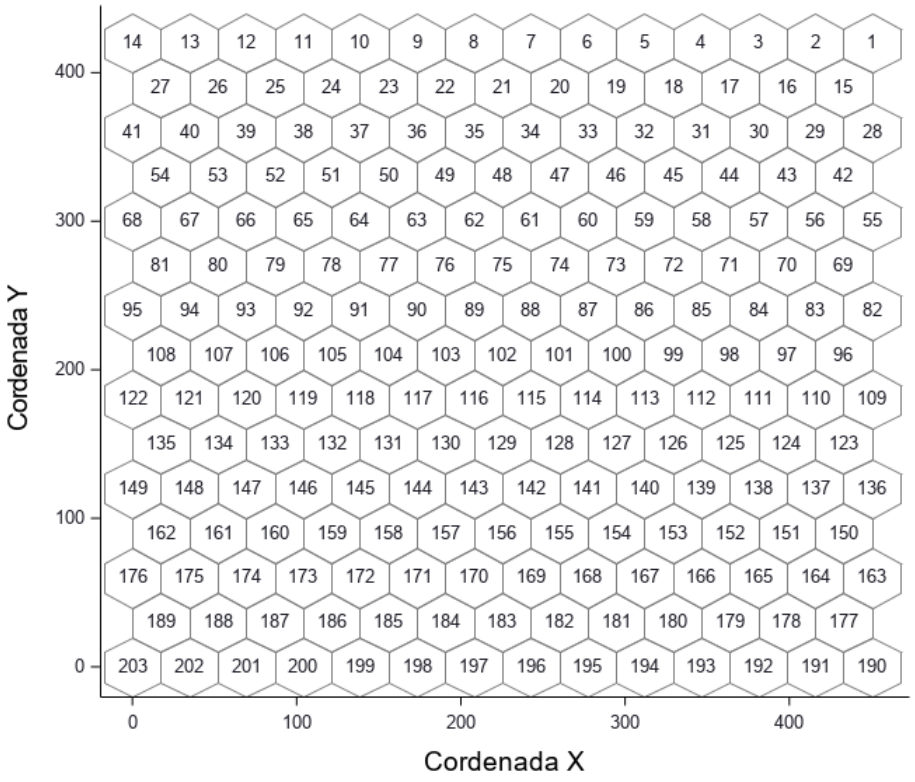


Figura 2.2: Mapa Hipotético com 203 regiões em forma de hexágonos

Os *clusters* árvore-espacial de cada cenário são formados por um conjunto de regiões(zonas) espacialmente diferentes e o nó do ramo 5 que é composto pelas folhas 10 e 11 (Figura 2.1). Dessa forma, em cada cenário tem-se um par do tipo:

$$\{Z = z_1, \cdots z_2, z_k, G = 5(10, 11)\}$$

A população foi distribuída uniformante sobre o mapa de forma que cada região possui $n_i = 1000$ indivíduos, onde $k = 203$, logo a população total em todo o mapa é $N = 203.000$. Em cada região tem-se $x_g(i)$ eventos, isto é, o número de casos de algum

evento de interesse na *i-ésima* região no ramo g. O número de casos de cada ramo da árvore foram distribuídos aleatoriamente sobre o mapa de forma que o número total de casos para cada ramo ao longo de todas as regiões fossem iguais a $C_g = 406$.

o número de casos do evento (por exemplo, quantidade de óbitos infantis, segundo classificação da CID-10) para cada ramo g na árvore, ao longo de cada região, foram distribuídos aleatoriamente sobre o mapa, de acordo com uma distribuição multinomial, sendo que as probabilidades são proporcionais ao risco relativo de cada região.

Os riscos relativos foram obtidos conforme descrito em (Kulldorff et al., 2003b) e levou-se em consideração a informação da árvore. Isto é, o número total de casos é dado pela soma do número de casos de cada ramo ao longo de todas as regiões. Dessa forma, os valores dos riscos relativos são mais altos para as regiões e ramos/nós da árvore que pertence ao *cluster* e menores para as regiões e ramos /nós que estão fora do *cluster*.

Os riscos relativos devem ser alto suficientes dentro das regiões condicionados ao número total de casos de cada ramos ao longo das regiões que pertencem ao *cluster*, para que a hipótese nula de ausência de *cluster* árvore-espacial seja rejeitada com probabilidade de $0,999$ em um teste binomial simples.

Considera-se n_z a população da zona z (conjunto de regiões) e N a população total do mapa e C_g o número total de casos de cada ramo g ao longo de todas as regiões.

Portanto, sob a hipótese nula de ausência de *cluster* árvore-espacial, isto é, o par $(z, g\ )$ e condicionando-se o número total de casos de cada ramo C_g, tem-se que sob H_0 o número de casos dentro do conjunto de regiões z e ramo g segue distribuição binomial com média $m_0 = C_g n_z/N$ e variância $v_0 = C_g n_z/N\,(N - n_z)$. Utilizando a aproximação da distribuição Binomial pela distribuição Normal padrão, tem-se que o valor crítico do número de casos k necessários para que um teste unilateral rejeite a hipótese nula ao nível $\alpha = 0,05$ de significância é um número k, tal que, $(k - m_0)/\sqrt{v_0}) = 1,645$.

Sob a hipótese alternativa H_a, as regiões e ramos da árvore que pertencem ao *cluster* possui um risco relativo r e o número de casos do ramo g ao longo das regiões seguem distribuição Binomial com média $m_a = (C_g n_z r)\,/\,(N - n_{z+n_z r})$ e variância $v_a = (C_g n_z r)\,/\,(N - n_z + n_z r)\,(N - n_z)\,/\,(N - n_z + n_z r)$. Usando novamente aproximação da distribuição Binomial para Normal, o risco relativo r é selecionado de tal forma que $(k - m_a)\,/\sqrt{v_a} = 3,09$ (Kulldorff et al., 2003b). Essa escolha permite um risco relativo r para cada um dos cenários de maneira a ter $\alpha = 0,05$ e um poder de $0,99$ ao utilizar um teste Binomial simples. Os valores do risco relativo para as regiões e ramos que não pertencem ao *cluster* são fixado como $r = 1$.

Resumidamente, os seguintes passos foram realizados para gerar os dados para cada uma das $H_a's$ de cada cenário:

- Para cada um dos quatro cenários, os *clusters* são formados por um conjunto de regiões espacialmente diferente e o ramo do nó 5 (folhas 10 e 11).

  1. Selecionar as folhas que não pertencem ao ramo do nó 5, ou seja, as folhas (8,

9, 12, 13, 14 e 15) na Figura 2.1. E gerar para cada uma dessa folhas $C_g = 406$ casos distribuídos aleatoriamente ao longo das regiões seguindo distribuição multinomial com probabilidades n_i/N, onde n_i é a população da *i-ésima* região e $N = \sum_{i=1}^{k} n_i$ é a população total. O número de casos de cada ramos é gerado conforme H_0 (ver passo 6, Seção 1.5.3).

2. Para cada folha que pertence ao ramo do nó 5 (folhas 10 e 11), gerar C_g=406 casos para cada folha ao longo das regiões supondo distribuição multinomial com probabilidades proporcionais ao risco relativo do conjunto de regiões que pertence ao *cluster*. Realizando os seguintes passos:
   - se a i-ésima região não pertence ao *cluster*, considere $n_i = p_i$;
   - se a i-ésima região pertence ao *cluster*, considere $n_i = rp_i$, onde (r é o risco relativo) e ;
   - em seguida calcule as probabilidades da multinomial como $p_i = n_i/N$, em que $N = \sum_{i=1}^{k} n_i$
3. Repetir os passos 1 e 2, ou seja, gerar os dados sob H_a 1000 vezes (número de simulações). Para cada uma das simulações deve-se verificar se um *cluster* detectado é significativo. E obter o poder do teste que é dado pela a proporção de vezes em que um *cluster* significativo foi detectado.

A significância dos *clusters* árvore-espacial detectados, foi verifica por meio da execução de 1000 simulações de Monte Carlo sob H_0, conforme descrito na Seção 1.5.3 Para obter o valor crítico da estatística do teste ao nível de $\alpha = 5\%$. Esse procedimento foi realizado apenas uma vez para cada cenário.

2.1.1 Cenários de simulação

O algoritmo estatística de varredura árvore-espacial foi utilizado para detectar os *clusters* artificiais dos cenários A, B, C e D. Os cenários foram definidos para avaliar o método com diferentes tipos de *clusters* quanto à estrutura espacial (forma e tamanho) e quanto à estrutura da árvore é utilizado o ramo do nó 5 da árvore.

As figuras abaixo mostram os cenários apenas com relação a estrutura espacial. A estrutura árvore-espacial ocorre quando adiciona-se o ramo do nó 5 que é formado pelas folhas 10 e 11 (Figura 2.1.1) no conjunto de região de cada cenário espacial. Dessa forma, existe um conjunto de regiões, na qual, a probabilidade de ocorrer casos associados ao ramo do nó 5, ou seja, nos cortes, $g_1 = [5, 10, 11]$ ou $g_2 = [10]$ e $g_3 = [11]$, é maior do que se estivesse fora desse conjunto de regiões.

A seguir, serão apresentados os cenários A, B, C e D.

- **Cenário A**: é formado por 19 regiões em formato circular e o ramo do nó 5 que é composto pelas folhas 10 e 11 (Figura 2.1).

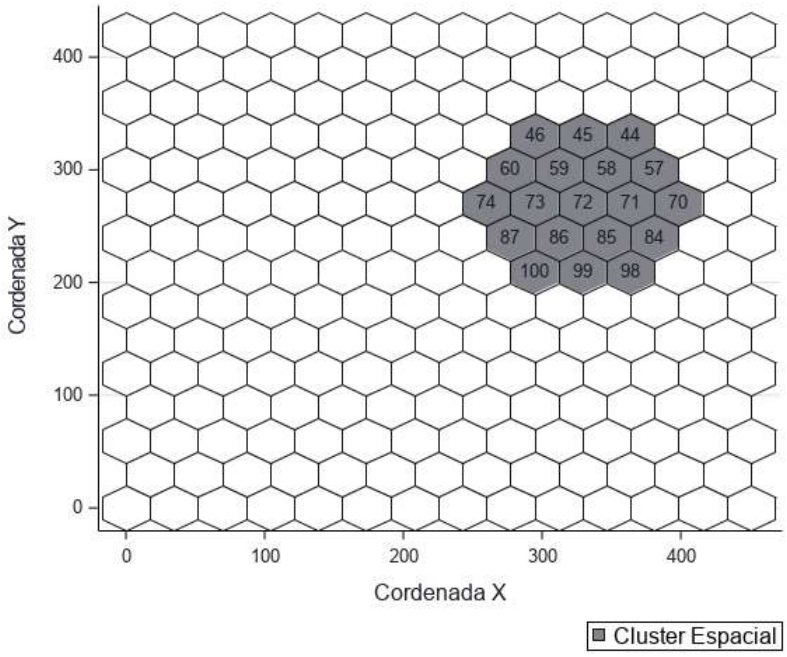


Figura 2.3: Localização espacial do cenário A

- **Cenário B**: possui o mesmo formato do cenário A, só difere quando à localização geográfica.

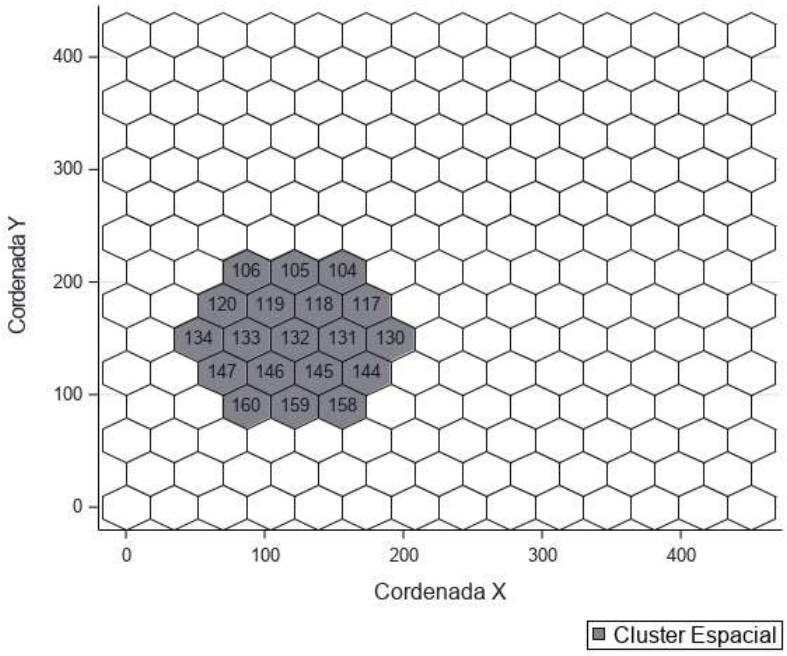


Figura 2.4: Localização espacial do cenário B

- **Cenário C**: possui formato circular, porém é formado apenas por 7 regiões e o ramo do nó 5 que é composto pelas folhas 10 e 11 (Figura 2.1).

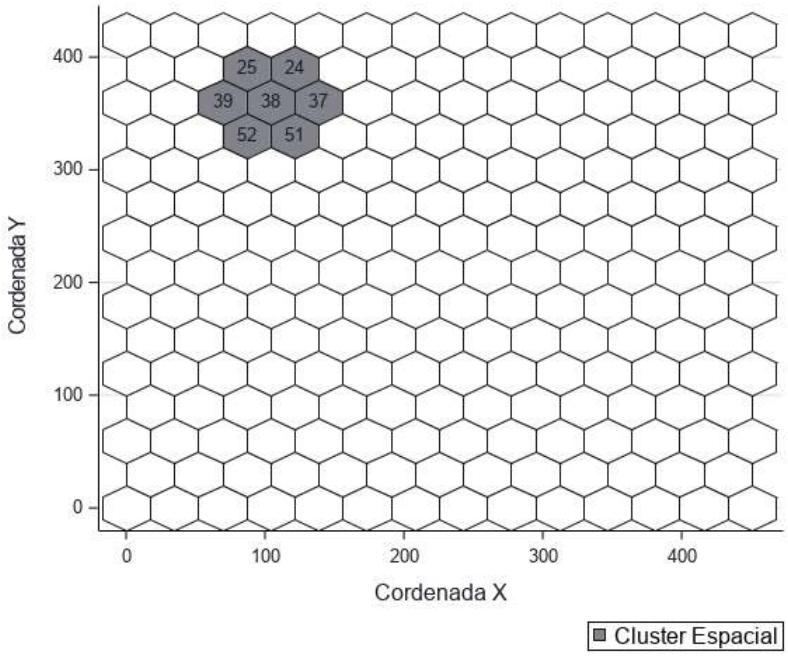


Figura 2.5: Localização espacial do cenário C

- **Cenário D**: possui formato aproximado de um L invertido, é formado por 17 regiões e o ramo do nó 5 que é composto pelas folhas 10 e 11 (Figura 2.1).

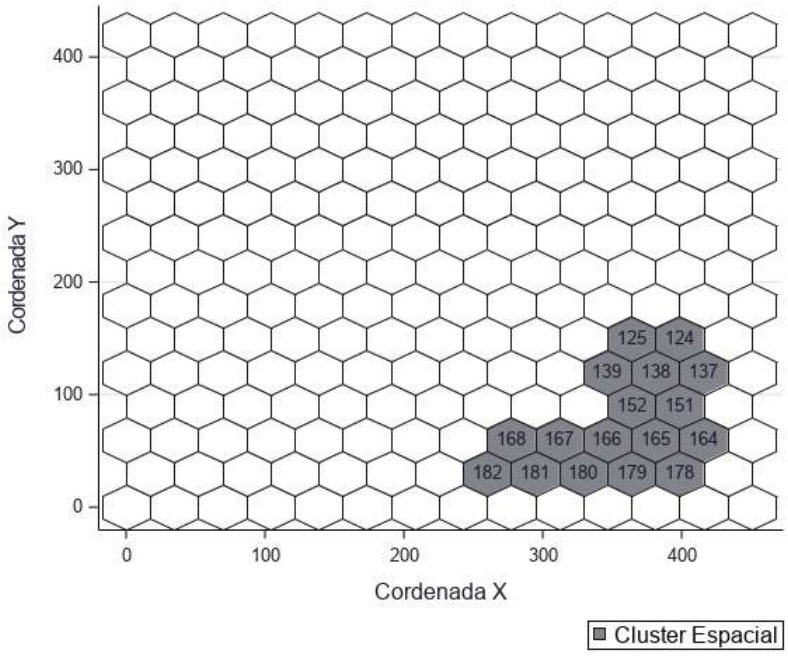


Figura 2.6: Localização espacial do cenário D

Tabela 2.2: Risco Relativo e número de casos esperados sob $H_0(m_0)$ e sob $H_a(m_a)$

Cenários	Tamanho *cluster* espacial	Nº Folhas	RR	m_0	m_a
A	19	2	1,696	38	60,506
B	19	2	1,696	38	60,506
C	7	2	2,210	14	29,700
D	17	2	1,7373	34	55,633

Verifica-se na Tabela 2.1.1 o risco relativo selecionado para cada cenário e o número de casos esperado do ramo 5 ao longo das regiões sob H_0 e H_a. O cenário C apresenta o maior risco relativo, conforme esperados, visto que contém a menor população exposta e número de regiões. Já os cenários A e D são próximos em tamanho espacial e risco relativo, mas possuem formatos diferentes.

O *Cluster* árvore-espacial acontece quando o risco relativo ultrapassa o valor 1 ao comparar a probabilidade dentro do *cluster* árvore-espacial e fora do *cluster* árvore-espacial.

2.1.2 Medidas de Desempenho ou eficiência do método

A acurácia ou eficiência do método em termos de detecção de *clusters* foi mensurada por meio das medidas: Sensibilidade, Valor Preditivo Positivo(PPV) e Poder do Teste. Essas medidas refletem o quanto o algoritmo está detectando um *cluster* árvore-espacial, quando este realmente existe. As seguintes definições são utilizadas no método de cálculo dessas métricas:

- ***Cluster Detectado***: é o *cluster* encontrado ao executar o algoritmo de varredura árvore-espacial;

- ***Cluster Verdadeiro***: é o *cluster* que foi artificialmente criado em cada um dos cenários (*cluster* real), considerando o conjunto de regiões e o ramo do nó 5 da árvore (Seção 2.1.1).

Poder do Teste

O Poder do teste é dado pela proporção de vezes que a hipótese nula é rejeita. Em outras palavras, é a probabilidade do algoritmo detectar um *cluster*, quando esse realmente existe.

Dessa forma, para cada um dos cenários (A, B, C e D) o algoritmo de varredura árvore-espacial foi repetido $M = 1000$ vezes sob H_a. Em cada uma das M simulações é obtido o valor da estatística de teste (razão de verossimilhança) $\lambda_g(z)$, e esse valor é

comparado com o valor crítico λ^* que é obtido por meio de simulações de Monte Carlos sob H_0.

O Poder é dado por:

$$Poder = \frac{\sum_{i=1}^{M} \lambda_i > \lambda^*}{M}, \tag{2.1}$$

onde $I(\lambda_i > \lambda^*) = 1$ se $\lambda_i > \lambda^*$ e 0 caso contrário. Isto é, a razão entre a quantidade de vezes que o método detecta o *cluster* verdadeiro e o total de simulações.

Sensibilidade

A sensibilidade indica a proporção da população/elementos do *cluster* verdadeiro que é encontrada no *cluster* detectado. Essa medida produz um ideia do quanto do *cluster* verdadeiro é detectado.

A sensibilidade é obtida pela média ao longo das M simulações de Monte Carlo da proporção da população da interseção entre o *cluster* detectado e o verdadeiro (real) em relação a população do *cluster* verdadeiro, conforme a Equação 2.2, onde *Pop* representa População.

$$Sensibilidade = \frac{Pop\left(ClusterVerdadeiro \cap ClusterDetectado\right)}{Pop(ClusterVerdadeiro)}. \tag{2.2}$$

Valor Preditivo Positivo

O Valor Preditivo Positivo(VPP) é a proporção da população/elementos do *cluster* detectado que pertence ao *cluster* verdadeiro e pode ser interpretado como o quanto do *cluster* detectado pertence ao verdadeiro.

O *VPP* é obtido por meio da média das M simulações de Monte Carlo da proporção da população da interseção entre o *cluster* detectado e o verdadeiro (real) em relação a população do *cluster* detectado (Equação 2.3).

$$VPP = \frac{Pop\left(ClusterDetectado \cap ClusterVerdadeiro\right)}{Pop(ClusterDetectado)}. \tag{2.3}$$

Tabela 2.3: Medidas de desempenho para os quatros cenários artificiais

Cenários	***Clusters***	**Poder**	**Sensibilidade**	**VPP**
A	Espacial	0,8140	0,8149	0,7989
	Árvore		0,8400	0,8930
B	Espacial	0,7970	0,8176	0,7937
	Árvore		0,8370	0,8910
C	Espacial	0.7980	0,8650	0,8035
	Árvore		0,8600	0,8960
D	Espacial	0,6550	0,6376	0,5734
	Árvore		0,7485	0,8280

Na Tabela 2.1.2 estão apresentadas as medidas de desempenho para cada um dos cenários artificias com base em 1000 simulações.

