Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Estatística de varredura espacial aplicada às ocorrências aeronáuticas na Aviação Civil no Território Nacional

por

Mariana Fehr Nicacio

Brasília, 25 de setembro de 2024

Estatística de varredura espacial aplicada às ocorrências aeronáuticas na Aviação Civil no Território Nacional

por

Mariana Fehr Nicacio

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. André Luiz Fernandes Cançado

Brasília, 25 de setembro de 2024

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. André Luiz Fernandes Cançado

Orientador, EST/UnB

Prof. Dr. Antônio Eduardo Gomes

EST/UnB

Prof. Dr. Fernando Luiz Pereira de Oliveira

DEEST/UFOP

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos ao meu orientador, Prof. Dr. André Luiz Fernandes Cançado, e aos professores do PPGEST/UnB, cuja orientação, paciência e incentivo foram fundamentais para a realização deste trabalho.

Agradeço profundamente à minha avó Bartira, à minha irmã Thaís, à minha tia Angélica, e à minhas amigas Fernanda, Fran e Luana, por todo o apoio, compreensão e carinho ao longo desses anos. Vocês foram minha base e fonte de força, sempre acreditando em mim, mesmo nos momentos mais desafiadores. Aos amigos do mestrado, meu sincero reconhecimento por cada momento compartilhado e pelas trocas enriquecedoras que contribuíram para o sucesso desta jornada.

Por fim, um agradecimento especial ao meu companheiro de vida, Vinnícius, por seu amor incondicional, paciência e apoio constante. Sua presença foi uma fonte de motivação e tranquilidade, permitindo que eu me dedicasse a este trabalho com confiança e serenidade.

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

Resumo

Utilizando os dados de ocorrências aeronáuticas da Aviação Civil no Brasil nos anos de 2013 a 2022, o trabalho aplica a estatística de varredura espacial (estatística Scan), proposta por Kulldorff (1997), para identificar clusters significativos dessas ocorrências. A análise envolve a construção de matrizes de distâncias entre cidades e a definição de zonas candidatas a clusters, utilizando janelas circulares através da ferramenta *SatScan*. Foram aplicados os modelos binomial, multinomial e ordinal, e por meio de simulações de Monte Carlo foram obtidos os p-valores, permitindo a avaliação da significância dos clusters detectados por cada modelo. Os resultados de cada modelo foram comparados, destacando a localização e a extensão dos clusters de acidentes, incidentes graves e incidentes, proporcionando *insights* valiosos para a prevenção e a investigação de acidentes aeronáuticos. Os resultados mostraram que o modelo binomial, que é o mais amplamente utilizado, nem sempre é o mais adequado para dados com mais de duas categorias.

Palavras-chave: Estatística. Estatística Scan. Cluster espacial. Aviação civil. Ocorrência aeronáutica. Análise geográfica.

Abstract

Using aviation occurrence data from Civil Aviation in Brazil from 2013 to 2022, this study applies spatial scan statistics, proposed by Kulldorff (1997), to identify significant clusters of these occurrences. The analysis involves constructing distance matrices between cities and defining candidate cluster zones using circular windows through the *SatScan* tool. Binomial, multinomial, and ordinal models were applied, and Monte Carlo simulations were used to obtain p-values, allowing for the assessment of the significance of the clusters detected by each model. The results of each model were compared, highlighting the location and extent of clusters of accidents, serious incidents, and incidents, providing valuable insights for the prevention and investigation of aviation accidents. The findings showed that the binomial model, while the most widely used, is not always the most suitable for data with more than two categories.

Keywords: Statistics, Scan Statistics, Spatial Cluster, Civil Aviation, Aviation Occurrence, Geographical Analysis.

Sumário

Lista de Tabelas

Lista de Figuras

Abreviações e Siglas

CENIPA	Centro de Investigação e Prevenção de Acidentes Aeronáuticos
EMV	Estimador de Máxima Verossimilhança
FDP	Função Densidade de Probabilidade
GAM	Máquina de Análise Geográfica
H_0	Hipótese nula
H_a	Hipótese alternativa
ICAO	*International Civil Aviation Organization*
MCA 3-6	Manual de Investigação do SIPAER 2017
MMV	Método da máxima verossimilhança
MSE	*Mean squared errors*
PAVA	Pool-Adjacent-Violators
P-valor	Nível descritivo de um teste de hipóeses
RR	Risco relativo
SIPAER	Sistema de Investigação e Prevenção de Acidentes Aeronáuticos
UF	Unidade Federativa

Capítulo 1

Introdução

O horizonte da aviação ocupa um espaço de grandes proporções no mundo, e no Brasil não é diferente. Com o fim da Segunda Guerra Mundial, o surgimento de diversas companhias aéreas e os investimentos governamentais em infraestrutura aeroportuária impulsionaram transformações profundas no setor aéreo brasileiro, moldando a indústria como a conhecemos hoje.

Originada da aviação militar, a aviação civil (comercial) expandiu significativamente, abrangendo companhias aéreas, aeroclubes e empresas especializadas, com destaque para os serviços de transporte de passageiros e carga, que são fundamentais para a economia e o desenvolvimento do país.

Dado o tamanho e a diversidade geográfica do Brasil, a aviação comercial tornou-se um dos principais meios de mobilidade, conectando importantes centros urbanos e regiões remotas. A extensa rede aeroportuária, que cobre praticamente todo o território nacional, impulsiona o turismo, o comércio e o desenvolvimento econômico, especialmente nas regiões mais distantes.

A aviação comercial tem sido marcada por constantes inovações tecnológicas e aprimoramentos ao longo dos anos. A credibilidade e a segurança das viagens aéreas são fatores fundamentais que impulsionam esse desenvolvimento e servem como norteadores para muitas dessas evoluções. Pensando na prevenção e segurança das pessoas que utilizam o serviço, o Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA) foi criado em 1971, por meio

do Decreto nº 69.565, como órgão do Comando da Aeronáutica responsável pelo Sistema de Investigação e Prevenção de Acidentes Aeronáuticos (SIPAER).

A criação do CENIPA trouxe uma nova filosofia para o país, substituindo a ideia de inquérito por uma abordagem voltada exclusivamente para a "prevenção de acidentes aeronáuticos". As investigações são embasadas no Anexo 13 à Convenção Internacional de Aviação Civil da ICAO –*International Civil Aviation Organization*, órgão de referência mundial, que normatiza as leis sobre aviação civil internacional.

A atividade desenvolvida pelo Centro baseia-se na análise técnico-científica do acidente ou incidente aeronáutico, de onde se retiram valiosos ensinamentos. Esse aprendizado, transformado em linguagem apropriada, é traduzido em recomendações de segurança específicas e objetivas para os fatos analisados, acarretando ao seu destinatário (proprietário, operador de equipamento, fabricante, piloto, oficina, órgão governamental, entidade civil, etc.) o cumprimento de ação ou medida que possibilitem o aumento da segurança (Força Aérea Brasileira, 2024).

Neste trabalho, foram analisados os dados de acidentes e incidentes aéreos notificados ao CENIPA nos anos de 2013 a 2022, visando identificar clusters espaciais – regiões onde o número de notificações severas é significativamente maior do que o esperado. Esta análise visa esclarecer quais regiões apresentam concentrações anômalas, permitindo uma melhor compreensão dos fatores que contribuem para a ocorrência dos eventos.

A análise utilizou três modelos para identificar regiões atípicas em todo o território nacional: binomial, multinomial e ordinal. Esses modelos foram aplicados aos dados usando a ferramenta de análise espacial *SaTScan*, que implementa testes de razão de verossimilhança para cada hipótese de interesse.

Em resumo, as extensões do teste de razão de verossimilhança para os modelos permitem identificar e avaliar estatisticamente a presença de clusters espaciais de eventos em uma determinada área geográfica. Essas extensões consideram as características específicas dos dados e suas classificações, proporcionando uma análise mais detalhada e precisa dos padrões espaciais

das ocorrências.

Os detalhes desses modelos estão apresentados no Capítulo 3.

1.1 Objetivo

A análise geoespacial do estudo consiste em comparar os modelos binomial, multinomial e ordinal utilizando dados da aviação civil, a fim de identificar regiões que apresentaram o número de ocorrências aeronáuticas, notificadas ao CENIPA, de caráter mais severo ao esperado.

A identificação das regiões críticas, como objetivo principal, tem o propósito de contribuir para a prevenção de novos acidentes, além de auxiliar na elaboração de políticas de segurança de voo.

De forma mais específica, os objetivos do trabalho são:

- comparar os resultados obtidos pelos modelos binomial, multinomial e ordinal quando aplicados aos dados de notificações de ocorrências aeronáuticas na aviação civil.

- identificar as regiões significativas encontradas pelos modelos, com especial foco nas áreas que apresentam um número de ocorrências aeronáuticas graves superior ao esperado.

Esses objetivos visam fornecer uma análise comparativa abrangente dos modelos estatísticos utilizados na identificação de clusters de ocorrências aeronáuticas e na avaliação de regiões críticas para a segurança de voo.

1.2 Organização do texto

O texto está organizado da seguinte forma: no Capítulo 2 é feita uma revisão bibliográfica, apresentando os principais trabalhos que foram utilizados como referência no presente estudo. No Capítulo 3 é apresentada uma descrição detalhada das versões da estatística de varredura es-

pacial para os modelos binomial, multinomial e ordinal encontrados na literatura. Os resultados e conclusões são discutidos nos Capítulos 4 e 5, respectivamente.

Capítulo 2

Revisão Bibliográfica

Neste capítulo fazemos uma breve revisão dos trabalhos que precederam e contribuíram para o desenvolvimento da estatística de varredura espacial de Kulldorff.

2.1 Mapas baseados em probabilidade de Choynowski, 1959

Choynowski, 1959, ao estudar a distribuição de tumores cerebrais, utilizou um tipo de mapa estatístico que permite estudar a distribuição espacial de um fenômeno sem o risco de tirar conclusões baseadas em variações aleatórias não significativas. Este método pode ser utilizado para qualquer fenômeno distribuído geograficamente.

No Artigo, Choynowski notou irregularidades geográficas marcantes, e percebeu que os conglomerados com desvios significativos, sem explicações claras, como diferenças na qualidade dos cuidados médicos ou composição etária, tinham populações pequenas. Consequentemente, mesmo pequenas diferenças em frequências absolutas resultavam em grandes diferenças nas taxas, o que poderia ser atribuído a variações amostrais.

Ele propôs um mapa que não apresentava a incidência de tumores, mas a probabilidade dessas incidências caso a verdadeira incidência fosse distribuída de forma uniforme em toda a área. Como os tumores cerebrais são relativamente raros, a distribuição de Poisson foi adequada

para calcular a probabilidade do número de tumores.

Antes da técnica de Choynowski, os mapas estatísticos eram utilizados apenas como uma ferramenta descritiva, apresentando frequências absolutas ou porcentagens para representar a distribuição geográfica do objeto de estudo. Embora essa abordagem seja útil para descrever os dados, ela apresenta limitações quando se deseja fazer inferências baseadas na distribuição espacial.

2.2 Distribuição do tamanho do cluster máximo de pontos em uma linha - Naus, 1965

O artigo explora a probabilidade de formação de agrupamentos em pontos distribuídos aleatoriamente ao longo de uma linha. O trabalho foca em identificar a probabilidade de que um subintervalo de um determinado comprimento contenha um certo número de pontos.

Portanto, N pontos independentes são extraídos da distribuição Uniforme $(0, 1)$. Denotado por $E(n \mid N; p)$, o evento de interesse é a existência de um subintervalo de $(0, 1)$ de tamanho p que contém pelo menos n dos N pontos. Naus encontra a probabilidade $P(n \mid N; p)$ de $E(n \mid N; p)$ para $n > N/2$.

Embora este caso trate de pontos em uma linha reta, ou seja, em apenas uma dimensão, essa técnica, pode ser aplicada à análise de clusters espaciais. Isso é possível se, em vez de utilizarmos as coordenadas (latitude, longitude), utilizarmos a soma dessas coordenadas, reduzindo o problema a uma única dimensão.

2.3 Agrupamento de pontos aleatórios em duas dimensões - Naus, 1965

Nesse artigo, Naus aborda a questão dos agrupamentos de pontos em duas dimensões. Ele investiga o problema de identificar clusters de pontos aleatórios dentro de um plano bidimensional, onde N pontos são escolhidos aleatoriamente de um quadrado unitário e representados por suas coordenadas. O evento de interesse é encontrar um subretângulo dentro desse quadrado, com lados u e v, que contenha pelo menos n dos N pontos escolhidos.

Naus desenvolve uma metodologia para determinar a probabilidade de ocorrência desse evento, apresentando teoremas e provas matemáticas que fornecem limites superiores e inferiores para essa probabilidade. Esses limites convergem à medida que u e v se aproximam de zero.

O artigo também discute as aplicações práticas dessa metodologia, como na análise de imagens microscópicas e na distribuição de navios no oceano. A técnica apresentada é relevante para diversas áreas, incluindo epidemiologia, onde a identificação de aglomerados de casos de doenças é essencial, bem como outras áreas que necessitam detectar padrões em distribuições espaciais aleatórias.

2.4 Investigação de clusters de Leucemia utilizando uma máquina de análise geográfica - Openshaw, 1988

O artigo de Stan Openshaw e Birch, 1988, detalha uma nova abordagem para identificar grupos de casos de leucemia infantil no norte da Inglaterra usando uma ferramenta de análise espacial chamada Máquina de Análise Geográfica (GAM).

A motivação para este estudo surgiu da preocupação pública sobre potenciais aglomerados de leucemia em torno da usina de reprocessamento nuclear de *Sellafield*. O estudo incluiu 853 crianças diagnosticadas com leucemia linfoblástica aguda entre 1968 e 1985.

O método GAM desenhou vários círculos sobrepostos de raios variados em toda a área de estudo para identificar grupos significativos de casos de leucemia. Cada círculo foi avaliado usando o teste de significância de Monte Carlo para comparar o número observado de casos com conjuntos de dados gerados aleatoriamente.

Os resultados sugerem que a distribuição dos casos de leucemia não é aleatória (distribuição de Poisson), indicando clusters verdadeiros. O artigo apresentou um método inovador para análise espacial de agrupamentos de doenças, destacando a importância da utilização de técnicas computacionais avançadas para analisar dados epidemiológicos complexos.

2.5 A detecção de clusters de doenças raras - Besag e Newell, 1991

Besag e Newell, 1991, realizaram um estudo para identificar agrupamentos de doenças raras. Eles dividiram a região de estudo em pequenas zonas administrativas, associando o número de observações em cada zona a um centroide. Em cada uma dessas zonas, eles realizaram testes de significância, onde a i-ésima zona era o centro, e outras zonas eram agregadas por distância até atingir um número pré-definido de casos. A estatística do teste foi o número mínimo de zonas necessárias para acumular pelo menos c casos, seguindo a ordenação por distância. O nível de significância do teste foi calculado de forma aproximada usando uma probabilidade dada pela distribuição Poisson.

O artigo discute as armadilhas comuns na aplicação de testes de agrupamento a dados epidemiológicos e também propõe uma técnica para a identificação de pequenos agrupamentos de doenças. Ele destaca que, enquanto os testes de agrupamento investigam se um padrão observado poderia ter surgido por acaso, os testes de detecção de agrupamentos examinam grandes regiões em busca de "*hot spots*" de doenças sem preconcepções sobre suas localizações prováveis.

Os métodos propostos por Besag e Newell têm menor carga computacional e uma base estatística mais sólida em comparação com métodos anteriores, como o de Stan Openshaw e Birch, 1988. Eles exemplificaram sua metodologia com dados de leucemia infantil no norte da Inglaterra entre os anos de 1975 e 1985, destacando que a detecção de agrupamentos é uma fase inicial crucial na determinação da etiologia de doenças raras.

Apesar das críticas sobre a produção de agrupamentos espúrios, Besag e Newell defendem que métodos de testes múltiplos de significância são essenciais para a identificação inicial de áreas que merecem investigação mais aprofundada. Uma boa medida de diagnóstico, segundo os autores, é criar um mapa com todos os círculos que foram significativos a um nível de 5%.

