UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Fernanda Buzza Alves Barros

COORDENADAS SINTÉTICAS EM BANCOS DE DADOS CONFIDENCIAIS: uma aplicação em dados de COVID-19

Belo Horizonte
2023

Fernanda Buzza Alves Barros

COORDENADAS SINTÉTICAS EM BANCOS DE DADOS CONFIDENCIAIS: uma aplicação em dados de COVID-19

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientadora: Profa. Dra. Thaís Paiva Galletti
Coorientador: Prof. Dr. Marcos Oliveira Prates

Belo Horizonte
2023

Barros, Fernanda Buzza Alves.

B277c Coordenadas sintéticas em bancos de dados confidenciais [recurso eletrônico]: uma aplicação em dados de covid-19 / Fernanda Buzza Alves Barros. – 2023.
1 recurso online (85 f. il, color.): pdf.

Orientadora: Thaís Paiva Galletti.
Coorientador: Marcos Oliveira Prates
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: f. 79-82.

1. Estatística – Teses. 2. Análise espacial (Estatística) – Teses. 3. Saúde pública – Estatística – Dados não estruturados - Teses. I. Galletti, Thaís Paiva. II. Prates, Marcos Oliveira. III Universidade Federal de Minas Gerais; Instituto de Ciências Exatas Departamento de Ciência da Estatística. IV. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

FOLHA DE APROVAÇÃO

"Coordenadas Sintéticas em Bancos de Dados Confidenciais: Uma Aplicação em Dados de Covid-19"

FERNANDA BUZZA ALVES BARROS

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Mestre em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 15 de agosto de 2023, pela banca constituída pelos membros:

Profa. Thais Paiva Galletti - Orientadora
DEST/UFMG

Prof. Marcos Oliveira Prates - Coorientador
DEST/UFMG

Prof. Victor Hugo Lachos Dávila
University of Connecticut

Prof. Vinícius Diniz Mayrink
DEST/UFMG

Belo Horizonte, 15 de agosto de 2023.

Dedicado aos meus familiares, em especial ao meu esposo Alex.

Agradecimentos

Agradeço a **Deus** por me ensinar a viver, apesar das dificuldades, dia após dia.
Agradeço a minha orientadora **Thaís Paiva**, pela dedicação a este trabalho.
Agradeço ao meu co-orientador **Marcos Prates**, pelas revisões e auxílios. E, a todos os **professores e funcionários** do Departamento de Estatística da UFMG. Também gostaria de agradecer ao **suporte financeiro** das agências de fomento **CAPES, CNPq e FAPEMIG** que auxiliam na infraestrutura do Programa de Pós-graduação e facilitaram a realização dessa dissertação.
Agradeço ao meu esposo **Alex**, por todo carinho, incentivo e suporte.
Agradeço aos meus pais **Ana Magali e Ilton**, por todos os ensinamentos.
Agradeço a minha querida avó **Conceição**, pelas palavras de sabedoria.
Aos **demais familiares**, agradeço por todo o apoio.
Aos meus **amigos**, agradeço pelas conversas descontraídas e por todas as palavras de ânimo.
Por fim, **agradeço a todos** que, de alguma forma, **contribuíram para mais uma conquista**.

"*Feliz o homem que acha sabedoria, e o homem que adquire conhecimento; porque melhor é o lucro que ela dá do que o da prata, e melhor a sua renda do que o ouro mais fino. Mais preciosa é do que pérolas, e tudo o que podes desejar não é comparável a ela.*"
(Provérbios 3:13-15)

Resumo

Muitos dados coletados por agências possuem características confidenciais e informações sensíveis, portanto as instituições de pesquisa devem obedecer protocolos legais e éticos para não divulgar tais informações de maneira indiscriminada. Este trabalho utiliza a metodologia de dados sintéticos e imputação múltipla que são técnicas desenvolvidas para a divulgação segura de dados sensíveis, uma vez que apresentam uma maior preservação da utilidade dos dados. Esse método substitui os valores originais por valores simulados utilizando distribuições de probabilidades ajustadas aos valores originais, podendo ser aplicado para substituir parcialmente ou completamente os dados originais. O modelo de [26] e atualizado por [25], utiliza essa metodologia para gerar coordenadas geográficas sintéticas, entretanto não existia no modelo a previsão de espaços não habitáveis, como por exemplo aeroporto e lagoas. Portanto, contribuímos com a inclusão de tais espaços e denominamos eles como áreas restritas (espaços em que não existem habitações de indivíduos). Para avaliar essa contribuição no modelo, utilizamos um banco de dados simulado e representamos graficamente os resultados da aplicação com e sem a inclusão das áreas restritas. Por fim, realizamos a aplicação em um banco de dados de casos de COVID-19 da cidade de Montes Claros - MG, e pudemos comprovar a importância da inclusão de espaços inabitáveis nos dados para geração das coordenadas sintéticas.

Palavras-chave: Dados Sintéticos. Confidencialidade. Coordenadas Geográficas Sintéticas. Estatística Espacial.

Abstract

Many data collected by agencies have confidential characteristics and sensitive information, so research institutions must obey legal and ethical protocols not to disclose such information indiscriminately. This work uses the methodology of synthetic data and multiple imputation, which are techniques developed for the safe disclosure of sensitive data, since they present a greater preservation of the usefulness of the data. This method replaces the original values with simulated values using probability distributions fitted to the original values, and can be applied to replace partially or completely the original data. The model by [26] and updated by [25], uses this methodology to generate synthetic geographic coordinates, however the model did not include the prediction of non-inhabitable spaces, such as airports and lakes. Therefore, we contribute to the inclusion of such spaces and call them restricted areas (spaces where individuals do not live). To evaluate this contribution in the model, we used a simulated database and graphically represented the results of the application with and without the inclusion of restricted areas. Finally, we carried out the application in a database of COVID-19 cases in the city of Montes Claros - MG, and we were able to prove the importance of including uninhabitable spaces in the data for the generation of synthetic coordinates.

Keywords: Synthetic Data. Confidenciality. Synthetic Geographical Coordinates. Spatial Statistics.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Diversas informações importantes para o desenvolvimento de políticas públicas e outras aplicações são coletadas através de realizações de pesquisas feitas por agências públicas e privadas de coleta de dados. Muitos desses dados a serem coletados possuem características confidenciais e informações sensíveis, portanto as instituições de pesquisa devem obedecer protocolos legais e éticos para não divulgar tais informações de maneira indiscriminada. Se houver a divulgação de tais características, podem ocorrer identificações de identidades de entrevistados por invasores. Os invasores são pessoas ou entidades que procuram identificar indivíduos específicos a partir das informações divulgadas ([9], [18] e [20]).

Observamos um aumento na discussão e preocupação com a divulgação de dados e armazenamento das informações de indivíduos. Para tratar desse assunto, existe a legislação *"Health Insurance Portability and Accountability Act"* (HIPAA), publicada pelo Departamento de Saúde e Serviços Humanos dos EUA ([17]), a União Europeia possui a legislação *"General Data Protection Regulation"* (GDPR) que entrou em vigor em 2018. Na legislação brasileira temos a Lei Geral de Proteção de Dados (LGPD - Lei nº 13.709/2018) que entrou em vigor em 2020 com o objetivo de proteger os direitos fundamentais de liberdade e privacidade dos indivíduos.

O ideal seria utilizar essas informações sem prejudicar a confidencialidade dos indivíduos envolvidos na pesquisa e garantir que análises estatísticas feitas com os dados tenham resultados válidos e confiáveis. Portanto, algumas técnicas são desenvolvidas pelas agências para a divulgação dos dados com o auxílio de métodos estatísticos.

Tais técnicas são chamadas de *Statistical Disclosure Control* (SDC) or *Statistical Disclosure Limitation* (SDL), sendo traduzidas para Controle Estatístico de Confidencialidade (CEC) ([37], [21], [18] e [32]). Entre as técnicas de CEC mais comumente empregadas, estão a desidentificação dos indivíduos, agregação de dados, supressão ou censura de valores, além de troca de dados e adição de ruído aleatório ([20]).

- Agregação: agrega valores em categorias mais abrangentes, como por exemplo bairros, cidades e estados. Por exemplo, um indivíduo que pode ter um risco de ser identificado pois as características dele a nível de cidade são únicas, entretanto pode estar protegido a nível de estado pois podem existir mais indivíduos com as

mesmas características. A desvantagem é que a agregação impossibilita análises e inferências em um nível mais refinado do que o agregado.

- Supressão: exclui valores ou variáveis que correm risco de serem descobertos. A principal desvantagem é que afeta as inferências, pois os dados são excluídos de maneira não aleatória.

- Troca de dados: troca de valores entre pares de informações similares. Por exemplo, trocar a cidade de duas casas que possuem a mesma quantidade de pessoas. A desvantagem é que afeta a relação entre todas as variáveis, tanto as que foram trocadas quanto as outras.

- Ruído aleatório: adiciona valores amostrados aleatoriamente aos valores observados. Por exemplo, a adição de uma variável aleatória com alguma distribuição estatística. A desvantagem é que, dependendo do tamanho da variância do ruído aleatório adicionado, ocorre distorção das distribuições marginais.

Além dos métodos de CEC citados anteriormente, temos também o método de dados sintéticos que apresentam uma maior preservação da utilidade dos dados ([33], [34]). O método de geração dos dados sintéticos é baseado na imputação múltipla que foi desenvolvida originalmente por [33] para imputar dados ausentes. Esta técnica substitui os valores originais por valores simulados utilizando distribuições de probabilidades ajustadas aos valores originais, podendo ser aplicada para substituir parcialmente ou completamente os dados originais.

Os dados sintéticos possibilitam uma maior preservação dos resultados de inferência quando comparado aos outros métodos. A desvantagem é que não são livres de risco, algumas inferências são dificultadas e existe a necessidade de divulgar vários bancos de dados sintéticos para possibilitar a estimação da variância corretamente ([20]).

Bancos de dados com informações geográficas, como por exemplo, local de residência de indivíduos, apresentam um desafio ainda mais complexo para a divulgação dos dados ao público. Em muitos casos, é impossível divulgar a informação de localização do entrevistado devido a questões de confidencialidade. Uma solução criada por [26] para problemas de divulgação de informações de localização dos indivíduos foi a simulação de coordenadas geográficas sintéticas baseada em modelos de mapeamento de doenças, e posteriormente estendida por [25] para incluir covariáveis contínuas.

O objetivo deste trabalho é utilizar a metodologia de simulação apresentada por [25], e estender para incluir restrições espaciais para regiões onde não há necessidade de gerar dados sintéticos. Além disso, a metodologia desenvolvida será aplicada em dados de COVID-19, para gerar coordenadas sintéticas, no município de Montes Claros/MG.

No Capitulo 2, trataremos das técnicas de CEC e a importância do método de dados sintéticos utilizado nesse estudo. No Capítulo 3, apresentaremos o modelo para geração

de localizações sintéticas, assim como sua atualização que é o objetivo principal desse trabalho. No Capítulo 4, mostraremos o impacto da mudança realizada na metodologia com o auxílio de dados simulados e a análise de Risco e Utilidade das mudanças no modelo. No Capítulo 5, traremos uma breve apresentação do pacote `syncoordinatesr` em desenvolvimento no `Software R`. No Capítulo 6, aplicaremos o novo modelo para gerar coordenadas sintéticas para os dados de COVID-19 do município de Montes Claros. E por fim, no Capítulo 7, traremos a conclusão deste trabalho.

Capítulo 2

Revisão Bibliográfica

2.1 Controle Estatístico de Confidencialidade

Os métodos de Controle Estatístico de Confidencialidade (CEC) são utilizados para garantir uma segura divulgação de dados coletados pelas agências através da modificação deles. [37], [21], [18] e [20] demonstram alguns métodos de CEC que as agências utilizam, sendo que é possível recorrer a mais de um método para proteger os dados. Os principais métodos são agregação, supressão e troca de dados. Entretanto, todos os métodos possuem algum nível de perda de informação.

É importante destacar que as agências também utilizam esses métodos para proteger os dados de possíveis invasores, indivíduos que utilizam as informações divulgadas para identificar características de um entrevistado ou de um grupo específico de entrevistados, e que possuem algum ganho ao identificar os indivíduos ([37]).

Segundo os autores, de modo geral, as técnicas de CEC aplicam o seguinte:

$$dados\ divulgados = f(dados\ originais), \tag{2.1}$$

onde, f = função que modifica os dados.

As técnicas de CEC abordadas por [37] são:

- Recodificação Local: combinação de categorias de variáveis em um registro por uma base de registro.
- Recodificação Global: combinação de uma ou mais categorias de uma variável em uma só. Ao realizar a redução do número de categorias de uma variável pode ocorrer perda de informação sobre ela.
- Supressão Local: exclusão de um valor em um registro e a substituição, desse valor, por um indicador de valor ausente.
- Supressão Local com Imputação: exclusão de um valor em um registro e a substituição, desse valor, por um valor imputado.

- Subamostragem: divulgação de apenas uma parte do banco de dados.
- Adicionando Ruído: adiciona valores amostrados aleatoriamente.
- Arredondamento: realiza o arredondamento de valores em variáveis quantitativas.
- Microagregação: aplicável em variáveis quantitativas, é uma técnica que envolve arredondamento e adição de ruído.
- Método Pós-randomização (PRAM): altera as pontuações de algumas variáveis em alguns registros para novas pontuações, de acordo com algum método de probabilidade descrito.
- Troca de Dados: troca de registros no banco de dados através da escolha aleatória de dois registros para servirem de troca.
- Microdados Sintéticos e Imputação Múltipla: divulgação de um banco de dados sintético gerado através de um modelo ajustado dos dados originais.

Note que o único método de CEC descrito acima que realiza a troca de todas as informações originais do banco de dados são os microdados sintéticos com imputação múltipla. E com isso, os dados sintéticos possibilitam uma maior preservação dos resultados de inferência ([20]).

A metodologia de imputação múltipla foi criada originalmente para tratar valores ausentes em bancos de dados. Esse método possui cinco principais vantagens, segundo [33] e [34]:

1. a capacidade de usar métodos de análise de dados completos;
2. a capacidade de incorporar o conhecimento do coletor dos dados originais;
3. a capacidade de aumentar a eficiência da estimação quando as imputações são geradas aleatoriamente tentando representar a distribuição dos dados;
4. a capacidade de obter inferências válidas combinando as inferências de dados completos de maneira direta;
5. a capacidade de analisar a sensibilidade de inferências a vários modelos através da repetição dos métodos de dados completos.

A utilização de dados sintéticos e imputação múltipla contribui para a divulgação segura de dados sensíveis, uma vez que preservam os dados originais. Além disso, podemos observar sua aplicação em dados geográficos. Segundo [36], dados geográficos são alvos da utilização dos métodos de CEC de agregação e supressão, entretanto seu uso pode ocasionar perda de refinamento geográfico dos dados e a qualidade da análise pode ficar

comprometida. Para contornar esses problemas, os autores propuseram a utilização da metodologia da imputação múltipla ([33], [34]) em dados geográficos. As vantagens dessa proposta são: maior preservação da relação entre os atributos geográficos e outros através da modelagem estatística, e a substituição de identificadores geográficos por imputações que dificultam que os invasores conheçam os reais valores dos dados originais ([36]).

A imputação múltipla em dados geográficos substitui as localizações reais dos indivíduos por localizações simuladas através de modelos estatísticos ajustados aos dados originais ([22], [31]). Além disso, os pesquisadores sugerem divulgar $m > 1$ versões dos bancos de dados gerados para considerar a incerteza incluída pela simulação.

Segundo [32], após a aplicação de técnicas de CEC nos dados a serem divulgados, as agências aplicam estratégias para analisar os riscos de exposição das informações. Portanto, as agências tentam determinar se os riscos são aceitáveis e se a utilidade dos dados é alta, sendo assim essas análises são chamadas na literatura de avaliação de Risco e Utilidade. Os processos para determinação dos riscos de exposição das informações podem ser realizados várias vezes e é encerrado quando as agências concluem que os riscos são aceitáveis e a utilidade dos dados é adequada ([5]).

2.2 Estatística Bayesiana

A estatística Bayesiana parte do Teorema de Bayes, a seguir, e combina a informação do especialista (distribuições *a priori*) com os modelos complexos de dados, resultando na distribuição *a posteriori* ([27]).

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}, \tag{2.2}$$

onde:

- $P(A_i)$, $i = 1, 2, \ldots, m$, são as probabilidades *a priori*, e;
- $P(A_i|B)$, $i = 1, \ldots, m$, são as probabilidades *a posteriori*.

Segundo [27], Laplace propôs que quando não existem informações *a priori*, assumimos que $P(A_i) = \frac{1}{m}$, onde $i = 1, \ldots, m$, e esses casos são chamados de distribuições *a priori* não informativas:

$$P(A_i|B) = \frac{P(B|A_i)}{\sum_i P(B|A_i)} \tag{2.3}$$

onde:

- $P(A_i|B)$, $i = 1, \ldots, m$, são as probabilidades *a posteriori*.