As medidas de desempenho sensibilidade e PPV foram mensuradas considerado a estrutura espacial (*cluster* espacial), isto é, o conjunto de regiões do mapa no qual a probabilidade do individuo pode vir a ser um caso relacionado ao ramo $g = 5(10, 11)$ da árvore é maior do que para esse mesmo ramo fora desse conjunto de regiões. Além disso, essas medidas são calculadas sobre a estrutura da árvore (hierárquica) onde a ideia é avaliar o método quanto à capacidade de detectar corretamente o ramo $g = 5(10, 11)$ da árvore, sendo considerado *cluster* verdadeiro quando o método detectar como corte mais verossímil o nó 5 e parcialmente quando for detectado pelo menos uma das duas folhas 10 ou 11 que são o nós filhos desse ramo.

Verifica-se na Tabela 2.1.2 que o algoritmo de varredura árvore-espacial apresenta um bom desempenho em relação a capacidade de detectar *clusters* árvore-espaciais, quando de fato este existe, visto que a proporção de vezes que o método detectou *clusters* ficou próximo a 80% para os cenário A, B e C. Já, quanto ao cenário D, é possível notar uma queda no poder, o que está relacionado ao formato não circular desse *cluster*.

A sensibilidade espacial média para os *clusters* foi superior a 80% para todos os cenários exceto para o caso do cenário D, para o qual novamente é possível observar uma queda nas medidas de desempenho causada pela irregularidade de forma. Como a construção das janelas é feita de forma circular, em geral, tem-se um ganho de desempenho no método *Scan* quando a estrutura espacial tem formato circular.

Para o *cluster* espacial obteve-se altos valores de sensibilidade e *VPP*. Portanto, tem-se evidências que o *cluster* detectado se aproxima bastante do *cluster* verdadeiro. Para todos os cenários, o Valor preditivo Positivo foi inferior a sensibilidade, logo o algoritmo detectou *clusters* com uma maior quantidade de regiões do que o real.

No caso do *cluster* relacionado à estrutura da árvore, tem-se em média uma proporção maior de ramos do *cluster* detectado que pertence ao verdadeiro (detectar

o corte que contém o ramo do nó 5). Nesse caso, os valores do *VPP* são maiores que a sensibilidade, visto que o *cluster* real é o corte que contém o nó do ramo 5, que é composto por duas folhas (10 e 11), portanto podem ser detectado os cortes $g_1 = [5, 10, 11]$ ou $g_2 = [10]$ e $g_3 = [11]$.

Nota-se que os valores da sensibilidade para estrutura da árvore são inferiores aos valores do *VPP*. A proporção de ramos do *cluster* verdadeiro que pertence ao *cluster* detectado são em média menores.

As medidas sensibilidade e *PPV* possuem valores altos para estrutura da árvore. Isso indica que o algoritmo detectou uma proporção maior do *cluster* real. Isto é, o algoritmo detectou exatamente o corte que contém o ramo do nó 5, e uma proporção menor dos cortes $g_2 = [10]$ e $g_3 = [11]$, que são apenas os nó folhas.

2.2 Aplicação em dados reais

Nessa Seção será apresentada uma aplicação da Estatística de Varredura Árvore-Espacial aos dados de mortalidade infantil segundo a Classificação Internacional de Doenças e Problemas Relacionados a Saúde-CID 10. O objetivo é encontrar um conjunto de doenças ou uma doença específica, ou seja, um nó da árvore (grupo de folhas), ou apenas uma folha específica, na qual em algum conjunto de regiões a incidência de óbitos infantis é significativamente maior. Isto é, para algum conjunto de regiões, a probabilidade de um excesso de casos ter ocorrido em algum ramo da árvore ao acaso é menor.

2.2.1 Visão Geral do Bancos de Dados

Para aplicação da metodologia proposta foram utilizados dados de mortalidade infantil segundo a Classificação Internacional de Doenças e Problemas Relacionados a Saúde-CID 10 do estado do Rio de Janeiro para o ano de 2016. Esses dados pertencem ao Sistema de Informação sobre Mortalidade-SIM e estão disponíveis gratuitamente no sítio do DATASUS. Além do Banco de dados do SIM, para o ano de 2016, foram utilizados outros dois bancos de dados. O primeiro corresponde à base populacional, que contém as estimativas populacionais realizadas para o TCU dos municípios do Rio de Janeiro para o ano de 2016, também disponíveis pelo sítio do DATASUS. E o segundo corresponde à malha municipal 2016 para o estado do Rio de Janeiro.

As bases de dados citadas acima podem ser acessadas pelo seguintes portal:

- **DOINF16.dbc**: corresponde à base de dados da declaração de óbitos infantis 1997-2017. Disponível por meio do portal *Acesso a Informação Serviços* $\longrightarrow$ Transferência/Download de Arquivos $\longrightarrow$ Acesse o SIM.
  O arquivo é disponibilizado em formato .dbc, mas pode ser descompactado por meio do *Software TABWIN* para os formatos .csv e .dbf.

  Sítio:`http://datasus.saude.gov.br/`.

- **POPTBR16.dbf**: corresponde à base de estimativas populacionismo TCU-1992-2018. Disponível em *Acesso a Informação Serviços* $\longrightarrow$ Transferência/Download de Arquivos $\longrightarrow$ Base Populacionais.

  Sítio:`http://datasus.saude.gov.br/`.

- **Malha Municipal 2016:** contém um conjunto de arquivo com as informações da malha territorial geográfica (latitude e longitude). É necessário acessar o portal *IBGE Downloads* $\longrightarrow$ *Geociências* $\longrightarrow$ *Organizacao do territorio* $\longrightarrow$ *Malhas Municipais* $\longrightarrow$ *Municipio Ano*, escolher ano e UF desejados.

  A malha municipal é composta por 4 aquivos, sendo que todos se inciam com o geocódigo da UF, o qual, para o RJ é igual a 33.

  `https://downloads.ibge.gov.br/downloads_geociencias.htm`

O banco de dados do Sistema de Informação sobre mortalidade-SIM (Doinf), armazena os registro de óbitos infantis.

O Sistema de Informação sobre mortalidade-SIM, criado pelo Mistério da Saúde em 1975, é responsável por coletar dados sobre mortalidade no Brasil e informações sociodemográficas relacionadas aos óbitos. O documento oficial utilizado para obter as informações sobre mortalidade infantil é o formulário da declaração de óbitos.

Para esse estudo de detecção de *cluster* árvore-espacial foram utilizadas apenas duas variáveis do Banco de dados DOINF2016.

1. CODMUNRES: variável que armazena o código de município de residência do falecido, conforme códigos IBGE (7 dígitos, sendo os dois primeiros o geocódigo da UF).

2. CAUSABAS: variável que armazena o código da causa básica, conforme a Classificação Internacional de Doenças- CID 10 (3 ou 4 caracteres, sendo uma letra seguida por dois ou três dígitos numéricos).

Essas duas variáveis são as de interesse no estudo. A primeira variável será utilizada para montar a estrutura espacial, pois possui o código do município de residência do falecido. Já a segunda, *CAUSABAS*, possui a informação que será utilizada para montar a estrutura da árvore, visto que contém a causa básica da morte segundo as categorias e subcategorias conforme a CID-10.

O SIM é de responsabilidade do Ministério da Saúde, do Centro Nacional de Epidemiologia e das secretarias de saúde municipais e estaduais. O Banco de dados sobre mortalidade passou a ser atualizado periodicamente mais recentemente devido às normas sobre o preenchimento e correta disponibilização da declaração de óbito. Mas os dados ainda demoram para serem disponibilizados, visto que o último banco de dados

disponibilizado é para o ano de 2016. Os dados para o ano de 2017 estão listados como preliminares.

O estado do Rio de Janeiro possui 92 municípios. No banco de dados tem-se registro para 90 municípios, considerado o código de municípios de residência, sendo que um desses códigos não pertence à malha municipal do Rio de janeiro para o ano de 2016. Esse, é um código genérico designado como município ignorado (Fonte Tabnet), ou seja, quando não é possível identificar o local de residência . A variável CAUSABAS contém informação para todos esses municípios.

O banco de dados também contém o código do município de ocorrência do óbito, CODMUNOCOR, que também seria de interesse do estudo. Contudo, optou-se pela variável CODMUNRES. A variável CODMUNOCOR, possui muitos municípios sem registro, visto que alguns municípios possuem uma maior concentração de registros.

Do Banco de dado POPTBR16 foram utilizadas as seguintes variáveis:

1. CD_GEOCMU: variável que contém o geocódigo do município
2. Ano: variável que armazena o ano da projeção
3. População: estimativas populacionais do IBGE enviadas para o TCU com data base de 01 de julho de 2016.

A variável CD_GEOCMU corresponde a variável *CODMUNRES* no banco de dados DOINF2016. O terceiro banco de dados do IBGE contém as informações georreferenciais para a malha municipal do Rio de janeiro. E as seguintes variáveis foram utilizada:

1. X: latitude, sistema georreferência SIRGAS2000.
2. Y: longitude, sistema georreferência SIRGAS2000.
3. CD_GEOCMU: geocódigo do Município.
4. NM_MUNICIP: nome dos municípios.

As variáveis X e Y correspondem às informações georreferencias latitude e longitude, respectivamente, para os municípios do Rio de Janeiro. Foram utilizadas para obter o centroide para cada polígino. O centroide de cada município foi obtido por meio da macro %*centroides* disponibilizada no *Software* SAS 9.04.01M5.

2.2.2 Construção da Estrutura Hierárquica

Nessa Seção será apresentado uma visão geral de como é a organização hierárquica da CID-10. Essa nomenclatura é utilizada para montar a estrutura da árvore.

2.2.3 Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - CID-10

A décima revisão da Classificação Internacional de Doenças e de Problemas Relacionados a Saúde é a última versão disponível desde o início das discussões em 1893. E a partir dessa versão passou a ser denominada de Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde, mas continua sendo vinculada a sigla CID-10 (de Classificação de Doenças CBCD, 2019).

A variável 'causa básica do óbito' está definida conforme um sistema de classificação de doença, apresentada em forma de categorias, que naturalmente estão dispostas de forma hierárquica. A Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde é uma sistema de categorização de doenças que atribui uma codificação para doenças e demais circunstâncias relacionadas a entidades mórbidas. Portanto, uma categoria é atribuída para cada condição de saúde, que recebe um código único. As categorias com doenças semelhantes formam grupos.

A estrutura da CID-10 é organizada de forma naturalmente hierárquica, conforme mostrado na Figura 2.2.3:

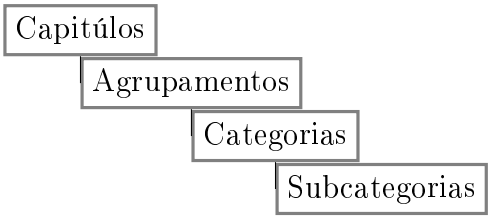


Figura 2.7: Estrutura Hierárquica da CID-10

- **Capítulos**: conjunto de vários agrupamentos semelhantes;
- **Agrupamentos**: conjunto de categorias com doenças semelhantes:
- **Categorias**: cada classificação atribuída para cada doença, corresponde a um código com uma letra e dois dígitos;
- **Subcategorias**: menor nível de classificação de uma doença, o código é composto pelo código da categoria com adição de um ponto e um algorismo de 0 ao 9 ou apenas os algarismos.

A estrutura Hierárquica da CID-10 oficial contém 4 níveis, respeitando a ordem mostrada na Figura 2.2.3. Mas nesse trabalho utilizamos uma árvore com 5 níveis, visto que foi adicionado um nó raiz (maior nível da árvore) identificado pelo pelo código RR, que representa o número total de óbitos, visto que é composto pela agregação de todos os capítulos.

No banco de dados *DOINF2016*, a causa básica do óbito armazena as classificações das doenças em categorias e subcategorias. Os registro são códigos formados por uma letra seguida de dois ou três dígitos numéricos. A quantidade de dígitos identifica quais classificações são categorias ou subcategorias. As codificações das subcategorias são definidas adicionando ao código da categoria a qual faz parte uma sequencia de 0 a 9. Existem algumas subcategorias que possuem um ponto que separa o código da categoria do digito de 0 a 9 (essa codificação é a descrita na documentação da CID -10)(Silva, 2017). Em geral, as subcategorias não são organizadas de maneira lógica, considerando por exemplo, uma ordenação crescente ou decrescente em relação ao dígitos que são adicionados aos códigos das categorias (Silva, 2017).

A Tabela 2.4 a seguir mostra a listagem tabular oficial da OMS da composição dos capítulos da CID-10.

Tabela 2.4: Descrição dos capítulos CID-10

Capítulos	**Categorias**	**Descrição Abreviada**
Capítulo I	A00-B99	Algumas doenças infecciosas e parasitárias
Capítulo II	C00-D48	Neoplasias [tumores]
Capítulo III	D50-D89	Doenças sangue órgãos hemat e transt imunitários
Capítulo IV	E00-E90	Doenças endócrinas
Capítulo V	F00-F99	Transtornos mentais e comportamentais
Capítulo VI	G00-G99	Doenças do sistema nervoso
Capítulo VII	H00-H59	Doenças do olho e anexos
Capítulo VIII	H60-H95	Doenças do ouvido e da apófise mastóide
Capítulo IX	I00-I99	Doenças do aparelho circulatório
Capítulo X	J00-J99	Doenças do aparelho respiratório
Capítulo XI	K00-K93	Doenças do aparelho digestivo
Capítulo XII	L00-L99	Doenças da pele e do tecido subcutâneo
Capítulo XIII	M00-M99	Doenças sist osteomuscular e tec conjuntivo
Capítulo XIV	N00-N99	Doenças do aparelho geniturinário
Capítulo XV	O00-O99	Gravidez
Capítulo XVI	P00-P96	Algumas afecções originadas no período perinatal
Capítulo XVII	Q00-Q99	Malformações congênitas
Capítulo XVIII	R00-R99	Sintomas e achados anormais de exames clínicos
Capítulo XIX	S00-T98	Lesões
Capítulo XX	V01-Y98	Causas externas de morbidade e de mortalidade
Capítulo XXI	Z00-Z99	Contatos com serviços de saúde
Capítulo XXII	U00-U9	Códigos para propósitos especiais

Verifica-se na Tabela 2.4 a listagem oficial da OMS, com a descrição do conjunto de categorias (primeira e última) que formam vários agrupamentos e, por sua vez,

compõem os capítulos. A listagem tabular da OMS que detalha como são formados os grupos e capítulos, além de todos os códigos das categorias e subcategorias, está disponível no sítio da CID-10/DATASUS, e são denominadas como *Tabelas da CID-10*. O acesso pode ser feito por meio do *Help Online* e arquivos (csv, xml, tabwin), por meio do endereço:
`http://www.datasus.gov.br/cid10/V2008/download.htm`.

As informações da listagem tabular da OMS são utilizadas para identificar os níveis superiores da árvore. Nesse trabalho foi utilizada a mesma estrutura hierárquica montada por (Silva, 2017), adicionando apenas um nó para armazenar todos os demais, que é denominado como nó raiz.

Oficialmente a listagem de agrupamentos possui 275 grupos, mas para facilitar a análise nesse trabalho serão considerados apenas 239 agrupamentos. A codificação dos grupos é dada por um prefixo de duas letras, *GG*, seguidas de uma sequência numérica de 1 a 239 (Silva, 2017). Esta padronização permite que cada grupo pertença a um conjunto de categorias únicas, visto que na composição da OMS, existem algumas categorias que pertencem a mais de um grupo, o que poderia trazer problemas à análise. Porém, a agregação customizada não altera os dados oficiais visto que é possível retornar para separação original da OMS, utilizando as informações das classificações das doenças. Esses grupos são formados pela junção de algumas categorias com classificações de doenças semelhantemente. Os capítulos são identificados por duas letras *CC* seguidos de uma sequência numérica de 1 a 22. Essa sequência representa a quantidade de capítulos oficiais, conforme a Tabela 2.4.

Tabela 2.5: Informações sobre a estrutura hierárquica da CID-10 oficial e a estrutura da CID-10 utilizada

Atributos	Quantidade	
	Estrutura Hierárquica	
	Árvore Original	Estrutura Utilizada
Nós	14494	713
Nós-raízes	22	1
Nós com filhos	2043	303
Nós folhas	12451	410
Níveis da árvore	4	5
Nós 1º nível	22	1
Nós 2º nível	239	16
Nós 3º nível	2045	91
Nós 4º nível	12188	220
Nós 5º nível	-	385

Verifica-se na Tabela 2.5 as informações referentes à estrutura hierárquica oficial da CID-10 (considerando os agrupamentos, conforme (Silva, 2017)) e à estrutura hierárquica da CID-10 que foi utilizada nesse trabalho. Ressalta-se que na estrutura construída, o primeiro nível é o nó raiz (RR), seguido dos capítulos, agrupamentos, categorias e subcategorias (5º nível).

Observa-se que a estrutura oficial possui 14494 ramos, sendo 12451 nós folhas (menor nível). A quantidade de nós demonstra o nível de complexidade da árvore, visto que o algoritmo árvore-espacial avalia todos os cortes possíveis da árvore para todos as zonas até um limite da população predefinido. Portanto, optou-se por utilizar apenas a quantidade de folhas que estava no registro da variável causa básica do óbito para o ano de 2016 no estado do Rio de Janeiro, eliminando as folhas com número de casos igual a zero. Dessa forma, a estrutura utilizada possui apenas 713 nós. A variável Causa Básica contém 410 classificações de doenças (categorias e subcategorias). Os níveis superiores da árvore foram construídos, utilizando todos os possíveis nós de níveis superiores, isto é, os níveis 2 e 3 associados aos níveis 5 e 4. Dessa forma, serão avaliados todos os possíveis cortes de uma árvore com 5 níveis contendo uma raiz, 410 nós folhas, sendo 16 dos 22 capítulos (2º nível), 95 agrupamentos (3º nível), 220 categorias e 385 subcategorias.

Os dados podem ser representados por um matriz, na qual as linhas contêm os nós da árvore e as colunas contêm os códigos de cada município do Rio de Janeiro. Portanto, tem-se uma matriz com 713 linhas e 92 colunas. Uma parte dessa matriz é mostrada na Figura 2.8. Os nós da árvore estão ordenados em ordem decrescente. O número total de casos de uma doença específica C_g é dado pela soma da linha.

ArvoreID	330010	330015	330020	330022	330023	330025	330030	330040	330045	330050
RR	22	4	23	4	8	11	15	21	120	7
CC01	1	0	3	0	0	0	1	0	9	0
CC02	0	2	0	0	0	0	0	0	1	0
CC03	0	0	0	0	0	0	0	0	0	0
CC04	0	1	0	0	0	0	0	1	0	0
CC06	0	0	0	0	0	0	0	0	3	0
CC08	0	0	0	0	0	0	0	0	0	0
CC09	0	0	0	0	0	0	0	0	1	0
CC10	1	0	1	0	0	0	1	1	17	1
CC11	0	0	0	0	0	0	0	0	1	0
CC12	0	0	0	0	0	0	0	0	0	0
CC13	0	0	0	0	0	0	0	0	0	0
CC14	0	0	0	0	0	0	0	0	0	0
CC16	11	0	14	3	3	5	9	8	43	6
CC17	6	1	2	1	3	6	4	9	33	0
CC18	0	0	2	0	0	0	0	0	1	0
CC20	3	0	1	0	2	0	0	2	11	0
GG001	1	0	0	0	0	0	0	0	3	0
GG002	0	0	0	0	0	0	0	0	1	0
GG004	0	0	3	0	0	0	1	0	3	0

Figura 2.8: Matriz do número de óbitos por município do RJ, segundo CID-10

O nó Raiz contém o número total de óbitos infantis para cada município do

Rio de Janeiro, visto que é composto por todos os capítulos sumarizados.