Como desvantagem, eles apontam que é necessário definir a priori um tamanho mínimo para o agrupamento. Contudo, a técnica desenvolvida por eles mostrou-se vantajosa por sua menor

carga computacional em relação ao método GAM e por sua fundamentação matemática mais robusta, permitindo um diagnóstico inicial eficiente dos clusters de doenças raras.

Capítulo 3

Metodologia

A análise de clusters, também conhecida como análise de agrupamento, é uma abordagem exploratória comumente usada para descobrir estruturas intrínsecas nos dados. Ela visa identificar padrões, grupos ou segmentos de observações semelhantes, agrupando-os de acordo com determinadas características ou medidas de similaridade.

No conjunto de dados, os casos são observados ao longo de um espaço geográfico. No presente contexto, cada caso pode ser, por exemplo, a ocorrência de um acidente, um incidente grave ou um incidente aeronáutico. A análise de cluster espacial consiste na utilização da distribuição geográfica desses dados para a detecção e identificação de regiões que apresentam incidência significativamente alta no número de casos. A estatística de varredura, ou do inglês, *scan*, apresentada na literatura por Naus em 1965 e Kulldorff em 1997, descreve essa abordagem da análise de pontos através de uma varredura ao longo do mapa em estudo.

A estatística *scan* de Kulldorf é utilizada para identificar áreas de alto risco ou com comportamentos anômalos, através de uma varredura no mapa de estudo. Essa técnica pode ser aplicada em diversas áreas. Ela detecta clusters em processos pontuais multidimensionais, sendo eficaz para dados categorizados. O método utiliza janelas circulares com centros e raios variáveis ao longo da região, comparando possíveis agrupamentos com o restante do mapa (Sant'Anna, 2020).

Para isso, a área geográfica de interesse é dividida em regiões menores, como áreas circulares, e a frequência de eventos em cada uma dessas regiões é comparada com a frequência esperada sob a hipótese nula.

Os modelos de distribuição dos dados mais populares utilizados são Bernoulli/Binomial e Poisson (Kulldorff, 1997), aplicados frequentemente a dados de contagem, como a prevalência de doenças epidemiológicas por região. Modelos com distribuição Exponencial (Huang, Kulldorff e Gregorio, 2007) e Weibull (Bhatt e Tiwari, 2014) também têm sido explorados na literatura. Em casos em que os dados apresentam múltiplas categorias sem uma ordem intrínseca entre elas, como em tipos específicos de variantes de doenças, o modelo multinomial (Jung, Kulldorff e Richard, 2010) pode ser usado para orientar as análises e identificar possíveis clusters. Quando há uma ordem entre as categorias, o modelo ordinal é o mais apropriado (Jung, Kulldorff e Klassen, 2007).

Nas próximas seções iremos detalhar os modelos e procedimentos necessários para a detecção e identificação de clusters espaciais. De modo geral, cada modelo leva a uma estatística de teste, baseada na razão de verossimilhança. Essa estatística deve ser calculada para cada candidato a cluster z, chamado *zona*, em um conjunto pré-estabelecido Z de zonas. O cluster mais verossímil é aquele que maximiza a estatística de teste, dentre todas as zonas $z \in Z$. Comumente, o conjunto Z é obtido através de janelas circulares, em um procedimento que será detalhado mais adiante.

Para avaliar a significância estatística do cluster candidato, é utilizado um procedimento baseado em simulações de Monte Carlo (Dwass, 1957), uma vez que a distribuição da estatística de varredura espacial sob a hipótese nula não pode ser obtida de forma analítica fechada. Assim, para realizar a inferência, é necessário gerar um grande número de conjuntos de dados aleatórios sob a hipótese nula. Em seguida, a estatística de teste é calculada para cada conjunto gerado aleatoriamente, obtendo-se assim uma amostra dessa estatística sob a hipótese nula (H_0).

Desta forma, a utilização da estatística *scan* envolve áreas diversas, como estatística espacial, estatística computacional e inferência não-paramétrica.

3.1 Estatística de varredura espacial para dados binomiais

A estatística de varredura espacial para dados binomiais compara a proporção de eventos observados dentro da área de varredura com a proporção esperada com base na distribuição espacial assumida pela hipótese nula, que sugere a igualdade dessas proporções. Os resultados obtidos por meio dessa análise espacial são valiosos para identificar áreas geográficas onde a ocorrência de eventos é estatisticamente significativa, indicando regiões que se destacam com altos índices.

A seguir, considere que dispomos de um mapa dividido em R regiões e seja z um subconjunto conexo de regiões próximas deste mapa, chamado de zona. Definimos Z como a coleção de todas as possíveis zonas. Consideremos as hipóteses

$$\begin{cases} H_0 : p_z = p_0, \forall z \in Z \\ H_a : \text{existe uma zona } z \text{ tal que } p_z > p_0, \end{cases}$$

em que p_z representa a probabilidade de que uma ocorrência em z seja um caso e p_0 representa essa probabilidade fora da zona z. Sob a hipótese nula, a probabilidade de ocorrência na zona z é considerada constante em todo o espaço geográfico, ou seja, não há diferença significativa na probabilidade dos casos de interesse em relação à probabilidade de referência p_0 na zona z.

A hipótese alternativa é de que há pelo menos uma zona z onde a probabilidade p_z de que uma ocorrência seja um caso é maior do que o valor p_0. Isso sugere a presença de um cluster espacial onde a probabilidade de ocorrer os casos é significativamente maior do que o esperado.

Portanto, cada zona z que atende à condição

$$\frac{c_z}{n_z} > \frac{C - c_z}{N - n_z}$$

se torna um cluster candidato, sendo identificados aglomerados com taxas elevadas (a desigualdade oposta se aplica quando se procura por conglomerados com taxas baixas). Assumindo

que, c_i são os casos e n_i o número total de ocorrências (ou população) na região i, $N = \sum_i n_i$ representa o total de ocorrências e $C = \sum_i c_i$ o total de casos em todo o mapa de estudo.

A região associada ao máximo da estatística de teste é definida como o cluster mais provável, e a significância estatística é determinada por testes de hipóteses de Monte Carlo (Kulldorff, 1997). A função de verossimilhança do modelo binomial sob a hipótese nula, em que $p_z = p_0$ é expressa por

$$L_0 = \left[\prod_{i=1}^{R} \binom{n_i}{c_i}\right] p_0^C (1-p_0)^{N-C}. \tag{3.1}$$

Aplicando o logaritmo em L_0, temos

$$\log(L_0) = \sum_i \log \binom{n_i}{c_i} + C\log(p_0) + (N-C)\log(1-p_0). \tag{3.2}$$

Dessa forma, para o ponto de máximo, deriva-se a equação 3.2, em relação a p_0 e, igualando a zero, encontramos o valor do parâmetro p_0 estimado ($\hat{p_0}$). Assim,

$$\frac{d}{dp_0}\left(\log(L_0)\right) = \frac{C}{p_0} - \frac{N-C}{1-p_0}$$

$$\frac{C}{p_0} - \frac{N-C}{1-p_0} = 0.$$

e multiplicando ambos os lados por $p_0(1-p_0)$ para simplificar, tem-se

$$C(1-p_0) - p_0(N-C) = 0$$

$$C - Cp_0 - p_0N + Cp_0 = 0$$

$$C - p_0N = 0$$

$$\hat{p_0} = \frac{C}{N}. \tag{3.3}$$

Portanto, o valor estimado $\hat{p}_0$ que maximiza a log-verossimilhança é expresso pela equação 3.3.

Seja $c_z = \sum_{i \in z} c_i$ o número de casos nas regiões i que pertencem à zona z, $c_{\bar{z}} = C - c_z$ o número de casos fora da zona z, n_z o total de ocorrências na zona z e $n_{\bar{z}}$ o total de ocorrências fora da zona z. A hipótese alternativa H_a consiste em uma zona z tal que $p_z > p_0$. A razão de verossimilhança é definida como:

$$\lambda = \frac{p_z^{c_z}(1-p_z)^{n_z - c_z} p_0^{c_{\bar{z}}}(1-p_0)^{n_{\bar{z}} - c_{\bar{z}}}}{p_0^C (1-p_0)^{N-C}}. \tag{3.4}$$

3.2 Estatística de varredura espacial para dados multinomiais

Ao aplicar a estatística de varredura espacial a dados multinomiais, considera-se a distribuição multinomial das categorias em vez de uma distribuição binomial, como é comum em dados de contagem. Esta metodologia revela-se de particular importância na identificação de áreas geográficas onde a distribuição dos casos é estatisticamente discrepante do valor esperado ao acaso. Fornecendo uma ferramenta valiosa para a compreensão da dinâmica espacial de eventos multinomiais, a estatística *scan* compara se as diferenças da frequência observada nas regiões de estudo com a frequência esperada para cada categoria da variável de interesse são significativas.

Portanto, o método envolve a varredura de uma área geográfica com diferentes janelas ou regiões, destacando áreas de formato circular que variam de tamanho, avaliando a probabilidade de observar um número específico de eventos de cada categoria dentro dessas janelas em comparação com o esperado ao acaso.

Suponha que tenhamos K categorias para os casos em uma área de estudo composta por I regiões. Seja c_{ik} o número de casos pertencentes à categoria k na região i ($k = 1, ..., K, i =$

$1, ..., I$). A probabilidade de estar na categoria k é a mesma em todas as partes da área de estudo, para todos os $k = 1, ..., K$.

Assim, a função de verossimilhança do modelo multinomial pode ser descrita como

$$L(Z, p_1, ..., p_K, q_1, ..., q_K) = \prod_{k=1}^{K} \left(\prod_{i \in z} p_k^{c_{ik}} \prod_{i \notin z} q_k^{c_{ik}} \right), \tag{3.5}$$

onde p_k e q_k são as probabilidades de estar na categoria k dentro e fora da zona da zona z, respectivamente ($k = 1, ..., K$). Observa-se que $\sum_k p_k = \sum_k q_k = 1$. Sob a hipótese nula, essas probabilidades são iguais para cada categoria. A hipótese alternativa é que existe pelo menos uma categoria na qual as probabilidades não são iguais, podendo ser maior ou menor em cada região. Portanto,

$$\begin{cases} H_0 : p_1 = q_1, \ldots, p_K = q_K, \ \forall z \in Z \\ H_1 : \text{existe uma zona } z \in Z \text{ tal que } p_k \neq q_k \text{ para pelo menos um valor de } k \in \{1, \ldots, K\}. \end{cases}$$

Seja $C_i (= \sum_k c_{ik})$ o total de casos na região i, $C_k (= \sum_i c_{ik})$ o número total de casos da categoria k, e $C (= \sum_k \sum_i c_{ik})$ o número total de casos em toda a área de estudo, a estatística do teste de razão de verossimilhança é expressa como

$$\lambda = \frac{\max_{Z, H_a} L(Z, p_1, ..., p_K, q_1, ..., q_K)}{\max_{Z, H_0} L(Z, p_1, ..., p_K, q_1, ..., q_K)} = \frac{\max_Z L(Z)}{L_0}. \tag{3.6}$$

sendo

$$L_0 = \prod_k \prod_i \hat{p}_{0k}^{c_{ik}} = \prod_k \left(\frac{C_k}{C} \right)^{\sum_i c_{ik}} = \prod_k \left(\frac{C_k}{C} \right)^{C_k}, \tag{3.7}$$

em que $\hat{p}_{0k} = \hat{q}_{0k} = C_k / C$ é o Estimador de Máxima Verossimilhança (EMV) de $p_k (= q_k)$ sob a hipótese nula e

$$L(Z) = \prod_k \left(\prod_{i \in z} \hat{p}_k^{c_{ik}} \prod_{i \notin z} \hat{q}_k^{c_{ik}} \right), \tag{3.8}$$

em que $\hat{p}_{0k}$ e $\hat{q}_{0k}$ é o EMV de p_k e q_k respectivamente sob a hipótese alternativa.

O EMV é a proporção do número de casos na categoria k sobre o número total de ocorrências dentro (p) e fora (q) da zona de varredura. Ou seja,

$$\hat{p}_k = \frac{\sum_{i \in z} c_{ik}}{\sum_k \sum_{i \in z} c_{ik}} = \frac{C_k(z)}{C(z)}, \text{ e}$$

$$\hat{q}_k = \frac{\sum_{i \notin z} c_{ik}}{\sum_k \sum_{i \notin z} c_{ik}} = \frac{C_k - C_k(z)}{C - C(z)}.$$

Note que L_0 é constante em todas as regiões de varredura, uma vez que depende apenas do número total de observações em cada categoria ($C_1, \ldots, C_K$).

Para cada zona z, calculamos o logaritmo na razão de verossimilhança e a zona associada ao maior valor do $\log \lambda_z$ é o cluster candidato com maior significância no número de casos.

$$\log \lambda_z = \sum_k \left\{ C_k(z) log \left(\frac{C_k(z)}{C(z)} \right) + (C_k - C_k(z)) log \left(\frac{C_k - C_k(z)}{C - C(z)} \right) \right\} - \sum_k C_k log \left(\frac{C_k}{C} \right). \tag{3.9}$$

3.3 Estatística de varredura espacial para dados ordinais

A estatística *scan* aplicada a dados ordinais representa uma metodologia analítica destinada à identificação de padrões espaciais em conjuntos de dados que aderem a uma escala ordinal. Essa abordagem encontra aplicação em estudos que envolvem variável ordinal, caracterizada por manter uma ordem específica.

A estratégia subjacente à estatística de varredura espacial para dados ordinais envolve a exploração de uma área geográfica por meio de janelas ou regiões, com a análise objetivando detectar agrupamentos espaciais significativos nas categorias ordinais. Este método permite não apenas a avaliação da presença de padrões, mas também a determinação da significância estatística da distribuição ordinal nas áreas geográficas de estudo.

O procedimento empregado assemelha-se à abordagem utilizada para dados multinomiais, no entanto, ajusta-se à natureza ordinal das variáveis em análise. A estatística de varredura espacial compara as frequências observadas nas janelas com as frequências esperadas, levando em consideração a ordenação das categorias.

Definindo a área de estudo, e sendo ela composta por I regiões e uma variável de interesse registrada em K categorias, então c_{ik} é o número de casos na localização i e na categoria k, onde $i = 1, ..., I$ e $k = 1, ..., K$.

As categorias são de natureza ordinal. Por exemplo, um valor maior de k reflete em uma ocorrência mais grave. Seja $C_i(= \sum_k c_{ik})$ o número total de casos na região i, $C_k(= \sum_i c_{ik})$ o número total de casos na categoria k, e $C(= \sum_k \sum_i c_{ik})$ o número total de casos em toda a área de estudo, tal qual o modelo multinomial. A função de verossimilhança do modelo ordinal pode ser descrita conforme a equação 3.5. Aplicando o logaritmo na função de verossimilhança do modelo ordinal, temos

$$\log L(Z, p_1, \ldots, p_K, q_1, \ldots, q_K) = \sum_k \left(\sum_{i \in z} c_{ik} \log(p_k) + \sum_{i \notin z} c_{ik} \log(q_k) \right) \tag{3.10}$$

onde p_k é a probabilidade de um caso dentro da região de varredura z pertencer à categoria k, e q_k é a probabilidade de um caso fora da zona de varredura z pertença à categoria k. Observa-se que $\sum_k p_k = 1$ e $\sum_k q_k = 1$.

A hipótese nula é que a probabilidade de pertencer à categoria k dentro da região é a mesma que fora da região e a hipótese alternativa é pelo menos uma desigualdade estrita, não decres-

cente,

$$\begin{cases} H_0 : p_1 = q_1, \ldots, p_k = q_k \\ H_a : \dfrac{p_1}{q_1} \leq \dfrac{p_2}{q_2} \leq \ldots \leq \dfrac{p_k}{q_k}, \text{ com pelo menos uma desigualdade estrita.} \end{cases}$$

A hipótese alternativa com desigualdade estrita garante que os clusters detectados representem uma área com taxas mais altas de categorias superiores na área circular. Esse tipo de restrição de ordem é chamado de ordenação por razão de verossimilhança.