[2] sugere uma outra forma de representar o Teorema de Bayes através de modelos hierárquicos. Seja a distribuição do modelo $f(\boldsymbol{y}|\boldsymbol{\theta})$ para os dados observados $\boldsymbol{y} = (y_1, \ldots, y_n)$, onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ é o vetor de parâmetros desconhecidos. $\boldsymbol{\theta}$ é uma quantidade aleatória amostrada da distribuição a *priori* $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, onde $\boldsymbol{\lambda}$ é o vetor de hiperparâmetros. Se $\boldsymbol{\lambda}$ é conhecido, inferência sobre $\boldsymbol{\theta}$ é baseada na distribuição a *posteriori*:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\lambda}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\lambda})}{p(\boldsymbol{y}|\boldsymbol{\lambda})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\lambda})}{\int p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\lambda}) d\boldsymbol{\theta}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta}}, \tag{2.4}$$

onde f é a função de verossimilhança (contribuição dos dados) e π é a distribuição a *priori* (contribuição dos conhecimentos prévios). Na prática, $\boldsymbol{\lambda}$ não é conhecido, então a distribuição de $h(\boldsymbol{\lambda})$ será necessária, assim teremos:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda}) d\boldsymbol{\lambda}}{\int\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda}) d\boldsymbol{\theta} d\boldsymbol{\lambda}}. \tag{2.5}$$

Podemos substituir $\boldsymbol{\lambda}$ por um estimador $\hat{\boldsymbol{\lambda}}$ obtido através da maximização da distribuição marginal $p(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta}$ que é uma função de $\boldsymbol{\lambda}$.

Algumas das vantagens na utilização de estatística Bayesiana são a sua abordagem unificada e sua capacidade de incorporar a informação prévia sobre os dados através da distribuição *a priori*.

Outras vantagens se referem às inferências, uma vez que se calcula a distribuição *a posteriori*, as inferências são simplesmente para extrair suas características, pois, como pode ser visto pelo Teorema de Bayes, a distribuição *a posteriori* resume todas as informações sobre os parâmetros do modelo em relação aos dados ([2]).

Para a inferência Bayesiana tem-se as estimativas pontuais e a estimativa intervalar. Na estimativa pontual temos:

- média a *posteriori*: $\hat{\theta} = E(\theta|y)$,
- mediana a *posteriori*: $\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|\boldsymbol{y}) d\theta = 0{,}5$,
- moda a *posteriori*: $\hat{\theta} : p(\hat{\theta}|\boldsymbol{y}) = \sup_{\theta} p(\theta|\boldsymbol{y})$.

Na estimativa intervalar, podemos computar qualquer quantil, sendo assim, por exemplo, se quisermos computar $\alpha/2$ e $(1 - \alpha/2)$ quantis de $p(\theta|\boldsymbol{y})$, teríamos:

- $\int_{\infty}^{q_L} p(\theta|\boldsymbol{y}) d\theta = \alpha/2$, e
- $\int_{q_U}^{\infty} p(\theta|\boldsymbol{y}) d\theta = 1 - \alpha/2$.

Desta forma, a confiança de que θ está entre (q_L, q_U) é $100 \times (1 - \alpha)\%$, sendo assim o intervalo de credibilidade é $100 \times (1 - \alpha)\%$ para θ.

Em muitos casos, as distribuições a *posteriori* dos dados são complexas, e para auxiliar em sua amostragem, muitos pesquisadores contribuíram, e contribuem atualmente, para a implementação de métodos computacionais para a estatística Bayesiana.

Os primeiros livros com foco na computação Bayesiana são os que utilizam métodos de *Markov Chain Monte Carlo (MCMC)* para realizar a amostragem das distribuições a *posteriori* complexas ([4], [10] e [14]). Entretanto, ainda temos os métodos que derivam do MCMC como o *Gibbs Sampler* ([12]), *Adaptive Rejection Sampling - ARS* ([13]), *Metropolis-Hastings* ([23], [16]) e *Slice Sampling* ([24]).

O algoritmo de *Gibbs Sampling* foi introduzido por [12] para a utilização, através da abordagem Bayesiana, na restauração de imagens degradadas. Entretanto, [11] propôs sua aplicação no cálculo da amostragem da distribuição a *posteriori*.

Nessa aplicação, quando a amostragem direta não é possível, o algoritmo a realiza para $p(\theta|y)$ com $\theta = (\theta_1, \theta_2, \ldots, \theta_d)'$, gerando sucessivamente a partir das distribuições condicionais, $p(\theta_i|\theta_{-i})$ com $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)'$.

2.3 Estatística Espacial

A estatística espacial é comumente utilizada em diversas áreas como epidemiologia, ecologia, saúde, climatologia. A utilização dela é devido aos tipos de análises a serem realizadas nas bases de dados que são coletadas por essas áreas. Esses dados geralmente são altamente multivariados (possuem muitas covariáveis e variáveis respostas importantes), possuem referência geográfica (muitas vezes são apresentados em mapas) e são temporalmente correlacionados (em estruturas longitudinais ou outras estruturas de séries temporais).

Os bancos de dados espaciais são classificados em três tipos básicos segundo [2]:

1. **Dados referenciados por pontos (*point-reference data*):** seja $\boldsymbol{Y}(\boldsymbol{s})$ um vetor aleatório no local $\boldsymbol{s} \in \mathbb{R}^r$, onde $\boldsymbol{s}$ varia continuamente sobre D, um subconjunto fixo de $\mathbb{R}^r$ que contém um retângulo r-dimensional de volume positivo;

2. **Dados de área (*areal data*):** seja D um subconjunto fixo, podendo ter forma irregular ou regular, sendo que é particionado em um número finito de unidades de área com limites definidos;

3. **Dados pontuais com padrão (*point pattern data*):** seja D aleatório, em que seu conjunto de índices fornece os locais de eventos aleatórios que são o padrão de pontos espaciais. Simplesmente podemos ter $\boldsymbol{Y}(\boldsymbol{s}) = 1$ para todo $\boldsymbol{s} \in D$ indicando

a ocorrência do evento, ou possivelmente fornecer alguma informação de covariáveis produzindo o processo de padrão de pontos marcados.

Além disso, segundo [2], um dos modelos amplamente utilizado para dados de área é o de mapeamento de doenças. Nesses modelos, temos a população em risco e o número de casos da doença. Seja n_i o número de pessoas em risco para a doença na área i, $i = 1, \ldots, G$, Y_i o número observado de casos da doença na área i, onde Y_i são variáveis aleatórias, e E_i o número esperado de casos da doença na área i, onde E_i são funções fixas e conhecidas de n_i. Podemos calcular a taxa da doença através da:

- padronização interna utilizando a taxa observada:

$$\overline{r} = \frac{\sum_i Y_i}{\sum_i n_i}, \quad \text{com} \quad E_i = P_i\,\overline{r}; \text{ ou} \tag{2.6}$$

- padronização externa utilizando uma tabela de taxas da doença.

Para calcular o risco relativo verdadeiro em cada área através das contagens de y_i, assumimos que $Y_i|\lambda_i \sim Poisson(E_i\lambda_i)$ e estimamos o risco λ_i. Assim temos,

$$Y_i|\lambda_i \sim Poisson(E_i\lambda_i), \tag{2.7}$$

$$log\lambda_i = X_i\boldsymbol{\beta} + \theta_i + \epsilon_i, \tag{2.8}$$

onde:

- X_i é o vetor p-dimensional com os valores das covariáveis para a área i;
- β contém os p coeficientes para cada covariável explicativa;
- θ_i é o efeito espacial específico da área;
- ϵ_i é o erro que captura as variações residuais do modelo Poisson.

Normalmente os pesquisadores utilizam distribuições a *priori* normais para os coeficientes em β e para ϵ, assim:

$$\beta_k \sim N(0, \sigma_\beta^2)\ para\ k = 1, \ldots p; \tag{2.9}$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)\ para\ i = 1, \ldots, G. \tag{2.10}$$

Por fim, segundo [2], para incluir associação espacial entre vizinhos de área para os efeitos espaciais $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$, os pesquisadores utilizam a distribuição a *priori* Auto-Regressiva Condicional Intrínseca (*Intrinsic Conditional Autoregressive - ICAR*),

sendo que as estruturas de vizinhança podem ser definidas tanto por distâncias entre os centroides das áreas ou por fronteiras. Assim temos a seguinte distribuição condicional, considerando que $j \sim i$ significa que a área j é vizinha da área i,

$$\theta_i|\boldsymbol{\theta}_{-i} \sim N(\hat{\theta}_i, \sigma^2_\theta/n_i), \tag{2.11}$$

onde:

- $\boldsymbol{\theta}_{-i}$ é o vetor com θ_j para todo $j \neq i$;
- $\hat{\theta}_i$ é a média de θ_j para todo $j \sim i$,
- n_i é o número de vizinhos da área i.

Note que $j \sim i$ significa que a área j é vizinha da área i. Por fim, as variâncias dos hiperparâmetros possuem distribuição a *priori* gama:

$$\frac{1}{\sigma^2_\epsilon} \sim Gama(a_\epsilon, b_\epsilon), \tag{2.12}$$

$$\frac{1}{\sigma^2_\beta} \sim Gama(a_\beta, b_\beta). \tag{2.13}$$

Capítulo 3

Metodologia

Para divulgar os dados geográficos sem incorrer em questões de quebra de confidencialidade, usaremos a metodologia proposta por [26]. O método utiliza a técnica de CEC de dados sintéticos para a aplicação em dados geográficos. Na Seção 2.1 a seguir, descrevemos o modelo proposto por [26] com covariáveis discretas, e na Seção 2.2 apresentamos a extensão proposta por [25] para incluir covariável contínua.

3.1 Modelo com covariáveis discretas

Suponha um banco de dados para n indivíduos denotado por $\boldsymbol{D} = (\boldsymbol{S}, \boldsymbol{X})$, em que S inclui as localizações de cada indivíduo $\boldsymbol{S} = (s_1, \ldots, s_n)^T$ e $\boldsymbol{X}$ é a matriz $n \times p$ dos atributos não espaciais dos indivíduos. Os p atributos são variáveis discretas $(X_1, \ldots, X_p)$. Podem ocorrer vários níveis em X_k para $k = 1, \ldots, p$; assim temos que $X_k \in (1, \ldots, d_k)$, em que d_k é o número de níveis em X_k. Seja b o índice das combinações únicas de atributos em X, então $b = 1, \ldots, B$, onde $B \leq \prod_{k=1}^{p} d_k$. Assim, para cada (b, k), $x_k^{(b)}$ é o valor de X_k na combinação b.

[26] apresenta um modelo para analisar onde as pessoas com certos atributos estão localizadas, e para isso é necessário dividir a área de interesse em uma grade regular G. Cada célula da grade é indexada por $i = 1, \ldots, G$. Então, para cada (i, b), temos que $c_i^{(b)}$ será o número de observações na célula i com a combinação de atributos b. É importante destacar que o tamanho da amostra de cada banco de dados possui um peso sobre o modelo, portanto se faz necessário incluir o parâmetro n como um *offset* na definição do modelo. Esse parâmetro pondera as intensidades de ocorrência dos dados de acordo com o tamanho de cada amostra.

Assim, temos o seguinte modelo proposto:

$$c_i^{(b)} \sim Poisson(n\lambda_i^{(b)}) \tag{3.1}$$

$$\log \lambda_i^{(b)} = \mu + \sum_{k=1}^{p} \alpha_k' \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \theta_i + \sum_{k=1}^{p} \phi_{ik}' \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \epsilon_i^{(b)}, \tag{3.2}$$

onde:

- μ: intercepto geral;
- $\alpha_k^{'} = (\alpha_{k1}, ..., \alpha_{kd_k})$: vetor de dimensão $d_k \times 1$ dos efeitos principais para o atributo k;
- $\mathbb{1}_{\left\{x_k^{(b)}\right\}}$: vetor indicador que tem dimensão $d_k \times 1$, que assume um na posição $x_k^{(b)}$ e zero caso contrário;
- θ_i: efeito espacial específico para cada célula da grade;
- $\phi_{ik}^{'} = (\phi_{ik1}, ..., \phi_{ikd_k})$ vetor de dimensão $d_k \times 1$ dos efeitos espaciais específicos de cada célula e para o atributo k;
- $\epsilon_i^{(b)}$: específico de cada célula e para cada combinação, que adiciona flexibilidade ao modelo e capta as variações residuais das taxas da distribuição de Poisson que não são explicadas pelas covariáveis.

O modelo possui componentes não espaciais e espaciais, sendo que por questões de identificabilidade cada α_{k1} e ϕ_{ik1} são iguais a zero. Nesse caso, a identificabilidade é por causa da presença do intercepto μ, não podemos ter, além dele, um parâmetro separado para cada um dos valores das covariáveis. Note que o modelo assume que as intensidades espaciais são homogêneas em cada célula da grade.

Além disso, assume-se que as células vizinhas são aquelas que possuem lados e vértices em comum. Dentro de uma abordagem Bayesiana, a correlação espacial entre os vizinhos das células da grade foi introduzida utilizando o modelo intrínseco autoregressivo condicional - (ICAR) [3]. Assim, a distribuição *a priori* para $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$, para todo i, é dada por

$$\theta_i | \boldsymbol{\theta}_{-i} \sim N\left(\bar{\theta}_i, \frac{\sigma_\theta^2}{n_i}\right), \tag{3.3}$$

sendo:

- $\boldsymbol{\theta_{-i}}$: inclui os valores de θ_j para todo $j \neq i$;
- $\overline{\theta}_i$: média dos n_i valores de θ_j, para as células j que são vizinhas da célula i;
- σ_θ^2: variância comum para todos os valores de θ.

Para $ikj : i = 1, \ldots, G;\ k = 1, \ldots, p;\ j = 2, \ldots, d_k$ a distribuição *a priori* assumida pelos autores é

$$\phi_{ikj} | \boldsymbol{\phi}_{-i,kj} \sim N\left(\overline{\phi}_{ikj}, \frac{\sigma_{\phi kj}^2}{n_i}\right), \tag{3.4}$$

em que os parâmetros $\boldsymbol{\phi}_{-i,kj}$, $\overline{\phi}_{ikj}$ e $\sigma^2_{\phi kj}$ possuem interpretações análogas aos parâmetros da Fórmula (3.3). Novamente, para garantir identificabilidade, os elementos de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}_{kj}$ foram restringidos para $\sum_{i=1}^{G} \theta_i = 0$ e $\sum_{i=1}^{G} \phi_{ijk} = 0$ para todo (kj); ver [2].

Para as distribuições *a priori* dos hiperparâmetros, [26] propuseram:

$$\epsilon_i^{(b)} \sim N(0, \sigma^2_\epsilon), \tag{3.5}$$

$$\mu \sim N(0, v_\mu), \tag{3.6}$$

$$\alpha_{kj} \sim N(0, v_{\alpha k}) \text{ para todo } (kj), \tag{3.7}$$

$$\frac{1}{\sigma^2_\theta} \sim Gama(a_\theta, b_\theta), \tag{3.8}$$

$$\frac{1}{\sigma^2_{\phi kj}} \sim Gama(a_{\phi k}, b_{\phi k}) \text{ para todo } (kj), \tag{3.9}$$

$$\frac{1}{\sigma^2_\epsilon} \sim Gama(a_\epsilon, b_\epsilon). \tag{3.10}$$

É recomendado pelos autores que as variâncias v sejam altas para que as distribuições *a priori* sejam vagas. A determinação das condicionais completas para cada um dos parâmetros do modelo podem ser verificadas em [25]. A amostragem utiliza *Markov Chain Monte Carlo* - (MCMC) com o algoritmo *Adaptive Rejection Sampling* desenvolvido por [13].

Após amostrar das distribuições *a posteriori* de $\boldsymbol{\lambda} = \{\lambda_i^{(b)}\}$, geram-se as localizações sintéticas para os n indivíduos. Primeiramente, seleciona-se um único valor de $\boldsymbol{\lambda}$, por exemplo $\lambda^{(l)}$, da sua distribuição *a posteriori*. Para todo (i, b), calcula-se

$$p_i^{(lb)} = \frac{\lambda_i^{(lb)}}{\sum_{i=1}^{G} \lambda_i^{(lb)}}. \tag{3.11}$$

Depois, amostra-se aleatoriamente e independentemente uma célula da grade para cada indivíduo com a combinação de atributos b, com probabilidade $(p_1^{(lb)},...,p_G^{(lb)})$. E assim, temos que:

- as células da grade amostradas podem servir como um conjunto de localizações sintéticas,
- ou, pode-se amostrar coordenadas mais precisas dentro da célula da grade, por exemplo, utilizar uma amostragem uniforme de localizações geográficas viáveis dentro da célula.