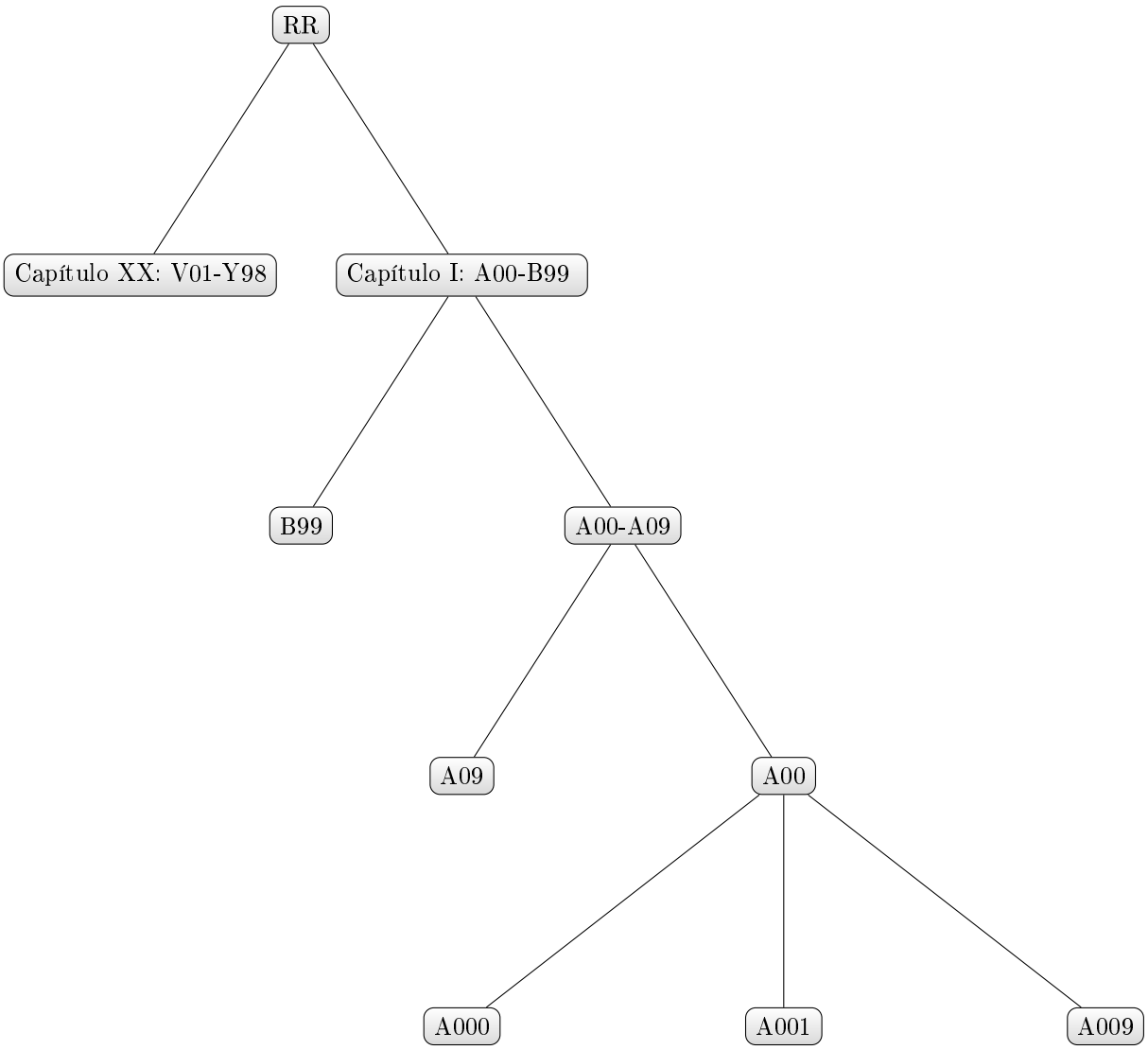


Figura 2.9: Estrutura Hierárquica CID-10

A Figura 2.2.3 ilustra apenas 2 ramos da estrutura hierárquica da CID-10 utilizada nesse estudo. Mas é possível ver como é realizada a identificação de uma doença de acordo com os níveis que vão das subcategorias (5º nível) até a raiz.

O Exemplo a seguir, ilustra o processo de identificação de acordo com a Figura 2.2.3.

1. **Raiz**: agrega o número de casos de todos os capítulos;
2. **Capítulo 1**: A00-B99 (Algumas Doenças Infecciosas e parasitárias);
3. **Agrupamentos**: A00-A09 (Doenças Infecciosas e Intestinais);

4. **Categorias**: A00 (Cólera);

5. **Subcategorias**: A000 (Cólera devida a Vibrio cholerae, biótipo cholerae Cólera clássic 0), A001 (Cólera devida a Vibrio cholerae 01, biótipo El Tor Cólera El Tor) e A009 (Cólera não especificada).

2.2.4 Análise Descritiva

Nessa Seção será apresentada uma breve análise exploratória sobre a distribuição dos óbitos infantis por município do Rio de Janeiro no ano de 2016, e será analisado o comportamento dos óbitos segundo a CID-10.

2.2.5 Distribuição de óbitos infantis no Rio de Janeiro em 2016 segundo a CID-10

A mortalidade infantil corresponde aos óbitos ocorridos até o primeiro ano de vida e pode ser categorizada em três fases: a primeira neonatal precoce (0 a 6 dias de vida), seguida pela neonatal tardio (7 a 27 dias) e pós-neonatal (28 dias ou mais).

O estado do Rio de Janeiro é divido em 92 municípios. A estimativa populacional é de 16.635.966 milhões de habitantes no ano de 2016. Além disso, apresentou 2.982 registros de óbitos infantis distribuídos nesses municípios. A variável causa básica do óbito contém 2.990 registros, mas 8 óbitos referem-se a municípios ignorados (quando o município de residência não é conhecido).

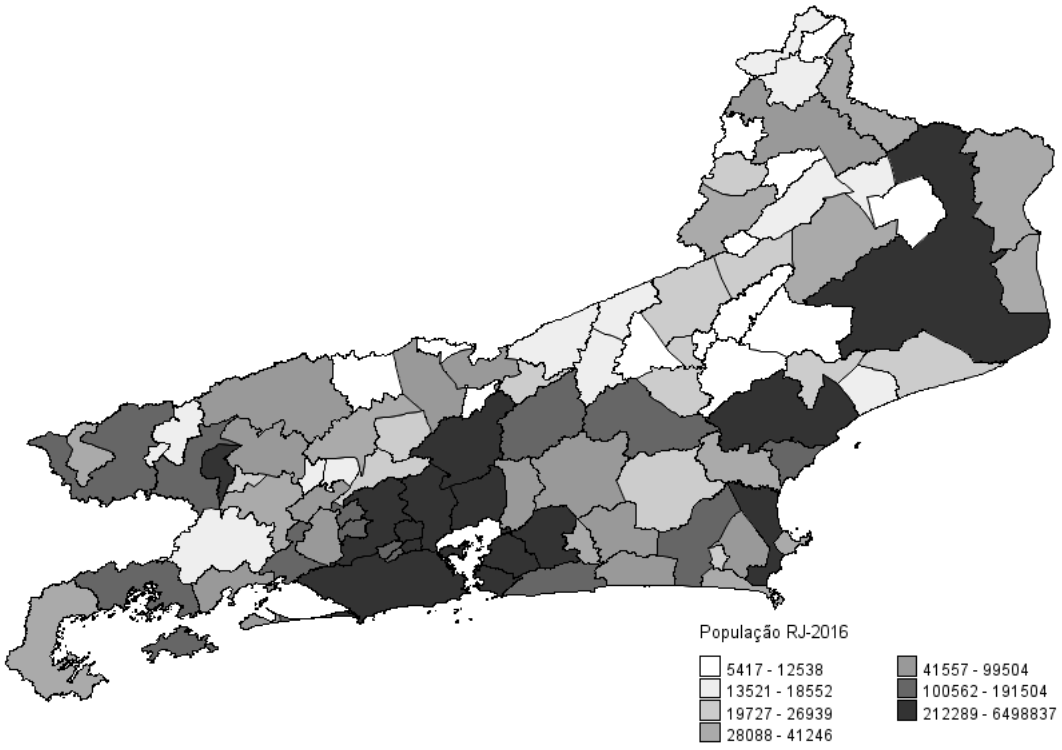


Figura 2.10: Distribuição da população do Rio de Janeiro no ano de 2016

Verifica-se na Figura 2.10 que existe uma pequena quantidade de municípios que pertencem às últimas duas faixas, para a qual o número de habitantes ultrapassa 1 milhão de pessoas. O município do Rio de Janeiro é o mais populoso, abrangendo 39,06%

da população (6.498.837), seguido por São Gonçalo, que possui 1.044.058 habitantes. Pelo menos 50% dos municípios tem população maior que 35.826 mil habitantes.

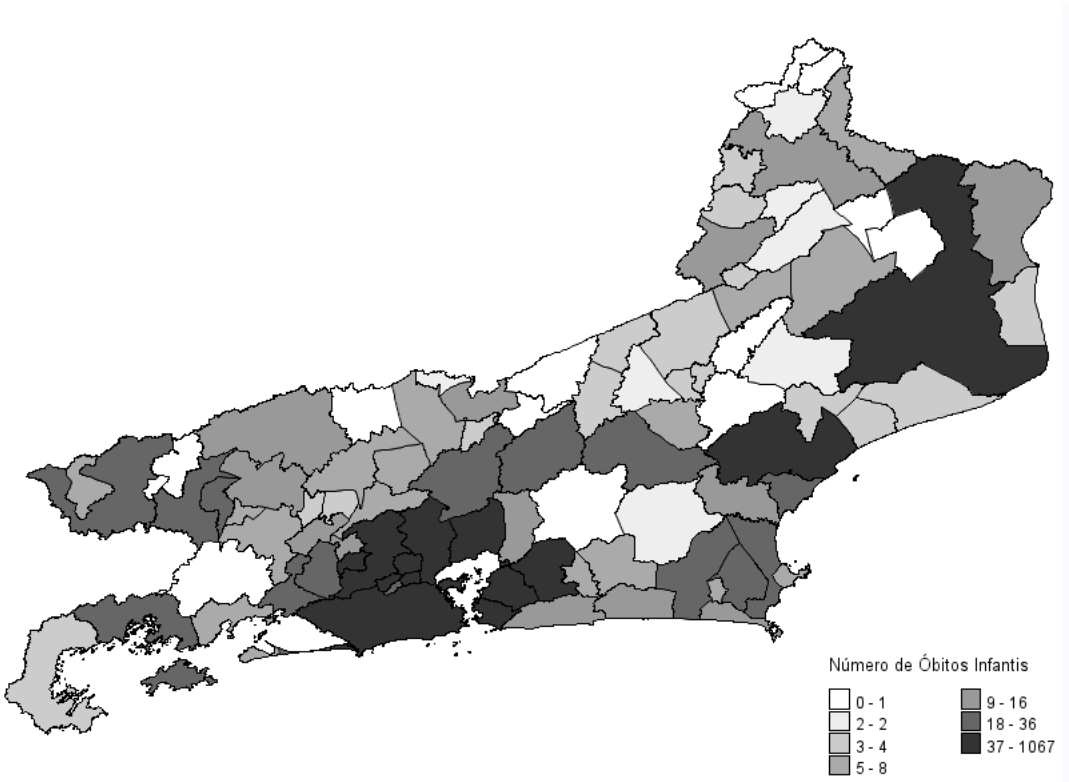


Figura 2.11: Distribuição do número de óbitos infantis segundo Munc RJ-2016

Tabela 2.6: Medidas resumo da variável causa básica do óbito

	Média	**Desv.Padrão**	**Coef.Var**	**1° Quartil**	**Mediana**	**3° Quartil**	**Mínimo**	**Máximo**
CAUSABAS	32,41	115,43	3,5614	3	7	22,5	0	1076

A Figura 2.11 mostra a distribuição do número de óbitos infantis, segundo os municípios do Rio de Janeiro no ano de 2016. Observa-se que a quantidade de óbitos por município oscilou de 0 a 1.067, com destaque para os municípios do Rio de Janeiro (1.067) óbitos seguidos por Duque de Caxias (214), São Gonçalo (162), Nova Iguaçu (155), Campo dos Goytacazes (149), São João de Meriti (81), Niterói(74) e Magé(74). As áreas mais claras do mapa evidenciam que alguns municípios não tiveram registro de óbitos ou houve apenas um caso de óbito infantil. Porém, não é possível afirmar que esses municípios não tiveram casos de óbitos, visto que pode existir subnotificação devido a problemas com o correto preenchimento do formulário da declaração de óbito. Existe uma concentração maior de caso nos municípios próximos à capital do Rio de Janeiro.

Verifica-se na Tabela 2.6, com base nas medidas de dispersão, que a variabilidade da quantidade óbitos infantis em torno da média é muito alta, visto que o coeficiente

de variação é (3,5614), ou seja, 356,14%, o que mostra que a distribuição dos óbitos é bem heterogênea). Tem-se que 50% dos número de óbitos são menores que 7 e 50% são superiores a 7 óbitos. Com base no terceiro quartil, observa-se que 25% da quantidade de óbitos infantis são maiores que 22,5.

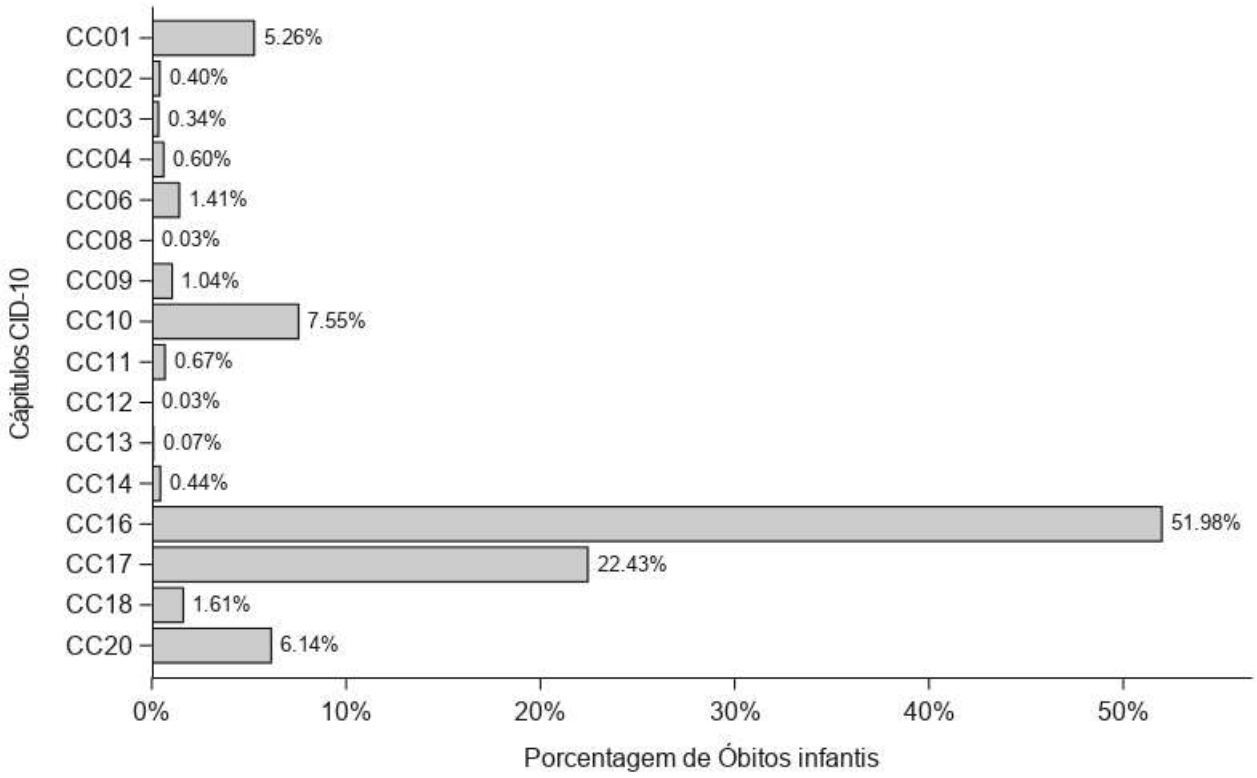


Figura 2.12: Distribuição do número de casos segundo capítulos da CID-10, RJ-2016

A Figura 2.12 e Tabela 2.2.5 apresentam a distribuição dos óbitos infantis segundo capítulos da CID-10 em 2016. Verifica-se que os capítulos XVI (51,98%) e XVII (22,43%) possuem o maior percentual de ocorrência de óbitos infantis, respectivamente. Em seguida têm-se os capítulos X, XX e I, em terceiro, quarto e quinto lugares, respectivamente. Esses possuem percentuais maiores que 5%. É possível notar que os capítulos V,VII, XV, XIX, XXI e XXII não possuíram registro, logo, não foi identificada nenhuma classificação de doença no registro da causa básica do óbito infantil para o estado do RJ no ano de 2016. Os demais capítulos apresentam percentuais de classificações inferiores a 2%.

Tabela 2.7: Distribuição do número de casos segundo capítulos da CID-10

Capítulos	**Frequência**	**Porcentagem**
Capítulo I	157	5,26%
Capítulo II	12	0,40%
Capítulo III	10	0,34%
Capítulo IV	18	0,60%
Capítulo VI	42	1,41%
Capítulo VIII	1	0,03%
Capítulo IX	31	1,04%
Capítulo X	225	7,55%
Capítulo XI	20	0,67%
Capítulo XII	1	0,03%
Capítulo XIII	2	0,07%
Capítulo XIV	13	0,44%
Capítulo XVI	1550	51,98%
Capítulo XVII	669	22,43%
Capítulo XVIII	48	1,61%
Capítulo XX	183	6,14%

Ressalta-se que o capítulo XVI armazena as classificações de agrupamentos de doenças relacionadas às categorias P00 até a P96, portanto, 51,98% dos óbitos infantis estão relacionados a algumas infecções originadas no período perinatal. Já o segundo maior registro, capítulo XVII, indica que 22,43% da causa dos óbitos infantis está associado a malformações congênitas, deformidades e anomalias cromossômicas (Tabela 2.4).

A terceira causa em destaque, capítulo X, com 7,55%, compreende as doenças do aparelho respiratório que são os agrupamentos de doenças que compreendem as categorias J00 a J99.

Os capítulos VIII e XII apresentaram apenas 1 registro e estão relacionado a doenças do ouvido e da apófise mastoide e doenças da pele e do tecido subcutâneo, respectivamente.

2.3 Resultado dos métodos

Nessa seção serão apresentados os resultados obtidos com o método Estatística de Varredura Árvore-Espacial e posteriormente os resultados da Estatística *Scan* Circular de Kulldorff e Estatística de Varredura baseada em árvore.

2.3.1 Resultados do método: Estatística de Varredura Árvore-Espacial

Os resultados da estatística de varredura árvore-espacial foram obtidos utilizando o modelo Poisson, conforme descrito na Seção 1.5.1. A implementação foi realizada por meio da macro *%treescancircular*, desenvolvida no *Software SAS* versão *9.04.01M5* por meio do módulo *SAS/IML* (*interactive matrix language*). Os módulos *SAS/STAT* e *SAS/GRAPH* também foram utilizados para as análises descritivas e confecção de gráficos e mapas.

A macro retorna um *data set* com informações do *cluster* árvore-espacial mais verossímil, tais como: a identificação do corte (ID do corte), ou seja, o código do nó da árvore; o código do conjunto de municípios que representam a estrutura espacial; a estatística do teste (logaritmo da razão de verossimilhança).

Outras medidas descritivas importante associadas ao *cluster* árvore-espacial são: número de casos esperados no corte/nó e na zona (conjunto de regiões); número de casos total do corte/nó e na zona (nesse caso, são desconsiderados o número de casos que pertencem ao nó, mas estão fora do conjunto de regiões); tamanho do *cluster*-estrutura espacial; o valor crítico e *P-valores*, mensurados com bases nas simulação de Monte Carlo, sob a hipótese de ausência de *cluster* árvore-espacial.

Também é possível listar e armazenar as informações de todos os *clusters* e em seguida identificar aqueles que são significativos com base no valor da estatística de razão de verossimilhança por meio do valor crítico e/ou *p-valor*.

A Estatística de varredura árvore-espacial foi aplicada aos dados de óbitos infantis segundo a CID-10, contendo 2892 registros distribuídos por municípios do Rio de Janeiro em 2016. Para análise fornecemos três conjuntos de dados, o primeiro contendo as informações das coordenadas geográficas para o Rio de Janeiro e a população para cada município, o segundo com os *ID*s dos nós filhos e os *ID*s dos nós pais, e o terceiro contendo os dados do número de óbito para cada nó da árvore e município, conforme Figura 2.8.

Na varredura espacial delimitou-se um limite máximo de 50% da população total do estado Rio de Janeiro. Dessa forma, o algoritmo de varredura árvore-espacial avaliou 713 possíveis cortes da árvore da CID-10 para cada subconjunto de municípios até o limite de 50% da população total.

Portanto, a função de verossimilhança é maximizada para todas as janelas geográficas e os possíveis cortes da árvore, de forma que o corte e a zona com o valor máximo de verossimilhança é dado como *cluster* mais verossímil. Logo, quanto maior for o valor do logaritmo da Razão de verossimilhança, mais forte será a evidência contra H_0 (ausência de *cluster*).

Além do *cluster* mais verossímil, o algoritmo também identifica *clusters* secundários, que são todos os demais pares g, z que possuem altos valores da razão de verossimilhança. Porém, muitos desses pares g, z secundários podem estar relacionados

ao *cluster* mais verossímil, tendo interseção nos nós da árvore e na região. Partindo do *cluster* árvore-espacial mais verossímil, é de interesse apenas os demais cortes e conjuntos de regiões que são significativos e pertencem aos seguintes cenários:

- não possuem nenhuma interseção com relação aos nós descendentes do corte mais verossímil e está localizados em regiões diferentes do *cluster* que possui o maior valor da RV;

- possui interseção com corte da árvore, mas está localizado em regiões diferentes das do *cluster* mais verossímil;

- está localizado no mesmo conjunto de regiões do corte mais verossímil, mas está em um nó da árvore diferente, que não é descendente do corte mais verossímil.

O nível de significância utilizado nesse estudo foi $\alpha = 5\%$. Inicialmente encontramos 19 *clusters* árvore-espaciais significativos. Ordenando os pares g, z pelo logaritmo da Razão de verossimilhança, identificam-se todas as sobreposições espaciais e na estrutura hierárquica, restando apenas dois *clusters* árvore-espaciais significativos distintos, conforme tabela 2.8.