A estatística do teste de razão de verossimilhança é descrita na equação 3.6 e a diferença entre os modelos multinomial e ordinal se dá sobre as condições necessárias para o cálculo da verossimilhança, conforme a hipótese alternativa, o que leva a estimadores diferentes.

No contexto ordinal, onde se deseja ajustar curvas monotônicas, há aplicação da regressão isotônica e do Teorema de Dykstra na análise de dados espaciais.

Teorema: (Boyle e Dykstra, 1986) Seja $C_1, C_2, \ldots, C_p$ conjuntos fechados e convexos de $\mathbb{R}^n$ tais que $C = \bigcap_{i=1}^{p} C_i \neq \emptyset$. Para qualquer $i = 1, 2, \ldots, p$ e qualquer $x_0 \in \mathbb{R}^n$, a sequência $\{x_k^i\}$ gerada pelas projeções sucessivas de um ponto nos conjuntos convexos, converge para $x^* = P_C(x_0)$, isto é, $\lim_{k\to\infty} \|x_k^i - x^*\| = 0$.

A regressão isotônica é usada para ajustar uma curva monotonicamente crescente ou decrescente a um conjunto de dados de forma a respeitar a ordem das variáveis em questão. O Teorema de Dykstra, fornece essa caracterização única para a regressão isotônica restrita (Dykstra, Kochar e Robertson, 1995).

Para obter os EMVs de p_k e q_k sob a hipótese alternativa assumimos que $W_k = \sum_{i\in z} c_{ik}$, $U_k = \sum_{i\notin z} c_{ik}$, $W = \sum_k W_k$ e $U = \sum_k U_k$. Assim, $C_k = W_k + U_k$ e $C = W + U$. Assim, a hipótese alternativa deseja detectar agrupamentos com número significativo de casos para as categorias mais graves.

Pelo Teorema de Dykstra, os EMVs de p_k e q_k são

$$\hat{p}_k = \left(\frac{W_k + U_k}{W}\right) E_{(\mathbf{W}+\mathbf{U})} \left(\frac{\mathbf{W}}{\mathbf{W}+\mathbf{U}}|\tau\right)_k \tag{3.11}$$

e

$$\hat{q}_k = \left(\frac{W_k + U_k}{U}\right) E_{(\mathbf{W}+\mathbf{U})} \left(\frac{\mathbf{U}}{\mathbf{W}+\mathbf{U}}|A\right)_k \tag{3.12}$$

onde, na expressão, $\tau = \{(\theta_1, \ldots, \theta_K); \theta_1 \leq \ldots \leq \theta_K\}$ e $\mathbf{A} = \{(\theta_1, \ldots, \theta_K); \theta_1 \geq \ldots \geq \theta_K\}$ em que

$$\theta_k = \frac{W_{p_k}}{(W_{p_k} + U_{q_k})}, \text{ para } k = 1, \ldots, K.$$

Conforme descrito por Barlow, a notação $E_v(\mathbf{B}|C)$ representa a regressão isotônica de $B = (B_1, \ldots, B_K)$ com pesos $v = (v_1, \ldots, v_K)$ em C.

É possível calcular $\hat{p}_k$ e $\hat{q}_k$ utilizando o algoritmo "Pool-Adjacent-Violators" (PAVA). Conforme descrito na literatura (Ayer et al., 1955), PAVA é uma técnica utilizada para realizar a regressão isotônica em conjuntos de dados unidimensionais (Barlow et al., 1972). O procedimento do PAVA pode ser formalmente descrito da forma como se segue.

Dado um conjunto de observações $Y = (y_1, y_2, \ldots, y_n)$, onde y_i representa a i-ésima observação, o PAVA busca uma sequência monotônica $X = (x_1, x_2, \ldots, x_n)$ tal que $x_1 \leq x_2 \leq \ldots \leq x_n$, minimizando a discrepância entre X e Y.

O algoritmo PAVA realiza iterações sobre a sequência Y, identificando pares adjacentes de observações y_i e y_{i+1} que violam a ordem desejada. Em caso de violação, os valores y_i e y_{i+1} são ajustados de forma a preservar a monotonicidade. O processo é repetido iterativamente até que não haja mais violações na ordem monotônica (Salgado, 2018).

Se a razão observada $(W_k/W)/(U_k/U)$ é não decrescente para $k = 1, \ldots, K$, então os estimadores $\hat{p}_k$ e $\hat{q}_k$ são os mesmos que os Estimadores de Máxima Verossimilhança irrestritos

$$\tilde{p}_k = \frac{W_k}{W} \text{ e } \tilde{q}_k = \frac{U_k}{U}.$$

Os clusters detectados não necessariamente terão taxas para todas as categorias exatamente em ordem crescente. Pode ocorrer um cluster significativo com uma taxa alta de casos na

categoria 1 em comparação com a 2 e 3 combinados, mesmo que a taxa da categoria 3 não seja tão alta quanto a da 1 ou da 2, etc.

Para uma zona específica Z, o PAVA calcula os EMVs da seguinte forma:

1. Com a notação de W_k, U_k, W e U os EMVs irrestritos $\tilde{p}_k$ e $\tilde{q}_k$ (com $k = 1, \ldots, K$), $\tilde{p}_k = W_k/W$ e $\tilde{q}_k = U_k/U$.

2. Temporariamente, definimos $\hat{p}_k = \tilde{p}_k$ e $\hat{q}_k = \tilde{q}_k$. Se $\hat{p}_k/\hat{q}_k$ for não decrescente para todos os $k = 1, ..., K$, então esses são os EMVs de p_k e q_k:

$$\hat{p}_k = \frac{W_k}{W},$$
$$\hat{q}_k = \frac{U_k}{U}.$$

3. Caso contrário, quando $\hat{p}_l/\hat{q}_l > \hat{p}_k/\hat{q}_k$, com $k = 1, \ldots, K$, $l = k - 1$ e $l > 0$. Sendo $j = l, \ldots, k$,

$$\hat{p}_j = \frac{\sum_{j=l}^{k} W_j}{W} \frac{C_j}{\sum_{j=l}^{k} C_j},$$

$$\hat{q}_j = \frac{\sum_{j=l}^{k} U_j}{U} \frac{C_j}{\sum_{j=l}^{k} C_j}.$$

   Note que $\hat{p}_j/\hat{q}_j$ é o mesmo para $j = l, \ldots, k$, que é a taxa observada para as categorias combinadas.

Neste processo, continuamos "agrupando violadores adjacentes" de $\hat{p}_k/\hat{q}_k$ e atualizando $\hat{p}_k$ e $\hat{q}_k$ conforme necessário até que $\hat{p}_k/\hat{q}_k$ se torne não decrescente para $k = 1, \ldots, K$. Assim, a estatística do teste da razão de verossimilhança para a zona Z específica são essas estimativas atualizadas e o EMV de p_k sob a hipótese nula ($\hat{p}_{0k} = C_k/C$).

3.4 Clusters candidatos

Definido o modelo a ser utilizado, precisamos definir um conjunto Z de zonas candidatas a cluster. Uma zona é definida como um conjunto conexo de regiões do mapa. Porém, é computacionalmente inviável que avaliemos todas as possíveis zonas. Para um conjunto com n elementos, o número de subconjuntos (excluindo o conjunto vazio) é $2^n - 1$. Desses subconjuntos ainda teríamos que testar quais são conexos para, em seguida, calcular a razão de verossimilhança para cada um deles.

Obviamente, para um mapa com um número moderado de cidades já seria impraticável avaliar todos os subconjuntos conexos de cidades. Sendo assim, devemos escolher um conjunto com uma quantidade de zonas candidatas que possam ser avaliadas em tempo razoável. A forma mais usual para definir as zonas candidatas é utilizando a varredura do mapa com janelas circulares (*circular scan*).

Inicialmente obtemos uma matriz de distâncias entre as cidades que compõem o mapa com base em suas latitudes e longitudes, gerando, no nosso caso, uma matriz 1.060 por 1.060. Em seguida, construímos uma matriz de índices de vizinhos mais próximos, a partir da ordenação das colunas da matriz de distâncias de forma crescente, ou seja, da menor distância até a cidade mais distante, sendo a primeira linha da matriz a própria cidade em questão. Por exemplo, a primeira coluna da matriz será formada, na primeira linha, pela própria cidade 1, seguida, na segunda linha, pela cidade mais próxima da cidade 1, na terceira linha pela segunda cidade mais próxima da cidade 1, e assim por diante. Analogamente, a segunda coluna é formada pela cidade 2 na primeira linha, pela cidade mais próxima da cidade 2 na segunda linha, etc, conforme exemplo na matriz I abaixo.

$$I = \begin{bmatrix} 1 & 2 & \dots & 530 & \dots & 1060 \\ 1006 & 999 & \dots & 559 & \dots & 84 \\ \dots & \dots & \ddots & \dots & \ddots & \dots \\ 607 & 651 & \dots & 44 & \dots & 323 \\ \dots & \dots & \ddots & \dots & \ddots & \dots \\ 296 & 870 & \dots & 687 & \dots & 870 \end{bmatrix}$$

Para construir o conjunto Z, consideramos inicialmente a zona $z = \{1\}$ formada apenas pela cidade (região) 1; em seguida $z = \{1, 1006\}$ formada pela cidade 1 e a cidade 1006, que é a cidade mais próxima da cidade 1; depois, a zona $z = \{1, 1006, 374\}$ formada pela cidade 1 mais as duas cidades mais próximas da 1, respectivamente, 1006 e 374. Através deste procedimento iterativo, as zonas são formadas pelas cidades mais próximas, até que a zona atinja o tamanho máximo de 50% do total de cidades e/ou 50% do total de ocorrências aeronáuticas notificadas no período. Em seguida, o procedimento é reiniciado começando da zona formada apenas pela cidade 2, e depois a partir de cada uma das demais cidades até a cidade 1060.

Como as cidades vão sendo agregadas por ordem de distância, este procedimento é equivalente a definir as zonas a partir de janelas circulares centradas em cada cidade e com raios variados. Assim, as zonas tendem a possuir formato aproximadamente circular.

Para cada zona $z \in Z$, calculamos o valor da estatística de teste λ_z e a zona z que maximiza λ_z é chamada de zona mais verossímil.

3.5 Simulação de Monte Carlo

Após identificar o cluster mais verossímil, sua significância deve ser determinada para que possamos concluir o teste de hipóteses. Como a distribuição da estatística do teste sob H_0 é desconhecida, é necessário executar uma simulação de Monte Carlo para que possamos obter uma amostra desta distribuição. Esse procedimento recria o mapa original, distribuindo o número de

casos aleatoriamente sob a hipótese nula, $p_z = p_0$.

As simulações são feitas distribuindo os C casos aleatoriamente de acordo com as distribuições Binomial (dicotomizada), multinomial e ordinal, sob H_0. Os valores de λ_z de cada zona z é obtido com base no novo banco de dados gerado e seu valor máximo é registrado, da mesma forma como explicado para os dados reais observados.

Esse procedimento de geração de casos ao acaso sob H_0 e a obtenção da razão de verossimilhança é repetido S vezes, de forma que, ao fim, teremos então uma amostra de tamanho S da estatística de teste sob H_0. Assim, o p-valor pode ser calculado a partir da ordenação da amostra de tamanho S. Deste modo, obtem-se o quantil amostral a que corresponde o valor obtido para os dados reais.

Para calcular os p-valores para os clusters candidatos, foram realizadas então simulações de Monte Carlo para gerar um número de replicações aleatórias do conjunto de dados sob a hipótese nula. Quando a razão de verossimilhança apresenta um valor alto é a evidência contra a hipótese nula e a favor da existência de clusters com casos significativos.

No *SaTScan*, a comparação entre os dados reais e aqueles gerados aleatoriamente é realizada por meio de três métodos. A ferramenta combina Monte Carlo Padrão, Monte Carlo Sequencial e a Aproximação de Gumbel. Geralmente, são feitas 999 replicações, o que assegura um poder estatístico adequado para todos os conjuntos de dados analisados.

A aplicação exata de cada método varia conforme o tipo de análise, uma vez que o Monte Carlo sequencial e a Aproximação de Gumbel não são adequados para todos os tipos de dados. A seguir, estão descritos os métodos utilizados.

3.5.1 Monte Carlo Padrão

A estatística do teste é calculada para cada replicação aleatória, bem como para o conjunto de dados reais, e se este último estiver entre os 5% maiores, então o teste é significativo ao nível de 5%. Se ele estiver entre os 1% mais elevados, o teste é significativo ao nível de 0,01, e assim por diante. Este é o chamado "teste de hipótese de Monte Carlo", e foi proposto pela primeira

vez por Dwass, 1957.

3.5.2 Monte Carlo Sequencial

Com mais replicações de Monte Carlo, o poder da estatística de varredura é maior, mas também é mais demorado de se executar. Quando o p-valor é pequeno, em geral o esforço vale à pena, mas para grandes valores de "p", geralmente é irrelevante.

O *SaTScan* fornece a opção de encerrar as simulações de Monte Carlo mais cedo quando o valor de "p" é grande, empregando o teste sequencial de Monte Carlo. Com esta opção, os cálculos do *SaTScan* terminarão logo que um número fixo de replicações de Monte Carlo tenha uma razão de verossimilhança maior do que a razão de verossimilhança do conjunto de dados reais. O valor padrão é de 50 réplicas. Se o número fixo nunca for alcançado, os cálculos continuarão até que o número máximo de réplicas de Monte Carlo tenha sido alcançado.

Com os valores padrão de 50 e 999, não há perda de poder ao nível de significância de 5%, quando se compara o teste de Monte Carlo sequencial com o teste padrão (Martin Kulldorff, 2024).

3.5.3 Aproximação de Gumbel

Como uma alternativa ao teste de hipótese de Monte Carlo, é possível empregar a distribuição de valores extremos de Gumbel para estimar os p-valores aproximados. Com esta abordagem, não existe um limite inferior para os p-valores resultantes. O método funciona através da geração de 999, ou algum outro número de réplicas aleatórias dos dados sob a hipótese nula. A razão da máxima verossimilhança de cada réplica é usada para ajustar uma distribuição de Gumbel para os dados usando métodos de estimações momentâneas.

A função densidade de probabilidade (FDP) de uma variável aleatória X com distribuição

Gumbel é dada por

$$f(x) = \frac{1}{\sigma} \exp\left(-\left(\frac{x-\mu}{\sigma}\right) - \exp\left(-\frac{x-\mu}{\sigma}\right)\right),$$

onde: $X \in \mathbb{R}$, μ é o parâmetro de localização, $\sigma > 0$ é o parâmetro de escala (Silva, 2021).

Uma vez que a distribuição de Gumbel que melhor se ajusta aos dados tenha sido obtida, o p-valor é calculado como a probabilidade de que esta distribuição gere um valor maior do que a razão da máxima verossimilhança observada para o conglomerado mais provável do conjunto de dados reais.

Para a estatística de varredura puramente espacial com os modelos de probabilidade discreto de Poisson e Bernoulli, demonstrou-se que a distribuição de Gumbel se ajusta muito bem aos dados e gera p-valores muito precisos (Martin Kulldorff, 2024).

3.6 Ferramenta *SaTScan*

SaTScan é uma ferramenta estatística desenvolvida para realizar análises espaciais, temporais ou espaço-temporais, sendo amplamente utilizada para a detecção de clusters e a realização de testes de hipóteses (Martin Kulldorff, 2024).

A ferramenta identifica agrupamentos significativos de eventos, como regiões onde a incidência de um fenômeno é maior ou menor do que o esperado. Ela é baseada em uma estatística de varredura espacial desenvolvida por Kulldorff em 1997.

O *SaTScan* utiliza a abordagem de janelas circulares (ou cilíndricas no caso de análises espaço-temporais) que são movidas ao longo da região de estudo para identificar possíveis clusters. Ele calcula a razão de verossimilhança para cada janela e compara a taxa de eventos dentro da janela com a taxa fora dela, utilizando testes de Monte Carlo, comparando o número de eventos observados com os valores esperados sob a hipótese nula, a ferramenta calcular os p-valores e determinar a significância dos clusters.