O resultado é um conjunto de localizações sintéticas, $\tilde{S}^{(l)} = (\tilde{s}_1^{(l)}, \ldots, \tilde{s}_n^{(l)})$, que quando combinadas às covariáveis X, obtém-se um banco de dados parcialmente sintético, $\tilde{D}^{(l)} = (\tilde{S}^{(l)}, X)$.

Para a redução do vício nas estimativas do modelo, geram-se m conjuntos de localizações sintéticas independentes, $\tilde{S} = (\tilde{S}^{(1)}, \ldots, \tilde{S}^{(m)})$, com seus respectivos bancos de dados $\tilde{D} = (\tilde{D}^{(1)}, \ldots, \tilde{D}^{(m)})$ que serão divulgados para o público.

Uma observação importante é que dois registros próximos nos dados originais não necessariamente ficarão próximos nos dados sintéticos, pois as localizações sintéticas (células de grade e coordenadas) são geradas independentemente pela estimação do modelo de Poisson.

Um dos pontos principais dessa metodologia é a escolha adequada do tamanho das células da grade. Se a grade possui muitas células, ou seja, as células são muito finas, as inferências são mais próximas das originais porém o risco de descobrimento é maior e podem gerar localizações sintéticas muito próximas das originais. Se a grade possui poucas células, ou seja o tamanho das células é grande, temos uma redução na qualidade dos dados, mas ocorrem melhorias na proteção das localizações originais e assim minimizam o risco de descobrimento.

3.2 Extensão do modelo com a inclusão de variável contínua

Seja $\boldsymbol{D} = (\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Z})$, em que $\boldsymbol{S}$ contém as coordenadas geográficas, $\boldsymbol{X}$ é o vetor de variáveis discretas $(X_1, \ldots, X_p)$, e $\boldsymbol{Z}$ é o vetor de variáveis contínuas $(Z_1, \ldots, Z_q)$. Considere também $b = 1, \ldots, B$ como a combinação de atributos de $\boldsymbol{X}$, e a grade regular G com células indexadas por $i = 1, \ldots, G$. Para cada variável contínua Z_j, [25] propôs a utilização da média dessa variável $\overline{Z}_{ij}^{(b)}$, pois ela representa a média de Z_j em cada célula da grade i e para cada combinação de atributo b. Novamente, temos a inclusão do tamanho da amostra n como *offset* no modelo.

Assim, com a inclusão de Z_j, temos:

$$c_i^{(b)} \sim Poisson(n\lambda_i^{(b)}), \tag{3.12}$$

$$\log \lambda_i^{(b)} = \mu + \sum_{k=1}^{p} \alpha_k' \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \theta_i + \sum_{k=1}^{p} \phi_{ik}' \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \sum_{j=1}^{q} \beta_{ij} \overline{Z}_{ij}^{(b)} + \epsilon_i^{(b)}, \tag{3.13}$$

em que:

- μ, α_k', θ_i, $\mathbb{1}_{\left\{x_k^{(b)}\right\}}$, ϕ_{ik}' e $\epsilon_i^{(b)}$ possuem a mesma interpretação e distribuições *a priori* mencionadas na Seção 3.1;

- β_{ij}: efeito da variável contínua j para cada célula da grade i;

- $\overline{Z}_{ij}^{(b)}$: média da variável Z_j em cada célula i e para cada combinação de atributo b.

Quando o parâmetro β_{ij} possuir uma interpretação similar a do parâmetro $\alpha_k^{'}$, sua distribuição *a priori* pode ser assumida como:

$$\beta_{ij} \sim N(0, v_\beta). \tag{3.14}$$

Recomenda-se que a variância v_β seja grande para que a distribuição *a priori* seja vaga. Caso exista estrutura de correlações espaciais entre os β_{ij}'s, o modelo ICAR pode ser usado como distribuição *a priori* de $\boldsymbol{\beta_{ij}} = (\beta_{1j}, \ldots, \beta_{Gj})$, para todo i e j,

$$\beta_{ij}|\boldsymbol{\beta}_{-ij} \sim N\left(\bar{\beta}_{ij}, \frac{\sigma_\beta^2}{n_{ij}}\right), \tag{3.15}$$

sendo:

- $\boldsymbol{\beta_{-ij}}$: os valores de β_{ij} para todo $j \neq i$;
- $\overline{\beta}_{ij}$: média dos n_{ij} valores de β_{ij} para as células j que são vizinhas da célula i;
- σ_β^2: variância comum para todos os valores de β.

Novamente, para garantir identificabilidade [2], os elementos de $\boldsymbol{\beta}$ foram restringidos para que $\sum_{i=1}^{G} \beta_i = 0$. A estimação das condicionais completas para β_i com distribuição *a priori* Normal independente e distribuição *a priori* ICAR podem ser verificadas em [25], e também a amostragem utilizando *Monte Carlo Markov Chain (MCMC)* com o algoritmo *Slice Sampling* desenvolvido por [24].

3.3 Atualização do modelo com a inclusão das áreas de restrições espaciais

Os modelos citados anteriormente não contemplam os casos em que a área de geração de coordenadas sintéticas possuem espaços não habitáveis, por exemplo lagoas, parques, aeroportos, entre outros. Nessa seção, iremos propor a inclusão da área restrita nos modelos com o objetivo de gerar coordenadas sintéticas somente em áreas de localizações habitáveis.

Para realizar a inclusão da área de restrição na função de geração de coordenadas sintéticas propomos dois algoritmos. No Algoritmo 1, verificamos a interseção entre a área de demarcação das células da grade e a área restrita (espaço em que não haverá geração de

coordenadas sintéticas). Para desenvolver tal verificação, utilizamos as funções `st_area` e `st_intersection` do pacote `sf` do *software* `R` ([28], [29]).

Note que, ao calcular a interseção entre as áreas de cada célula da grade e a área de restrição, utilizaremos o complementar desse resultado. Uma vez que a interseção tem como resultado a proporção da área restrita que está dentro da célula da grade, e para o cálculo que iremos realizar do novo valor de $\boldsymbol{\lambda} = \{\lambda_i^{(b)}\}$, necessitamos da proporção da célula da grade que não está na área restrita.

Algorithm 1 Algoritmo de interseção das áreas

Input: *coords_area_restrita*: coordenadas das áreas restritas (lon, lat)
lon_vector: vetor com as divisões da grade na longitude
lat_vector: vetor com as divisões da grade na latitude
grid: tamanho do grid em cada dimensão
Output: a: matriz de tamanho ($grid \times grid$) com as proporções da interseção entre as células da grade e as áreas restritas

$a \leftarrow$ matriz vazia de tamanho $grid \times grid$
poligonos_areas_restritas $\leftarrow$ lista com polígonos das áreas restritas criada a partir do objeto *coords_area_restrita*

for $j = 1$ **to** tamanho do *lat_vector* **do**
  for $i = 1$ **to** tamanho do *lon_vector* **do**
    coords_celulas_grade$[i, j]$ = concatena os quatro vértices da célula iniciando em (*lon_vector*$[i]$, *lat_vector*$[j]$)
    poligono_celulas_grade$[i, j]$ = transforma em polígono o conjunto de coordenadas (*coords_celulas_grade*$[i, j]$)
    $a[i, j]$ = 1 - (área da interseção entre *poligono_celulas_grade*$[i, j]$ e *poligonos_areas_restritas*) / área do *poligono_celulas_grade*$[i, j]$
  end for
end for

No Algoritmo 2, descrevemos a atualização do procedimento do cálculo da Equação 3.11. Uma vez que armazenamos as proporções de cada célula da grade em a_i que será utilizada no cálculo do novo valor de $\boldsymbol{\lambda}$, temos a seguinte atualização no cálculo da probabilidade. Para todo (i, b),

$$p_i^{(lb)} = \frac{\lambda_i^{(lb)} \times a_i}{\sum_{i=1}^{G} \lambda_i^{(lb)} \times a_i}. \tag{3.16}$$

Algorithm 2 Algoritmo de cálculo do novo λ

Input: a: transformação da matriz de tamanho ($grid \times grid$) com as proporções da interseção entre as células da grade e as áreas restritas em vetor
$lambda\_celula$: array contendo o λ de cada célula da grade por cada combinação e para cada um dos bancos de dados sintéticos
m: quantidade de bancos sintéticos
B: número de combinações
Output: p: array contendo as probabilidades de λ calculadas para cada célula da grade por cada combinação e para cada um dos bancos de dados sintéticos

$G \leftarrow$ grid $\times$ grid
for $k = 1$ **to** m **do**
  for $j = 1$ **to** B **do**
    soma_lambda = somar os lambdas de lambda_celula[, k, j]
    for $i = 1$ **to** G **do**
      p[i, k, j] = (lambda_celula[i] $\times$ a[i]) / (soma_lambda $\times$ a[i])
    end for
  end for
end for

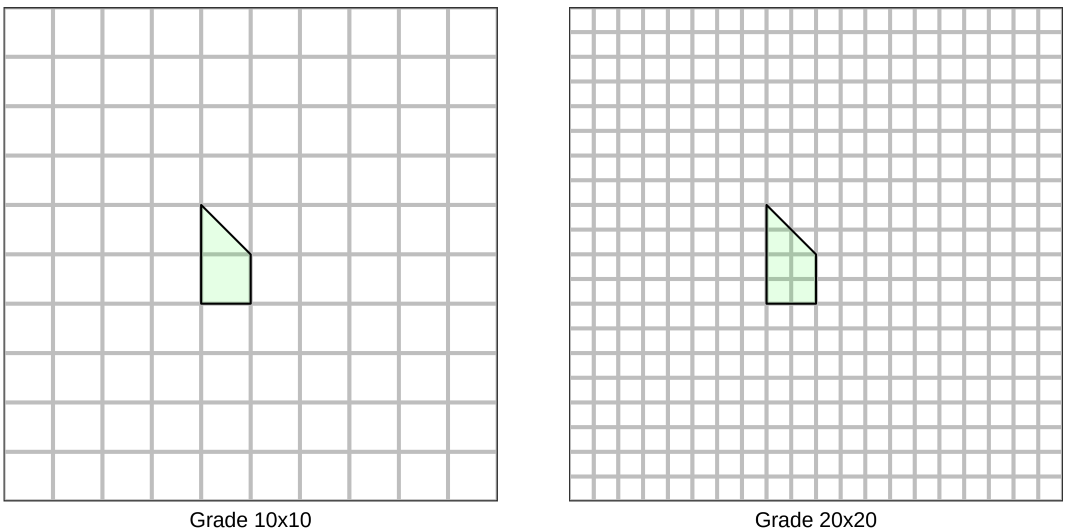
É importante ressaltar que implementamos um passo de rejeição em que é verificado se a coordenada sintética gerada está dentro da área restrita, caso esteja é gerado uma nova coordenada sintética. Não houve alteração nos outros procedimentos que são realizados para gerar as coordenadas sintéticas.

Na Figura 3.1 exemplificamos o cálculo do a_i, temos o espaço sombreado que é a área restrita e a divisão das células da grade. Os valores de a_i para esse exemplo são:

- **Grade 10x10**: $a_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0,5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$

- **Grade 20x20**: $a_i = \begin{bmatrix} 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&0,5&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&0&0,5&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&0&0&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&0&0&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \end{bmatrix}$

Figura 3.1: Exemplo de cálculo do a_i com as células da grade de 10x10 e de 20x20, e a área de restrição espacial que não possuem coordenadas geográficas.



Fonte: Elaborado pela autora.

Por fim, continuamos definindo o modelo da seguinte forma, seja $\boldsymbol{D} = (\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Z})$, em que $\boldsymbol{S}$ contém as coordenadas geográficas, $\boldsymbol{X}$ é o vetor de variáveis discretas $(X_1, \ldots, X_p)$, e $\boldsymbol{Z}$ é o vetor de variáveis contínuas $(Z_1, \ldots, Z_j)$. Considere $b = 1, \ldots, B$ como a combinação de atributos de $\boldsymbol{X}$, a grade regular G com células indexadas por $i = 1, \ldots, G$.

Sendo n, o parâmetro que atua como *offset* da amostra de dados, e a_i, a ponderação de cada célula da grade, indexadas por $i = 1, \ldots, G$, na área que será utilizada para gerar as coordenadas sintéticas.

Como não houve alteração no modelo em si, entretanto houve a criação da ponderação a_i, o modelo continua sendo definido do seguinte modo:

$$c_i^{(b)} \sim Poisson(n\lambda_i^{(b)}), \tag{3.17}$$

$$\log \lambda_i^{(b)} = \mu + \sum_{k=1}^{p} \alpha_k^{'} \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \theta_i + \sum_{k=1}^{p} \phi_{ik}^{'} \mathbb{1}_{\left\{x_k^{(b)}\right\}} + \sum_{j=1}^{q} \beta_{ij} \overline{Z}_{ij}^{(b)} + \epsilon_i^{(b)}, \tag{3.18}$$

em que:

- μ, $\alpha_k^{'}$, θ_i, $\mathbb{1}_{\left\{x_k^{(b)}\right\}}$, $\phi_{ik}^{'}$ e $\epsilon_i^{(b)}$ possuem a mesma interpretação e distribuições *a priori* mencionadas na Seção 3.1;
- β_{ij} possui a mesma interpretação e distribuições mencionadas na Seção 3.2;
- $\overline{Z}_{ij}^{(b)}$ possui a mesma interpretação mencionada na Seção 3.2.

Capítulo 4

Avaliação do Modelo

Neste capítulo, avaliaremos a inclusão da restrição espacial nos modelos apresentados no Capítulo 3. Utilizaremos dados simulados que conhecemos a área onde estarão as coordenadas e a área que não terá coordenadas, esta última chamaremos de área restrita. Para fins de identificação chamaremos os dados que foram gerados através de simulação de banco de dados originais simulado, e o conjunto de dados que são gerados através da função das coordenadas sintéticas de bancos de dados sintéticos, note que o banco de dados sintéticos contempla mais de um banco.

4.1 Dados Simulados

Utilizamos um método similar à [25] na geração do banco de dados simulados, com as covariáveis discretas $x_1 \in \{1, 2\}$, $x_2 \in \{1, 2, 3\}$ e $y \in \{0, 1\}$, a covariável contínua z e ainda as coordenadas de localização dos indivíduos $s = (s_1, s_2)$. Para as simulações a seguir, consideramos um banco simulado com 500 observações. Como o intuito dessa simulação também compreende avaliar a inclusão de uma área de restrição de geração dos dados, limitamos a geração das coordenadas para os dados simulados no intervalo entre $(0, 10) \times (0, 10)$.

As localizações s são geradas a partir dos valores das covariáveis. Para isso, primeiro geramos os valores de x_1 de acordo com as probabilidades da Tabela 4.1. A partir de x_1 geramos os valores para x_2 com distribuição de probabilidade condicional, ou seja, $x_2|x_1$, em que as probabilidades podem ser observadas na Tabela 4.2.

Tabela 4.1: Distribuição de probabilidade de x_1

x	$p(X_1 = x)$
1	0,6
2	0,4

Fonte: [25].

Tabela 4.2: Distribuição de probabilidade de $x_2|x_1$

x	$p(X_2 = x\|X_1 = 1)$	$p(X_2 = x\|X_1 = 2)$
1	0,333	0,6
2	0,333	0,1
3	0,333	0,3

Fonte: [25].

Para as probabilidades de y, utilizamos a seguinte regressão logística:

$$logit(p(y = 1)) = \beta_0 + \beta_1 \mathbb{1}_{(x_1=2)} + \beta_2 \mathbb{1}_{(x_2=2)} + \beta_3 \mathbb{1}_{(x_2=3)}, \tag{4.1}$$

onde $\beta_0 =$ -1, $\beta_1 =$ 1,5, $\beta_2 =$ -0,5 e $\beta_3 =$ 0,5. Os valores dos coeficientes aqui foram fixados arbitrariamente, apenas para gerar valores de y de diferentes distribuições de acordo com os valores de x_1 e x_2.

E, para gerar a variável contínua z utilizamos:

$$\overline{Z}^{(b)} = \beta_0 + \beta_1 \mathbb{1}_{(y=1)} + \beta_2 \mathbb{1}_{(x_1=2)} + \beta_3 \mathbb{1}_{(x_2=2)} + \beta_4 \mathbb{1}_{(x_2=3)}, \tag{4.2}$$

onde $\beta_0 = 50$, $\beta_1 =$ -3,5, $\beta_2 = 1$, $\beta_3 =$ -1,5 e $\beta_4 = 2$. Temos a média de Z para cada combinação, a partir dessa média geramos um valor de Z para cada indivíduo de uma distribuição Normal($\overline{Z}^{(b)}$, 10^2).

Finalmente, o último passo é gerar as coordenadas geográficas a partir de densidades de normais bivariadas. Para cada combinação possível dos valores já simulados de x_1, x_2 e y, geramos os valores de (s_1, s_2) de acordo com as distribuições bivariadas descritas na Tabela 4.3.