Tabela 2.8: *Clusters* árvore-espaciais identificados

ID Corte	**ID das regiões**	$\log RV$	**Casos Nó e Regiões**	**Casos Nó**	**Casos Esperados**	**População Região**	**Tamanho** *Cluster* **Espacial**	**P-valor**
P209	11,14,15,17,19,31,32,33,36, 46,58,62,69,71,72,74,76,90	39,62	27	59	3,22	908.037	18	$< 10^{-3}$
RR	9,25,39,43,44,47,50,75	21,30	724	2982	578,89	3.229.547	8	$< 10^{-3}$

A Tabela 2.8 resume os dois *clusters* árvore-espaciais identificados por meio da estatística de varredura árvore-espacial. O primeiro *cluster* é formado pelo par ($z = 11, 14, 15, 17, 19, 31, 32, 33, 3646, 58, 62, 69, 71, 72, 74, 76, 90$ e $g = P209$). Logo, a estrutura espacial é composta de 18 municípios do Rio de Janeiro e a estrutura hierárquica é formada pelo corte P209, que é um nó folha, ou seja, uma subcategoria de classificação de doença (menor nível). Já o segundo é formado por 8 municípios espacialmente diferentes do primeiro, e o corte é identificado pelo código RR, correspondente ao nó raiz, que é o maior nível da árvore e armazena todos os demais nós.

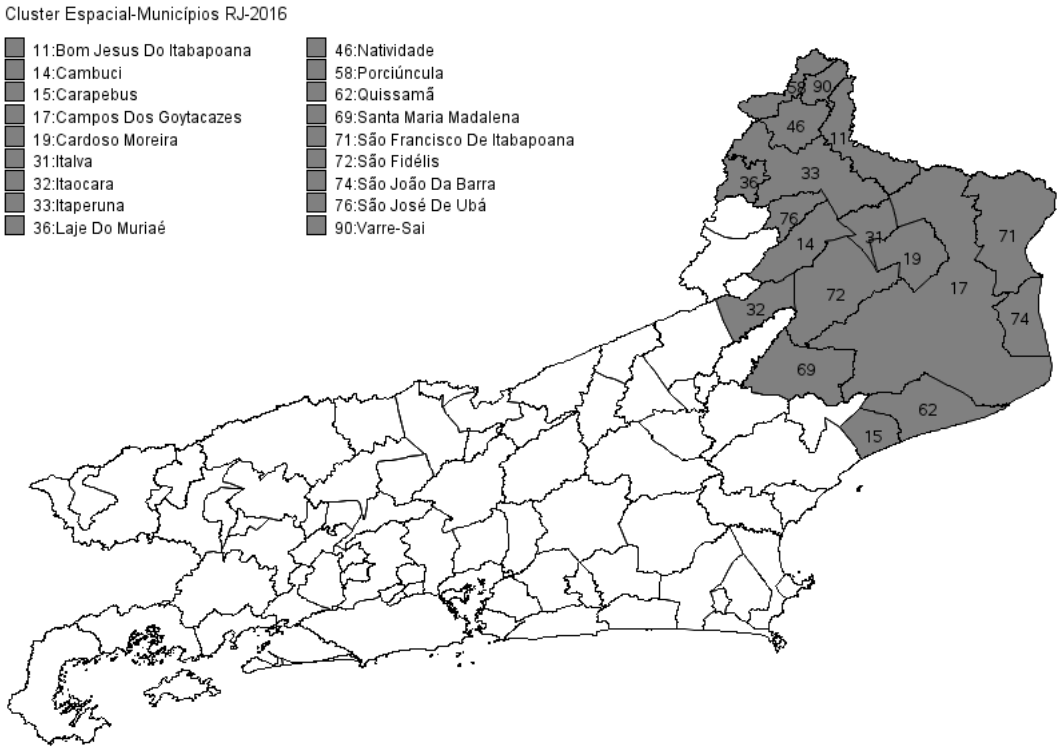


Figura 2.13: *Cluster* Espacial relacionado ao nó P209

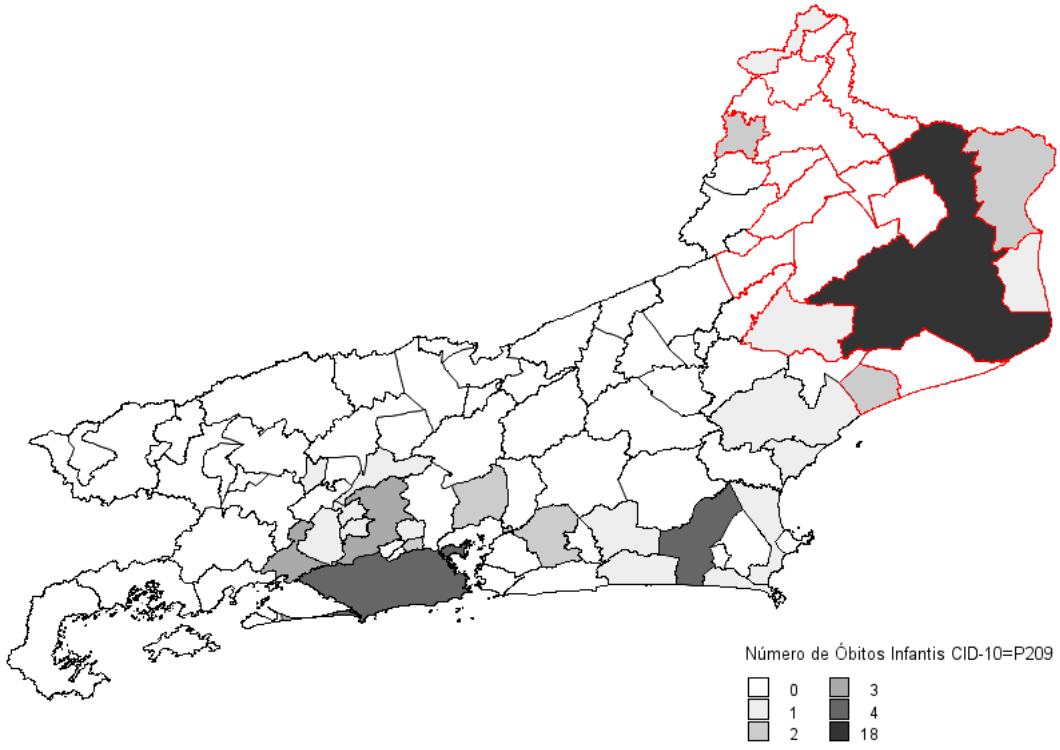


Figura 2.14: Distribuição do número de óbitos associados ao nó **P209** por municípios do RJ em 2016

A Figura 2.13 mostra a localização espacial do primeiro *cluster*. Esse conjunto de municípios possui 908.037 habitantes. Na Figura 2.14 tem-se a distribuição do número de óbitos infantis associado à subcategoria de doenças P209 para o estado do RJ. Nota-se que o número de óbitos infantis associados essa folha oscilou entre 0 e 18 e existe uma maior concentração de óbitos em algumas das regiões identificadas na estrutura espacial. É possível notar que o município 17 (Campos dos Goytacazes) apresentou o maior número de óbitos para a causa básica P209. Além disso, algumas regiões dentro do *cluster* espacial não apresentam ocorrência de óbitos associados a subcategoria P209, mas entraram na janela de varredura circular, visto que estão rodeadas pelos municípios com os maiores número de óbitos associados à classificação P209.

Verifica-se que o número esperado de ocorrências óbitos do nó P209 nos 18 municípios acima é $\theta_g(z)$= 3,27 casos, enquanto o número observado de casos nesse nó nos 18 municípios é $C_g(z) = 27$, ou seja, é aproximadamente 8 vezes o valor esperado sob a hipótese de homogeneidade de casos desse ramo em todos os municípios. Esse fato fortalece a hipótese que os óbitos infantis relacionado a P209 (hipóxia intra-uterina não especificada) são mais incidentes dentro desse conjunto de municípios do que esse mesmo nó fora dessas regiões (área destacada em vermelho, Figura 2.14).

A Figura 2.14 ilustra a ideia principal do algoritmo, ou seja, a detecção do ramo P209 (anomalia) em algum lugar no espaço (conjunto de municípios). O comportamento ilustrado nesses dois mapas produzem evidências de que o número de óbitos associados ao nó do ramo P209 nesses municípios não está distribuído totalmente ao acaso. Esse fato é confirmado com base no *P-valor* que é inferior a 0,01%, ou seja, de fato rejeitam fortemente a hipótese nula de ausência de *cluster* árvore-espacial. Assim, ao nível $\alpha = 5\%$, há evidências para se rejeitar o fato de que a probabilidade da ocorrência de óbitos infantis associado à categoria é a mesma em qualquer município.

Observa-se na Figura 2.15 a distribuição empírica do logaritmo da razão de verossimilhança obtida por meio de 1000 simulações de Monte Carlo. Nota-se que o valor crítico foi 12.006, portanto a estatística do teste (máximo da RV) pertence à região crítica, evidenciando que a hipótese nula pode ser rejeita para o primeiro par (z, g) mostrado na Tabela 2.8.

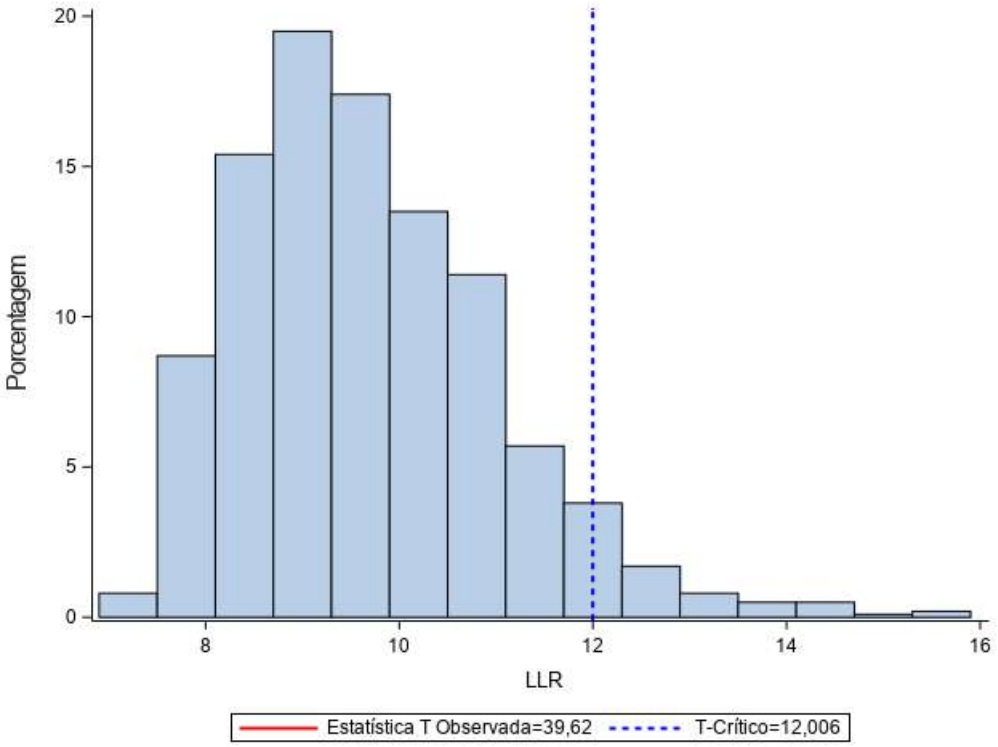


Figura 2.15: Distribuição da estatística de razão de verossimilhança empírica sob H_0

Com relação à estrutura hierarquia o nó P209 (hipóxia intra-uterina não especificada) é uma subcategoria da árvore (menor nível) e pertence à categoria P20 que está associada à hipóxia intra-uterina (4º nível), que por sua vez está no mesmo ramo do capítulo XVI (Algumas afecções originadas no período perinatal), que é o capítulo onde houve o maior percentual de registro de óbitos. Ressalta-se que muitos *clusters* identificados com altos valores da RV estão altamente ligados com relação à estrutura hierárquica, ou seja, normalmente são localizados nos ramos/nós descendentes do *cluster* mais verossímil.

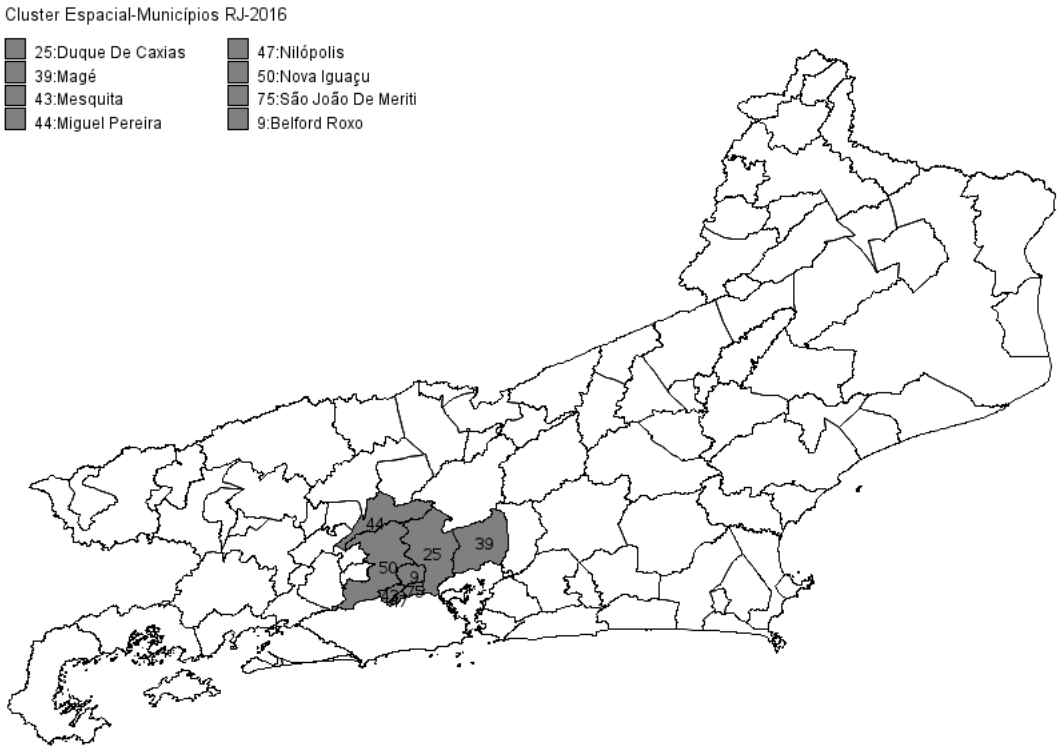


Figura 2.16: *Cluster* Espacial relacionado ao nó RR

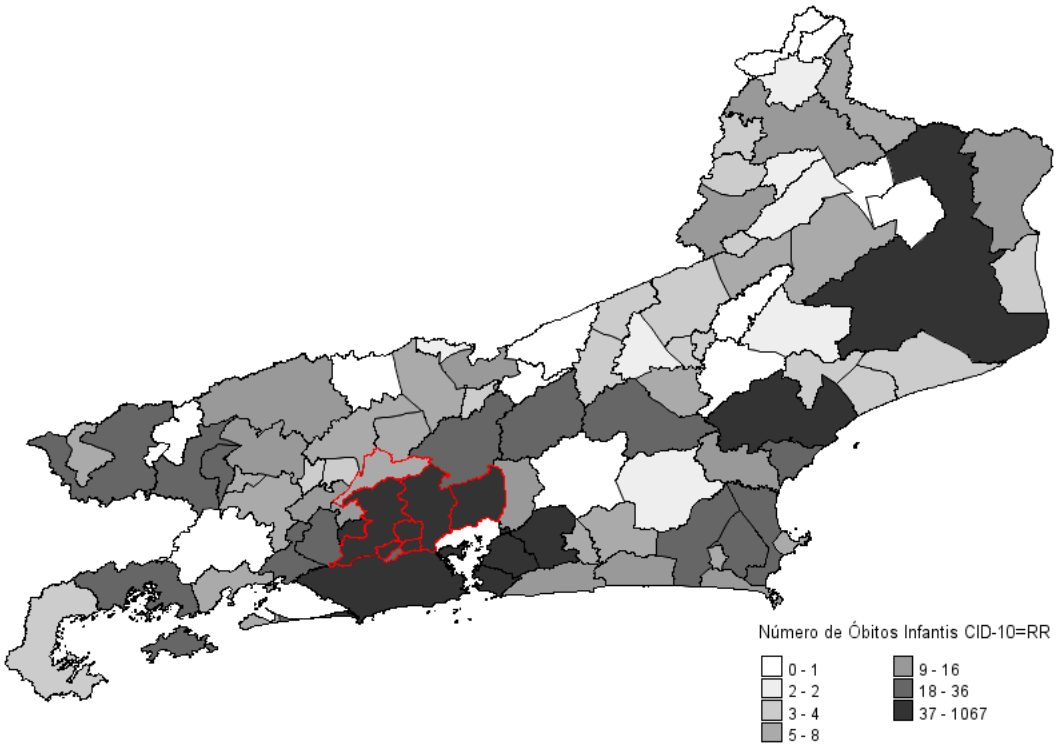


Figura 2.17: Distribuição do número de óbitos associados ao nó **RR** por municípios do RJ em 2016

O segundo *cluster* é formado pelo nó Raiz e um conjunto de oitos municípios.

Destaca-se na Figura 2.17 que o corte RR, é um ramo do nível mais alto da árvore (1º nível), portanto, armazena a informação da quantidade de óbitos infantis para o estado do Rio de Janeiro em 2016. Logo, é possível concluir que o *cluster* identificado por esse corte é puramente espacial, visto que apenas a estrutura espacial é utilizada, e que o nó RR está agregando todos os capítulos de doenças (2º nível) da CID-10 por municípios. O número esperado de óbitos infantis dentro desse conjunto de municípios (Duque de Caxias, Magé, Mesquita, Miguel Pereira, Nilópolis, Nova Iguaçu, São João De Meriti e Belford Roxo) que contém 3.229.547 habitantes (área destacada em vermelho no mapa) corresponde a 578,89. Já o número total de caso dentro dessa zona é 724, ou seja, é 25,06% maior do que o esperado supondo que os óbitos infantis são distribuídos igualmente sobre esse mapa.

Com base na Tabela 2.8, tem-se que o *p-valor* é menor que 0,01%, então existem evidências para rejeitar a hipótese nula com $\alpha = 5\%$. Portanto, pode-se concluir que esse conjunto de zonas forma um *cluster* atípico.

Com base nessas análises, nota-se que quando a estrutura hierárquica contém um único nó raiz, o corte correspondente à raiz que compreende toda a árvore torna-se equivalente a analisar o número total de óbitos infantis segundo as regiões do mapa. Portanto, como a estatística de varredura árvore-espacial verifica todos os possíveis cortes da árvore para todos os conjuntos de zonas, a detecção de um *cluster* que compreende toda a árvore equivale à detecção de um *cluster* puramente espacial, sendo equivalente ao processo realizado pela estatística *Scan* Circular de kulldorff.

O resultado do algoritmo *Scan* circular, conforme descrito na Seção 3.2, será apresentado para fins de comparação.

A tabela 2.9 apresenta alguns dos *clusters* significativos retornados pela estatística de varredura árvore-espacial, mas todos possuem algum tipo interseção no espaço e ou na árvore. Porém, a ideia é evidenciar algumas características observadas do método.

Tabela 2.9: Clusters árvore-espaciais significativos.

ID Corte	**ID das regiões**	$\log RV$	**Casos Nó e Regiões**	**Casos Nó**	**Casos Esperado**	**População Região**	**Tamanho *Cluster* Espacial**	**P-valor**
P209	11,14,15,17,19,31,32,33,36, 46,58,62,69,71,72,74,76,90	39,62	27	59	3,22	908037	18	$< 10^{-3}$
GG143	17,71	29,53	59	548	17,40	528426	2	$< 10^{-3}$
CC16	17,71	25,60	106	1550	49,23	528426	2	$< 10^{-3}$
RR	9,25,39,43,44,47,50,75	21,30	724	2982	578,89	3229547	8	$< 10^{-3}$
CC10	9,25,39,50,75	19,60	78	225	38,88	2875353	5	$< 10^{-3}$
W849	9,25,43,47,50,61,75	16,94	36	74	13,84	3112898	8	$< 10^{-3}$
J18	9,25,39,43,44,47,50,54,75	16, 85	58	142	27,79	3256486	9	$< 10^{-3}$
GG088	9,25,39,43,44,47,50,54,75	16, 21	60	151	29,55	3256486	9	$< 10^{-3}$
W84	9,25,39,43,44,47,50,54,75	15,51	40	88	16,95	3204692	9	$< 10^{-3}$
J189	9,25	15,28	31	125	10,37	1381058	2	$< 10^{-3}$
GG001	2,11,14,15,17,18,19,20,22,23,31,32, 33,36,37,38,45,46,58,62,69,70,71,72, 76,79,87,90	12,28	14	39	3.042	1298002	28	0.003

Uma característica importante observada no algorítimo de varredura árvore-espacial está ligada à identificação da estrutura espacial associada ao ramo, visto que permite identificar um nó de menor nível, isto é, uma subcategoria (P209) com apenas 27

casos em um conjunto de zonas, e apresenta um maior valor para a estatística do teste do que um agrupamento de doenças que engloba essa categoria e possui mais ocorrências $GG143$, mas estão localizados em um subconjunto dessas regiões (Tabela 2.9).