3.6.1 Modelos implementados

- Modelo Binomial: usado para dados categorizados em dois grupos (por exemplo, casos e controles).
- Modelo Poisson: útil para dados de contagem, comparando o número de casos em uma região com o número de casos esperados, baseado em uma taxa de incidência.
- Modelo Multinomial e Ordinal: são usados para dados com mais de duas categorias.

SaTScan é amplamente utilizado em áreas como saúde pública, meio ambiente, segurança e estudos espaciais para fornecer padrões de eventos e apoiar a tomada de decisões.

3.7 Definições

Nessa Seção são apresentadas as definições dos parâmetros utilizados na análise dos resultados dos modelos.

3.7.1 Conceitos

- Coordenadas: são as coordenadas do centro do aglomerado.
- Raio: utilizando a latitude e longitude o raio do círculo é dado em quilômetros (km).
- População: quantidade de ocorrências observadas nos dados.
- Caso: quantidade de ocorrências em determinada categoria observada nos dados.
- Caso esperado: quantidade esperada de casos nos dados.
- Risco relativo (RR): é o risco estimado dentro do aglomerado dividido pelo risco estimado fora do aglomerado. Calculado como:

$$RR = \frac{Observado_z / Esperado_z}{Observado_{\bar{z}} / Esperado_{\bar{z}}}.$$

3.7.2 Tamanho Máximo do Aglomerado Espacial

O *SaTScan* fará a varredura para aglomerados de tamanho geográfico entre zero e algum limite superior definido. O limite superior pode ser especificado em porcentagem da população empregada na análise.

O limite superior especificado como um percentual da população sob risco, foi 50% dos casos. Um conglomerado de tamanho maior indicaria áreas com taxas excepcionalmente baixas fora do círculo, ao invés de uma área com taxa excepcionalmente alta dentro do círculo.

Visto que também utilizamos as coordenadas geográficas em latitude e longitude, estabelecemos um limite máximo de 500 km para o raio do cluster. Isso foi feito para melhorar a detecção de áreas críticas e evitar a formação de clusters excessivamente grandes.

Capítulo 4

Resultados

4.1 Análise das ocorrências

Os dados utilizados para realizar o trabalho são de domínio público e divulgados pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA) e correspondem às ocorrências aeronáuticas, no âmbito da Aviação Civil, notificadas ao CENIPA, no período de 2013 a 2022 em todo o território Nacional (Governo Federal, 2024).

As ocorrências aeronáuticas notificadas são classificadas de acordo com o MCA 3-6 (*Manual de Investigação do SIPAER* 2017). Para o trabalho, foram utilizadas as ocorrências classificadas como *acidente*, *incidente grave* ou *incidente*. Na classificação, os acidentes são as ocorrências mais severas, em que ocorrem fatalidade e/ou aeronaves destruídas. Já os incidentes graves são as ocorrências que indicaram um elevado risco de acidente relacionado à operação de uma aeronave com a intenção de voo. A diferença entre o incidente grave e o acidente está apenas nas consequências. Os incidentes são as ocorrências mais leves, que não geraram grandes danos às aeronaves ou vítimas fatais mas que afetaram a segurança da operação de uma aeronave com intenção de voo (*Manual de Investigação do SIPAER* 2017).

Para o georeferenciamento foi utilizada a informação de latitude e longitude da cidade cadastrada para cada uma das ocorrências. Quando a cidade não estava informada (nove casos)

na notificação ao Centro, usou-se a Unidade Federativa (UF) da região. Sete ocorrências sem a UF foram retiradas da análise.

Após o tratamento, o conjunto de dados ficou composto por um total de 5.136 ocorrências identificadas em 1.060 cidades distintas em todo o território brasileiro, sendo 1.595 (31,1%) acidentes, 697 (13,6%) incidentes graves e 2.844 (55,4%) incidentes. A distribuição geográfica dessas ocorrências pode ser visualizada no mapa da Figura 4.1.

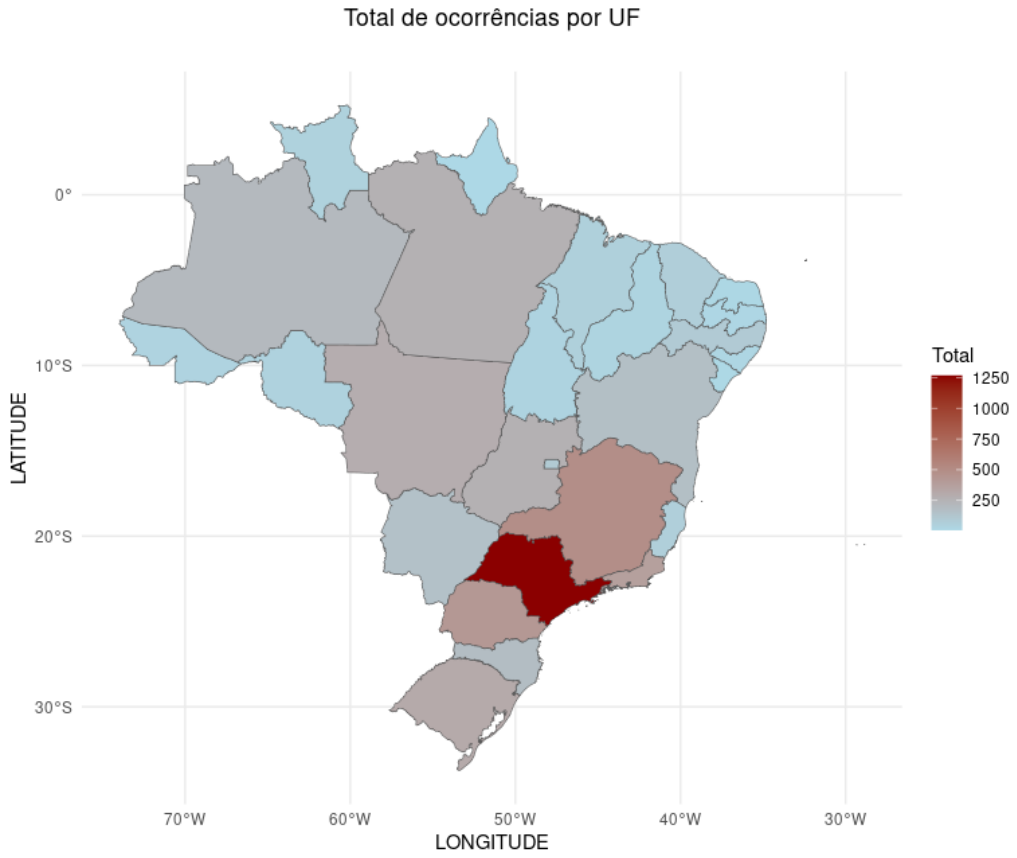

Figura 4.1: Mapa das ocorrências aeronáuticas da Aviação Civil no período de 2013 a 2022 por UF.

O estado de São Paulo (SP) foi o que apresentou o maior número de ocorrências notificadas, 1.265, representando 24,6% do total, seguido por Minas Gerais (MG), com 492 (9,6%) e Paraná (PR), com 426 (8,3%). O estado do Rio de Janeiro (RJ) aparece na quarta posição, com 379 (7,4%). Os quatro estados juntos representam cerca de 50% do total de ocorrências no estudo.

Os números dos demais estados podem ser encontrados no Anexo A.1.

Note que, SP tem mais que o dobro das ocorrências de MG. Essa diferença representa 15,0% do total e identifica que SP possui uma malha aérea intensa, sendo o estado chave para chegadas e partidas de inúmeros voos nacionais e internacionais. Outro indicativo desse alto volume da movimentação aérea do estado é o consumo de querosene de aviação. Segundo a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), SP representou mais de 50% de todo o consumo desse combustível em 2022.

O querosene de aviação, medido em m^3, é um tipo de combustível utilizado em aeronaves. Ele geralmente é de uma qualidade maior do que os outros combustíveis e contém mais aditivos para reduzir o risco de congelar ou explodir em altas temperaturas (ANP, 2023). O estado de SP aparece com o maior volume de querosene para o período de 2013 a 2022, sendo 47,2% do total, seguido por RJ e Distrito Federal (DF), com 14,6% e 6,3% respectivamente. MG aparece em quarto com 4,1% e PR em oitavo com 2,5%.

Observa-se que SP se destaca significativamente como o principal estado tanto no número de ocorrências quanto no consumo de querosene de aviação. Por outro lado, MG e PR exibem um consumo de combustível relativamente menor em comparação ao número de ocorrências registradas. Em contraste, o estado do RJ apresenta uma proporção consideravelmente maior de consumo de combustível em relação ao número de ocorrências, indicando uma demanda de querosene de aviação elevada comparada à quantidade de eventos notificados.

As informações do consumo de combustível de aviação por estado, para efeito de referência da movimentação aérea, está descrito no Anexo A.2.

Com relação às cidades, Rio de Janeiro-RJ , São Paulo-SP e Campinas-SP estão no topo das ocorrências, com 250 (4,9%), 246 (4,8%) e 185 (3,6%) notificações. Em relação aos acidentes, as principais cidades foram Itaituba-PA, Rio de Janeiro-RJ e Manaus-AM, com 28 casos (1,8%), 23 (1,4%) e 20 (1,3%), respectivamente. Para os incidentes graves foram Goiânia-GO, 37 (5,3%), Rio de Janeiro-RJ, 21 (3,0%) e São Paulo-SP, 19 (2,7%). Os incidentes apresentaram São Paulo-SP, 209 (7,3%), Rio de Janeiro-RJ, 206 (7,2%) e Campinas-SP, 176 (6,2%). A relação

das principais cidades está descrita no Anexo A.1.

A Figura 4.2 apresenta o mapa das ocorrências classificadas como acidentes na Aviação Civil. É importante destacar que esses acidentes ocorreram em 803 cidades, representando 52,1% das localidades identificadas no estudo. Embora os acidentes representem 31,1% do total de ocorrências, sua distribuição geográfica foi mais ampla em relação às demais classificações analisadas. As principais regiões foram o Sudeste, Centro-Oeste e Norte.

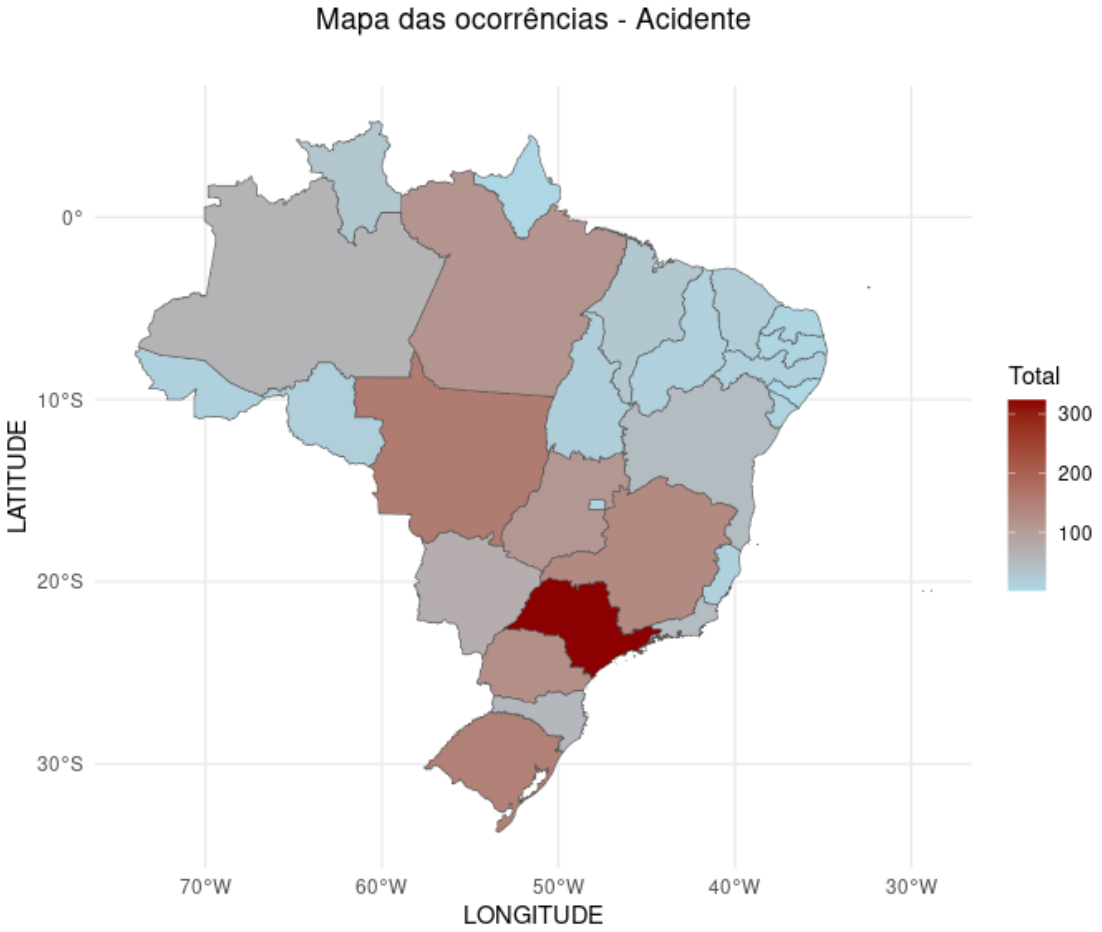

Figura 4.2: Mapa das ocorrências classificadas como acidentes da Aviação Civil no período de 2013 a 2022.

A Figura 4.3 apresenta o mapa das ocorrências classificadas como incidentes graves. As notificações dessa classificação foram registradas em 357 cidades diferentes, representando 23,2% do total. Embora a distribuição geográfica dos incidentes graves seja semelhante à dos aciden-

tes, o número de incidentes graves foi duas vezes menor.

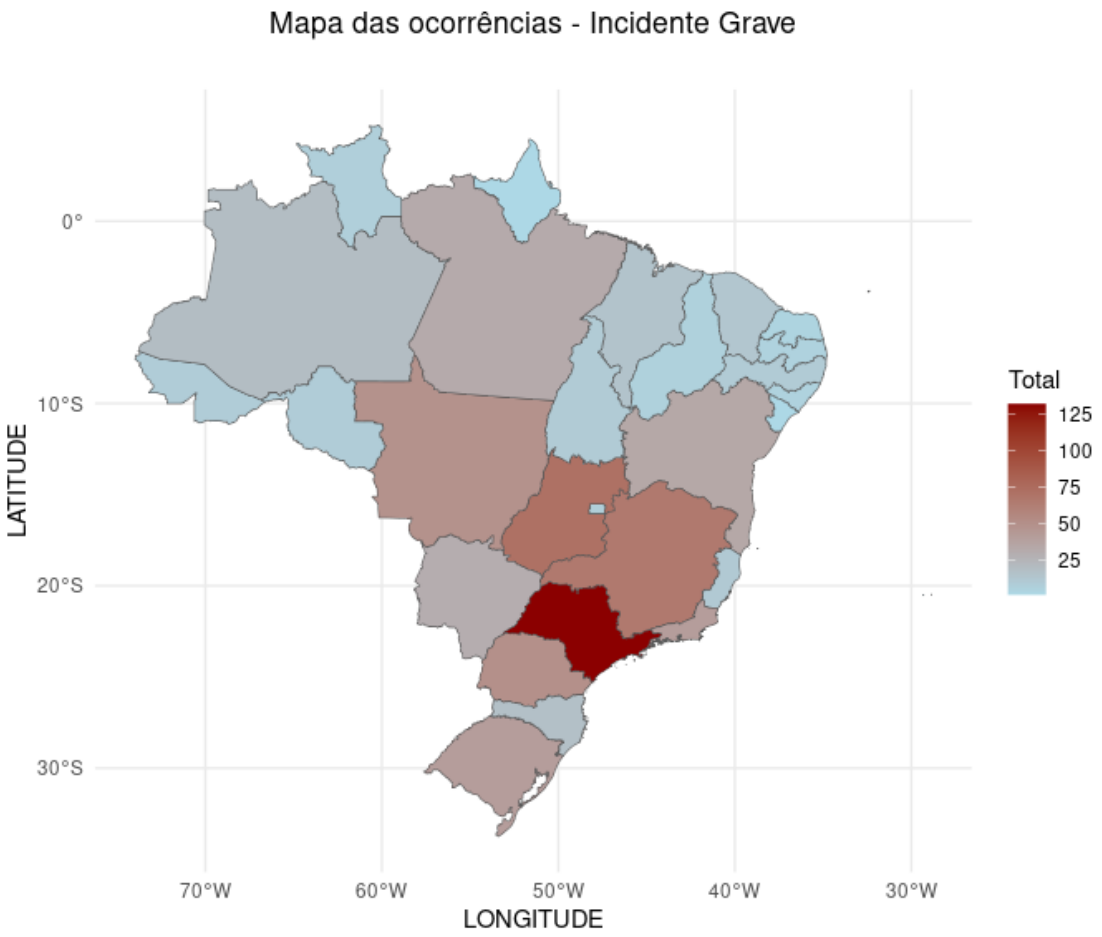

Figura 4.3: Mapa das ocorrências classificadas como incidentes graves da Aviação Civil no período de 2013 a 2022.