Segundo [25], as distribuições da Tabela 4.3 demonstram diferentes padrões espaciais para as combinações. Sendo assim, utilizando essas intensidades, conseguimos gerar as coordenadas de maneira variada dependendo dos valores das covariáveis.

Seja n_b o número de pontos que serão gerados para cada combinação b. Como realizamos a geração dos valores de y, x_1, x_2 e z, já sabemos os números de pontos que serão para cada combinação. E para gerarmos os n_b, para cada combinação b utilizamos a probabilidade das densidades normais escolhidas na Tabela 4.3, sendo que as coordenadas são geradas exatamente das normais.

Na Figura 4.1 temos a distribuição espacial dos 500 dados originais simulados, de acordo com o valor das variáveis y, x_1, x_2 e z. Note que fixamos o limite das localizações dos dados entre (0,10) no eixo horizontal e (0,10) no eixo vertical, e também incluímos um retângulo $((4 \times 6), (4 \times 5))$ que corresponde a área de restrição espacial, que será removida da geração de localizações sintéticas. Note que esse retângulo também foi removido na geração das coordenadas geográficas originais simuladas, ou seja, não existem localizações dentro desse espaço nos dados simulados originais.

Tabela 4.3: Distribuições dos dados simulados para s

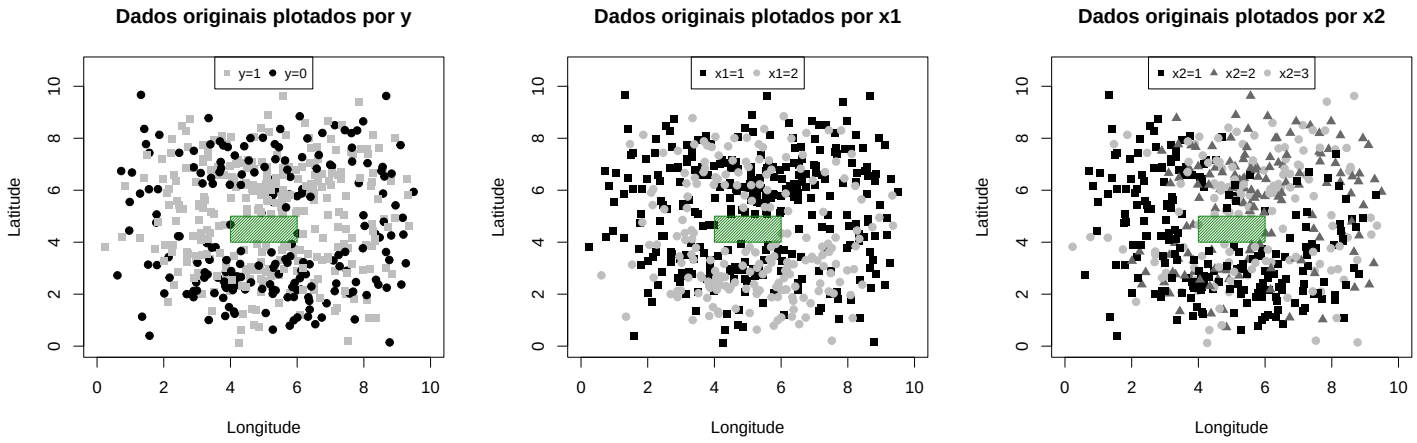
y	x_1	x_2	Distribuição de $\begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$
0	1	1	$N_2\left[\begin{pmatrix} 3 \\ 7 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$ & $N_2\left[\begin{pmatrix} 7 \\ 3 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$
		2	$N_2\left[\begin{pmatrix} 3 \\ 3 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$ & $N_2\left[\begin{pmatrix} 7 \\ 7 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$
		3	$N_2\left[\begin{pmatrix} 5 \\ 5 \end{pmatrix}; \begin{pmatrix} 0,3 & 0 \\ 0 & 4,5 \end{pmatrix}\right]$
	2	1	$N_2\left[\begin{pmatrix} 3 \\ 5 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$ & $N_2\left[\begin{pmatrix} 7 \\ 5 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$
		2	$N_2\left[\begin{pmatrix} 5 \\ 3 \end{pmatrix}; \begin{pmatrix} 0,8 & 0 \\ 0 & 0,8 \end{pmatrix}\right]$ & $N_2\left[\begin{pmatrix} 5 \\ 7 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$
		3	$N_2\left[\begin{pmatrix} 7 \\ 3 \end{pmatrix}; \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right]$
1	1	1	$N_2\left[\begin{pmatrix} 1,5 \\ 5 \end{pmatrix}; \begin{pmatrix} 0,3 & 0 \\ 0 & 4,5 \end{pmatrix}\right]$
		2	$N_2\left[\begin{pmatrix} 5 \\ 5 \end{pmatrix}; \begin{pmatrix} 1,5 & 0,9 * \sqrt{(1,5 * 2)} \\ 0,9 * \sqrt{(1,5 * 2)} & 2 \end{pmatrix}\right]$
		3	$N_2\left[\begin{pmatrix} 8,5 \\ 5 \end{pmatrix}; \begin{pmatrix} 0,3 & 0 \\ 0 & 4,5 \end{pmatrix}\right]$
	2	1	$N_2\left[\begin{pmatrix} 5 \\ 2 \end{pmatrix}; \begin{pmatrix} 2,5 & 0 \\ 0 & 0,7 \end{pmatrix}\right]$
		2	$N(5,5); N_2\left[\begin{pmatrix} 5 \\ 5 \end{pmatrix}; \begin{pmatrix} 1,5 & -0,9 * \sqrt{(1,5 * 1,5)} \\ -0,9 * \sqrt{(1,5 * 1,5)} & 2 \end{pmatrix}\right]$
		3	$N(5;7,5); N_2\left[\begin{pmatrix} 5 \\ 7,5 \end{pmatrix}; \begin{pmatrix} 2,5 & 0 \\ 0 & 0,7 \end{pmatrix}\right]$

Fonte: [25].

Na Figura 4.2, observamos as intensidades verdadeiras fixadas pelas distribuições da Tabela 4.3, e para cada combinação de atributos temos a amostra de localizações geradas nos dados originais simulados. Além disso, temos também a área restrita que foi incluída nos dados originais simulados. Notamos que todos os pontos estão fora da área restrita mas ainda é esperado a ocorrência de pontos ao redor da área restrita.

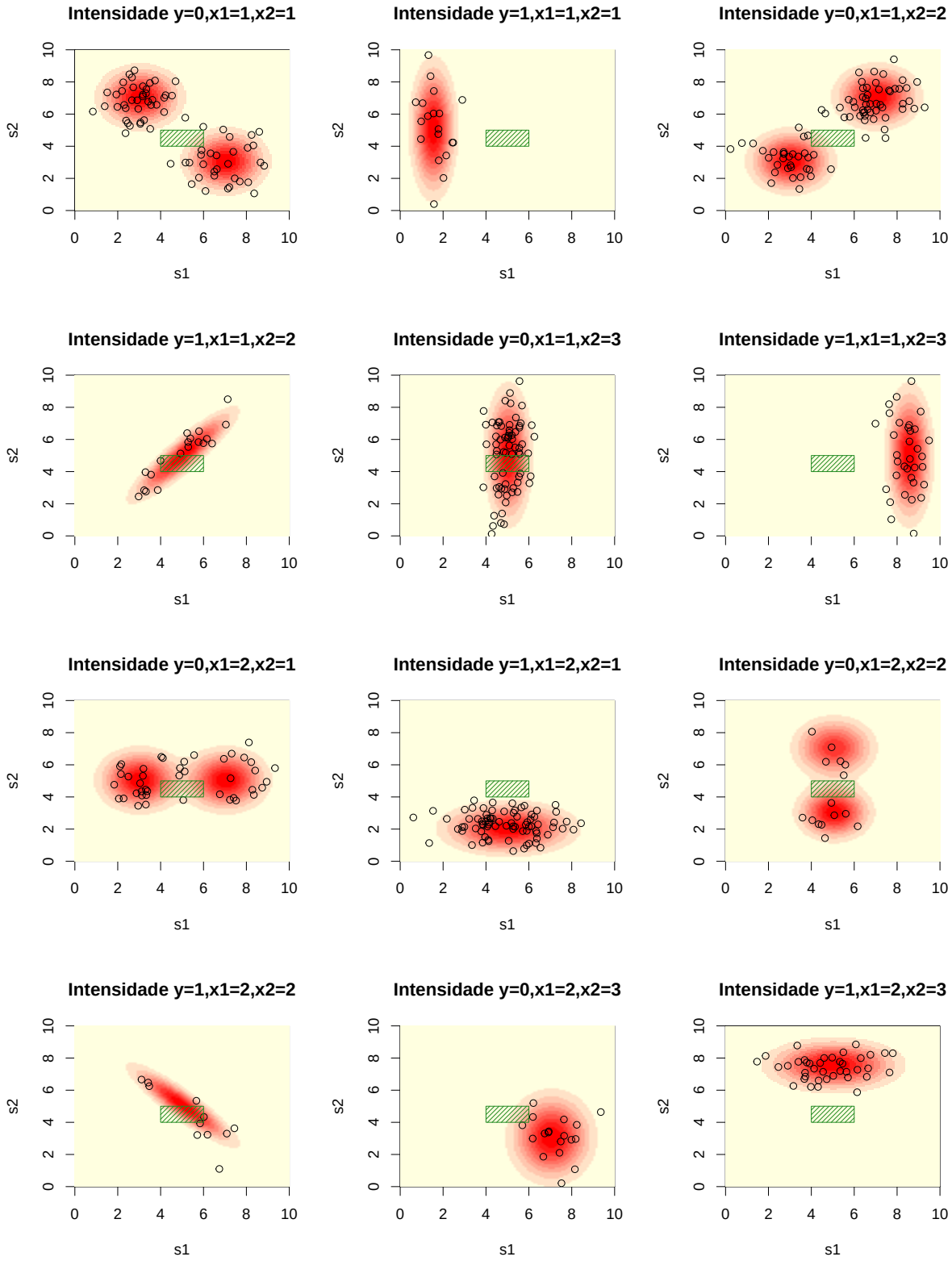
Por fim, realizamos a análise da variável contínua, observamos a Figura 4.3 que demonstra como estão distribuídos os valores dessa variável através das longitudes e latitudes originais simuladas. Os pontos estão bastante espalhados porém existe uma quantidade maior de valores no intervalo entre 30 e 60, e observamos alguns valores espalhados para abaixo de 20 e acima de 60 também. Já a Figura 4.4 demonstra os box-plots da z pelas combinações possíveis. Note que o mínimo e máximo da variável contínua são 16 e 80, respectivamente, sendo que a média está entre 48-49.

Figura 4.1: Dados Originais Simulados.



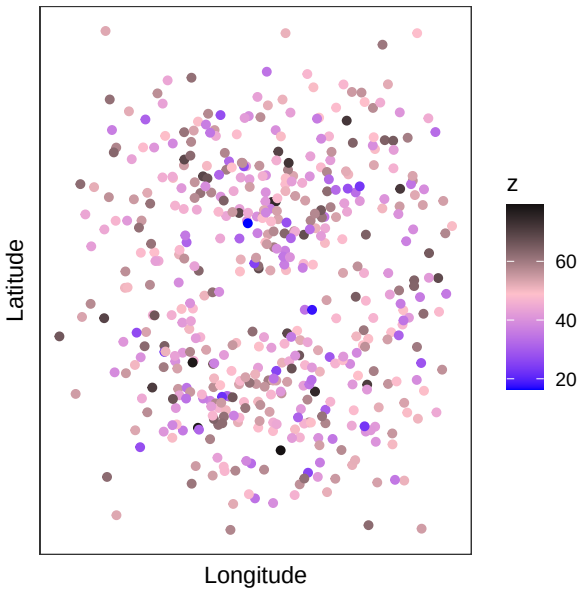
Fonte: Elaborado pela autora.

Figura 4.2: Intensidades fixadas e coordenadas originais simuladas para cada uma das combinações com a inclusão da área restrita.



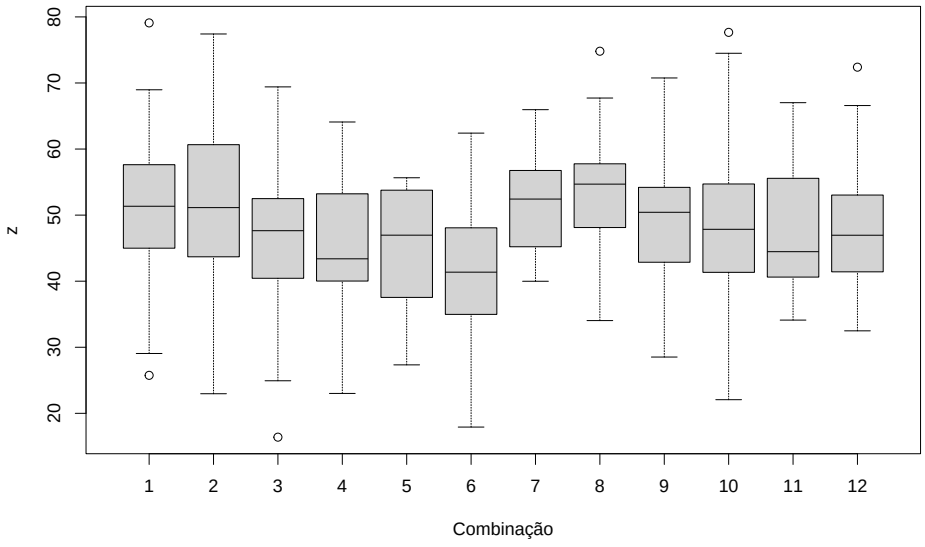
Fonte: Elaborado pela autora.

Figura 4.3: Faixas da variável contínua (z) dos dados originais simulados através dos pontos de longitude e latitude.



Fonte: Elaborado pela autora.

Figura 4.4: Box-plots da variável contínua dos dados originais simulados para cada uma das combinações.



Fonte: Elaborado pela autora.

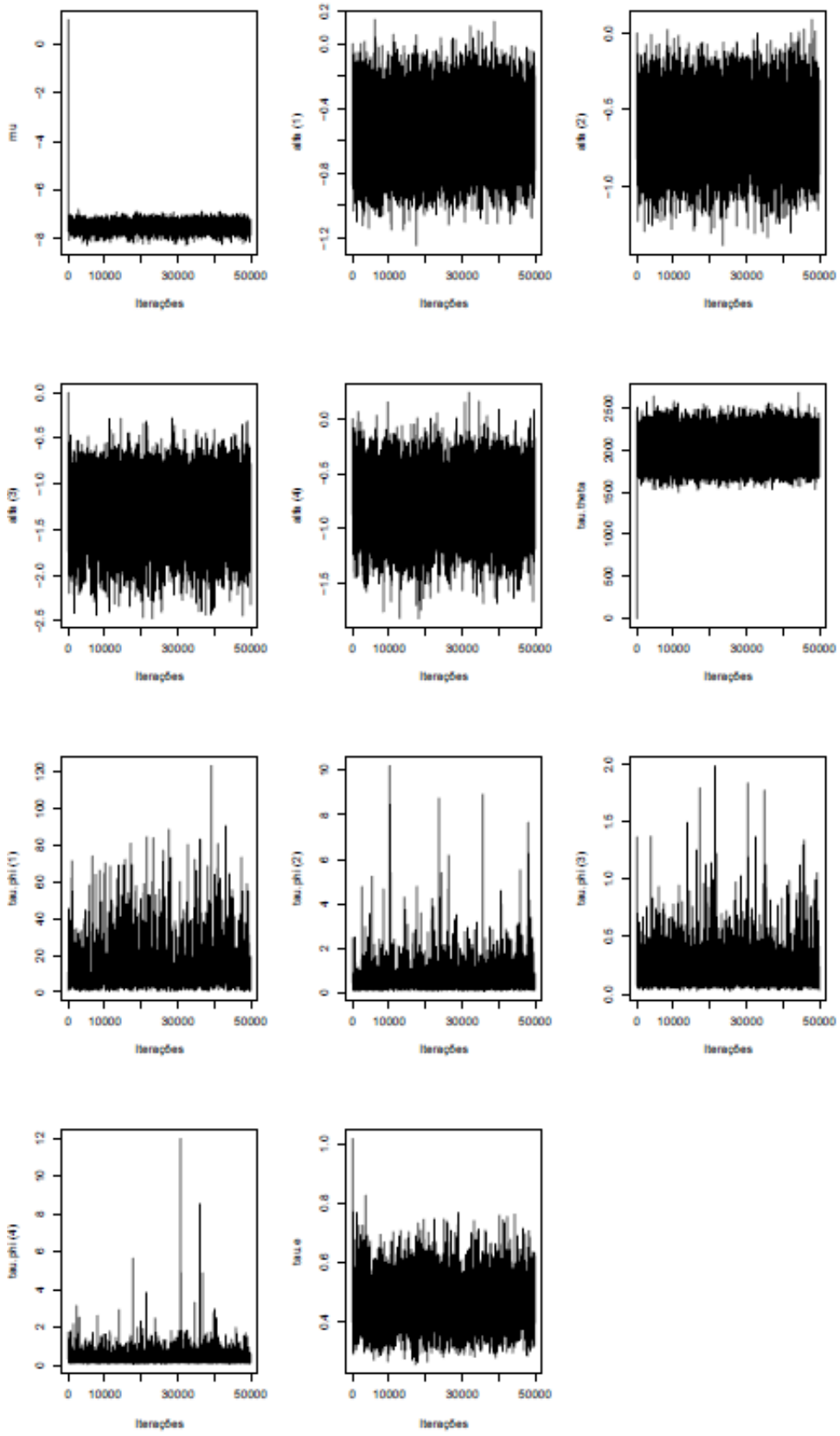
4.2 Avaliando o ajuste do modelo

Nessa seção gostaríamos de verificar como está a convergência dos parâmetros do modelo. Portando, após gerarmos as coordenadas originais simuladas, utilizamos $D = (y, x_1, x_2, z, s_1, s_2)$ para estimar as intensidades de acordo com o modelo, e gerar as coordenadas sintéticas. Utilizando a função `syn_mcmc` (que será descrita no Capítulo 5) do pacote em desenvolvimento para o `Software R` ([30]) geramos $S = 50.000$ iterações do MCMC. Descartamos, no período de aquecimento, 5.000 iterações e utilizamos uma grade regular 10×10, sendo assim $G = 100$.