Verifica-se na Tabela 2.9, com relação à estrutura da árvore, que muitos *clusters* são identificados no mesmo ramo da árvore, mas cada um contribuiu de maneira significativa para obtenção dos maiores valores da estatística do teste. Analisando os 5 primeiros cortes é possível ver que o primeiro e mais verossímil é descendente dos três cortes imediatamente abaixo. Visto que a folha P209 pertence ao capitulo XXVI que é composto dos agrupamentos GG140 a GG149, logo contempla o corte GG143. Além do aspecto hierárquico, os conjuntos de regiões para cada corte são os mesmos conjuntos de municípios do corte mais verossímil, ou o nó descendente é identificado em um conjunto de município que possui alguma sobreposição. Mas isso não impede de algum corte ser capturado em um conjunto diferente de regiões que, no caso do segundo *cluster* mais verossímil, o corte RR, que por ser a raiz da árvore, tem intersecção com todos os demais nós, mas está em un conjunto de municípios diferentes do primeiro, por isso foi selecionado.

Observa-se que o corte GG001 é o único depois do mais verossímil que não possui nós descendentes diretos, com exceção do nó raiz, visto que pertence ao ramo do capítulo I. Com base nesses aspectos, nota-se que os *clusters* árvore-espaciais encontrados são coerentes.

O método de varredura árvore-espacial é bastante intensivo computacionalmente. Para essa estrutura, uma árvore com 713 ramos (possíveis cortes) e 92 regiões, o tempo de execução foi de aproximadamente 8 minutos. Para as simulações sob H_0 foram utilizadas 5 máquinas com as seguintes configurações: *Windows Server* 2008 (64 bits), 4 cores 48GB RAM. O tempo de cada réplica oscilou entre 8 e 10 minutos.

2.3.2 Resultados da Estatística Scan Espacial de Kulldorff

O algoritmo *Scan* circular foi aplicado aos 2982 casos de óbitos infantis distribuídos nos 92 municípios do Rio de Janeiro. Utilizou-se a estatística*Scan* circular com base no modelo Poisson. O método foi executado por meio da macro *%ScanCircualr* desenvolvida em linguagem SAS.

Tabela 2.10: *Cluster* espacial identificado

ID das regiões	$\log RV$	**Casos Observado**	**Casos Esperados**	**População Região**	**Tamanho *Cluster* Espacial**	**P-valor**
9,25,39,43,44,47,50,75	21,30	724	578,89	3.229.547	8	$< 10^{-3}$

Verifica-se na Tabela 2.10 que o algoritmo *Scan* de Kulldorff retornou como *cluster* mais verossímil a mesma zona encontrada pelo algoritmo árvore-espacial para o corte relativo ao nó raiz RR (ver Tabela 2.8).

A estatística do teste para essa zona com 8 municípios foi $LRV = 21,30$, o número de caso esperado supondo a não existência de *cluster* foi de 578,89 óbitos infantis. A localização espacial desse conglomerado e a distribuição de óbitos infantis são apresentadas nas Figuras 2.16 e 2.17, respectivamente.

Para averiguar a significância do *cluster* foi gerada a distribuição empírica da estatística do teste com base em 1000 simulações de Monte Carlo.

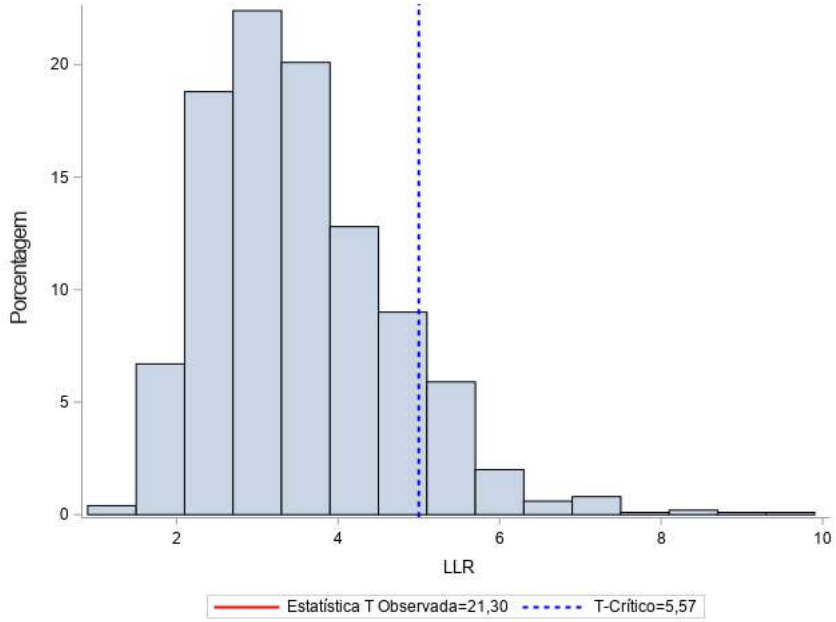


Figura 2.18: Distribuição empírica sob H_0

Observa-se na Figura 2.18 que o valor crítico, percentil 95, é igual a 5,57, portanto, ao nível $\alpha = 5\%$, têm-se evidências para rejeitar a hipótese nula de ausência de *cluster* espacial. Além disso, o p-valor empírico foi menor que 0,01%.

Ressalta-se que a distribuição empírica sob a hipótese de ausência de *cluster* árvore-espacial (Figura 2.15) é levemente mais assimétrica do que a distribuição empírica sob ausência de *cluster* espacial (2.18). No primeiro cenário são gerados casos sob H_0 de acordo com a distribuição multinomial condicionada ao número de casos de cada folha da árvore.

2.3.3 Resultado da Estatística de Varredura baseada em árvore

A estatística de varredura baseada em árvore foi aplicada aos dados de óbitos infantis para o estado do Rio de Janeiro. Para aplicação do algoritmo considerou-se a mesma estrutura hierárquica apresentada da Tabela 2.5. Portanto, analisaram-se todos os possíveis cortes de uma árvore com 713 nós, sendo 410 nós folha.

O método foi executado por meio da macro *%treescan* desenvolvida em linguagem SAS. A estatística de varredura baseada em árvore foi utilizada com base no modelo Poisson condicionado ao número de casos. Portanto, considerou-se n_g como o número

esperado de casos em cada ramo sob H_0, em uma árvore com 410 folhas e 2982 óbitos infantis.

Na Tabela 2.11 são apresentados os *clusters* encontrados somente na estrutura hierárquica, isto é, correspondentes aos ramos da árvore da causa básica dos óbitos infantis, onde a probabilidade de ocorrer óbitos infantis nessa classificação é maior que nos demais. Ressalta-se que o método retornou mais de 50 cortes significativos ao nível de 5% de confiança, considerando os nós descendentes nos vários níveis.

O Corte mais verossímil é uma subcategoria (P369), que está associada à categoria P36, e pertence ao ramo do capítulo XVI. Esse capítulo engloba os agrupamentos GG140 a GG149 e compreende XVI (algumas afecções originadas no período perinatal). A subcategoria P369 corresponde especificamente às classificações de óbitos segundo Septicemia bacteriana não especificada do recém-nascido. Todos os cortes possuem *p-valor* inferiores a 0.01%.

Tabela 2.11: Lista de *clusters* encontrados com base na estatística baseada em árvore

ID Corte	$\log RV$	**Casos Observados**	**Casos Esperados**	***P-valor***
P369	601,49	236	7,27	$< 10^{-}3$
J189	240,14	125	7,27	$< 10^{-}3$
Q249	203,38	112	7,27	$< 10^{-}3$
W849	105,69	74	7,27	$< 10^{-}3$
P209	72,23	59	7,27	$< 10^{-}3$
A41	70,13	58	7,27	$< 10^{-}3$
R99	35,64	40	7,27	$< 10^{-}3$

Ressalta-se que o nó P209 é o corte mais verossímil encontrado pela estatística de varredura árvore-espacial, também foi identificado como um corte significativo pela estatística baseada em árvore, mas olhando para o aspecto apenas hierárquico é uma informação redundante, visto que o corte mais verossímil é o nó folha P360 que está no mesmo ramo capítulo XVI.

Em seguida, destaca-se um corte do quinto nível, o nó J189, que pertence ao capítulo X-Doenças respiratórias. Essa subcategoria está relacionada a causas de Pneumonia não especificadas. O capítulo X (nó de 2º nível) abrange as categorias de J00 a J099.

Ressalta-se que todos os *clusters* mostrados na Tabela 2.11 na estrutura hierárquica são nós do último nível (onde são armazenadas as informações sobre o evento de interesse), portanto o número de casos esperados são iguais 7,27 óbitos, considerando que a probabilidade de classificação da causa básica do óbito fosse igual.

O terceiro corte identificado (Q249) pertence ao capítulo XVII que possui causa de óbitos relacionadas a Malformações congênitas, deformidades e anomalias cromossômi-

cas. Esse capítulo contém as categorias Q00 a Q99. A subcategoria Q249 está relacionada à Malformação não especificada do coração.

Os dois últimos *clusters* identificados pertencem aos capítulos I e XVIII, respectivamente. O corte identificado pelo código R99 é uma categoria e está relacionada a Outras causas mal definidas e as não especificadas de mortalidade. O capítulo XXIII-Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte, contempla as categorias RR00 a R99 (de Classificação de Doenças CBCD, 2019). Esse corte revela um aspecto interessante, visto que o nó R99 armazena qualquer causa quando o terno especifico não tiver relacionado a um órgão para os casos de cirurgias, ou não for possível identificar. Isso pode indicar que ainda existe um percentual elevado de classificações mal definidas.

2.4 Uma aplicação da Estatística de Varredura Árvore-Espacial considerando o número de nascidos vivos como população em risco

Inicialmente, no estudo de identificação de *clusters* árvore-espacias foi considerado a população em risco em cada região como a população do Rio de Janeiro em 2016 por município. Nessa Seção serão apresentados os resultados dos métodos árvore-espacial e o *scan* circular considerando como população em risco o número de nascidos vivos segundo municípios do Rio de Janeiro em 2016. A ideia é verificar se existe mudanças significativas se comparado ao resultados da Seção 2.3.1. Nesse cenário, é possível analisar a taxa de mortalidade infantil associada a cada ramo da árvore da CID-10(grupo de doença ou doença específica).

Os dados sobre o número de nascidos vivos faz parte do Sistema de Informações sobre Nascidos Vivos (SINASC) desenvolvido pelo DATASUS. O banco de dados é disponibilizado pelo sítio do DATASUS. Os dados são obtidos por meio do formulário da Declaração de Nascidos Vivos (DN).

Tabela 2.12: Medidas resumo da variável número de nascidos vivos

	Média	**Desv.Padrão**	**Coef.Var**	**1° Quartil**	**Mediana**	**3° Quartil**	**Mínimo**	**Máximo**
NASVIVO	2381,78	8877,17	3,72	200,5	482	1847	88	83166

Verifica-se na Tabela 2.12 e Figura 2.19 que o número de nascidos vivos oscilou entre 0 e 83.166 mil e 75% do número de nascimentos são inferiores a 1847.

Na Figura 2.19 tem-se a distribuição do número de nascimentos segundo o código do município de residência da mãe.

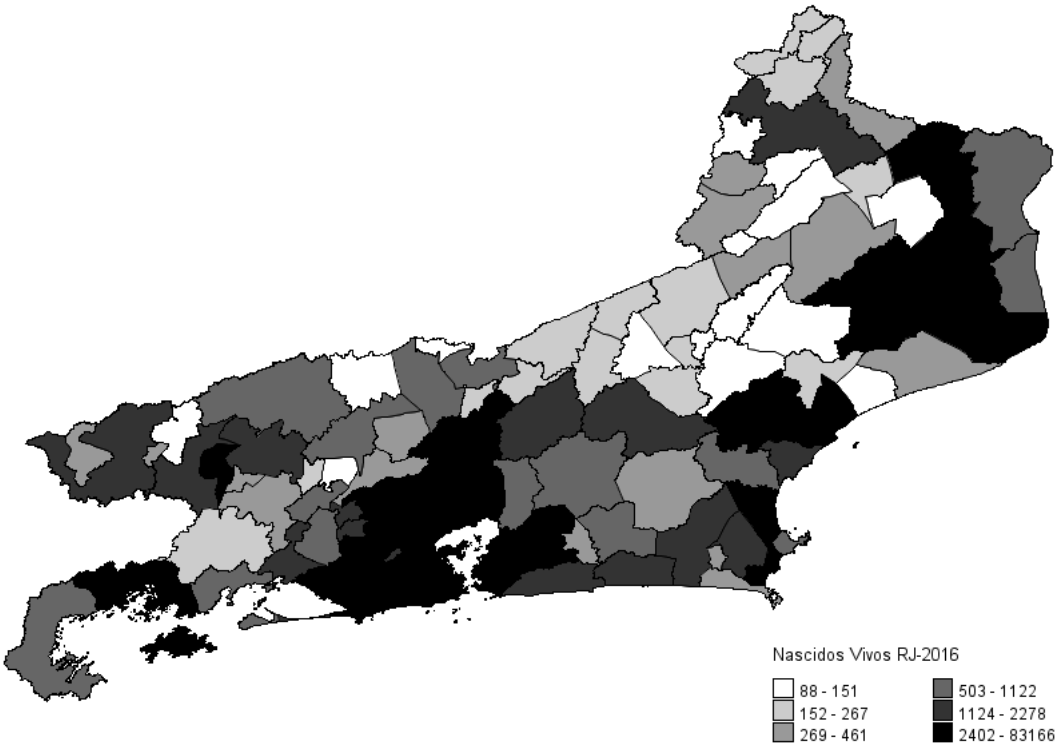


Figura 2.19: Distribuição do número de nascidos vivos segundo municípios do Rio de Janeiro-2016

No estado do Rio de Janeiro em 2016 registrou-se 219.124 mil nascimentos. Ressalta-se que no banco de dados DNRJ2016.dbc obteve-se 219.129 mil registros de nascimentos para 2016, mas cinco desses registros correspondem a código de município de residência ignorado (quando município de residência da mãe não é conhecido). Portanto não foram considerados na análise.

2.5 Resultados: Método Árvore-Espacial

Verifica-se na Tabela 2.13 que foram identificados somente dois candidatos a *cluters* árvore-espaciais distintos. Considerando um nível $\alpha = 5\%$, apenas para o primeiro par (g, z) tem-se fortes evidências contra a hipótese de homogeneidade de casos do ramo para todos os municípios, visto que o *p-valor* é menor que 0,001. O segundo *cluster* identificado é formado pelo par ($g = CC10$ e $z = 9, 25$) e apresenta *p-valor*=0,0049, está na fronteira do $\alpha = 0,05$ (nível de confiança escolhido). Mas, consideraremos como um *cluster* árvore-espacial significativo. Ressalta-se, que poderia ser necessário realizar um número maior de simulações para averiguar a significância desse *cluster*.

O *cluster* árvore-espacial mais verossímil obtido considerando o número de nascidos vivos como a população em risco em cada município do Rio de Janeiro é formado pelo par: ($g = P209$ e $z = 11, 14, 15, 17, 19, 31, 32, 33, 3646, 58, 62, 69, 71, 72, 74, 76, 90$) que corresponde ao mesmo ramo da árvore da CID-10 (subcategoria de doença específica) e o

conjunto de municípios obtidos na Seção 2.3.1.

A localização do *cluster* e a distribuição do número de óbitos associados ao ramo P209 podem ser analisados nas Figuras 2.13 e 2.14.

Tabela 2.13: *Clusters* árvore-espaciais identificados considerando número de nascidos vivos

ID Corte	**ID das regiões**	$\log RV$	**Casos Nó e Regiões**	**Casos Nó**	**Casos Esperados**	**Nº Nascidos Vivos**	**Tamanho *Cluster* Espacial**	**P-valor**
P209	11,14,15,17,19,31,32,33,36, 46,58,62,69,71,72,74,76,90	38,48	27	59	3,37	12.519	18	$< 10^{-3}$
CC10	9,25	15,01	49	225	21,35	20.798	2	0,0049

Verifica-se que o número de óbitos infantis esperados para o ramo P209 nos 18 municípios é $\theta_g(z)$= 3,37 casos, enquanto o número observado de casos desse nó nos 18 municípios é $C_g(z) = 27$, ou seja, é aproximadamente 8 vezes o valor esperado sob a hipótese de homogeneidade de casos desse ramo em todos os municípios.

Os resultados demonstram que o número de óbitos infantis relacionados a folha P209 (hipóxia intra-uterina não especificada) são mais incidentes dentro desse conjunto de municípios do que para esse mesmo ramo(P209) fora dessas regiões.

Observa-se na Tabela 2.13 que o segundo *cluster* árvore-espacial mais verossímil é formado pelo ramo CC10 da árvore da CID-10 e os municípios (9: Duque de Caxias e 25: Belford Rocho). A localização do *cluster* espacial é ilustrada na Figura 2.20.

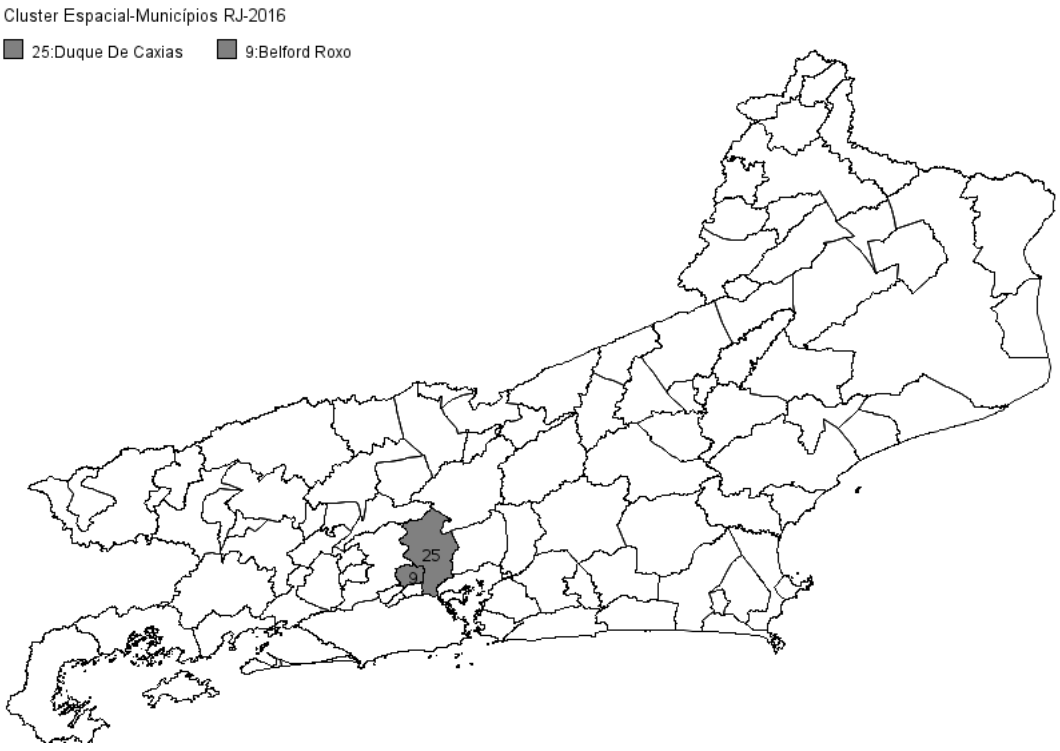


Figura 2.20: Localização espacial do *cluster* relacionado ao nó CC10

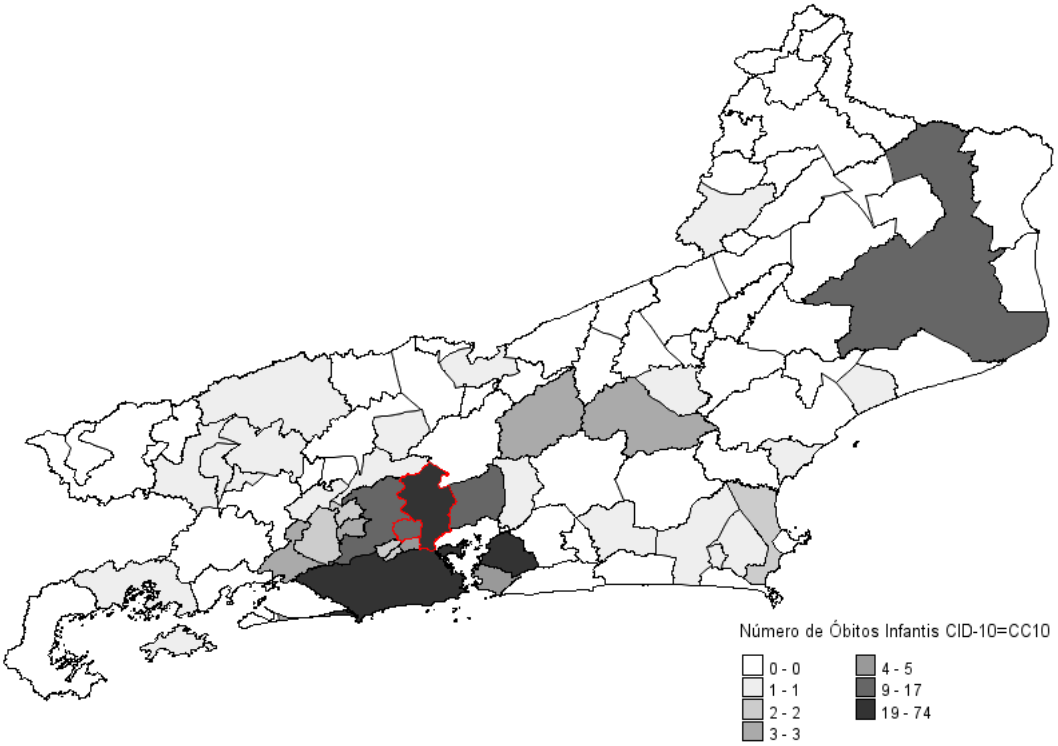


Figura 2.21: Distribuição do número de óbitos infantis relacionados ao nó CC10

Verifica-se na Figura 2.21 que o número de óbitos infantis associados ao nó CC10 oscilou entre 0 e 74 casos. O nó CC10 corresponde ao capítulo X e armazena os agrupamentos de doenças do sistema respiratório. Esses dois municípios possuem 20.798 nascidos Vivos.