A Figura 4.4 apresenta o mapa das ocorrências classificadas como incidentes, registradas em 381 cidades diferentes, correspondendo a 24,7% do total. Nota-se que a região sudeste ganha destaque no volume de notificações em ambas as classificações. Apesar dos incidentes representarem mais da metade das ocorrências notificadas, a quantidade pequena de locais reforça a concentração dessas ocorrências em cidades específicas, principalmente na região Sudeste.

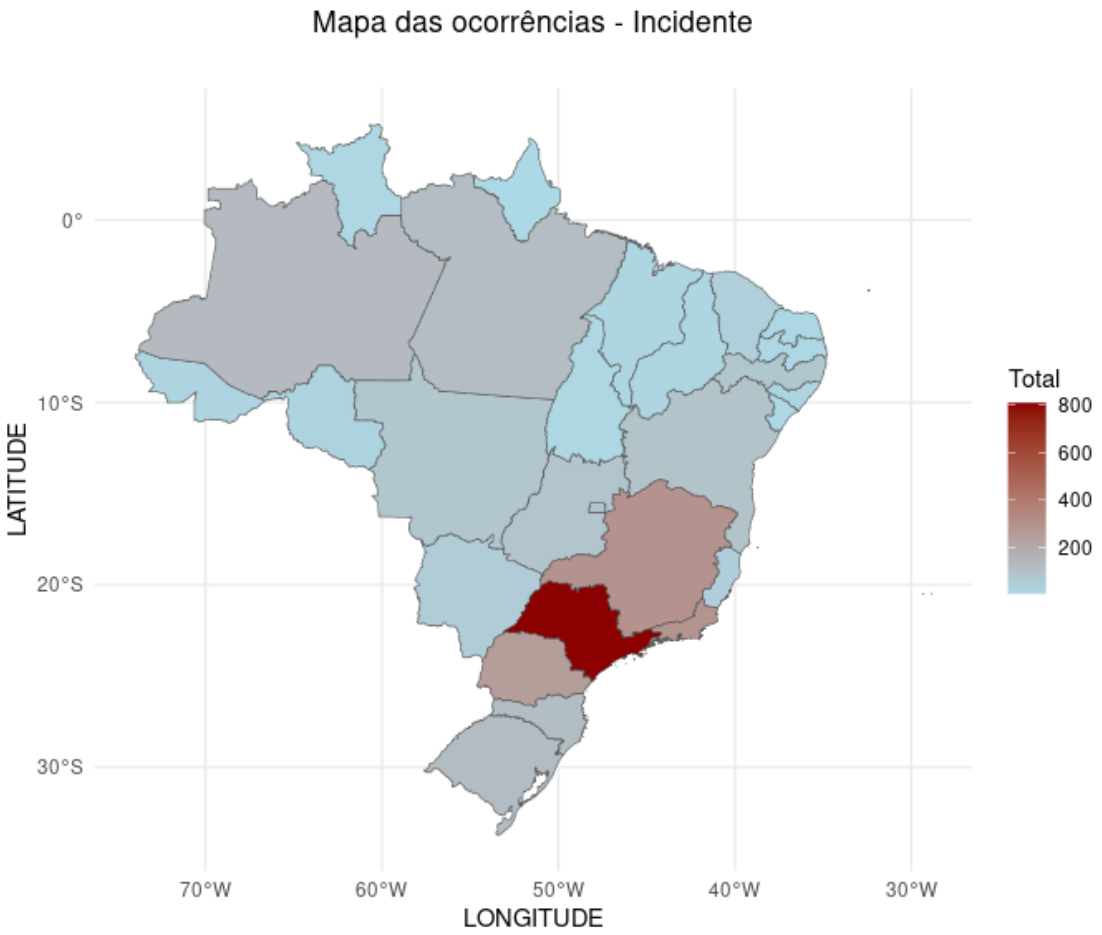

Figura 4.4: Mapa das ocorrências classificadas como incidentes da Aviação Civil no período de 2013 a 2022.

4.2 Análise dos modelos

4.2.1 Modelo binomial

Como o modelo binomial comporta apenas duas categorias (casos $vs.$ não-casos), é necessário que os dados sejam dicotomizados. Dessa forma, são apresentados os resultados para as três dicotomizações possíveis.

Casos de Acidentes

Aqui, os acidentes são considerados casos, enquanto os não-casos são os incidentes somados aos incidentes graves. O modelo binomial para os casos classificados como acidente, identificou 12 clusters significativos ao nível de 5% de confiança (Figura 4.5). Dentre essas 12 regiões, o cluster de número 1, mais significativo, registrou 102 acidentes quando o número esperado foi de 41,61. O risco relativo (RR) de uma ocorrência ser um acidente na região, com raio de 386 km e centro na cidade Sant'ana do livramento-RS, foi de 2,55.

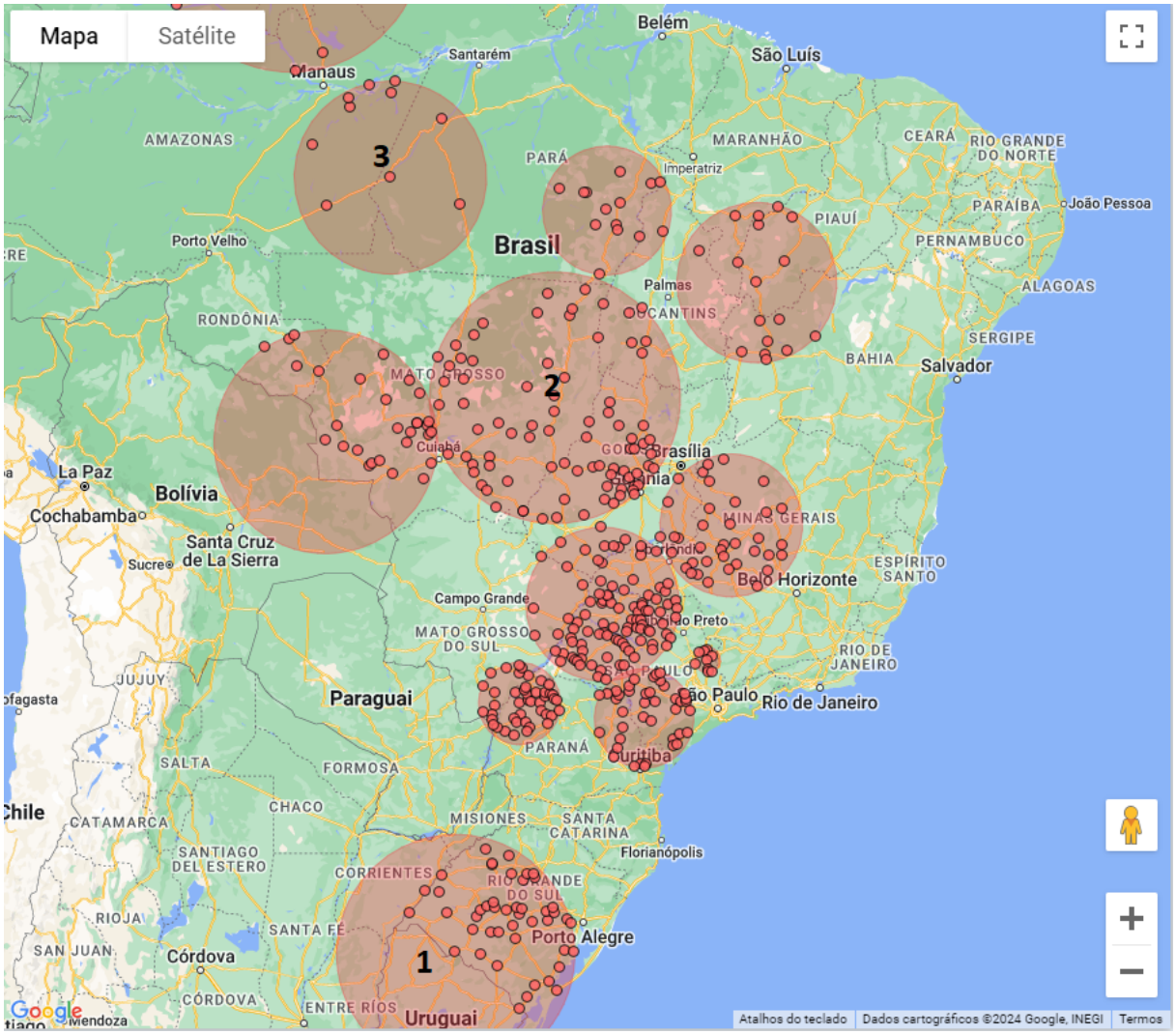

Figura 4.5: Mapa dos clusters significativos para o modelo binomial nos casos de acidente.

O cluster 2, apresentou 127 casos de acidentes e número esperado de 60,25. O RR foi de 2,20 com centro aproximado na cidade de Gaúcha do Norte-MT e com raio de 460,62 km. O cluster 3, apresentou o maior RR, de 2,79, mas a quantidade de acidentes foi menor. O número de casos na região foi 51 e o número esperado foi de 18,63. A cidade Jacareacanga-PA foi o centro e o raio foi de 361,59 km, abrangendo a região Norte, os estados de Amazonas e Pará, e um pedaço da região Centro-Oeste, do estado de Mato Grosso.

Tabela 4.1: Tabela dos clusters do modelo binomial para os casos de acidentes.

Cluster	Centro	Raio (km)	Casos	Casos esperados	RR	Cidades	p-valor
1	Sant'ana do livramento-RS	386,01	102	41,61	2,55	47	$< 1,0 \cdot 10^{-17}$
2	Gaúcha do Norte-MT	460,62	127	60,25	2,20	84	$< 1,0 \cdot 10^{-17}$
3	Jacareacanga-PA	361,59	60	18,63	2,79	10	$1,2 \cdot 10^{-13}$
4	Jales-SP	272,31	126	66,46	1,97	98	$3,9 \cdot 10^{-13}$
5	Vila Bela da Santíssima Trindade-MT	407,11	56	23,60	2,42	29	$3,2 \cdot 10^{-10}$
6	Ivaté-PR	142,54	52	21,74	2,44	44	$1,4 \cdot 10^{-9}$
7	Itaberá-SP	171,85	58	27,33	2,16	41	$1,6 \cdot 10^{-7}$
8	João Pineiro-MG	256,55	47	21,74	2,20	37	$4,0 \cdot 10^{-6}$
9	Monte Alegre do Piauí-PI	297,78	38	16,77	2,30	17	$2,0 \cdot 10^{-5}$
10	Aguaí-SP	44,96	17	5,59	3,06	9	$6,80 \cdot 10^{-5}$
11	Caracaraí-RR	497,67	37	18,63	2,01	16	0,0042
12	Bannach-PA	241,36	30	14,60	2,08	14	0,016

Casos de Incidentes Graves

Neste ponto, os incidentes graves são os casos e as demais categorias são consideradas não-casos. O modelo binomial para os casos de incidentes graves identificou dois clusters significativos ao nível de 5%. O cluster 1 tem seu centro na cidade de Paraúna-GO, com um raio de 177,93 km e um RR de 2,56. O número de casos observados foi de 62, enquanto o número esperado era de aproximadamente 26. O p-valor encontrado foi de $5,0 \cdot 10^{-8}$.

A cidade de Getúlio Vargas-RS foi o centro do cluster 2, com raio de 84,91 km, 11 cidades e p-valor de 0,035, apresentou 14 casos enquanto esperava-se 4. O RR foi de 3,39. A Tabela 4.2 detalha os clusters encontrados pelo modelo.

Tabela 4.2: Tabela dos clusters do modelo binomial para os casos de incidentes graves.

Cluster	Centro	Raio (km)	Casos	Casos esperados	RR	Cidades	p-valor
1	Paraúna-GO	177,93	62	25,65	2,56	31	$5,0 \cdot 10^{-08}$
2	Getúlio Vargas-RS	84,91	14	4,07	3,49	11	0,035

A Figura 4.6 mostra o mapa dos casos de incidente grave identificados pelo modelo.

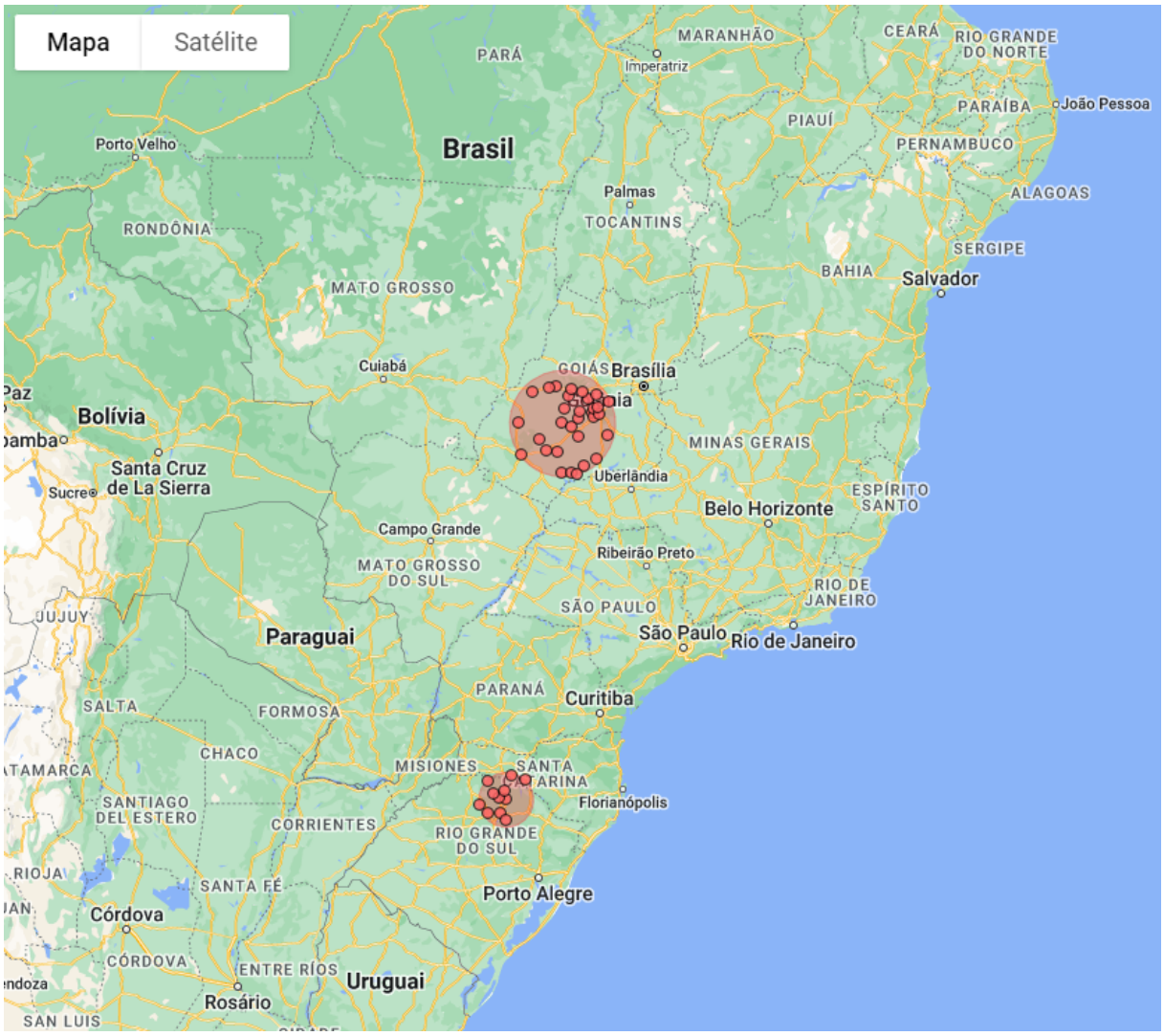

Figura 4.6: Mapa dos clusters significativos para os casos de incidente grave.