Na Figura 4.5 podemos observar a convergência dos parâmetros (μ, α, τ_θ, τ_ϕ e τ_ϵ) do modelo aplicado nos dados originais simulados. Os parâmetros que possuem dependência espacial (θ_i e ϕ_{ik}) não estão apresentados no gráfico porque teríamos um gráfico para cada i, $i = 1, \ldots, G$. Observamos que todos os parâmetros convergiram e, por padrão do modelo, μ e τ_θ iniciaram em zero mas, em seguida, convergiram para os seus respectivos valores.

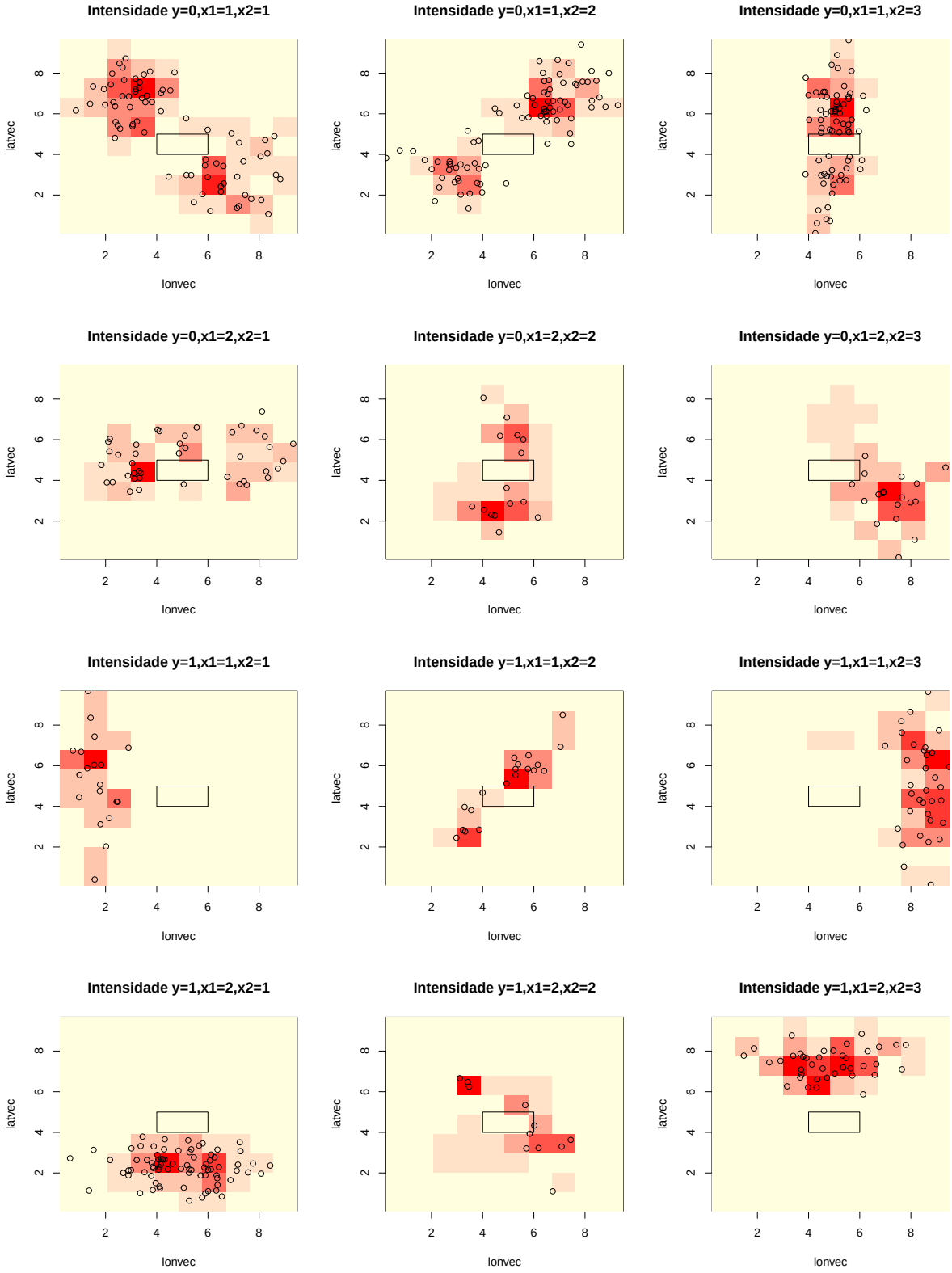
As intensidades médias estimadas pelo modelo desenvolvido por [26] e atualizado por [25] podem ser observadas na Figura 4.6. Notamos que as intensidades estimadas condizem com os pontos das coordenadas originais simuladas.

Figura 4.5: Traceplots dos parâmetros gerados através do modelo com S = 50.000, período de aquecimento de 5.000 e G = 100.



Fonte: Elaborado pela autora.

Figura 4.6: Intensidades estimadas e coordenadas originais simuladas para cada uma das combinações com o destaque da área restrita.



Fonte: Elaborado pela autora.

4.3 Análise do modelo nos dados simulados com a inclusão da área restrita

Utilizamos a função `syncoordinates` do pacote `syncoordinatesr` que está em desenvolvimento para o `Software R` ([30]) para gerar, através do modelo, as coordenadas sintéticas dos dados originais simulados. Para analisar a inclusão da área restrita, geramos $m = 4$ bancos de coordenadas sintéticas, com e sem a inclusão da área restrita. Optamos por realizar $S = 10.000$ iterações pois verificamos a convergência dos parâmetros e nessa quantidade de iterações eles já convergiam. Com isso temos os seguintes cenários de análise, com $n = 500$, $S = 10.000$, $burn - in = 1.000$ para todos eles:

- **Caso 1:** $grid = 10 \times 10$, com área restrita, com variável contínua;
- **Caso 2:** $grid = 10 \times 10$, com área restrita, sem variável contínua;
- **Caso 3:** $grid = 10 \times 10$, sem área restrita, com variável contínua;
- **Caso 4:** $grid = 10 \times 10$, sem área restrita, sem variável contínua;
- **Caso 5:** $grid = 20 \times 20$, com área restrita, com variável contínua;
- **Caso 6:** $grid = 20 \times 20$, com área restrita, sem variável contínua;
- **Caso 7:** $grid = 20 \times 20$, sem área restrita, com variável contínua;
- **Caso 8:** $grid = 20 \times 20$, sem área restrita, sem variável contínua;

Como especificado anteriormente, a área restrita considerada é um retângulo de (4×6), (4×5) e os dados simulados possuem o limite implementado de geração de coordenadas sintéticas entre (0×10),(0×10). Criamos esse limite para os casos em que o usuário das funções do pacote desejem restringir a área de geração de coordenadas sintéticas.

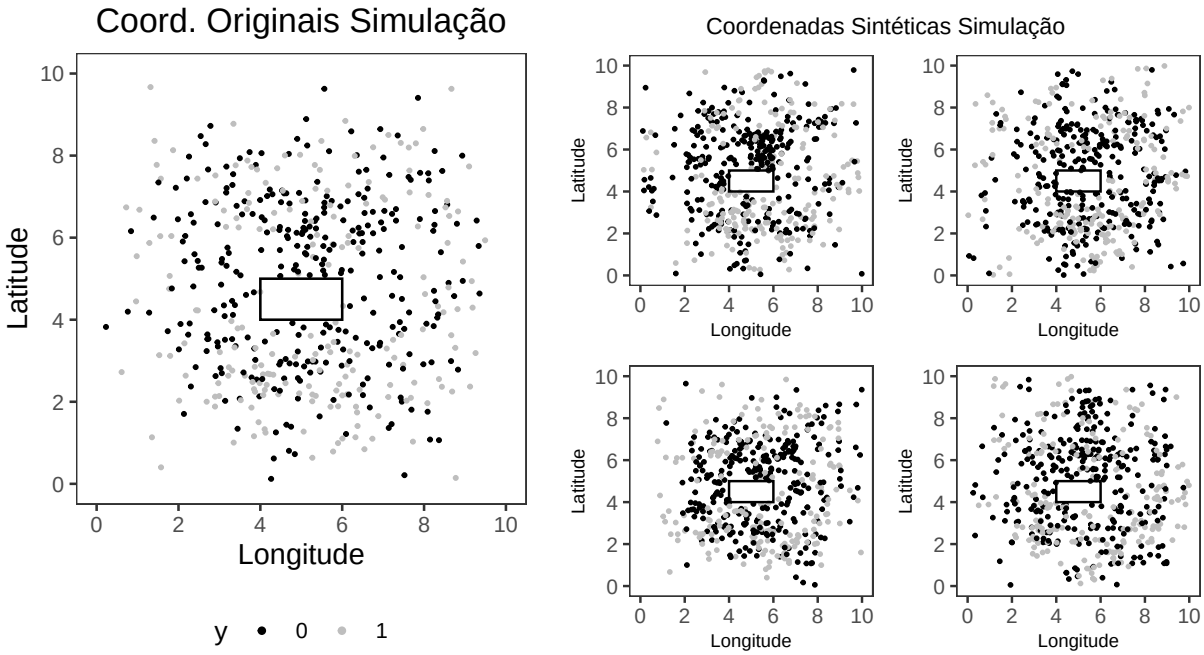
Nas Figuras 4.7 a 4.14, podemos observar a comparação entre as coordenadas originais da simulação e os $m = 4$ bancos de dados de coordenadas sintéticas gerados para cada um dos casos especificados acima.

Observamos claramente a diferença entre os casos que possuem e os casos que não possuem a implementação da área restrita no modelo. No espaço vazio, sem dados de localizações, do banco de dados original simulado, notamos que a função de geração das coordenadas sintéticas não realiza tratamentos para esses casos, e mesmo assim são geradas localizações sintéticas nesse espaço. Entretanto, com a inclusão da área restrita a função não mais gera valores de coordenadas sintéticas para o espaço vazio dos dados

originais simulados. Com esse ajuste na função, o usuário poderá determinar as áreas restritas em que não serão geradas as coordenadas sintéticas.

Notamos algumas diferenças entre $G = 100$ e $G = 400$, como por exemplo um maior refinamento nas coordenadas sintéticas, uma vez que a área foi subdividida em mais células da grade. Também é importante destacar que os painéis referentes às coordenadas sintéticas seguem um padrão similar de nuvem de pontos entre si e entre o painel das coordenadas originais. Entretanto, como já mencionado, observamos pequenas alterações devido às oscilações de valores que ocorrem nas coordenadas sintéticas.

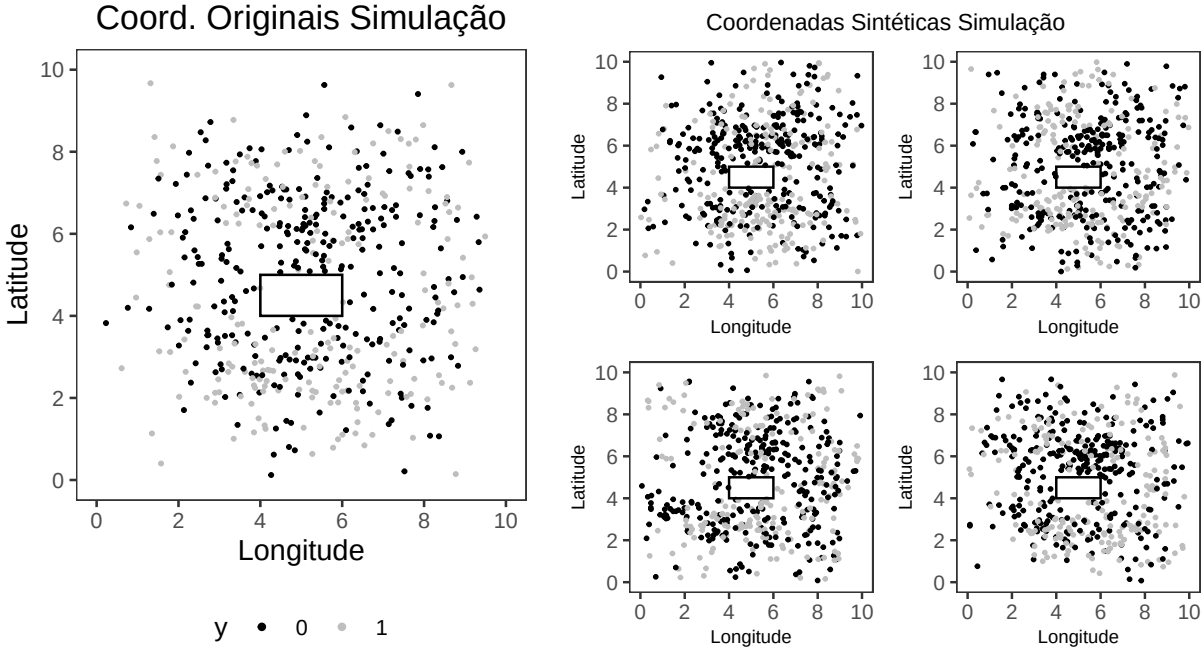
Figura 4.7: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 1 ($grid = 10\times10$, com área restrita, com variável contínua).



Fonte: Elaborado pela autora.

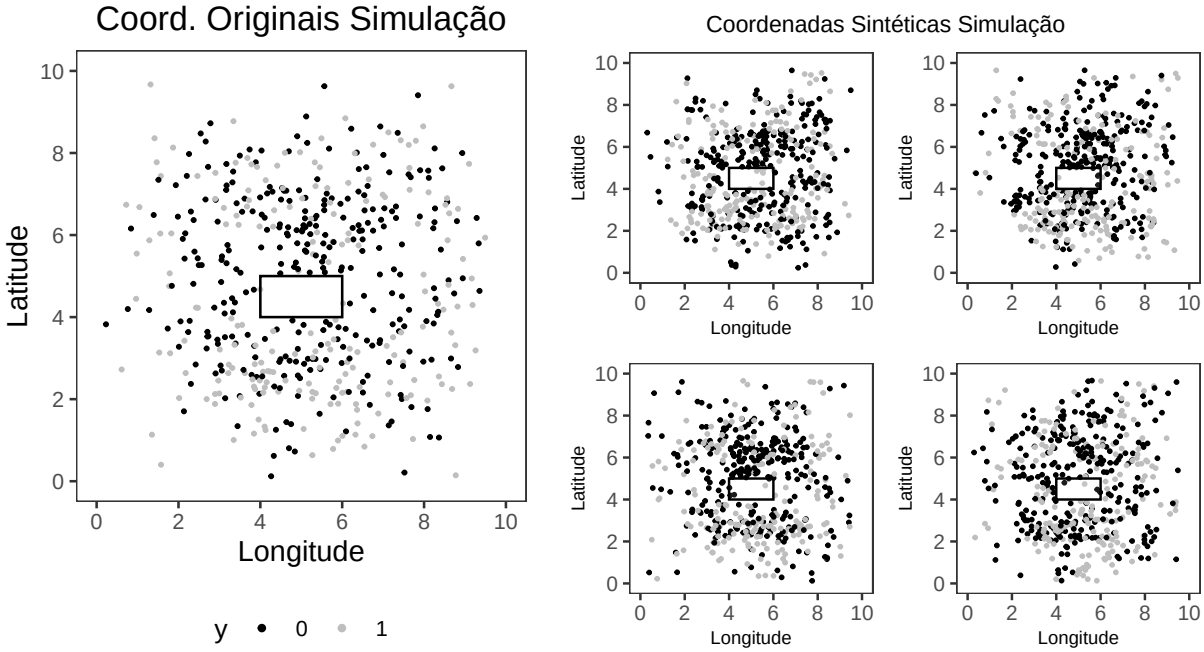
No Apêndice A estão as nuvens de pontos das longitudes e latitudes originais e sintéticas para cada um dos casos analisados. Observamos que as coordenadas sintéticas estão marginalmente distribuídas na mesma região do espaço que as coordenadas originais.