O número esperado de óbitos infantis do nó CC10 nos municípios (17 e 71) é igual a 21,31. Já o número de óbitos infantis observados desse nó nos dois municípios é igual a 49, ou seja, é aproximadamente 2,3 vezes o número de casos esperado sob H_0 (ausencia de *cluster* árvore-espacial). Nota-se que número de óbitos infantis associados a doenças do aparelho respiratório é mais frequente para esses dois municípios do que para esse mesmo ramo (CC10) fora desses municípios.

Ressalta-se que o estudo considerando o número de nascidos vivos permitiu identificar apenas um *cluster* árvore-espacial secundário diferente dos detectados considerando a população do Rio de Janeiro como população em risco. O par ($g = CC10$ e $9, 25$) produz uma informação importante para as políticas voltadas a indicadores de mortalidade infantil. Portanto, pode-se investigar os fatores ou variáveis que podem estar relacionados a taxa de mortalidade infantil associado a doenças do sistema respiratório nos municípios 9 e 25 (Figura 2.21).

2.6 Resultados: Estatística *Scan* Circular

O método *Scan* circular foi aplicado aos 2982 casos de óbitos infantis, equivalente ao número de óbitos do nó raiz (RR, ver Seção 2.3.1), considerando a população em risco com o número de nascidos vivos. Utilizou-se a estatística *Scan* circular com base no modelo Poisson, e limitou-se a 50% do total de nascidos vivos (109.562).

Tabela 2.14: *Cluster* espacial identificado (número de nascidos vivos)

ID das regiões	$\log RV$	**Casos Observado**	**Casos Esperados**	**Nº Nascidos Vivos**	**Tamanho *Cluster* Espacial**	**P-valor**
17,71	10,15	159	109.6	8.059	2	$< 10^{-3}$

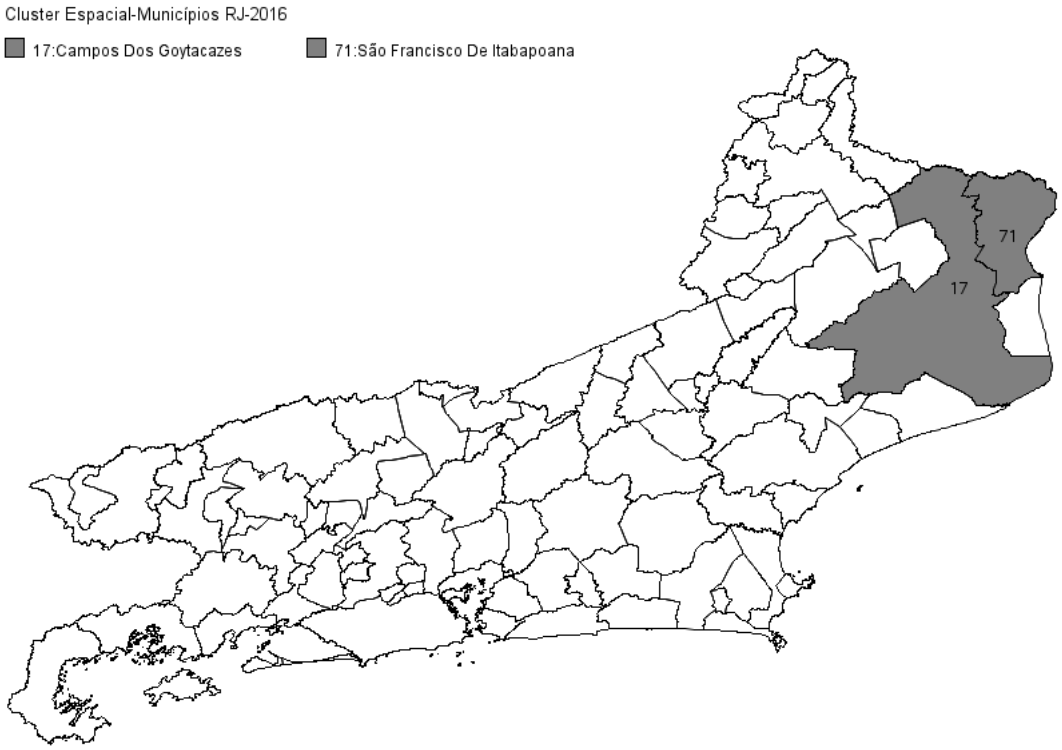


Figura 2.22: Localização do *cluster* espacial

Verifica-se na Tabela 2.14, que o método *scan* circular detectou um *cluster* espacial significativo ao nível de 5%, com base no *p-valor*(<0,001). O número de óbitos infantis é 45,07% maior que o número de casos esperados sob a hipótese nula (homogeneidade de casos no mapa). Isso traz evidência que os dois municípios detectados formam um conglomerado atípico.

Na Figura 2.22 tem-se a localização do *cluster* detectado, que é formado pelos municípios Campos Dos Goytacazes e São Francisco de Itabapoana. O número de nascidos vivos dentro do *cluster* é de 8.059.

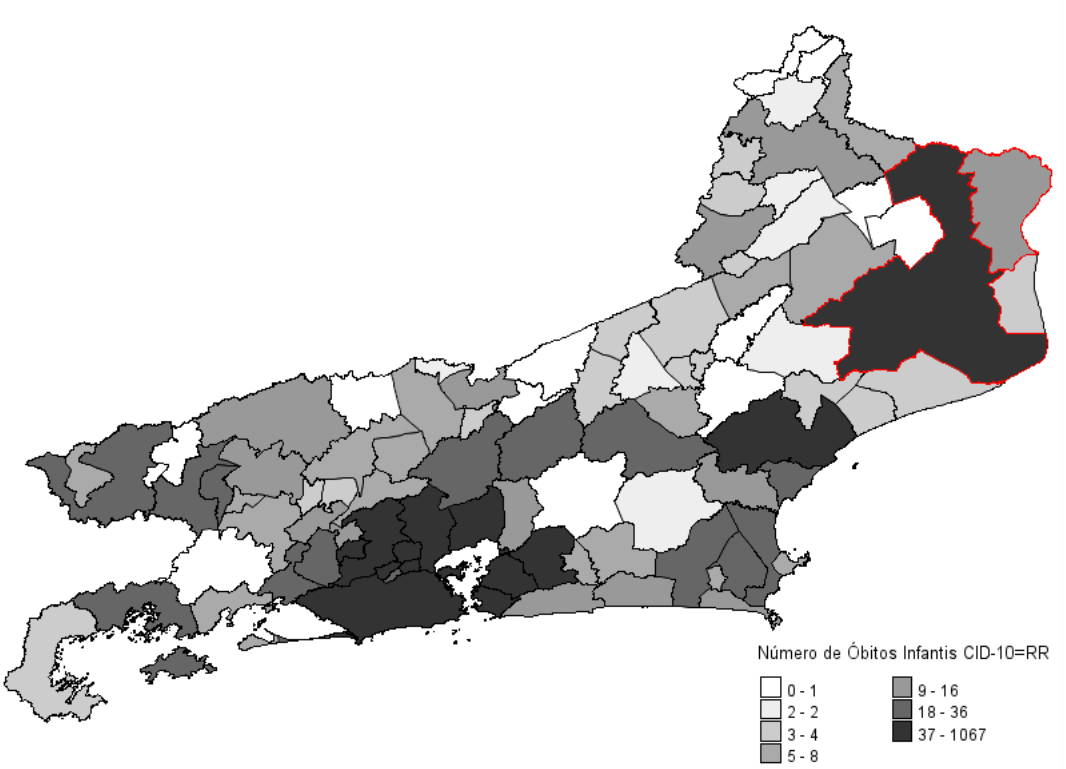


Figura 2.23: Distribuição do número de óbitos infantis para Rio de Janeiro 2016

Observa-se na Figura 2.23, a distribuição de óbitos infantis. Campo de Goytacazes é o município com maior número de óbitos (valor absoluto).

O *cluster* espacial encontrado difere do obtido considerando a população total do Rio de Janeiro (Seção 2.3.2) como população em Risco. O par ($g = RR$ e $Z = 17,71$) não é um *cluster* árvore-espacial significativo.

Capítulo 3

Conclusão e Discussões Finais

O algoritmo proposto nesse trabalho, a estatística de varredura árvore-espacial, mostrou-se eficiente na detecção de *clusters* nas situações em que se pode fazer a combinação de uma estrutura hierárquica e uma estrutura espacial. Dessa forma, as medidas de desempenho propostas produziram bons resultados quanto à capacidade de detectar conjuntos de regiões, quando de fato essas regiões pertencem ao *cluster* verdadeiro e identificaram corretamente o nó da árvore, na qual a probabilidade de um excesso de casos nessas regiões ter ocorrido ao acaso é menor.

Considerando o estudo de simulação, é possível dizer que a técnica proposta pode ser utilizada para incorporar a informação espacial do evento de interesse na estatística de varredura baseada em árvore desenvolvida por Kulldorff et al. (2003a).

O algoritmo mostrou benefícios na aplicação aos dados de óbitos infantis segundo a CID-10 para o estado do Rio de Janeiro em 2016, pois detectou *clusters* árvore-espaciais significativos. Isso trouxe uma informação nova que permite identificar qual a localização de algum grupo da causa básica do óbito ou uma classificação específica (uma folha da árvore), na qual dentro das regiões detectadas tem-se uma incidência de casos no corte detectado maior que o esperado para esse mesmo corte da árvore, fora dessas regiões.

O primeiro *cluster* detectado mostra um conjunto de 18 municípios em que a incidência de óbitos infantis relacionados à subcategoria P209 é muito maior que o esperado. De fato, foi possível identificar uma anomalia em algum lugar no espaço no qual o excesso de casos para as regiões não deve ser atribuído apenas ao acaso. Já o segundo *cluster* permitiu identificar um conjunto de municípios diferentes do primeiro para os quais a incidência de óbitos infantis é bem maior que o esperado.

Ressalta-se que o segundo resultado foi possível pois a estrutura hierárquica contém um único nó raiz que armazena o número de casos total da árvore e a estatística de varredura árvore-espacial avalia todos os possíveis cortes para todas as janelas geográficas. Nesse cenário, a estatística de varredura árvore-espacial equipara-se à estatística *Scan* Circular de Kulldorff.

Outra característica interessante da metodologia é a identificação do *cluster*

mais verossímil como um nó especifico da árvore (menor nível), para um conjunto de regiões ao invés de um nó do mesmo ramo (maior nível), e maior número de casos.

Observou-se que, além disso, ao analisar apenas a estrutura hierárquica, o corte mais verossímil (P369) difere daquele identificado em um conjunto de regiões específicas. Porém, o corte P369 pertence à categoria P36 (sepse bacteriana do recém-nascido) e está associado ao grupo de doenças 'Infecções específicas do período perinatal' que compreende as categorias P35-P39, incluindo as infecções adquiridas no útero ou durante o parto. O corte detectado pelo método árvore-espacial P209 (Hipóxia intra-uterina não especificada) pertence às categorias P20-P29 -Transtornos respiratórios e cardiovasculares específicos do período perinatal. Portanto, os dois cortes pertencem ao mesmo ramo da árvore, que é o capítulo XVI (categorias P00 a P96).

Um aspecto interessante notado na estrutura hierárquica do corte P209 é que essa subcategoria compreende os casos de mortalidade perinatal, ainda podendo ser casos específicos de morte fetal. E a categoria P20 contém três nós filhos: P200 Hipóxia intra-uterina diagnosticada antes do início do trabalho de parto; P201 Hipóxia intra-uterina diagnosticada durante o trabalho de parto e a P209 Hipóxia intra-uterina não especificada (de Classificação de Doenças CBCD, 2019).

Não são atendidos os detalhes específicos das definições. Apesar disso, observa-se uma maior incidência de causas de Hipóxia intra-uterina, que não foi possível identificar antes do trabalho de parto ou durante o trabalho do parto. Esse fator pode ser relevante para as políticas públicas ao se averiguar as variáveis que contribuem para a não identificação da causa, visando torná-la evitável.

De acordo com a OMS, a mortalidade perinatal é dada como o número de óbitos fetais e óbitos de recém-nascidos até a primeira semana de vida. A mortalidade fetal refere-se à ocorrência de óbitos *in utero* em fetos com 28 ou mais semanas de gestação (Nogueira et al., 2013).

A seguir serão apresentados alguns aspectos que podem ser aprimorados ou revisados na técnica proposta:

- A capacidade do método em detectar *clusters* árvore-espaciais, quando eles de fato estão presentes, foram verificados em diferentes cenários por meio das medidas conhecidas para avaliar detecção de *clusters* quanto à estrutura espacial: Sensibilidade, VPP e Poder.

  Destaca-se que apesar de considerarmos que um *cluster* árvore-espacial é definido como um par (g, z), a verificação do desempenho do método foi mensurada para as estruturas hierárquica e espacial separadamente, visto que foi uma maneira de mensurar, para cada estrutura, o quanto foi detectado corretamente. Porém, vale considerar um estudo mais detalhado visando averiguar a possibilidade de mensurar a capacidade de detecção do método de maneira conjunta, isto é, avaliando o par (g, z) por meio dessas medidas já utilizadas ou de outras.

- Outra possibilidade é a construção de outros tipos de cenários quanto à estrutura espacial e hierárquica. Por exemplo, considerar cenários que contenham dois *clusters* espaciais em formatos diferentes e cada um associado a um nó da árvore em ramos distintos. Ressalta-se que como o método varre todos os possíveis cortes, a construção de cenários considerando diferentes configurações hierárquicas não deve trazer mudanças significativas.

- Analisar com mais detalhes a distribuição dos óbitos de acordo com a CID-10 com relação a distribuição geográfica, pois é comum existir classificações que não ocorreram em muitos municípios, e quando esses estão ao redor dos que possui maior quantidade de casos para uma classificação específica, podem entrar na janela de varredura circular. Estudar a possibilidade de outros tipos de janelas de varredura que podem considerar as especificidades desses dados.

- Otimizar as macros desenvolvidas em linguagem *SAS*, visando melhorar o tempo computacional e integrar o código que verifica todas as sobreposições de regiões e as intersecções entre os nós dos *clusters* significativos retornados pelo método.

O entendimento da causa dos óbitos e sua distribuição geográfica mostrados nesse estudo abrem a possibilidade de estender a análise para maiores abrangências geográficas e a possibilidade de trabalhar com a população em risco de cada ramo da árvore.

Diante da adequação da técnica de varredura árvore-espacial é possível pensar em estudos futuros para incorporar a informação temporal na estrutura árvore-espacial, com o intuito de encontrar alguma anomalia em algum lugar no espaço em uma determinada janela temporal.

Referências Bibliográficas

Araújo, T. C. (2012). *Extensão da estatística Scan para detecção de conglomerados espaço-temporais em dados com excesso de zeros.* Master's thesis, Universidade de Brasília. 6

Choynowski, M. (1959). Maps based on probabilities. *Journal of the American Statistical Association*, 54(286):385–388. 2, 7

Câmara, G., Druck, S., Carvalho, M. S., e Monteiro, A. M. (2000). *Análise espacial de dados geográficos.* Embrapa Cerrados. 1

de Classificação de Doenças CBCD, C. B. (Acesso em: 09 de jun. de 2019). *CID-10 Apresentacão.* Sao Paulo, Centro Brasileiro de Classificação de Doenças. `http://www.datasus.gov.br/cid10/V2008/cid10.htm`. 45, 65, 72

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187. 17, 19, 28

Fernandes, L. B. (2015). *Uma estatística scan espacial bayesiana para dados com excesso de zeros.* Master's thesis, Universidade de Brasília. 15, 16, 17

Johnson, R. A. e Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc. 5, 6

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496. 2, 9, 14, 23, 28

Kulldorff, M. (2018). *TreeScan User Guide v1.4.* Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics Harvard Medical School and Brigham and Women?s Hospital. 21

Kulldorff, M., Fang, Z., e Walsh, S. J. (2003a). A tree based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331. 1, 2, 20, 22, 24, 25, 71

Kulldorff, M. e Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. 14:799–810. 9, 17

Kulldorff, M., Tango, T., e Park, P. J. (2003b). Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, 42(4):665 – 684. 35

Marshall, R. (1991). A review of methods for the statistical analysis of spatial patterns of disease. A 154:421–441. 7

Mingoti, S. (2005). *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG. 5, 6

Naus, J. I. (1965a). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60(310):532–538. 7, 9

Naus, J. L. (1965b). Clustering of random points in two dimensions*. *Biometrika*, 52(1-2):263–266. 8, 16

Nogueira, P. J., Costa, A. J., Pinto, C. S., Alves, M. I., e Rosa, M. V. (2013). Estudo comparativo do número de número de óbitos e causas de morte da mortalidade infanti le suas componentes (2009-2011). *Direção de Serviços de Informação e Análise Direção-Geral da Saúde*. 72

Openshaw, S., Charlton, M., Craft, A., e Birch, J. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet*, 331(8580):272 – 273. Originally published as Volume 1, Issue 8580. 2, 8, 16

Prates, M. O. (2008). *Método Scan Flexivel para detecção em árvore hierarquicas*. Master's thesis, Universidade Federal de Minas Gerias. 20

Silva, H. S. (2017). *Estudo sobre causas de mortalidade infantil no Brasil através da Estatística de Varredura Árvore - Temporal*. Monografia de graduação, Universidade de Brasília. 46, 47, 48

Apêndice A

Códigos SAS

A.1 Algoritmo *Scan* Circular

A macro **%scancircular**, pode ser executada, passando os parâmetros conforme chamada abaixo;

```
1 %scancircular(data=pop_reg_centroide_RJ,casos=Ncasos, pop=populacao
    , casos=Ncasos, cordx=cord_x ,cordy=cord_y , dist=POISSON, nsim
    =1000,pmax=0.5, alpha=0.05,seed=1234);
```

Os seguintes parâmetros devem ser informados para executar o algoritmo da estatística Scan de Kulldorff.