Casos de Incidentes

Finalmente, aqui consideram-se casos as ocorrências de incidentes, enquanto os acidentes e os incidentes graves são não-casos. Os clusters identificados no modelo binomial para os casos de incidentes totalizaram sete significativos ao nível de 5%. O cluster 1, o mais significativo (p-valor $< 1,0 \cdot 10^{-17}$), apresentou uma quantidade de 1.212 incidentes e o número esperado de casos foi de aproximadamente 889. O RR foi de 1,63 e o centro do círculo foi a cidade de Niterói-RJ, abrangendo também, cidades dos estados de Minas Gerais e São Paulo, com um raio de 405 km. O cluster 1 destacou-se em relação aos demais pelo número de casos observados elevado enquanto as outras regiões significativas ficaram com o número abaixo de 220 casos.

O cluster 2, apresentou seu centro na cidade de Canoas-RS, o raio foi de 14 km, abrangendo um total de três cidades. O RR foi de 1,72 e foram registrados 80 casos de incidentes. a quantidade de incidentes foi quase 15 vezes menor em relação ao primeiro cluster. O número esperado do conglomerado foi de 47 registros.

O cluster 3, foi localizado exclusivamente na cidade de Brasília-DF. Na região, ocorreram 88 incidentes, enquanto o número esperado foi de 56 casos, resultando um RR de 1,58.

Os clusters podem ser melhor analisados na tabela 4.3.

Tabela 4.3: Tabela dos clusters do modelo binomial para os casos de incidentes.

Cluster	Centro	Raio (km)	Casos	Casos esperados	RR	Cidades	p-valor
1	Niterói-RJ	405,40	1.212	888,75	1,63	135	$< 1,0 \cdot 10^{-17}$
2	Canoas-RS	13,88	80	47,07	1,72	3	$6,7 \cdot 10^{-11}$
3	Brasília-DF	0,00	88	56,48	1,58	1	$3,9 \cdot 10^{-07}$
4	Sertanópolis-PR	41,57	91	59,25	1,55	6	$9,1 \cdot 10^{-07}$
5	Camaragibe-PE	11,5	60	37,10	1,63	3	$1,4 \cdot 10^{-05}$
6	Chapadão do Lageado-SC	242,74	218	175,54	1,26	53	$3,3 \cdot 10^{-03}$
7	Lauro de Freitas-BA	41,24	52	35,44	1,48	3	0,048

A Figura 4.7 destaca os clusters identificados.

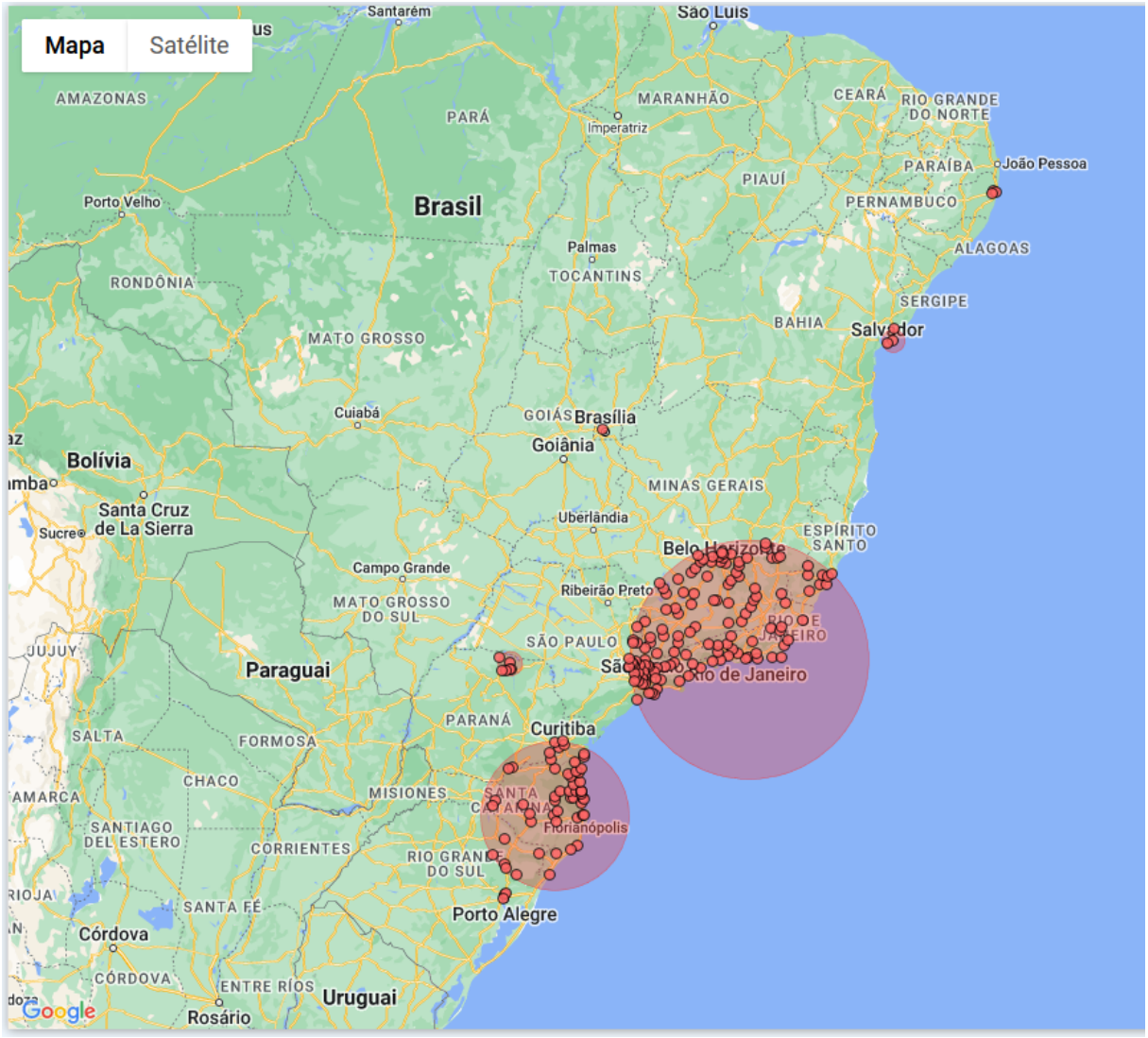

Figura 4.7: Mapa dos clusters significativos para os casos de incidente

4.2.2 Modelo multinomial

O modelo multinomial utilizou as categorias acidente, incidente grave e incidente. Foram identificados 14 clusters significativos ao nível de 5%. O cluster 1, mais significativo, foi composto por ocorrências do tipo incidente. Seu centro foi na cidade de Niterói-RJ, com raio de aproximadamente 405,40 km. Essa região apresentou 1.605 casos dentre as categorias analisadas, e o número observado de incidentes (1.212) superou o esperado em aproximadamente 324 casos. O RR apresentado foi de 1,63.

O cluster 5 também foi identificado como um cluster de incidente pelo modelo. A quantidade de casos (80) foi menor em relação ao cluster 1. A cidade de Canoas-RS foi o centro, com o raio do círculo de aproximadamente 14 km. Esperava-se 56,48 casos de incidente e, portanto, a região apresentou um RR de 1,72, com 33 casos a mais do que o esperado. Ambos os clusters (1 e 5) de incidentes também foram identificados pelo modelo binomial.

O cluster 7 e 9 foram apresentados pelo modelo como regiões de casos de incidente grave e incidente. Os centros foram em Corumbá de Goiás - GO e Aracaju - SE, respectivamente.

Para as regiões identificadas com casos de acidente, o modelo registrou 3 clusters. O cluster 8, com p-valor de $6,9 \cdot 10^{-7}$, exclusivamente na cidade de Itaituba-PA, apresentou 28 casos observados e 9,63 esperados, resultando um RR de 2,94. O cluster 11 do modelo, também identificou uma região de acidentes, com o centro em Guarda-Mor-MG e raio de 176,84 km, teve RR de 2,44, com 27 casos observados e 11,18 esperado.

Por fim, o cluster 13, com centro na cidade de Caracaraí-RR, teve RR de 2,00, com 18 casos esperados e 36 casos observados. A maior incidência das categorias foram de acidente e incidente grave. A Tabela 4.4 apresenta os clusters mais significativos do modelo multinomial.

Tabela 4.4: Tabela dos clusters do modelo multinomial.

Cluster	Centro	Raio (km)	Categoria									p-valor
			Acidente			Incidente Grave			Incidente			
			Casos	Casos esperados	RR	Casos	Casos esperados	RR	Casos	Casos esperados	RR	
1	Niterói - RJ	405,40	238	498,44	0,39	155	217,81	0,71	1.212	888,75	1,36	$< 1,0 \cdot 10^{-17}$
2	Camapuã - MS	483,60	214	115,53	1,85	68	50,48	1,35	90	205,99	0,44	$< 1,0 \cdot 10^{-17}$
3	Sant'ana do Livramento - RS	423,40	118	52,48	2,35	29	22,93	1,28	22	93,58	0,23	$< 1,0 \cdot 10^{-17}$
4	Colíder - MT	482,60	103	47,83	2,23	29	20,90	1,40	22	85,28	0,25	$< 1,0 \cdot 10^{-17}$
5	Canoas - RS	13,88	3	26,40	0,11	2	11,54	0,17	80	47,07	1,70	$3,3 \cdot 10^{-10}$
6	Riachão das Neves - BA	496,20	66	36,33	1,85	28	15,88	1,8	23	64,79	0,35	$6,0 \cdot 10^{-10}$
7	Corumbá de Goiás - GO	96,18	29	73,60	0,38	52	32,16	1,67	156	131,24	1,20	$1,5 \cdot 10^{-7}$
8	Itaituba - PA	0,0	28	9,63	2,94	1	4,21	0,24	2	17,17	0,12	$6,9 \cdot 10^{-7}$
9	Aracaju - SE	398,26	23	60,87	0,37	28	26,60	1,05	145	108,53	1,35	$6,7 \cdot 10^{-6}$
10	Candói - PR	236,68	52	30,43	1,73	20	13,30	1,52	26	54,27	0,47	$4,0 \cdot 10^{-4}$
11	Guarda-Mor - MG	176,84	27	11,18	2,44	4	4,89	0,82	5	19,93	0,25	$9,8 \cdot 10^{-4}$
12	Boa viagem - CE	201,04	13	7,14	1,83	9	3,12	2,91	1	12,74	0,078	0,0034
13	Caracaraí - RR	492,69	36	18,01	2,02	7	7,87	0,89	15	32,12	0,46	0,019
14	Santa Rosa do Purus - AC	295,57	10	5,90	1,70	8	2,58	3,13	1	10,52	0,095	0,045

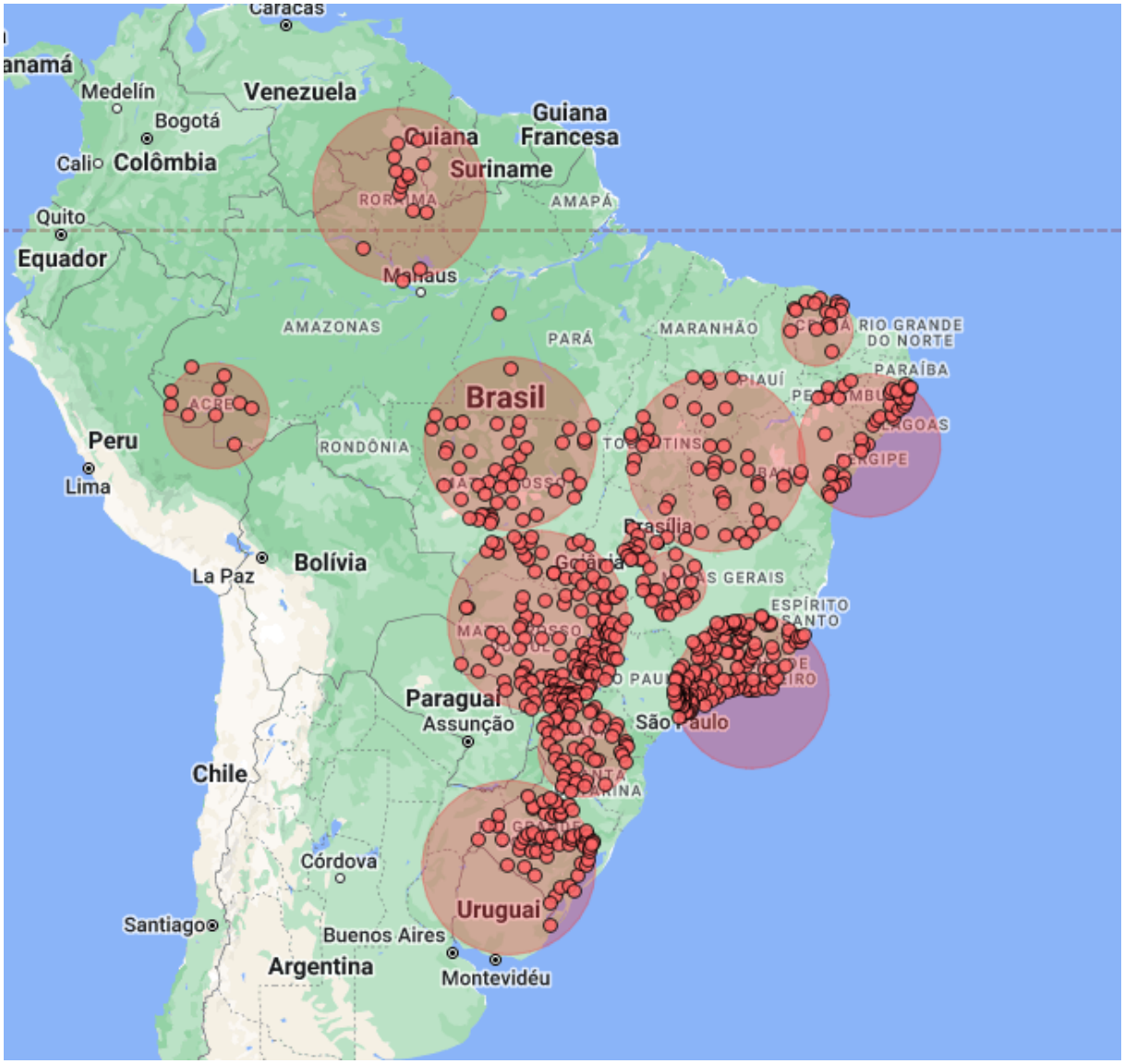

Figura 4.8: Mapa dos clusters significativos para o modelo multinomial.

4.2.3 Modelo ordinal

O modelo ordinal identificou 14 clusters significativos ao nível de 5%. Os cinco primeiros cluster foram de regiões que o número de acidentes e incidentes graves foi maior que o esperado, indicando regiões com casos de maior gravidade.

Destacam-se as regiões dos clusters 2 e 3, com centro na cidade de Sant'ana do Livramento-BA e Colíder-MT, respectivamente, que apresentaram um RR duas vezes maior nos casos ob-

servados em relação ao esperado.

O cluster 6, com p-valor de $3,2 \cdot 10^{-7}$, apresentou exclusivamente a cidade de Itaituba-PA, onde o número de acidentes (28) foi quase três vezes maior que o esperado (10). Esta mesma cidade também foi identificada pelo modelo multinomial como o cluster 8.