Figura 4.8: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 2 ($grid = 10\times10$, com área restrita, sem variável contínua).



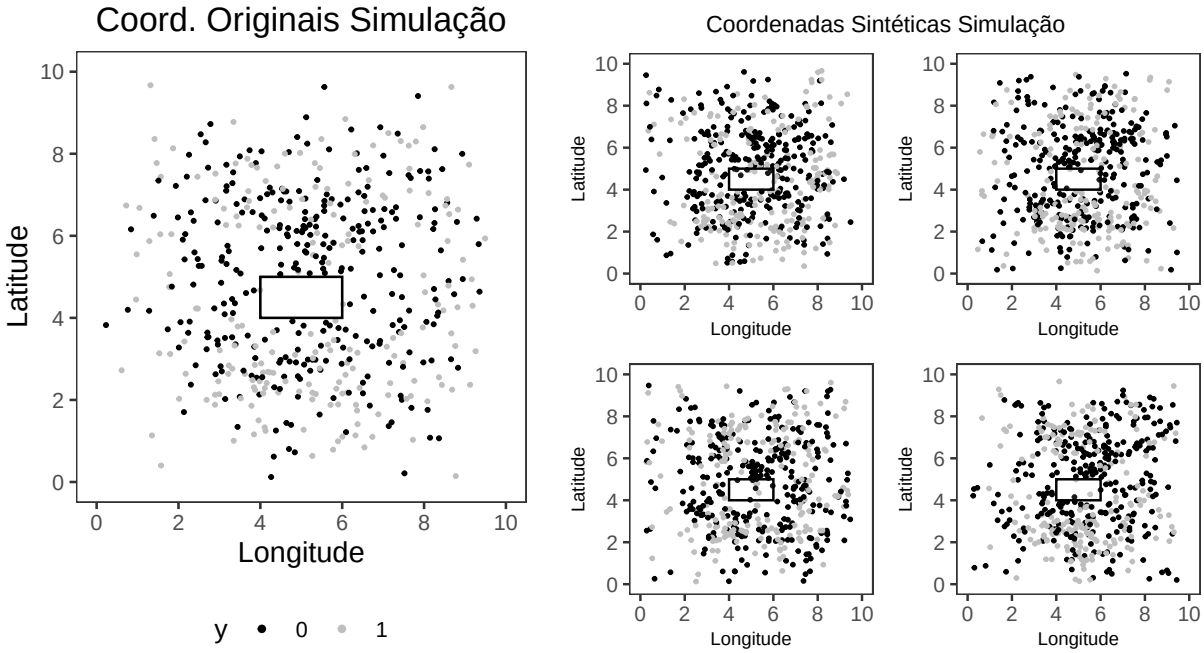
Fonte: Elaborado pela autora.

Figura 4.9: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 3 ($grid = 10\times10$, sem área restrita, com variável contínua).



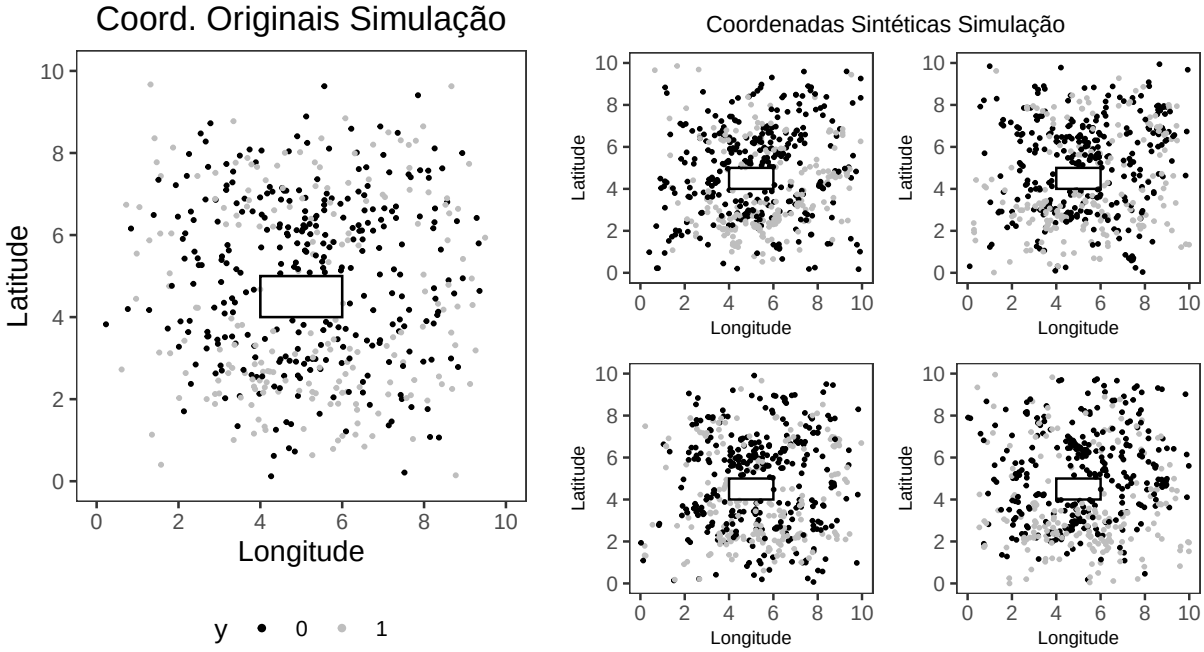
Fonte: Elaborado pela autora.

Figura 4.10: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 4 ($grid = 10\times10$, sem área restrita, sem variável contínua).



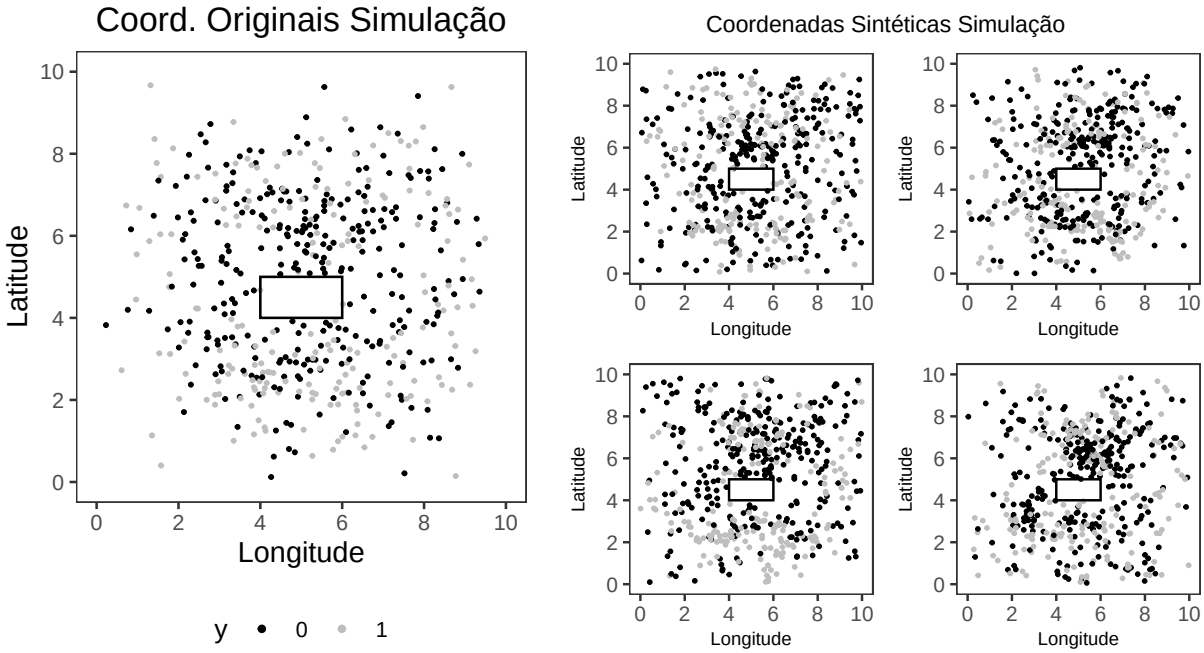
Fonte: Elaborado pela autora.

Figura 4.11: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 5 ($grid = 20\times20$, com área restrita, com variável contínua).



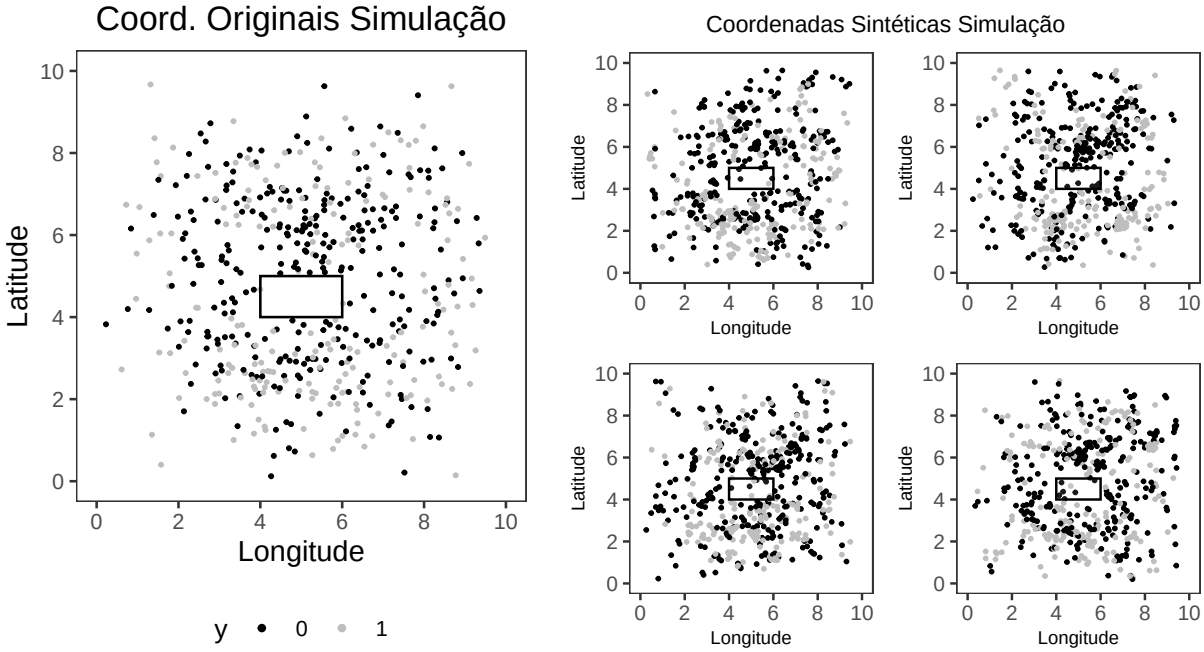
Fonte: Elaborado pela autora.

Figura 4.12: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 6 ($grid = 20\times20$, com área restrita, sem variável contínua).



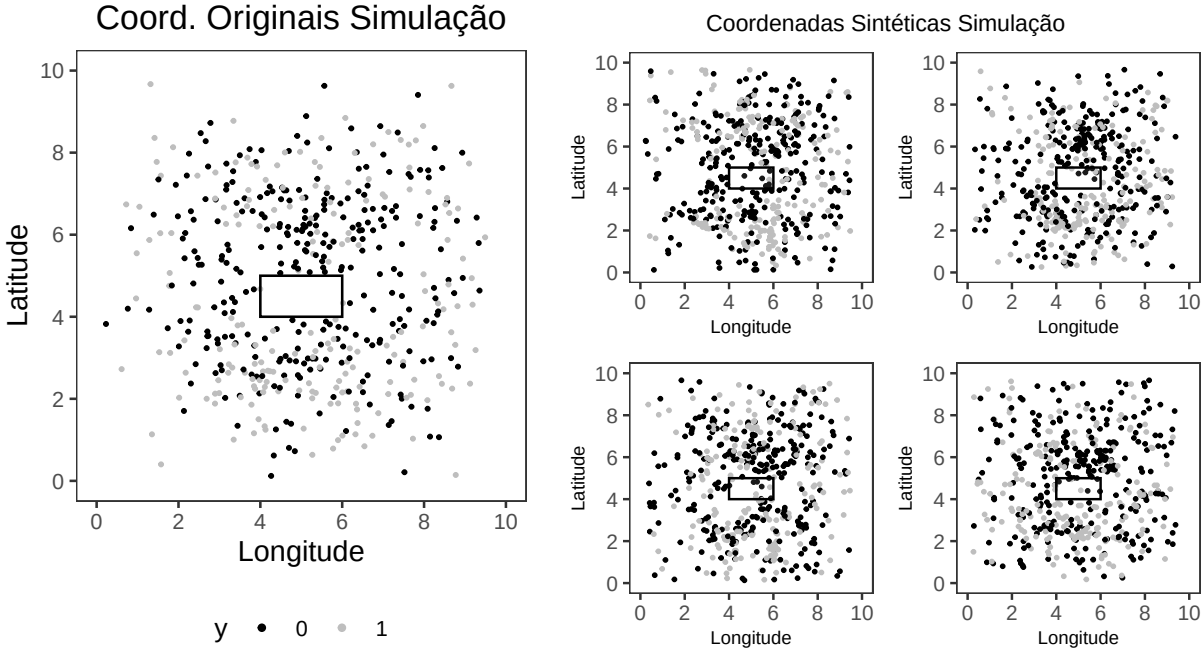
Fonte: Elaborado pela autora.

Figura 4.13: Coordenadas originais dos dados simulados e coordenadas sintéticas geradas para os dados simulados - Caso 7 ($grid = 20\times20$, sem área restrita, com variável contínua).



Fonte: Elaborado pela autora.

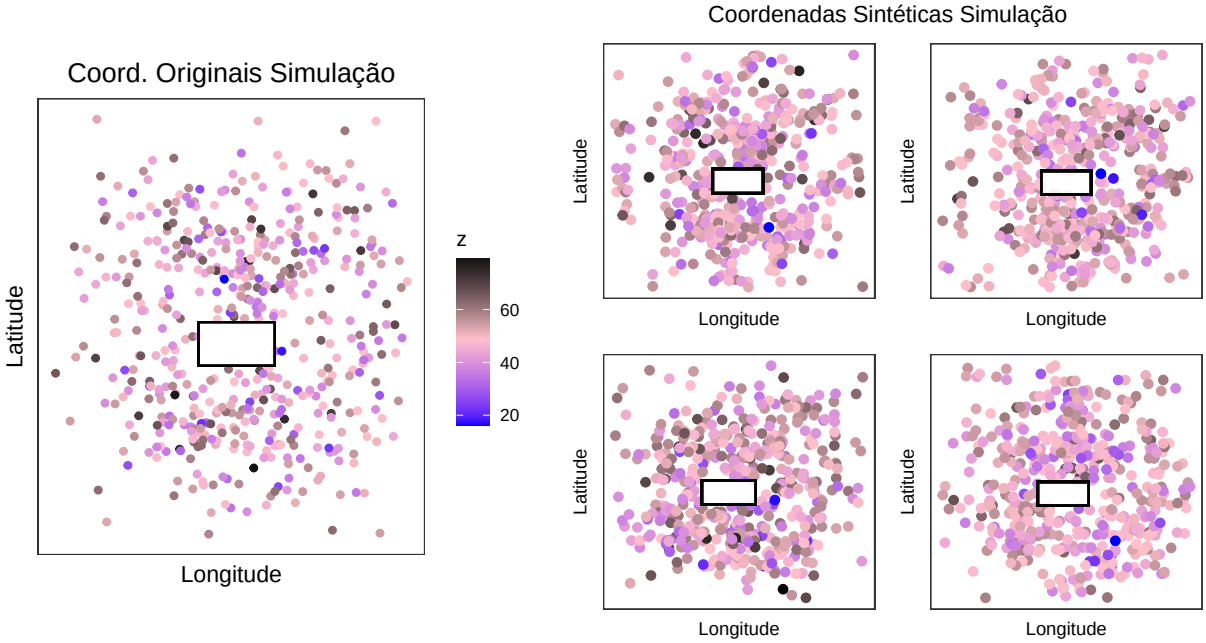
Figura 4.14: Coordenadas originais dos dados Simulados e Coordenadas Sintéticas Geradas para os Dados Simulados - Caso 8 ($grid = 20 \times 20$, sem área restrita, sem variável contínua).