- data: Nome do *data set* que contém as informações do número de casos, coordenadas geográfica e população de cada região.
- casos: Nome da variável que armazena o número de casos do evento de interesse.
- cord_x: Nome da variável que armazena o centroide da coordenada X.
- cord_Y: Nome da variável que armazena o centroide da coordenada Y.
- pop: Nome da variável que contém a população de cada região.
- dist: Nome do modelo de probabilidade, sendo dois possíveis; *POISSON* ou *BINOMIAL* .
- nsim: Número de simulações de Monte Carlo.
- seed: Um número inteiro positivo, usado para criar uma semente aleatória inicial.
- alpha: É o nível de confiança.

```
1
2  %macro scancircular(data=pop_casos_cord, pop=pop_regiao, casos=
       ncasos, cordx=cord_x ,cordy=cord_y , dist=POISSON, nsim=100,pmax
       =, alpha=, seed=);
3
4  %if %bquote(&data.) = %then %do;
5  %put ERROR: A base de dados que contém as cordenadas geográficas e
       a população não foi informda;
6  %end;
7  %else %if %bquote(&pop.) = %then %do;
8  %put ERROR: O nome da variável que guarda os dados da população não
       foi informado;
9  %end;
10 %else %if %bquote(&cordx) = %then %do;
11 %put ERROR: O nome da variável  que comtém os dados da coordenada X
       (latitude) não foi informado;
12 %end;
13 %if %bquote(&cordy.) = %then    %do;
14 %put ERROR: O nome da variável  que comtém os dados da coordenada Y
       (longintude) não foi informado;
15 %end;
16 %if %bquote(&dist.) = %then %do;
17 %put ERROR: O nome da distribuição:(POISSON,BINOMIAL) não foi
       informado;
18 %end;
19 %else %if %bquote(&nsim.) = %then %do;
20 %put ERROR: A quantidade de simualçoes não foi informado;
21 %end;
22 %else %if %bquote(&pmax.) = %then %do;
23 %put ERROR: O limite Máximo  do total de população dentro da zona
       não foi informado;
24 %end;
25 %else %if %bquote(&alpha.) = %then %do;
26 %put ERROR: O nível de confianca não foi informado;
27 %end;
28 %else %if %bquote(&seed.) = %then %do;
29 %put ERROR: a semente não foi informada;
30 %end;
31 %else %do;
32 proc iml;
33 use &data.;
34 names={&pop., &casos. , &cordx. , &cordy.};
35 read all  var{&pop., &casos. , &cordx. , &cordy.}  into pcxy[
```

```
    colname=names];
n=nrow(pcxy);/*número total de zonas*/
/*PRINT pcxy;*/
/* variaveis auxiliares */
P=sum(pcxy[,1]); /*N- número (tamanho) total da populacao*/
C=sum(pcxy[,2]); /*N- número total de casos */
/*Propoção Máxima da populacao*/
Pmax=%sysevalf(&pmax. )*P;
/*Passo 2:Matriz de distancias*/
/*Alocando  tamnaho da mátriz de distancias*/
d=j(n, n,0);
z=j(n,n,0);
d=distance(pcxy[, 3:4], "L2" );
/*inicializando variaveis*/
popz=0;
/*número de casos zona z*/
nxz=0;
/*Verossimilhanca*/
mllr=.;
mcluster=j(1,n,0);
do j=1 to n;
/*vertor de indices (idx)*/
/*Ordenando a matriz para cada coluna (regiaão ii) */
call sortndx(idx, d[,j], 1);
/*guarda asposicoes da distancai ordendas*/
z[,j]=idx;
mcluster=j(1,n,0);
/*inicializando variaveis*/
popz=0;
/*número de casos zona z*/
nxz=0;
/*numero esperado de casos*/
theta=C/P;
do k=1 to n;
/*numeros de casos em cada zona*/
nc=idx[k,];
/*populacao zona j*/
npz=popz + pcxy[nc,1];
popz=npz;
/*populacao da zona z é menor que o tamanho maximo do clustering*/
if npz <= pmax then do;
mcluster[1, idx[1,1]]=1;
mcluster[1, idx[k,1]]=1;
```

```
78 /*print mcluster;*/
79 ncz=nxz+pcxy[nc,2];
80 nxz=ncz;
81 /*casos esperado dentro da zona z*/
82 theta_1=(C*     npz)/P;
83 nxz_bar=C-nxz;
84 popz_bar=P-popz;
85 theta_0= nxz_bar/popz_bar;
86 theta_z= nxz / npz;
87 dist={"POISSON"};
88 if dist="&dist." then do;
89 if  nxz >  theta_1 then do;
90 if nxz_bar >0 then
91 llr= nxz*(log( nxz) - log(theta_1))  +( nxz_bar) * (  log(nxz_bar)
      - log(C - theta_1));
92 else llr= nxz*(log( nxz) - log(theta_1))  +(    nxz_bar) * ( - log(
      C - theta_1));
93 end;
94 else do;
95 llr=0;
96 end;
97 end;
98
99 if dist ^=  "&dist." then do;
100 if  nxz >  theta_1 then
101 llr=nxz*log( theta_z)+( npz - nxz)*log (1- theta_z)+nxz_bar*log(
      theta_0)+(popz_bar -nxz_bar)*log (1- theta_0)-
102 C*log( theta )-(P-C)*log (1- theta );
103 else llr=0;
104 end;
105
106 /*Selecionando Apenas o cluster + verossimil*/
107 if   (llr > mllr ) &  (npz <= pmax) then do;
108 /* Regioes selecionadas*/
109 cluster=mcluster;
110 /*Razao Log verossimilhanca*/
111 mllr=llr;
112 /*Número de casos Obervados dentro da zona z(conjunto de regioes
      selecionadas)*/
113 mncz=ncz;
114 /*Ncasos Esperado*/
115 nc_esperado=theta_1;
116 /*Populacao dentro da zona z(conjunto de regioes selecionadas)*/
```

```
117 mpopz=npz;
118 /*Tamanho do Cluster Av Selecionado*/
119 tmclust=sum(cluster[1,]);
120 /**/
121 cluster_esc=loc(cluster);
122 end;
123 end;
124 end;
125 end;
126 /*1-Data set que armazena a informação do ID das regioes detectados
      */
127 create cluster_Detectado from cluster_esc;
128 append from cluster_esc;
129 names={"Estatística T" "população_cluster"   "tamanho_cluster"  "
      casos_esperados" "casos_observados" };
130 results= mllr||mpopz||ncol(cluster_esc)||nc_esperado||mncz;
131
132 /*2-Estatisticas relacionadas aos cluster detectado*/
133 create estat_cluster  from results[colname=names];
134 append from results;
135
136 /*Iniciando a simulação de monte carlo */
137 start ProbCasos(x);
138 m=nrow(x);/*número total de zonas*/
139 /*Total da populacao*/
140 N=sum(x[,1]);
141 PR=j(m,1,0);
142 do i = 1  to m;
143 PR[i,1]=x[i,1]/N;
144 end;
145 return PR;
146 finish ProbCasos;
147
148 /*Probabilidades */
149 PR=ProbCasos(pcxy);
150 prob=t(PR[,1]);
151 llr=j(1,1,0);
152
153 /*alocando dim função de  verossimilhança*/
154 vllr=j(&nsim., 1,0);
155
156 /* Definido a maxtriz casos  simulaco  com base na multinomial */
157 ncs=j(&nsim., 1,0);
```

```
158
159 /*Semente*/
160 call randseed(&seed.);
161 ncs=RandMultinomial(&nsim.,C, prob);
162 casos=t(ncs);
163 do sm=1 to &nsim.;
164 /*Atribuindo zero para a colun matriz que guarda caso matriz
       original*/
165 pcxy[,2]=0;
166 /*Tempo de exceução*/
167 t0 = time();
168 /*incializando as variaveis para o calculo da llr empirica */
169 mllr=.;
170 mcluster=J(1,n,0);
171 popz=0;
172 nxz=0;
173
174 do j=1 to n;
175 pcxy[j,2]=casos[j,sm];
176
177 /*       end;*/
178 /*       end;*/
179 /*print pcxy;*/
180 pcxy=pcxy[,1:4];
181 call sortndx(idx, d[,j], 1);
182
183 /*criando a matriz que guarda os casos em cada zona*/
184 mcluster=j(1,n,0);
185
186 /*inicializando variaveis*/
187 popz=0;
188
189 /*número de casos zona z*/
190 nxz=0;
191
192 /*numero esperado de casos*/
193 theta=C/P;
194
195 do k=1 to n;
196 /*numeros de casos em cada zona*/
197 nc=z[k,j];
198 /*populacao zona j*/
199 npz=popz + pcxy[nc,1];
```

```
200 popz=npz;
201 /*populacao da zona z é menor que o tamanho maximo do clustering*/
202 if npz <= pmax then
203 do;
204 mcluster[1, z[1,j]]=1;
205 mcluster[1, z[k,j]]=1;
206 ncz=nxz+pcxy[nc,2];
207 nxz=ncz;
208 /*casos esperado dentro da zona z*/
209 theta_1=(C*     npz)/P;
210 nxz_bar =C-nxz;
211 popz_bar=P-popz;
212 theta_0= nxz_bar/popz_bar;
213 theta_z= nxz/npz;
214 dist={"POISSON"};
215 if dist="&dist." then do;
216 if  nxz >  theta_1 then do;
217 if nxz_bar >0 then
218 llr= nxz*(log( nxz) - log(theta_1))  +( nxz_bar) * (  log(nxz_bar)
      - log(C - theta_1));
219 /*x*log(x) tende para zero*/
220 else llr= nxz*(log( nxz) - log(theta_1))  +(    nxz_bar) * ( - log(
      C - theta_1));
221 end;
222 else do;
223 llr=0;
224 end;
225 end;
226
227 if dist ^= "&dist." then do;
228 if  nxz >  theta_1 then
229 llr=nxz*log( theta_z)+( npz - nxz)*log (1- theta_z)+nxz_bar*log(
      theta_0)+(popz_bar -nxz_bar)*log (1- theta_0)-
230 C*log( theta )-(P-C)*log (1- theta );
231 else llr=0;
232 end;
233
234 /*armazena o maximo da verossimilhanca*/
235 if   (llr > mllr ) &  (npz <= pmax) then do;
236 cluster=mcluster;
237 mllr=llr;
238 end;
239 end;
```

```
240 end;
241 end;
242
243 /*Verossimilhaca calculada para cada iteracao*/
244 vllr[sm,1]=mllr;
245 end;
246 /*Calcula p_vqloe*/
247 t=time() -t0;
248
249 /*Tempo de execução*/
250 print t[colname={"Tempo de Execução"}];
251 /*Guardar os valores da verossimilhanca empirica*/
252 create llremp  from vllr[colname="LLR"];
253 append from vllr;
254 quit;
255
256 %end;
257 /*Calculado o percentil de acordo com o nivel de confiancao*/
258 proc univariate data=llremp NOPRINT;
259 var llr;
260 output out=percentis pctlpts=%sysevalf((1 -&alpha.)*100)  pctlpre=P
       ;
261 run;
262
263 /*valor critico-Teste de hipotese*/
264 data teste_hipotese;
265 merge ESTAT_CLUSTER(keep='Estatística T'n)  percentis ;
266 attrib Decisao_Teste  length=$32. format=$32.;
267 if  'Estatística T'n > P%sysevalf((1 -&alpha.)*100) then  Decisao_
       Teste = "Rejeita Ho";
268 else Decisao_Teste = "Não Rejeita Ho";
269 run;
270
271 /*Guarda a informação da estatistica do teste*/
272 data _null_;
273 set teste_hipotese;
274 call symputx("Estat_T","Estatística T"n );
275 call symputx("nc",_N_);
276 run;
277 %put &Estat_T.;
278
279
280 proc iml;
```

```
281 use llremp;
282 read all var{llr} into llr;
283 close ;
284 use teste_hipotese;
285 read all var _NUM_ into LLR_OBS;
286 close;
287 nsup=j(&nc.,1,0);
288 do i=1 to 1;
289 IDX=loc(llr[,1]>=LLR_OBS[i,1]);
290 nsup[i,1]=ncol(IDX);
291 end;
292 create NSUP_LLR from nsup[Colname={"N_sup_LlR"}];
293 append from nsup;
294 quit;
295
296 data Etatistica_Cluster_detectado ;
297 merge teste_hipotese NSUP_LLR;
298 p_valor=(1+N_sup_LlR )/(&nsim.+1);
299 alpha=(&alpha.);
300 run;
301
302 /*Valor critico*/
303 data _null_ ;
304 set Teste_Hipotese;
305 call symputx("P95",P95);
306 run;
307 /*Grafico Distribuicao Empririca*/
308 title "Distribuicao  da estatistica Log RV";
309 /*ods listing gpath="&file.\";*/
310 /*ods graphics on / imagename="DistEmp" reset=index noborder
      imagefmt=png ;*/
311
312 proc sgplot data=LLREMP noborder;
313 histogram LLR ;
314 refline &Estat_T. / axis=x  lineattrs= (color=red thickness=2)
315 name="Est" legendlabel="Estatística T Observada=&Estat_T.";
316 refline &p95.   /axis=x lineattrs=(color=blue pattern=2 thickness=2
       )
317 name="T-Crítico" legendlabel="T-Crítico=&p95."   labelattrs=(color=
      dabr  size=12pt) ;
318 yaxis label="Porcentagem"  labelattrs=(color=dabr  size=12pt)  ;
319 run;
320
```

```
321
322 %mend scancircular;
323
324 %scancircular(data=, pop=POPULACAO, casos=Ncasos, cordx=cord_x ,
      cordy=cord_y , dist=POISSON, nsim=1000,pmax=0.5, alpha=0.05,seed
      =1234);
```

Apêndice B

Códigos do Algoritmo Estatística de varredura baseada em árvore

A macro **%treescan**, pode ser executada, passando os parâmetros conforme chamada abaixo;

```
%treescan(data=NCASOS_ARVORE_&UF._CI10, nsim=100, seed=1234, alpha
    =0.05);
```

Os seguintes parâmetros devem ser informados para executar o algoritmo de varredura baseada em árvore:

- data: Nome do *data set* que contém as informações do número de casos, população ou casos esperado e os ID da arvore (nós) e ID dos nós pais.
- nsim: Número de simulações de Monte Carlo.
- seed: Um número inteiro positivo, usado para criar uma semente aleatória inicial.
- alpha: É o nível de confiança.

```
/*Nome base de dados*/
%let data=NCASOS_ARVORE_&UF._CI10;
%global N;
data  &data.;
set NCASOS_ARVORE_&UF._CI10 end=last;
IDP=_N_;
if last then call symputx('N', _N_);
run;
%put o números de nós é igual &N.;

/*salvando e var macro nome das variaveis numericas*/
proc sql;
```

```
13 select quote(left(trim(name)), " ") into :namevarnun separated by "
       "
14 from dictionary.columns
15 where   memname = UPCASE("&data.") and type="num";
16 quit;
17
18 proc sql;
19 select left( trim(name)) into :names_var_char separated by " "
20 from dictionary.columns
21 where   memname = upcase("&data.") and type="char";
22 quit;
23
24
25 PROC SORT DATA=&data. out=nopai(keep=PnodeID) nodupkey ;
26 BY PnodeID;
27 RUN;
28
29 data _null_;
30 set nopai end=last;
31 if last then call symputx("Nopai",_n_);
32 run;
33 %put &Nopai.;
34
35
36 /*Selecionado apenas nos pais */
37 proc sql;
38 create table Nospais as
39 select distinct PNODEID into :NosPais separated by " "
40 from &data.
41 where  not missing(PNODEID)  ;
42 quit;
43
44 /*Selecionando os nos filhos */
45 proc sort data=&data. out=nosfilhos(keep=nodeID) nodupkey ;
46 by nodeID;
47 where NODEID is not missing  ;
48 run;
49
50 /*Selecionado apenas os nos filhos*/
51 proc sql;
52 create table Nosfilhos as
53 select distinct Nodeid
54 from &data
```

```
except
select pNodeid from  Nospais;

quit;
/*salvando em variavel macro a quantidade de nos filhos*/
data _null_;
set Nosfilhos end=last;
if last then call symputx("Nofilho",_n_);
run;
%put A árvore tem &Nofilho. folhas ;



/*%FREE libera memoria;*/
%PUT &N.;
%let data=NCASOS_ARVORE_&UF._CI10;
%let pmax=0.5;
%let nsim=1000;
%let max_no_simples=7;
%LET SEED=1234;
%macro treescan(data=, nsim=, seed, alpha=);
proc iml;
/*Lendo data set com as informaçoes da árvore*/
use &data.;
varnames={&names_var_char.};
numnames={&namevarnun.};
/*Codigos dos ramos que são texto*/
read   ALL var varnames into x[colname=varnames];
/*Codigos dos ramos que são numericos*/
read all var numnames  into y[colname=numnames];
close;
/* ID Folhas     */
NF=x[,1];
/*ID Nos pais*/
NP=x[,2];
N=sum(y[,2]); /*N- número (tamanho) total da populacao*/
C=sum(y[,1]); /*C- número total de casos */
/*Quantidade de nos */
qdno=nrow(y);
NoRaiz=nf[loc(missing(np[,1]))];
qdr=ncol(NoRaiz);
QNfilhos=&Nofilho.;
Print NoRaiz  QNfilhos;
```

```
/*quantidades de colunas*/
nc=ncol(y);
print N C qdno ;
/*incializando a funcao max similhnaca*/
mllr=j(1,1,0);
mcluster=J(1,qdno,0);
theta=C/N;
/*selecionado apenas a interseção( ou seja, todos os cortes pais)*/
popg=0;
idg = xsect(X[,1],np);
ip=loc(element(x[,1],idg));
/*Tabela que armazena os nos */
cut=j(&N.,6,0);
g=j(1,nc,0);
NF=x[,1];
NP=x[,2];
theta=C/N;
/*Matriz para armazenar os cortes */
/*Criando modulo- Calcula estatistica de máxima verossimilhança*/
/*Pegando todos os possivéis cortes (a árvore prescisa ter dois
   niveis) */
do i=1 to  &N. while(%eval(&N.)>1 );
m=%eval(&N.)-i+1;
/*V*/
if  loc(ip=m) then do;
idz=loc(element(x[,2], x[m,1]  ));
igd=x[idz,1];
ig=loc(element(x[,1],igd));
y[m,1]=sum(y[ig,1]);
y[m,2]=sum(y[ig,2]);
end;
end;

do i=1 to &n.;
g=j(1,nc,0);
g=y[i,];
/*      populacao no ramo g*/
popg=0;
npg=g[1,2];
popg=npg;
/*Numero de casos no ramo g*/
ncg=g[1,1];
nxg=ncg;
```

```
140 /*Número de casos esperado dentro do corte g;*/
141 theta_1=(C*npg)/N;
142 /* Número casos fora ramo g*/
143 nxg_bar=C-nxg;
144 /*Numeros de casos do ramo g*/
145 popg_bar=N-popg;
146 /*Propoção de casos fora do corte g*/
147 theta_0= nxg_bar/popg_bar;
148 /*Propoção de casos dentro do corte g*/
149 theta_g= nxg / npg;
150 if  (nxg >  theta_1 )  then DO;
151 IF nxg_bar>0 THEN llr= nxg*(log(nxg) - log(theta_1))  +(nxg_bar) *
       (log(nxg_bar) - log(C - theta_1));
152 else llr= nxg*(log(nxg) - log(theta_1))  +(nxg_bar) * ( - log(C -
       theta_1));
153 end;
154 else do;
155 llr=0;
156 end;
157 /* Armazenados os valores de todos os possiéveis cortes */
158 cut[i,]=g||nxg||theta_1||llr;
159 /*  Criando um data set que armazena ;*/
160 create Cortes_avores from cut[colname={&names_var. nxg
       CasosEsperdo LLR }];
161 append from cut;
162 close Cortes_avores;
163 end;
164 start ProbCasos(x);
165 /*Total da populacao*/
166 N=sum(x[,2]);
167 max_no_simples=%EVAL(&Nofilho.);
168 P =j(max_no_simples ,1,0);
169 do i = 1 to max_no_simples ;
170 P[i,1]=x[i,2]/N;
171 end;
172 return P;
173 finish;
174 npf= remove(NF,loc(element(nf,np)));
175 idf=&N.-(QNfilhos-qdr);
176 npt=IDF:&n.;
177 ipf=loc(element(t(npf),x[,1]));
178 P=ProbCasos(y[npt,]);
179 create sim from P;
```

```
180 append from P;
181 prob=t(P[,1]);
182 pr=sum(prob);
183 mllr=j(&n.,&nsim.,0);
184 llr=j(1,1,0);
185 /*alocando dim função de  verossimilhança*/
186 vllr=j(&nsim., 1,0);
187 k=j(%eval(&Nofilho.), 1, 0);
188 /* Definido a maxtriz de simulaco  */
189 d=j(&nsim.,&Nofilho. ,0);
190 call randseed(&seed.);
191 d=RandMultinomial(&nsim.,C, prob);
192 casos=t(d);
193 /*      print casos;*/
194 cut[,1]=0;
195 create casos from casos;
196 append from casos;
197 /*Incializando o lopp da simulação*/
198 do j=1 to &nsim.;
199 /* criando lista com a posição  dos cortes simples ( não inclui
       raiz)     */
200 ns=1:%eval(&Nofilho.);
201 cut[npt,1]=casos[ns,j];
202 cut=cut[,1:4];
203 do i=1 to &N. while(%eval(&N.) >1);
204 m=%eval(&N.)-i+1;
205 if  loc(ip=m) then do;
206 idz=loc(element(x[,2], x[m,1]  ));
207 igd=x[idz,1];
208 ig=loc(element(x[,1],igd));
209 cut[m,1]=sum(cut[ig,1]);
210 cut[m,1]=sum(cut[ig,2]);
211
212 end;
213 end;
214
215
216 do i=1 to &N. while(%eval(&N.) >1);
217 g=j(1,nc,0);
218 g=cut[i,];
219 /*      populacao no ramo g*/
220 popg=0;
221 npg=g[1,2];
```

```
222 popg=npg;
223 /*Numero de casos no ramo g*/
224 ncg=g[1,1];
225 nxg=ncg;
226 /*Número de casos esperado dentro do corte g;*/
227 theta_1=(C*npg)/N;
228 /* Número casos fora ramo g*/
229 nxg_bar=C-nxg;
230 /*Numeros de casos do ramo g*/
231 popg_bar=N-popg;
232 /*Propoção de casos fora do corte g*/
233 theta_0= nxg_bar/popg_bar;
234 /*Propoção de casos dentro do corte g*/
235 theta_g= nxg / npg;
236 /* funcao indicadora   */
237 if  (nxg >  theta_1 )  then DO;
238 IF nxg_bar>0 THEN llr= nxg*(log(nxg) - log(theta_1))  +(nxg_bar) *
       (log(nxg_bar) - log(C - theta_1));
239 else llr= nxg*(log(nxg) - log(theta_1))  +(nxg_bar) * ( - log(C -
       theta_1));
240 end;
241 else do;
242 llr=0;
243 end;
244
245 mllr[i,j]=llr;
246
247 end;
248 vllr[j,1]=max(mllr[,j]);
249 create MLLR_SIM_ARVORE from vllr[colname={LLR}];
250 append from vllr;
251 close MLLR_SIM_ARVORE;
252 end;
253
254 /*Calculo do P-valor*/
255 %let alpha=0.05;
256 proc univariate data=MLLR_SIM_ARVORE;
257 var llr;
258 output out=percentis pctlpts=%sysevalf((1 -&alpha.)*100) pctlpre=P;
259 RUN;
260
261
262 data _null_;
```

```
263 set percentis;
264 call symputx("pctl",p95);
265 run;
266
267 %put &PCTL.;
268 data teste_hipotese;
269 merge Cortes_avores &data.(keep=&names_var_char.)  ;
270 attrib Decisao_Teste  length=$32. format=$32.;
271 if  LLR > &PCTL. then  Decisao_Teste = "Rejeita Ho";
272 else Decisao_Teste = "Não Rejeita Ho";
273 run;
274
275 /*Cortes ordenados por maior log RV*/
276 proc sort data=teste_hipotese(where=(Decisao_Teste="Rejeita Ho"));
277 by descending llr ;
278 run;
279
280 %mend treescan;
281
282 %treescan(data=NCASOS_ARVORE_&UF._CI10, nsim=100, seed=1234, alpha
      =0.05);
```

Apêndice C

Códigos do Algoritmo de Estatística varredura árvore-espacial

A macro **%treescancircular**, pode ser executada, passando os parâmetros conforme chamada abaixo;

```
1 %treescancircular(data_1=pop_reg_centroide_RJ, data_2=NCASOS_ARVORE
    _REG_RJ, data_3=ARVORE_ID,var_geo=
2 POPULACAO cord_x cord_y ,Arvore_id=NODEID nsim=1000,pmax=0.50,
    alpha=0.05,seed=1234)
```

Os seguintes parâmetros devem ser informados para executar o algoritmo árvore-espacial.