Em quatro conglomerados identificados pelo modelo, foram agrupados casos de acidente e incidente grave. Os clusters 7, 10, 12 e 14 registraram a incidência de casos mais severos quando consideradas essas categorias juntas.

No caso do cluster 13, identificado com centro em Campina Grande do Sul-PR, o modelo agrupou os casos de incidente grave e incidente. Esta região apresentou o maior RR, 3,23 com 10 acidentes observados comparados a um esperado de 3,22. A tabela 4.5 apresenta as informações mais detalhadas dos clusters identificados pelo modelo ordinal.

Tabela 4.5: Tabela dos clusters do modelo ordinal.

Cluster	Centro	Raio (km)	Categoria									p-valor
			Acidente			Incidente Grave			Incidente			
			Casos	Casos esperados	RR	Casos	Casos esperados	RR	Casos	Casos esperados	RR	
1	Camapuã - MS	483,60	214	115,53	1,85	68	50,48	1,35	90	205,99	0,44	$< 1,0 \cdot 10^{-17}$
2	Sant'ana do Livramento - RS	423,40	118	52,48	2,35	29	22,93	1,28	22	93,58	0,23	$< 1,0 \cdot 10^{-17}$
3	Colíder - MT	482,60	103	47,83	2,23	29	20,9	1,4	22	85,28	0,25	$< 1,0 \cdot 10^{-17}$
4	Riachão das Neves - BA	496,20	66	36,33	1,85	28	15,88	1,8	23	64,79	0,35	$3,8 \cdot 10^{-10}$
5	Macatuba - SP	169,39	103	58,07	1,83	32	25,38	1,27	52	103,55	0,49	$5,2 \cdot 10^{-10}$
6	Itaituba - PA	0,00	28	9,63	2,94	1	4,21	0,24	2	17,17	0,12	$3,2 \cdot 10^{-7}$
7	Vargem Bonita - MG	174,40	54	28,11	1,83	-	-	-	12	34,89	0,34	$8,7 \cdot 10^{-5}$
8	Juquiá - SP	87,34	17	5,9	2,90	2	2,58	0,78	0	10,52	0	$1,1 \cdot 10^{-4}$
9	Candói - PR	236,68	52	30,43	1,73	20	13,3	1,52	26	54,27	0,47	$1,4 \cdot 10^{-4}$
10	Boa viagem - CE	201,04	22	10,26	2,15	-	-	-	1	12,74	0,078	0,0017
11	Caracaraí - RR	492,69	36	18,01	2,02	7	7,87	0,89	15	32,12	0,46	0,0059
12	Socorro - SP	60,32	66	42,39	1,57	-	-	-	29	52,61	0,55	0,01
13	Campina Grande do Sul - PR	17,12	10	3,11	3,23	0	6,89	0	-	-	-	0,014
14	Santa Rosa do Purus - AC	295,57	18	8,48	2,13	-	-	-	1	10,52	0,095	0,025

Os mapas geográficos com a identificação dos clusters significativos dos modelos estão disponíveis também no apêndice.

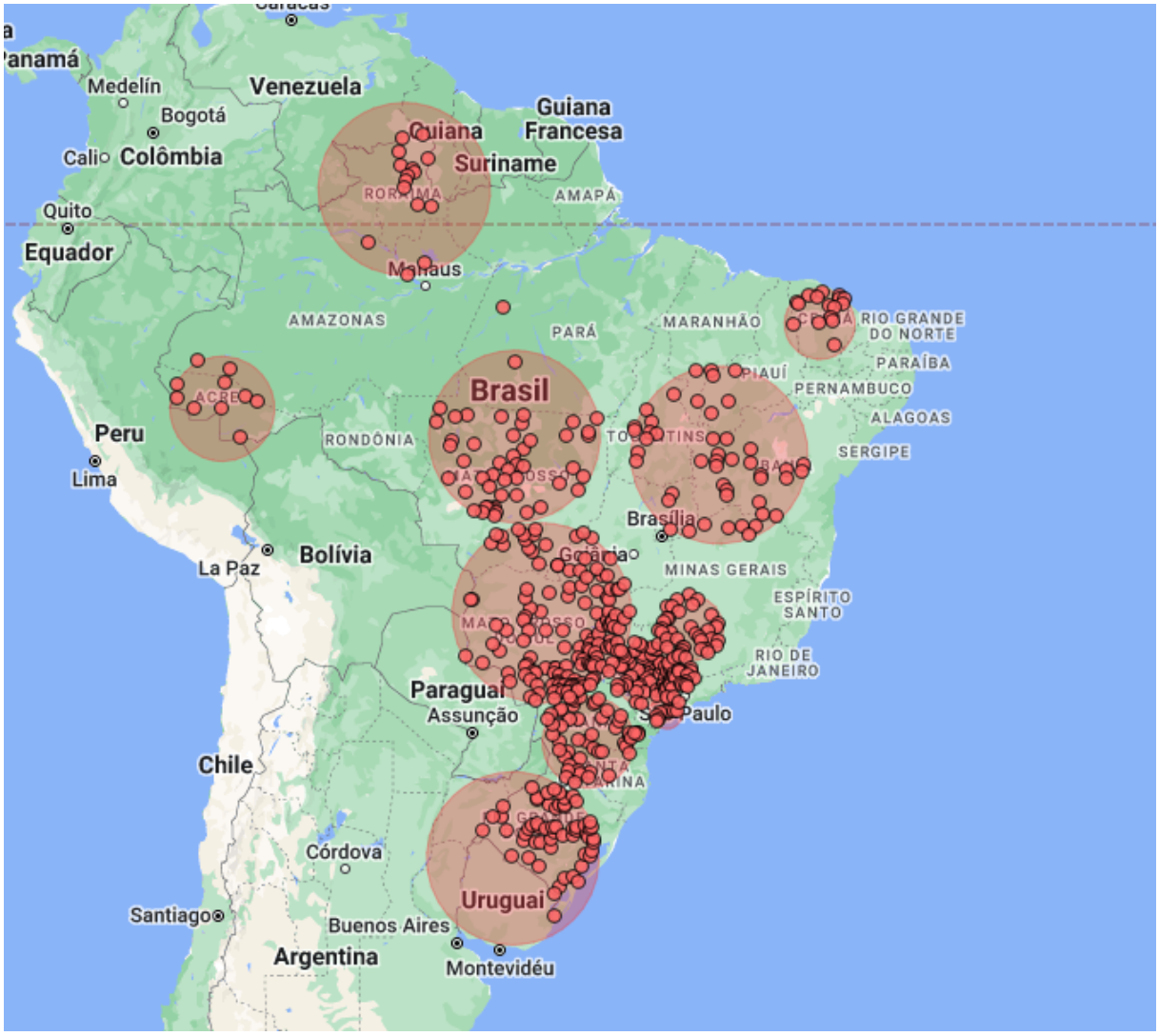

Figura 4.9: Mapa dos clusters significativos para o modelo ordinal.

Capítulo 5

Conclusões

5.1 Discussão

Ao aplicar os três modelos na análise dos dados da aviação civil, observamos comportamentos distintos entre os modelos ordinal, multinomial e binomial.

O modelo multinomial identificou um maior número de localidades (594) em comparação ao modelo ordinal (514), destacando regiões com maior risco para incidentes, particularmente nos clusters 1 e 5, e casos de incidentes graves e incidentes nos clusters 7 e 9. O modelo ordinal, por outro lado, mostrou melhor desempenho na identificação de regiões com ocorrências mais graves, sugerindo que a ordem dos casos tem um papel significativo na análise.

O modelo binomial, que compara o número de casos com um conjunto de controles, identificou clusters significativos para incidentes de maneira coerente com o modelo multinomial. O estado de Goiás (GO), por exemplo, foi identificado como uma região de casos de incidente grave e incidente, tanto pelo modelo multinomial quanto pelo modelo binomial, apesar dos centros dos clusters serem diferentes. No entanto, essas regiões não foram observadas no modelo ordinal.

A principal contribuição do modelo binomial foi sua capacidade de identificar clusters significativos para incidentes graves e incidentes, alinhando-se em parte com o modelo multinomial.

A identificação de regiões críticas por ambos os modelos multinomial e binomial, mesmo com diferentes centros e raios, ressalta a sensibilidade do modelo multinomial às categorias analisadas sem uma ordem estabelecida. Isso sugere que o modelo multinomial pode ser mais eficaz em cenários onde as categorias de eventos são variadas e não necessariamente ordenadas.

Contudo, a dicotomização das categorias, utilizada pelo modelo binomial, pode levar à perda de informações e à potencial distorção dos resultados. Nos casos classificados como acidentes, o modelo binomial apresentou padrões de identificação de regiões distintos dos modelos ordinal e multinomial, indicando que a análise individual das categorias pode preservar mais informações e permitir interpretações mais precisas. Essa abordagem diferenciada do modelo binomial pode ser vantajosa em determinadas situações, mas pode também limitar a compreensão completa dos dados.

Os modelos ordinal e multinomial apresentaram resultados similares, identificando 14 clusters significativos, dos quais nove coincidiram. Essas regiões mostraram uma incidência superior ao esperado para os casos de acidentes e incidentes graves. Apesar do número de incidentes ser mais expressivo no estudo (55,4%), com uma maior concentração de ocorrências leves em comparação com outras classificações, o modelo ordinal teve bom desempenho em ressaltar regiões com alta incidência de casos graves.

5.2 Conclusão

A análise dos dados da aviação civil revela insights importantes sobre a distribuição e a concentração das ocorrências. Ao aplicar os três modelos na análise dos dados, observamos comportamentos distintos, mas complementares, entre os modelos ordinal, multinomial e binomial. Com base nas considerações, várias observações relevantes podem ser feitas para direcionar uma discussão aprofundada.

Ambos os modelos, ordinal e multinomial, identificaram clusters significativos em regiões classificadas como acidentes e incidentes graves. A similaridade entre os dois modelos, sugere

robustez na identificação de regiões críticas. Entretanto, as diferenças encontradas indicam variações na sensibilidade do modelo binomial.

O modelo multinomial identificou 80 localidades a mais do que o modelo ordinal. Esta diferença pode ser atribuída à sensibilidade do modelo multinomial em captar variações dentro de categorias não ordenadas, podendo ser particularmente útil em contextos onde a ordem das categorias não é predefinida. Já o modelo ordinal, com seu foco em casos graves, é ideal para análises que visam identificar áreas de maior criticidade.

Os resultados indicam que ambos os modelos têm seus méritos e limitações. O modelo multinomial é vantajoso para uma visão mais abrangente e diversificada das áreas de risco, enquanto o modelo ordinal proporciona uma análise focada nas áreas de maior gravidade. A escolha entre os modelos deve ser guiada pelos objetivos específicos da análise e pela natureza das intervenções de segurança que se pretende implementar. Integrar as vantagens de ambos os modelos pode oferecer uma abordagem mais holística e eficaz na gestão de riscos na Aviação Civil.

A identificação de áreas com ocorrências mais graves pode ser crucial para a alocação de recursos e implementação de medidas de segurança mais rigorosas nas regiões de maior criticidade.

5.3 Limitações e sugestões

A análise dos dados notificados ao CENIPA entre 2013 e 2022 proporcionou um ponto de partida relevante para compreender os padrões das ocorrências aeronáuticas no Brasil. No entanto, a profundidade da investigação foi limitada pela falta de detalhamento em algumas variáveis cruciais.

Para trabalhos futuros, recomenda-se uma segmentação mais precisa das ocorrências, levando em consideração tipos específicos de aeronaves, como de pequeno e grande porte, bem como diferentes segmentos da aviação, como a aviação comercial, privada, agrícola e expe-

rimental. Essa abordagem permitiria uma análise mais robusta e detalhada dos fatores que influenciam os clusters de acidentes, incidentes graves e incidentes aeronáuticos.

Apêndice A

Tabelas

Tabela A.1: Tabela das principais cidades por ocorrência.

Cidade	UF	Ocorrências	%	Acidente	%	Incidente Grave	%	Incidente	%
Rio de Janeiro	RJ	250	4,9	23	1,4	21	3,0	206	7,2
São Paulo	SP	246	4,8	18	1,1	19	2,7	209	7,3
Campinas	SP	185	3,6	5	0,3	4	0,6	176	6,2
Belo Horizonte	MG	156	3,0	11	0,7	15	2,2	130	4,6
Guarulhos	SP	134	2,6	4	0,3	2	0,3	128	4,5
Goiânia	GO	111	2,2	14	0,9	37	5,3	60	2,1
Manaus	AM	103	2,0	20	1,3	8	1,1	75	2,6
Brasília	DF	102	2,0	6	0,4	8	1,1	88	3,1
Londrina	PR	102	2,0	9	0,6	6	0,9	87	3,1
Curitiba	PR	84	1,6	6	0,4	10	1,4	68	2,4
Porto Alegre	RS	82	1,6	3	0,2	2	0,3	77	2,7
Recife	PE	65	1,3	3	0,2	4	0,6	58	2,0
Belém	AL	65	1,3	7	0,4	5	0,7	53	1,9
Jundiaí	SP	65	1,3	8	0,5	7	1,0	50	1,8
Salvador	BA	62	1,2	1	0,1	11	1,6	50	1,8
Ribeirão Preto	SP	59	1,1	3	0,2	3	0,4	53	1,9
Florianópolis	SC	59	1,1	5	0,3	4	0,6	50	1,8
Bragança Paulista	SP	59	1,1	18	1,1	18	2,6	23	0,8
Confins	MG	58	1,1	2	0,1	1	0,1	55	1,9
Campo Grande	AL	57	1,1	11	0,7	9	1,3	37	1,3
Uberlândia	MG	42	0,8	3	0,2	5	0,7	34	1,2
São José dos Campos	SP	41	0,8	4	0,3	4	0,6	33	1,2
Fortaleza	CE	40	0,8	3	0,2	3	0,4	34	1,2
Vitória	ES	36	0,7	-	0,0	1	0,1	35	1,2
Macaé	RJ	34	0,7	2	0,1	3	0,4	29	1,0
Campos dos Goytacazes	RJ	34	0,7	3	0,2	2	0,3	29	1,0
Cuiabá	MT	34	0,7	2	0,1	5	0,7	27	0,9
Maringá	PR	32	0,6	7	0,4	8	1,1	17	0,6
Itaituba	PA	31	0,6	28	1,8	1	0,1	2	0,1
Santarém	PA	28	0,5	6	0,4	2	0,3	20	0,7
São José dos Pinhais	PR	26	0,5	1	0,1	2	0,3	23	0,8
Navegantes	SC	23	0,4	3	0,2	-	0,0	20	0,7
Porto Velho	RO	23	0,4	2	0,1	4	0,6	17	0,6
Sorocaba	SP	23	0,4	6	0,4	2	0,3	15	0,5
Marabá	PA	22	0,4	-	0,0	2	0,3	20	0,7
Total parcial		2.573	50,1	247	15,5	238	34,1	2.088	73,4
Outras cidades		2.563	49,9	1.348	84,5	459	65,9	756	26,6
Total		5.136	100	1.595	100	697	100	2.844	100

Fonte: Compilado pelo autor

Tabela A.2: Tabela do total de ocorrências e querosene de aviação por UF de 2013 a 2022.