Fonte: Elaborado pela autora.

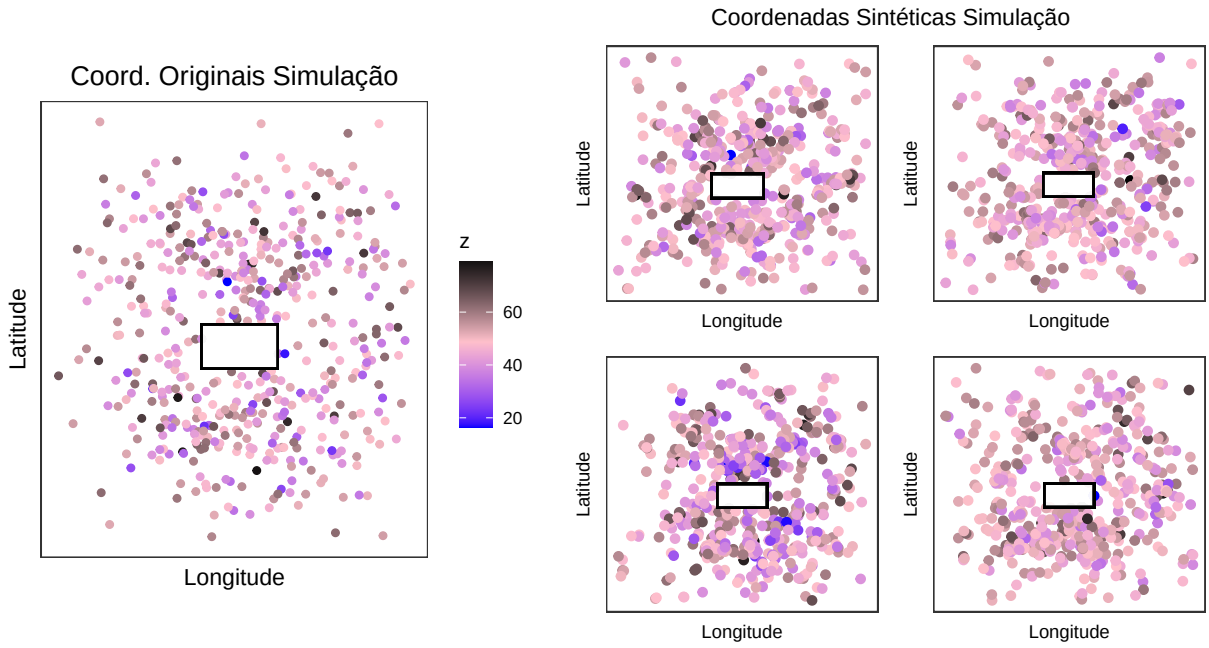
Por fim, selecionamos os casos 1 e 5 para demonstrar como ficaram as distribuições da variável contínua através da longitude e latitude, comparando entre os valores de z do banco original simulado com os valores de z nos $m = 4$ bancos de dados sintéticos para cada um dos casos. Optamos por selecionar esses dois casos porque eles possuem tanto a variável contínua quanto a área restrita em sua descrição. Observamos que os valores originais e os simulados possuem padrões similares de distribuição dos pontos, como pode ser verificado através das Figuras 4.15 e 4.16.

Figura 4.15: Distribuição da variável contínua dos dados simulados e dos dados sintéticos através dos pontos de longitude e latitude - Caso 1 ($grid = 10 \times 10$, com área restrita, com variável contínua).



Fonte: Elaborado pela autora.

Figura 4.16: Distribuição da variável contínua dos dados simulados e dos dados sintéticos através dos pontos de longitude e latitude - Caso 5 ($grid = 20 \times 20$, com área restrita, com variável contínua).



Fonte: Elaborado pela autora.

Capítulo 5

Pacote `syncoordinatesr`

Nesse capítulo, falaremos do pacote desenvolvido para o `Software R` ([30]) que contém as funções dos modelos utilizados neste trabalho. O pacote é chamado `syncoordinatesr` e atualmente está disponível para download no site https://github.com/leogalhardo/syncoordinatesr. Possui três principais funções para a geração das coordenadas sintéticas e seus autores são Thaís Paiva, Letícia Silva Nunes, Leonardo de Mattos Galhardo, Fernanda Buzza Alves Barros e Camila Artur de Souza Plácido Teixeira.

5.1 Função `prepare_data`

Função auxiliar para gerar objetos úteis para as outras funções do pacote. O objetivo da função é receber o banco de dados do usuário e gerar variáveis importantes que serão usadas no MCMC. E, no fim, para gerar as coordenadas sintéticas. Na entrada, a função recebe os parâmetros: `dataset`, `coord`, `limits`, `grid` e `continuous` que serão descritos a seguir.

Uso:

```
prepare_data(dataset, coord, limits = c(), grid = 10,
             continuous = FALSE)
```

Argumentos:

`dataset` : Banco de dados com todas as covariáveis discretas e contínuas, exceto as coordenadas.

`coord` : Objeto com duas colunas indicando a longitude e a latitude, respectivamente, dos elementos no conjunto de dados.

`limits` : Objeto que é um vetor das dimensões onde serão criadas as células da grade informadas pela sequência de *xmin*, *xmax*, *ymin*, *ymax*. O padrão é usando o máximo e o mínimo do objeto `coords`.

`grid` : Tamanho da divisão, em cada dimensão, da grade a ser criada na região dos dados. O valor *default* é `grid = 10`.

`continuous` : Objeto que indica quais colunas do conjunto de dados correspondem às variáveis contínuas. O valor *default* do argumento é `FALSE`, o que significa que não há nenhuma variável contínua. Caso contrário, o argumento deve receber um vetor numérico com os índices das colunas correspondentes.

Valor:
Uma lista contendo objetos úteis para a `syn_mcmc`.
Exemplo:

```
prepare_data(dataset = my_database, coord = my_coords,
             limits = c(0,10,0,10), grid = 10,
             continuous = FALSE)
```

5.2 Função `syn_mcmc`

Essa função executa o MCMC para ajustar o modelo das intensidades que será necessário obter as coordenadas sintéticas posteriormente. A função `syn_mcmc` recebe o banco de dados, e junto com a saída da função `prepare_data`, realiza o MCMC. Ao final desta função, obtemos as estimativas do parâmetro λ que será necessário na geração das coordenadas sintéticas.
Uso:

```
syn_mcmc(dataset, coord, limits = c(), grid = 10,
         S = 5000, burn = 1000, continuous = FALSE,
         spatial_beta = FALSE, return_parameters = FALSE)
```

Argumentos:

`dataset` : Banco de dados com todas as covariáveis discretas e contínuas, exceto as coordenadas.

`coord` : Objeto com duas colunas indicando a longitude e a latitude, respectivamente, dos elementos no conjunto de dados.

`limits` : Objeto que é um vetor das dimensões onde serão criadas as células da grade informadas pela sequência de *xmin*, *xmax*, *ymin*, *ymax*. O padrão é usando o máximo e o mínimo do objeto `coords`.

`grid` : Tamanho da divisão, em cada dimensão, da grade a ser criada na região dos dados. O valor *default* é `grid = 10`.

`S` : Quantidades de MCMC que serão feitas. O *default* é `S = 5000`.

`burn` : Número de simulações que serão descartadas do período de aquecimento do MCMC. O *default* é `burn = 1000`.

`continuous` : Objeto que indica quais colunas do conjunto de dados correspondem às variáveis contínuas. O valor *default* do argumento é `FALSE`, o que significa que não há nenhuma variável contínua. Caso contrário, o argumento deve receber um vetor numérico com os índices das colunas correspondentes.

`spatial_beta` : Argumento indicando se o usuário deseja especificar uma distribuição a *priori* espacial (como descrito na seção 3.2), para os parâmetros β. Se o usuário deseja especificar essa distribuição espacial para todos os β's, pode usar o argumento lógico `TRUE`. Caso contrário, o usuário deve inserir um vetor indicando qual β deverá contar com a priori espacial.

`return_parameters` : Argumento lógico indicando se a função deve retornar ou não os valores de todas as iterações de todos os parâmetros do modelo, além dos valores de λ.

Valor:
Dependendo do parâmetro `return_parameters`, esta função pode retornar apenas os valores simulados do parâmetro λ ou também todos os outros parâmetros individuais do modelo (μ, α, θ, ϕ, ϵ, τ_θ, τ_ϕ, τ_ϵ).
Exemplo:

```
syn_mcmc(dataset = my_database, coord = my_coords, S = 2500,
         burn = 500, return_parameters = TRUE)
```

5.3 Função syncoordinates

Essa função gera as coordenadas sintéticas a partir do resultado do MCMC. A função `syncoordinates` recebe o banco de dados, o parâmetro λ e a quantidade de dados sintéticos que o usuário deseja gerar. E a função retorna os bancos de dados sintéticos contendo as coordenadas sintéticas.

Uso:

```
syncoordinates(dataset, coord, grid = 10, continuous = FALSE,
               restricted_area = FALSE, coord_restricted_area,
               list_mcmc, n.syn = 5)
```

Argumentos:

`dataset` : Banco de dados com todas as covariáveis discretas e contínuas, exceto as coordenadas.

`coord` : Objeto com duas colunas indicando a longitude e a latitude, respectivamente, dos elementos no conjunto de dados.

`limits` : Objeto que é um vetor das dimensões onde serão criadas as células da grade informadas pela sequência de *xmin*, *xmax*, *ymin*, *ymax*. O padrão é usando o máximo e o mínimo do objeto `coords`.

`grid` : Tamanho da divisão, em cada dimensão, da grade a ser criada na região dos dados. O valor *default* é `grid = 10`.

`continuous` : Objeto que indica quais colunas do conjunto de dados correspondem às variáveis contínuas. O valor *default* do argumento é `FALSE`, o que significa que não há nenhuma variável contínua. Caso contrário, o argumento deve receber um vetor numérico com os índices das colunas correspondentes.

`restricted_area` : Argumento lógico que indica se existem áreas restritas onde não deverão ser geradas coordenadas geográficas sintéticas. O *default* é `FALSE`, o que significa que não há área restrita.

`coord_restricted_area` : Objeto com duas colunas indicando a longitude e a latitude dos pontos que formam as áreas restritas. Por *default*, para o objeto ser considerado um polígono é necessário que os primeiros pontos de longitude e latitude sejam iguais aos últimos pontos de longitude e latitude.

`list_mcmc` : Saída da função `syn_mcmc`.

`n.syn` : Números de bancos de dados sintéticos que serão gerados.

Valor:
O retorno dependerá do argumento `continuous`. Se `continuous = FALSE`, a função retornará um objeto da classe `data.frame` contendo todas as novas coordenadas sintéticas. Porém, se `continuous != FALSE`, além do `data.frame` com as coordenadas sintéticas, a função retornará novos dados sintéticos para cada variável contínua indicada no argumento `continuous`.
Exemplo:

```
syncoordinates(dataset = my_database, coord = my_coords,
               grid = 10, continuous = FALSE,
               restricted_area = FALSE,
               list_mcmc = my_mcmc, n.syn = 5)
```

5.4 Conclusão

Apresentamos neste capítulo todas as funções utilizadas nesse trabalho para geração das coordenadas sintéticas. Ressaltamos as principais contribuições deste trabalho no pacote `syncoordinatesr`:

1. Inclusão de área limite: incluímos o parâmetro `limits` na função `prepare_data` em que o usuário pode escolher limites de área em que as coordenadas sintéticas serão geradas. A função possui uma mensagem de aviso sobre quantos pontos das coordenadas geográficas originais existem fora da área limite especificada pelo usuário;

2. Inclusão da área restrita: incluímos os parâmetros `restricted_area` e `coord_restricted_area` na função `syncoordinates` em que o usuário pode incluir as áreas em que não existem dados de coordenadas geográficas do banco original, para que não haja a geração das coordenadas sintéticas nesses espaços.

Por fim, ressaltamos que o pacote também inclui as funções implementadas do MCMC ([35]) para utilização através do `NIMBLE` ([8], [6], [7]) no `Software R` ([30]).

Capítulo 6

Aplicação ao banco de dados de COVID-19

6.1 Banco de Dados

Para a implementação da metodologia, utilizamos o banco de dados sobre o coronavírus disponibilizado em parceria com a Prefeitura de Montes Claros. A coleta das informações ocorreu entre Março de 2020 e Agosto de 2021. Inicialmente, o banco possuía 82 variáveis e 166.553 observações de notificações sobre indivíduos com suspeita de infecção por COVID-19, causada pelo vírus denominado SARS-CoV-2[1].

Realizamos uma análise exploratória dos dados observando valores ausentes e que não correspondiam como resposta viável a algumas variáveis, por exemplo, algumas idades eram uma sequência aleatória de números e não idades em anos inteiros.

Após essa primeira análise, o banco de dados reduziu para 150.927 observações. Devido ao objetivo da pesquisa, algumas variáveis como sintomas e dados cadastrais não demonstravam relevância para a aplicação porque não possuíam informações espaciais. Portanto, desconsideramos 30 variáveis.

Verificando as variáveis que continham informações sobre comorbidades, observamos duplicidade sobre essa informação. O mesmo tipo de comorbidade aparecia em duas variáveis. Dessa forma foram descartadas as variáveis duplicadas, obtendo um banco de dados com 44 variáveis.

Como o objetivo do trabalho é criar bancos com coordenadas espaciais sintéticas para proteger a confidencialidade dos dados, dessa forma, fez-se necessário transformar as variáveis correspondentes aos endereços em longitude e latitude. Para realizar essa transformação, utilizamos a função `geocode` pertencente ao pacote `ggmap` ([19]) do `Software R` ([30]), e a ferramenta do *Google* denominada *Cloud*[2].

Ao transformar os endereços em coordenadas geográficas, encontramos 670 observações em que não foi possível determinar as longitudes e latitudes. Essas observações

[1] https://www.paho.org/pt/covid19
[2] https://cloud.google.com/

foram codificadas como *NA* e excluídas do banco de dados. Após a obtenção das coordenadas, foi feita uma validação se os pontos criados pertenciam à delimitação do município determinada pelo *Instituto Brasileiro de Geografia e Estatística (IBGE)*[3]. Encontramos observações que excediam os limites do município, as quais foram excluídas do banco. Portanto, ao final dessa etapa, obtivemos 149.982 observações para o banco de dados.

Devido ao banco de dados possuir entradas relativas a suspeitas de COVID-19, é possível que haja indivíduos com duas ou mais entradas no banco. Para verificar a existência de duplicidade de indivíduos, utilizamos as variáveis `sexo`, `nascimento`, `longitude` e `latitude`, e a função `duplicated` pertencente ao pacote `base` do `Software R`. Após rodar essa função em todo o banco de dados, foi retornado que 25.750 observações referiam-se a indivíduos que apareciam mais de uma vez no banco de dados. O critério utilizado para a seleção das entradas duplicadas foi de deixar somente a primeira entrada de cada indivíduo no banco de dados, eliminando assim qualquer viés.

Após essas verificações, obtivemos ao final 123.232 observações. Para a análise do resultado de teste positivo ou negativo para COVID-19, utilizamos a variável `sindrome` que possuía a classificação em: *COVID*, *COVID + H1N1 + Síndrome Gripal*, *COVID + H1N1*, *Gestante/COVID*, *Não Identificadas*, *Síndrome Gripal*, *Síndrome Gripal Inespecífica em Acompanhamento*, e *Síndrome Gripal + H1N1*. Utilizamos essas respostas categóricas para criar a variável binária `resultado` (positivo e negativo), indicando quais pacientes que estavam com COVID-19 no banco de dados.

O próximo passo realizado foi de verificar os limites do contorno do município e da sua área urbana. Um arquivo do tipo *shapefile* contém as coordenadas geográficas que geram o polígono do contorno da cidade. O arquivo *shapefile* do município de Montes Claros foi obtido no site do IBGE, e foram analisadas se todas as coordenadas geográficas do banco estavam contidas dentro do limite municipal. As observações que não pertenciam ao limite municipal foram excluídas, e ao final, o banco de dados ficou com 123.184 observações. Todos esses pontos foram classificados, através das informações disponibilizadas pelo IBGE sobre as divisões de setores censitários, se pertenciam à área rural ou urbana do município, e essa divisão foi incluída na variável `area`, sendo assim o banco de dados final possui 48 variáveis.

[3] https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais.html

6.2 Análise Descritiva

O banco de dados final que será analisado neste trabalho possui 123.184 observações e 48 variáveis. Realizamos a análise para todo o município e também separadamente para a área urbana e rural. Entretanto, pela distribuição dos dados, observamos que as medidas estatísticas são bem próximas quando comparamos o município e suas áreas. Note que, com relação à área rural temos 410 observações, enquanto para a área urbana temos 122.774 observações.