- data_1: Nome do *data set* que contém as informações do número de casos, coordenadas geográficas e população de cada região.
- data_2: Nome do *data set* que contém o número de casos para cada ID da árvore e regiões(colunas) .
- data_3: Nome do *data set* que contém as variáveis ID da árvore e ID dos nós pais.
- arvore_id: Nome da variável que armazena as informações do dos nós (ID) da árvore.
- var_geo: Informar o nome das variáveis geográficas separada por espaço: Populacao cordx cordy.
- nsim: Número de simulações de Monte Carlo.
- seed: Um número inteiro positivo, usado para criar uma semente aleatória inicial.
- alpha: É o nível de confiança.

```
1
2  /*Cria variaveis macro com algumas infomacoes de variaveis,
       quantidade de nos da arvore, quantidade de nos filhos, nos pais*
       /
3  /*1-Banco de dados  com os dados dos centroides de cada região e
       populacao, codigo de cada região*/
4  %let data_1=pop_reg_centroide_RJ;
5  /*2-Banco de dados com número de casos de cada nó folha ao longo de
        todas regiãos (linhas=ID Arvore, Coluna=ID Regiao)*/
6  %let data_2=NCASOS_ARVORE_REG_RJ;
7  /*3-Banco de dados com todos os corte -2 Varivies (ID ARVORE , ID
       ARVORE NOS Pais)*/
8  %let data_3=ARVORE_ID;
9  %global N;
10 data  &data_3;
11 set &data_3 end=last;
12 /*     Agrupando pelos id dos nos */
13 by  pnodeid;
14 IDP=_N_;
15 /*Guardando em variável macro a quantidade de nos */
16 if last then call symputx('N', _N_);
17 run;
18 %put o números de nós é igual &N.;
19
20 /*Ordenando pelo ID dos ramos*/
21 proc sort data=&data_3  out=&data_3;
22 by PNODEID;
23 run;
24
25 /*Selecionado apenas nos pais */
26 proc sql;
27 create table Nospais as
28 select distinct PNODEID into :NosPais separated by " "
29 from &data_3
30 where  not missing(PNODEID)  ;
31 quit;
32
33 /*Selecionando os nos filhos */
34 proc sort data=&data_3 out=nosfilhos(keep=nodeID) nodupkey ;
35 by nodeID;
36 where NODEID is not missing  ;
37 run;
38
```

```
39 /*Selecionado apenas os nos filhos*/
40 proc sql;
41 create table Nosfilhos as
42 select distinct Nodeid
43 from &data_3
44 except
45 select pNodeid from  Nospais;
46 quit;
47
48 /*salvando em variavel macro a quantidade de nos filhos*/
49 data _null_;
50 set Nosfilhos end=last;
51 if last then call symputx("Nofilho",_n_);
52 run;
53 %put A árvore tem &Nofilho. folhas ;
54
55 proc sql;
56 select distinct NODEID into :Nosfilhos separated by " "
57 from &data_3. ;
58 quit;
59
60 data _null_;
61 set Nospais end=last;
62 if last then call symputx("Nopai",_n_);
63 run;
64 %put &Nopai.;
65
66 /*salvando em macro variaveis  o nome/ codigo do municipios*/
67 proc sql;
68 select quote(left(trim(name)), " ") into :names_var_casos separated
       by " "
69 from dictionary.columns
70 where   memname = UPCASE("&data_2.") and type="num";
71 quit;
72 %put &names_var_casos.;
73
74 /*salvando e var macro nome dos ids da árvore- character*/
75 proc sql;
76 select distinct quote(left(trim(name)), " ") into :names_id_av
      separated by " "
77 from dictionary.columns
78 where   memname = UPCASE("&data_3.") and type="char";
79 quit;
```

```
80
81 %put &names_id_av.;
82
83 %macro treescancircular(data_1=&datas_1, data_2=, data_3=,var_geo=,
      Arvore_id=NODEID nsim=100,pmax=, alpha=,seed=);
84 %if %bquote(&data_1) = %then %do;
85 %put ERROR: A base de dados que contém as cordenadas geográficas e
       a população não foi informda;
86 %end;
87 %else %if %bquote(&data_2) = %then %do;
88 %put ERROR: A base de dados que contém Os ID da árvore e Id dos
      Pais não foi informadp;
89 %end;
90 %else %if %bquote(&data_3) = %then %do;
91 %put ERROR: A base de dados que contém o número de casos do event
      ode intersse para cada ID ARVORE(folha) segundo regioes (
      colunas) nao foi informada;
92 %end;
93 %else %if %bquote(&var_geo) = %then %do;
94 %put ERROR: as variaveis Populacao, cord_x e cord_y não foram
      identificadas ;
95 %end;
96 %else %if %bquote(&Arvore_id) = %then %do;
97 %put ERROR: nome da variavei que corresponde ao ID da árvore nao
      foi informado;
98 %end;
99 %else %if %bquote(&pmax.) = %then %do;
100 %put ERROR: Limite maximo populacao nao foi informado;
101 %end;
102 %else %if %bquote(&pmax.) = %then %do;
103 %put ERROR:nivel de cofiança nao foi infromado;
104 %end;
105 %else %if %bquote(&pmax.) = %then %do;
106 %put ERROR: a semente não foi informada;
107 %end;
108
109 %else %do;
110 proc iml;
111 names={&var_geo.};
112 use &data_1.;
113 read all  var{&var_geo.}  into pcxy[colname=names];
114 n=nrow(pcxy);/*número total de zonas*/
115 close;
```

```
/*Lendo data set com as informaçoes da árvore*/
use &data_3.;
varnames={&NAMENUM.};
/*     Codigos dos ramos que são texto*/
read   ALL var varnames into x[colname=varnames];
/*       Codigos dos ramos que são numericos*/
close;
/*       print x;*/
nf=x[,1];
/*       print nf;*/
np=x[,2];
/*Lendo a base com a informação do numero de casos para cada no da
   arvore e regiao*/
use &data_2.;
read all  var{&Arvore_id.} into nodes_arvore;
use &data_2.;
nreg={&names_var_casos.};
read all  var _num_ into nx[colname=nreg];
mattrib nx rowname= nodes_arvore colname=nreg;
close;
nodes=t(1:&n.);
nxzg=nodes||nx;
nvarc={"&Arvore_id." &names_var_casos.};
mattrib nxzg rowname= nodes_arvore colname=nvarc;
/*print nxzg;*/
/* variaveis auxiliares */
P=sum(pcxy[,1]);/*N- número (tamanho) total da populacao*/
Pmax=%sysevalf(&pmax.)*P;
/*Passo 2:Matriz de distancias*/
/*Alocando  tamnaho da mátriz de distancias*/
d=j(n, n,0);
d=distance(pcxy[, 2:3], "L2" );
/*inicializando variaveis*/
popz=0;
/*número de casos zona z*/
nxz=0;
/*Verossimilhanca*/
mllr=.;
mcluster=j(1,n,.);
/*Quantidade de nos */
qdno=nrow(y);
/*Indentificando o Codigo do no raiz*/
NoRaiz=nf[loc(missing(np[,1]))];
```

```
158 QNfilhos=&Nofilho.;
159 Print NoRaiz  QNfilhos;
160 /*quantidades de colunas*/
161 nc=ncol(y);
162 /*incializando a funcao max similhnaca*/
163 mllr=j(1,1,.);
164 /*mcluster=J(1,qdno,0);*/
165 /*selecionado apenas a interseção( ou seja, todos os cortes pais)*/
166 idg = xsect(nodes_arvore[,1],np);
167 /*Indentifica apenas os ID */
168 ip=loc(element(nodes_arvore[,1],idg));
169 /*Cotem apenas os nos filhos */
170 npf= remove(nf,loc(element(nf,np)));
171 idf=loc(element(x[,1],npf));
172 idff=x[idf,2];
173 ipf=loc(element(idg,idff));
174 do j=1 to n;
175 /*Extraindo os */
176 do i=1  to &N. while(%eval(&N.)>1 );
177 /*  Criando uma sequencia decrescente */
178 m=%eval(&N.)-i+1;
179 if  loc(ip=m) then do;
180 idz=loc(element(x[,2], nodes_arvore[m,1]  ));
181 igd=x[idz,1];
182 ig=loc(element(nodes_arvore[,1],igd));
183 nxzg[m,j+1]=sum(nxzg[loc( element(nxzg[,1],ig )),j+1]);
184 end;
185 end;
186 end;
187 create nzx from nxzg;
188 append from nxzg;
189 Cg=j(&N.,1);
190 z=j(n,n,0);
191 do j=1 to n;
192 /*vertor de indices (idx)*/
193 /*Ordenando a matriz para cada coluna (regiaão ii) */
194 call sortndx(idx, d[,j], 1);
195 mcluster=j(1,n,0);
196 /*inicializando variaveis*/
197 popz=0;
198 /*Armazena todas as possiveis zonas[Ordenazação decrescente]*/
199 z[,j]=idx;
200 /*número de casos zona z*/
```

```
201 /*numero esperado de casos*/
202 do k=1 to n;
203 /*numeros de casos em cada zona*/
204 nc=idx[k,];
205 /*populacao zona j*/
206 npz=popz + pcxy[nc,1];
207 popz=npz;
208 /*populacao da zona z é menor que o tamanho maximo do clustering*/
209 if npz <= pmax then do;
210 mcluster[1, idx[1,1]]=1;
211 mcluster[1, idx[k,1]]=1;
212 ncg=loc(mcluster);
213 /*Matriz para armazenar os cortes */
214 /*Pegando todos os possivéis cortes (a árvore prescisa ter dois
       niveis) */
215 do l=1 to  &N. while(%eval(&N.)>1 );
216 g=nxzg[l,];
217 C=sum(nxzg[l,2:n+1]);
218 Cg[l,]=C;
219 nxz_g=nxzg[,2:n+1];
220 nxz=sum(nxz_g[l,ncg]);
221 /*casos esperado dentro da zona z*/
222 theta_1=(C*     npz)/P;
223 nxz_bar=C-nxz;
224 popz_bar=P-popz;
225 theta_0= nxz_bar/popz_bar;
226 theta_z= nxz/npz;
227 if ( nxz >  theta_1) then do;
228 if (nxz_bar >0)  then    llr= nxz*(log( nxz ) - log(theta_1))  +(
       nxz_bar) * (  log(nxz_bar) - log(C - theta_1));
229 else llr=  llr= nxz*(log( nxz ) - log(theta_1))  +(nxz_bar) * ( -
       log(C - theta_1));
230 end;
231 else do;
232 llr=0;
233 end;
234 if   (llr > mllr ) &  (npz <= pmax) then do;
235 cluster=mcluster;
236 mllr=llr;
237 mncz=nxz;
238 mpopz=npz;
239 regioes=ncg;
240 nca=nxz_g;
```

```
241 tmclust=sum(cluster[1,]);
242 cluster_id=loc(cluster);
243 cluster_arvore=g;
244 nc_esperado=Theta_1;
245 end;
246 end;
247 end;
248 end;
249 end;
250
251 /*Definindo o Id do corte */
252 ID_ARVORE=nodes_arvore[cluster_arvore[1,1],1];
253 create ID_ARVORE from ID_ARVORE[colname={ID_ARVORE}];
254 append from ID_ARVORE;
255
256 corte_xgz=nca[cluster_arvore[1,1],regioes];
257 corte_xz_g=cluster_arvore[1,1]||corte_xgz;
258 create cluster_xgz from corte_xz_g;
259 append from corte_xz_g;
260 /*Clusteres esperados*/
261 create cluter_esc from cluster_id;
262 append from cluster_id;
263 create nca from nca;
264 append from nca;
265 names={ "Estatística T" "população_cluster"   "tamanho_cluster"  "
      casos_esperados" "casos_observados"};
266 results= mllr||mpopz||ncol(cluster_id)||nc_esperado||mncz;
267 /*print results[colname=names];*/
268 create estat_cluster_av  from results[colname=names];
269 append from results;
270
271 /*Simulacao sob H0*/
272 start ProbCasos(x);
273 m=nrow(x);/*número total de zonas*/
274 /*Total da populacao*/
275 N=sum(x[,1]);
276 PR=j(m,1,0);
277 do i = 1  to m;
278 PR[i,1]=x[i,1]/N;
279 end;
280 return PR;
281 finish ProbCasos;
282 PR=ProbCasos(pcxy);
```

```
/*print PR;*/
create sim from PR;
append from PR;
prob=t(PR[,1]);
/*alocando dim função de  verossimilhança*/
llr=j(1,1,.);
/*alocando dim função de  verossimilhança*/
vllr=j(&nsim., 1,.);
/* Definido a maxtriz de simulaco  */
nxzg_g=nxzg;
nt=ncol(nxzg_g);
nxzg=j(&N.,     nt ,0);
ncf=t(idf);
do sm=1 to &nsim.;
/* Gerando casos apenas para as folhas  */
do  i=1 to &nofilho.;
/*Número de casos total do ramo g em todas as regioes k*/
C_g=sum(nxzg_g[ncf[i,1],2:n+1]);
/*       print C;*/
/*Semente*/
nct=1:&N.;
nxzg[,1]=t(nct);
call randseed(&seed.);
nxzg[ncf[i,1],2:n+1]=RandMultinomial(1,C_g, prob);
end;
/*  print nxzg;*/
do j=1 to n;
/*Extraindo os */
do i=1  to &N. while(%eval(&N.)>1 );
/*  Criando uma sequencia decrescente */
m=%eval(&N)-i+1;
if  loc(ip=m) then do;
idz=loc(element(x[,2], nodes_arvore[m,1]  ));
igd=x[idz,1];
ig=loc(element(nodes_arvore[,1],igd));
nxzg[m,j+1]=sum(nxzg[loc( element(nxzg[,1],ig )),j+1]);
end;
end;
end;
mllr=.;
mcluster=J(1,n,0);
popz=0;
nxz=0;
```

```
/*start simulacao;*/
do j=1 to n;
/*vertor de indices (idx)*/
/*Ordenando a matriz para cada coluna (regiaão ii) */
mcluster=j(1,n,0);
/*inicializando variaveis*/
popz=0;
/*número de casos zona z*/
nxz=0;
/*numero esperado de casos*/
do k=1 to n;
/*numeros de casos em cada zona*/
nc=z[k,j];
/*              print nc;*/
/*populacao zona j*/
npz=popz + pcxy[nc,1];
popz=npz;
/*              print npz;*/
/*populacao da zona z é menor que o tamanho maximo do clustering*/
if npz <= pmax then do;
mcluster[1, z[1,j]]=1;
mcluster[1, z[k,j]]=1;
ncg=loc(mcluster);
do l=1 to  &N. while(%eval(&N.)>1 );
g=nxzg[l,];
C=sum(nxzg[l,2:n+1]);
/*Cg  total de caso do ramo  G ao longo de todas as regioes */
nxz_g=nxzg[,2:n+1];
ncz=sum(nxz_g[l,ncg]);
nxz=ncz;
/*casos esperado dentro da zona z*/
theta_1=(C*     npz)/P;
nxz_bar=C-nxz;
popz_bar=P-popz;
theta_0= nxz_bar/popz_bar;
theta_z= nxz/npz;
if ( nxz >  theta_1) then do;
if (nxz_bar >0)  then    llr= nxz*(log( nxz ) - log(theta_1))  +(
   nxz_bar) * (  log(nxz_bar) - log(C - theta_1));
else llr=  nxz*(log( nxz ) - log(theta_1))  +(nxz_bar) * ( - log(C
   - theta_1));
/*                                      print llr;*/
end;
```

```
367 else do;
368 llr=0;
369 end;
370 /*                                          print llr;*/
371 if   (llr > mllr ) &  (npz <= pmax) then do;
372 mllr=llr;
373
374 end;
375 end;
376 end;
377 end;
378 end;
379 /*Verossimilhaca calculada para cada iteracao*/
380 vllr[sm,1]=mllr;
381 end;
382 /*Guardar os valores da verossimilhanca empirica*/
383 create llremp  from vllr[colname="LLR"];
384 append from vllr;
385 quit;
386
387 %end;
388
389
390 /*Orgazinado os resultados */
391 /*Informacoes do cluster*/
392 data estat_cluster_av;
393 merge ID_ARVORE   estat_cluster_av ;
394 run;
395
396 /*Regioes Busca a infomacão das regiões*/
397 proc transpose data=cluter_esc out=cluster_reg(rename=(col1=ID_REG)
       );
398 run;
399
400 data _Null_;
401 length cluster_reg $1000;
402 retain cluster_reg;
403 set cluster_reg end=fim;
404 cluster_reg = catx(",",trim(left(cluster_reg)) ,left(ID_REG));
405 if fim then call symputx("reg_DET",  cluster_reg);
406 run;
407
408 /*Informacoes do cluster*/
```

```
409 data estat_cluster_AV_ESP;
410 merge ID_ARVORE   estat_cluster_av ;
411 format ID_REG_DETECTADAS $200.;
412 ID_REG_DETECTADAS="&reg_DET.";
413 run;
414
415
416 /*Teste de Hipotese e Pvalor*/
417
418 /*Calculado o percentil de acordo com o nivel de confiancao*/
419 proc univariate data=llremp NOPRINT;
420 var llr;
421 output out=percentis pctlpts=%sysevalf((1 -&alpha.)*100)  pctlpre=P
      ;
422 run;
423
424 /*valor critico-Teste de hipotese*/
425 data teste_hipotese;
426 merge estat_cluster_AV_ESP(keep='Estatística T'n)  percentis ;
427 attrib Decisao_Teste  length=$32. format=$32.;
428 if  'Estatística T'n > P%sysevalf((1 -&alpha.)*100) then  Decisao_
      Teste = "Rejeita Ho";
429 else Decisao_Teste = "Não Rejeita Ho";
430 run;
431
432 /*Guarda a informação da estatistica do teste*/
433 data _null_;
434 set teste_hipotese;
435 call symputx("Estat_T","Estatística T"n );
436 call symputx("nc",_N_);
437 run;
438 %put &Estat_T.;
439
440
441 proc iml;
442 use llremp;
443 read all var{llr} into llr;
444 close ;
445 use teste_hipotese;
446 read all var _NUM_ into LLR_OBS;
447 close;
448 nsup=j(&nc.,1,0);
449 do i=1 to 1;
```

```
IDX=loc(llr[,1]>=LLR_OBS[i,1]);
nsup[i,1]=ncol(IDX);
end;
create NSUP_LLR from nsup[Colname={"N_sup_LlR"}];
append from nsup;
quit;

data Etatistica_Cluster_detectado ;
merge teste_hipotese NSUP_LLR;
p_valor=(1+N_sup_LlR )/(&nsim.+1);
alpha=(&alpha.);
run;

/*Valor critico*/
data _null_ ;
set Teste_Hipotese;
call symputx("P95",P95);
run;
/*Grafico Distribuicao Empririca*/
title "Distribuicao  da estatistica Log RV";
/*ods listing gpath="&file.\";*/
/*ods graphics on / imagename="DistEmp" reset=index noborder
   imagefmt=png ;*/
proc sgplot data=LLREMP noborder;
histogram LLR ;
refline &Estat_T. / axis=x  lineattrs= (color=red thickness=2)
name="Est" legendlabel="Estatística T Observada=&Estat_T.";
refline &p95.   /axis=x lineattrs=(color=blue pattern=2 thickness=2
    )
name="T-Crítico" legendlabel="T-Crítico=&p95."   labelattrs=(color=
   dabr  size=12pt) ;
yaxis label="Porcentagem"  labelattrs=(color=dabr  size=12pt)  ;
run;


%mend treescancircular;

treescancircular(data_1=pop_reg_centroide_RJ, data_2=NCASOS_ARVORE_
   REG_RJ, data_3=ARVORE_ID,var_geo=
POPULACAO cord_x cord_y ,Arvore_id=NODEID nsim=1000,pmax=0.50,
   alpha=0.05,seed=1234)
```