	Estado	Sigla	Ocorrências	%	Querosene de Aviação (m^3)*	%
1	Acre	AC	42	0,8	98.078	0,2
2	Alagoas	AL	26	0,5	465.705	0,7
3	Amapá	AP	10	0,2	42.368	0,1
4	Amazonas	AM	211	4,1	1.283.733	2,0
5	Bahia	BA	175	3,4	2.282.161	3,6
6	Ceará	CE	76	1,5	1.798.978	2,8
7	Distrito Federal	DF	102	2,0	3.986.110	6,3
8	Espírito Santo	ES	68	1,3	305.610	0,5
9	Goiás	GO	266	5,2	726.987	1,1
10	Maranhão	MA	63	1,2	410.818	0,6
11	Mato Grosso	MT	287	5,6	573.118	0,9
12	Mato Grosso do Sul	MS	158	3,1	271.372	0,4
13	Minas Gerais	MG	492	9,6	2.579.713	4,1
14	Pará	PA	266	5,2	1.180.726	1,9
15	Paraíba	PB	21	0,4	406.081	0,6
16	Paraná	PR	426	8,3	1.573.139	2,5
17	Pernambuco	PE	99	1,9	2.231.565	3,5
18	Piauí	PI	38	0,7	216.364	0,3
19	Rio de Janeiro	RJ	379	7,4	9.261.286	14,6
20	Rio Grande do Norte	RN	18	0,4	792.833	1,2
21	Rio Grande do Sul	RS	310	6,0	1.548.052	2,4
22	Rondônia	RO	49	1,0	234.769	0,4
23	Roraima	RR	52	1,0	82.536	0,1
24	Santa Catarina	SC	183	3,6	892.382	1,4
25	São Paulo	SP	1.265	24,6	29.975.382	47,2
26	Sergipe	SE	13	0,3	251.225	0,4
27	Tocantins	TO	41	0,8	73.976	0,1
Total			5.136	100	63.545.066	100

*Fonte: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis - ANP, conforme Resolução ANP n° 729/2018. Vendas, pelas distribuidoras, dos derivados combustíveis de petróleo por Unidade da Federação e produto - 2013-2022 (m³).

Apêndice B

Mapa das Ocorrências

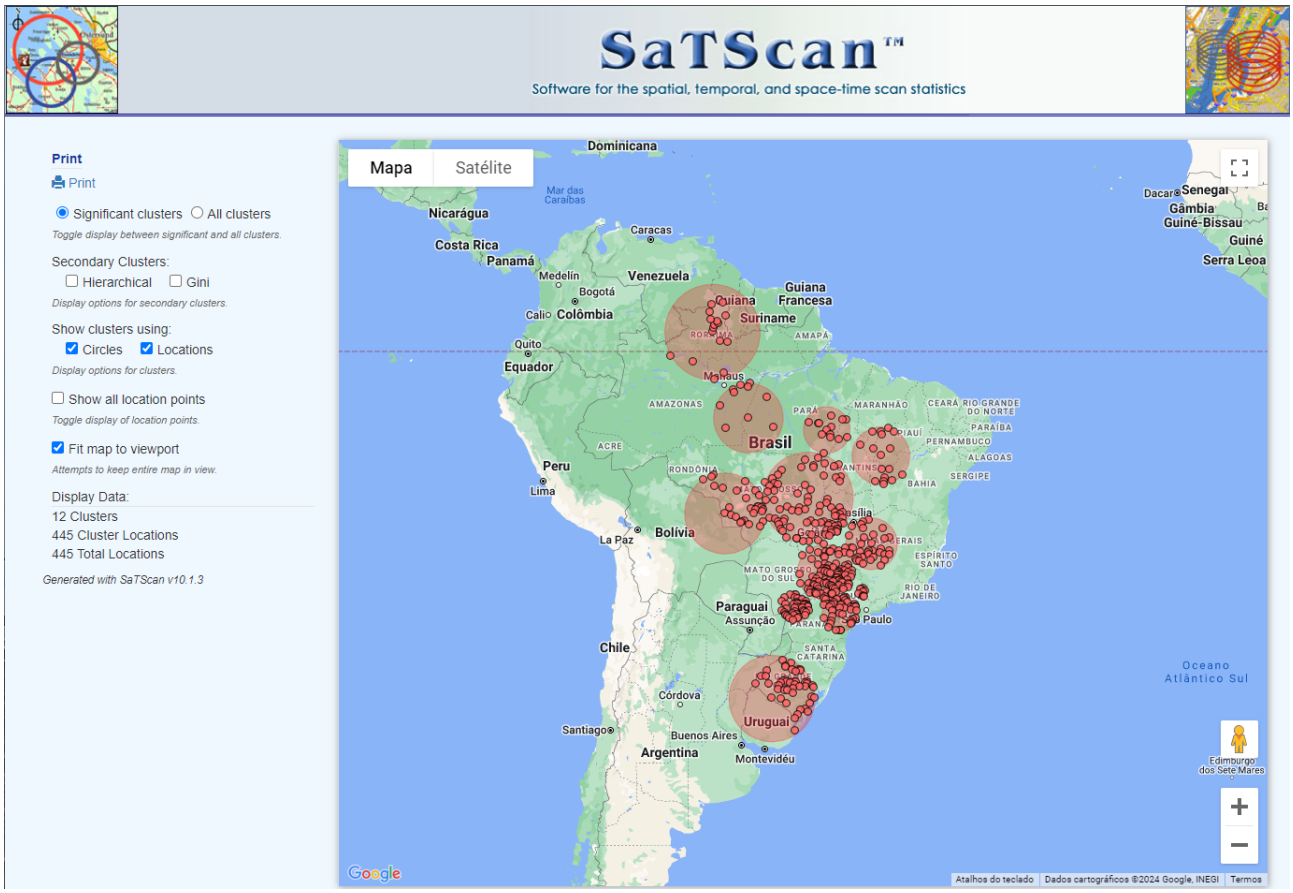

Figura B.1: Modelo binomial - Acidente.

Figura B.2: Modelo binomial - Incidente grave.

Figura B.3: Modelo binomial - Incidente.

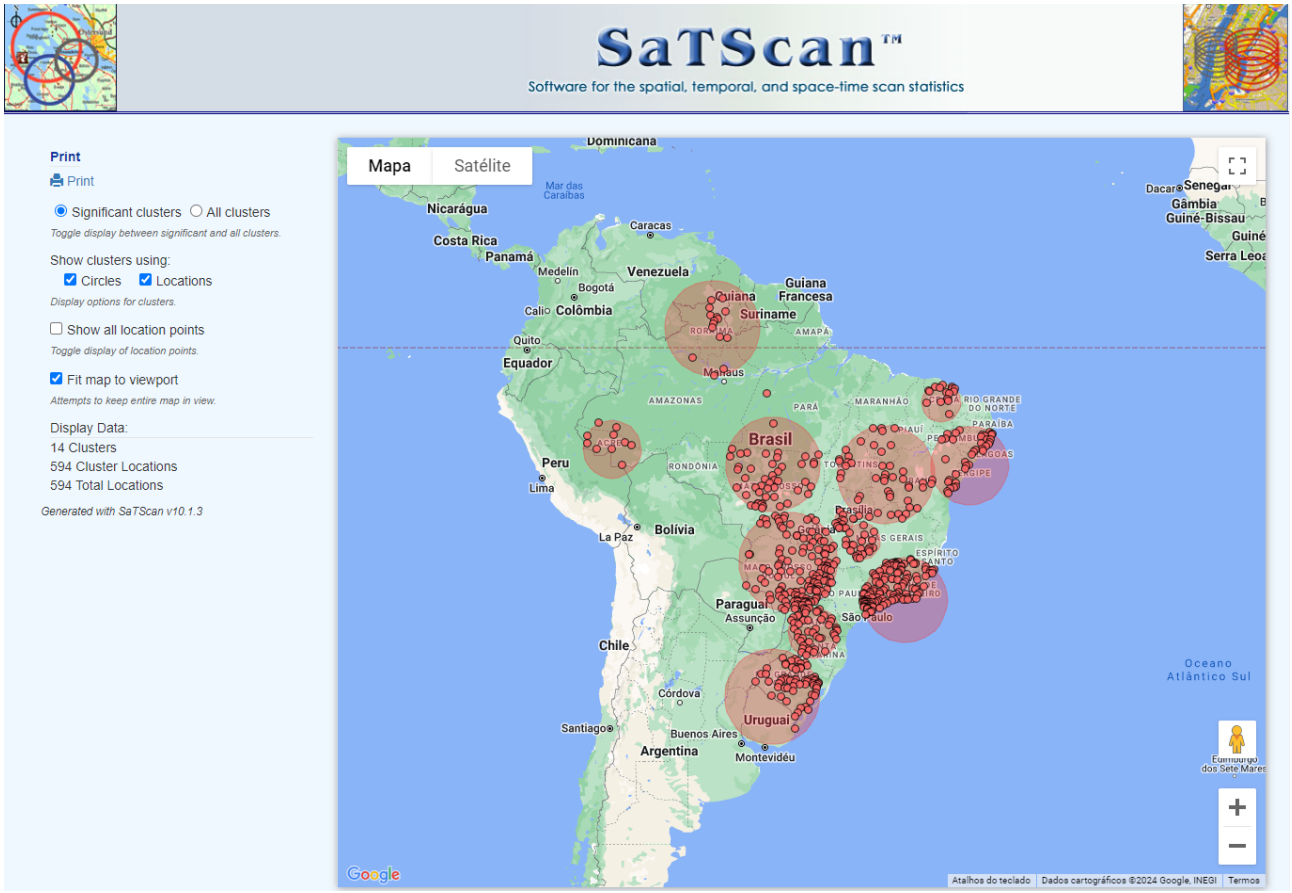

Figura B.4: Modelo multinomial.

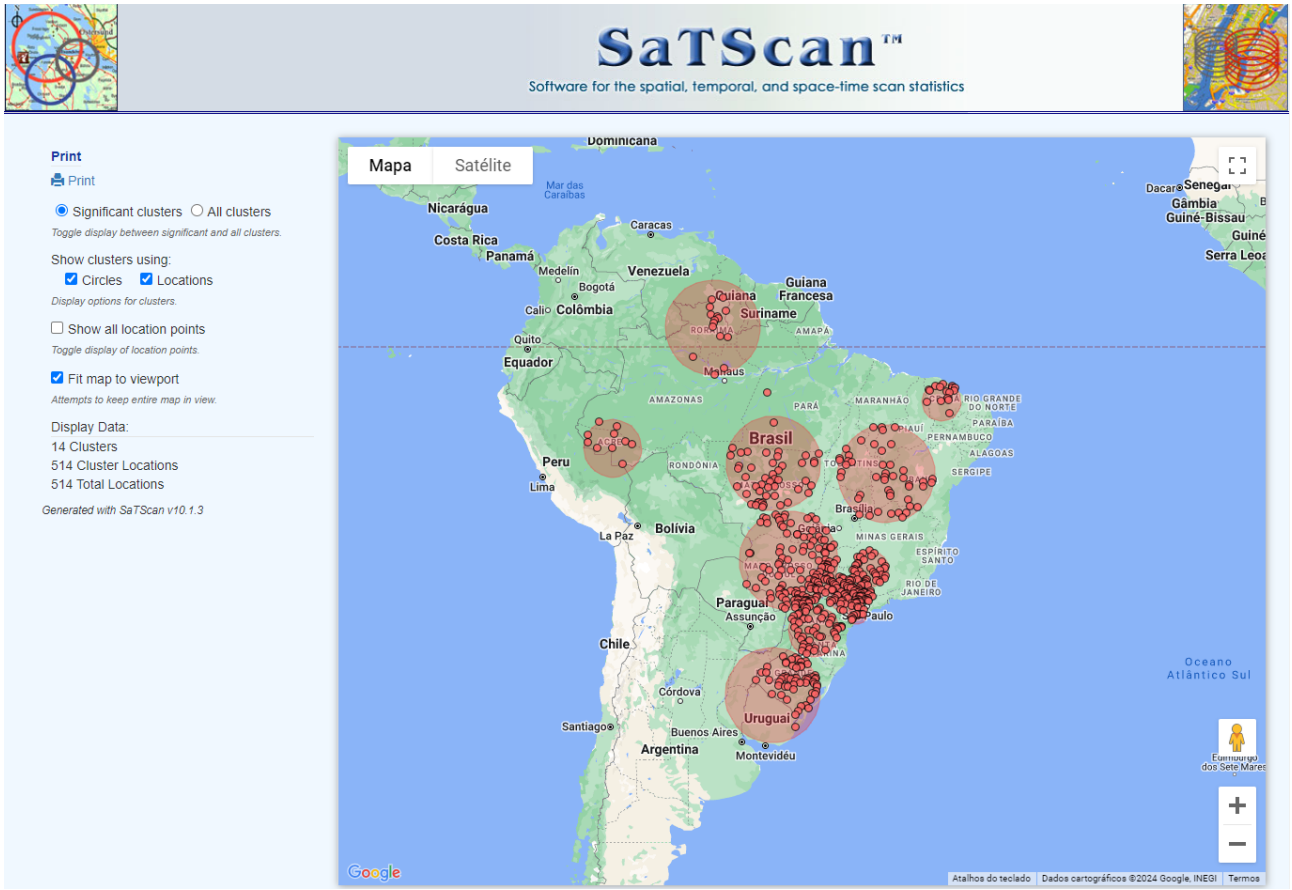

Figura B.5: Modelo ordinal.

Referências Bibliográficas

Ayer, Miriam et al. (1955). "An empirical distribution function for sampling with incomplete information". *The annals of mathematical statistics*, pp. 641–647.

Barlow, Richard E. et al. (1972). *Statistical inference under order restrictions : the theory and application of isotonic regression*. Books on Demand.

Besag, Julian e Newell, James (1991). "The detection of clusters in rare diseases". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154.1, pp. 143–155.

Bhatt, Vijaya e Tiwari, Neeraj (2014). "A spatial scan statistic for survival data based on Weibull distribution". *Statistics in medicine* 33.11, pp. 1867–1876.

Boyle, James P. e Dykstra, Richard L. (1986). "A method for finding projections onto the intersection of convex sets in Hilbert spaces". *Lecture Notes in Statistics*, pp. 28–47.

Choynowski, Mieczyslaw (1959). "Maps Based on Probabilities". *Journal of the American Statistical*, pp. 385–388.

Dwass, Meyer (1957). "Modified randomization tests for nonparametric hypotheses". *The Annals of Mathematical Statistics*, pp. 181–187.

Dykstra, Richard, Kochar, Subhash e Robertson, Tim (1995). "Inference for likelihood ratio ordering in the two-sample problem". *Journal of the American Statistical Association* 90.431, pp. 1034–1040.

Força Aérea Brasileira, FAB (2024). *História do Centro de Investigação e Prevenção de Acidentes Aeronáuticos*. Disponível em: `https://www2.fab.mil.br/cenipa/index.php`. Acesso em: 17 de abril de 2024.

Governo Federal (2024). *Portal de Dados Abertos*. Disponível em: `https://dados.gov.br/dados/conjuntos-dados/ocorrencias-aeronauticas-da-aviacao-civil-brasileira`. Acesso em: 19 de setembro de 2024.

Huang, Lan, Kulldorff, Martin e Gregorio, David (2007). "A spatial scan statistic for survival data". *Biometrics* 63.1, pp. 109–118.

Jung, Inkyung, Kulldorff, Martin e Klassen, Ann C (2007). "A spatial scan statistic for ordinal data". *Statistics in medicine* 26.7, pp. 1594–1607.

Jung, Inkyung, Kulldorff, Martin e Richard, Otukei John (2010). "A spatial scan statistic for multinomial data". *Statistics in medicine* 29.18, pp. 1910–1918.

Kulldorff, Martin (1997). "A spatial scan statistic". *Communications in Statistics-Theory and methods* 26.6, pp. 1481–1496.

Manual de Investigação do SIPAER (2017). 2ª ed. Força Aérea Brasileira. Brasília.

Martin Kulldorff (2024). *SaTScan™*. Disponível em: `https://www.satscan.org/`. Acesso em: 19 de setembro de 2024.

Salgado, Alfredo Moreira (2018). "Uso de regressão isotônica na escolha de itens em testes adaptativos computadorizados". Diss. de mestr. Brasília-DF: Universidade de Brasília.

Sant'Anna, Juliano Cézar (2020). "Estatística de varredura espacial Touchard baseada em expectância". Diss. de mestr. Brasília-DF: Universidade de Brasília.

Silva, Eduarda Bahiense Machado da (2021). "Distribuição Gumbel Bimodal: propriedades e estimação". Diss. de mestr. Brasília-DF: Universidade de Brasília.

Stan Openshaw Martin Charlton, Alan William Craft e Birch, JM (1988). "Investigation of Leukaemia Clusters by Use of a Geographical Analysis Machine". *The Lancet, Issue 8580* 331, pp. 272–273.