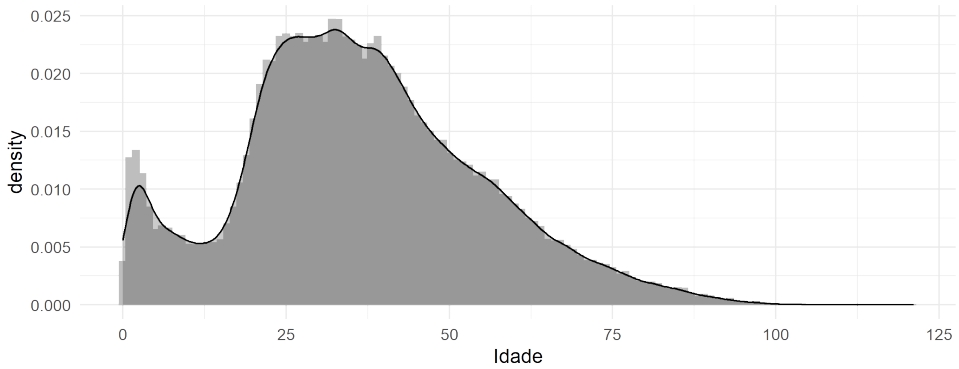
Inicialmente, analisamos a variável `sexo` e, como pode ser observado pela Tabela 6.1, há uma maior predominância de pessoas do sexo feminino no banco de dados. E, notamos que o mesmo efeito ocorre na área rural e na área urbana.

Tabela 6.1: Frequências Absolutas e Relativas da Variável Sexo

	Sexo Feminino		Sexo Masculino	
Área	Freq. Absoluta	Freq. Relativa	Freq. Absoluta	Freq. Relativa
Município	67.604	54,9%	55.580	45,1%
Rural	218	53,2%	192	46,8%
Urbana	67.386	54,9%	55.388	45,1%

Para a variável `idade`, observamos que a idade mínima é zero e a máxima é 121 anos. A idade média dos indivíduos do banco de dados é de aproximadamente 37 anos. As distribuições das idades podem ser verificadas nas Figuras 6.1, 6.2 e 6.3. Note que as distribuições do município e da área urbana são similares e apresentam um formato bimodal, diferentemente da distribuição da área rural.

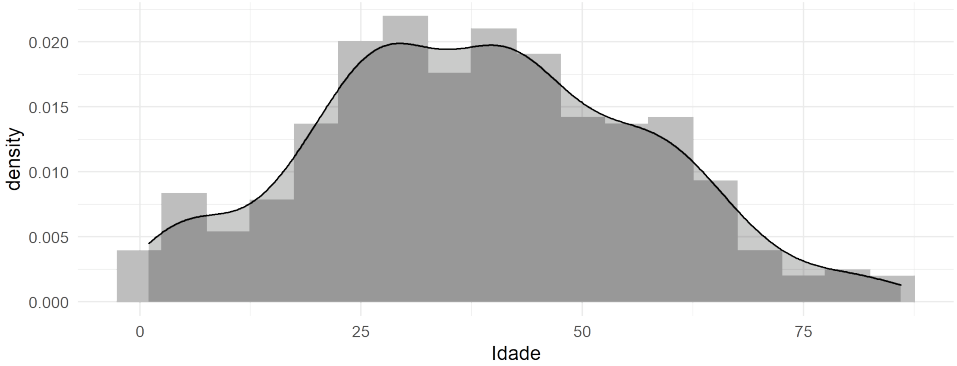
Figura 6.1: Distribuição da Idade para os Dados do Município de Montes Claros.



Fonte: Elaborado pela autora.

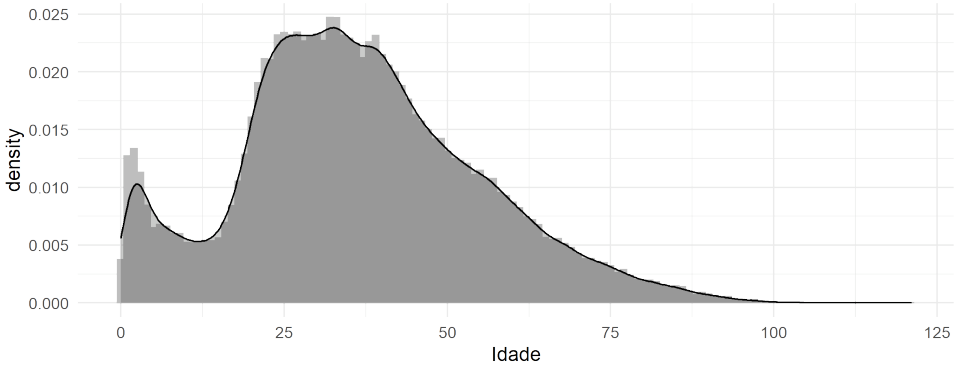
Analisando a variável `resultado`, referente ao resultado do teste diagnóstico para COVID-19 (positivo ou negativo), notamos que a distribuição de casos é similar para o

Figura 6.2: Distribuição da Idade para os Dados da Área Rural do Município de Montes Claros.



Fonte: Elaborado pela autora.

Figura 6.3: Distribuição da Idade para os Dados da Área Urbana do Município de Montes Claros.



Fonte: Elaborado pela autora.

município e para a área urbana, sendo que a maior parte de casos do banco de dados são positivos para a doença. Já na área rural do município, observamos que a porcentagem de casos positivos é maior do que quando comparado com o município e a área urbana como pode ser visto na Tabela 6.2.

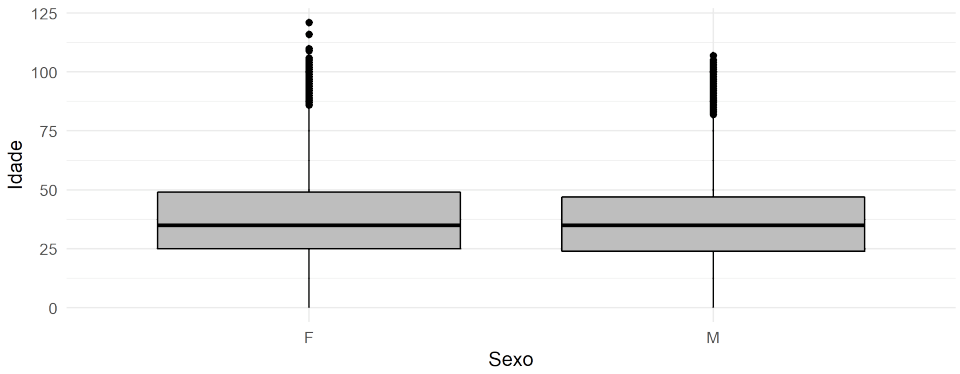
Tabela 6.2: Frequências Absolutas e Relativas da Variável Resultado

	Resultado Negativo		Resultado Positivo	
Área	Freq. Absoluta	Freq. Relativa	Freq. Absoluta	Freq. Relativa
Município	21.794	18%	101.390	82%
Rural	51	12%	359	88%
Urbana	21.743	18%	101.031	82%

Analisando as variáveis em conjunto, observamos que a idade média para toda a cidade e para a área urbana é de aproximadamente 37 anos para as mulheres e de 36 anos para os homens, como pode ser visto nas Figuras 6.4 e 6.6. Já para a área rural do

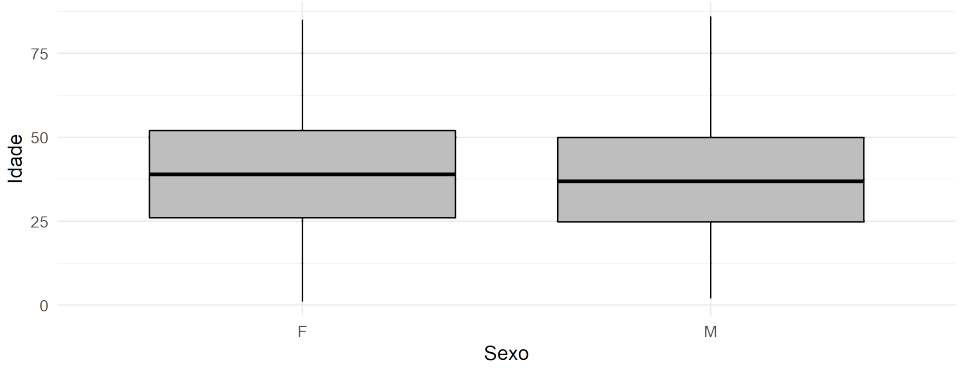
município, a idade média é de aproximadamente 38 anos tanto para as mulheres quanto para os homens (Figura 6.5).

Figura 6.4: Box Plots da Distribuição Conjunta da Idade e do Sexo para os Dados do Município de Montes Claros.



Fonte: Elaborado pela autora.

Figura 6.5: Box Plots da Distribuição Conjunta da Idade e do Sexo para os Dados da Área Rural do Município de Montes Claros.

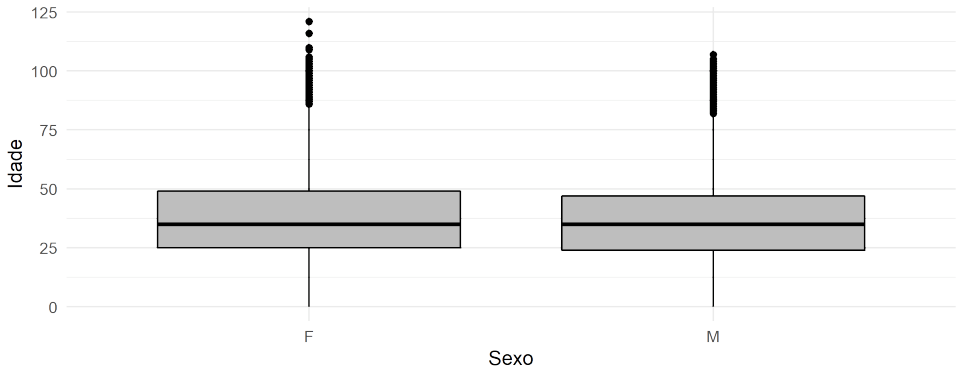


Fonte: Elaborado pela autora.

Na Figura 6.7, temos a pirâmide etária para os dados do município com o destaque dos resultados positivos para a doença. Essa figura reflete a porcentagem anteriormente descrita de que 82% dos dados são referentes aos casos positivos. Além disso, a figura demonstra a distribuição de casos positivos em cada idade em anos inteiros.

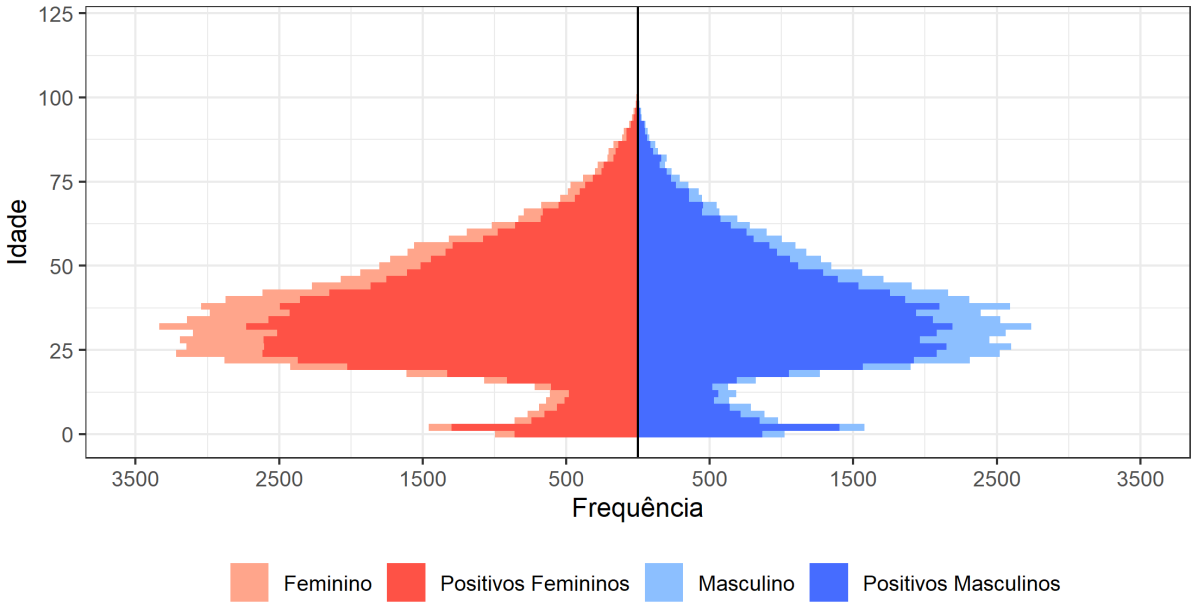
Ao realizar a análise descritiva dos dados, observamos uma similaridade entre o município completo e a área urbana. Notamos também que a quantidade de observações pertencentes à área rural corresponde a uma porcentagem bastante baixa de todo o banco de dados. Portanto, devido a questões de confidencialidade, decidimos continuar as análises somente para a área urbana do município de Montes Claros - MG.

Figura 6.6: Box Plots da Distribuição Conjunta da Idade e do Sexo para os Dados da Área Urbana do Município de Montes Claros.



Fonte: Elaborado pela autora.

Figura 6.7: Pirâmide Etária com Destaque nos Resultados Positivos de COVID-19 para os Dados do Município de Montes Claros.



Fonte: Elaborado pela autora.

6.3 Análise Espacial

Para essa análise, utilizamos algumas ferramentas para a divulgação segura das coordenadas geográficas do banco de dados. Incluímos ruído aleatório e omissões de localizações para alguns indivíduos.

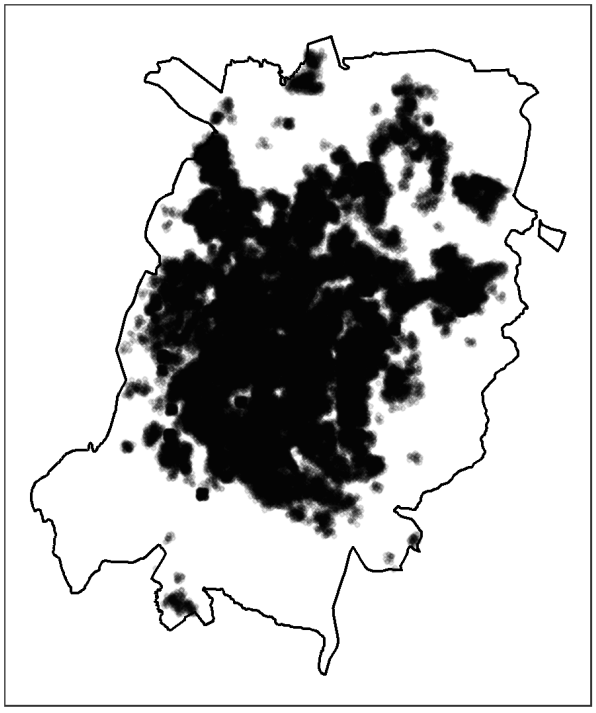
O ruído cria uma distorção, de tamanho escolhido pelo pesquisador, em cada ponto no gráfico. Uma das vantagens do uso do ruído aleatório em coordenadas geográficas é a identificação de pontos sobrepostos (localizações de várias observações no mesmo ponto do gráfico). Entretanto, uma desvantagem é que somente a utilização dessa ferramenta não é suficiente para evitar a identificação de observações mais afastadas ao se divulgar os mapas.

Devido à limitação do ruído aleatório para proteção dos indivíduos no banco, utilizamos outro método para identificar pontos que estão mais isolados de outras observações, e portanto, teriam maior risco de terem sua identidade revelada com a publicação dos mapas. Para isso, usamos a média das distâncias para as k observações mais próximas de cada coordenada do mapa. As coordenadas com as maiores médias serão consideradas mais isoladas. Para fazer isso de maneira computacionalmente eficiente, usamos a função `knn.dist` do pacote `FNN` ([1]) do `Software R`. A função implementa o método *k-nearest neighbours* (em tradução livre, k vizinhos próximos) que faz uso de algoritmos de busca pelos vizinhos próximos daquela coordenada geográfica. Utilizamos $k = 10$ para o cálculo da média dos 10 vizinhos mais próximos de cada observação, e assim, após esses cálculos, utilizamos o quantil de 99.5% como ponto de corte para omitir nos gráficos as coordenadas geográficas cujas médias de distâncias estão acima desse valor. Escolhemos esse quantil com o intuito de minimizar a omissão das coordenadas, com isso omitimos os 857 pontos mais isolados do mapa.

Para melhor visualização das coordenadas geográficas, utilizamos o recurso de transparência de cores para a confecção dos gráficos. A transparência melhora a visualização da quantidade de pontos juntos no mesmo espaço. Portanto as cores nos pontos e nas legendas das Figuras 6.9 e 6.10 possuem a aplicação da transparência.

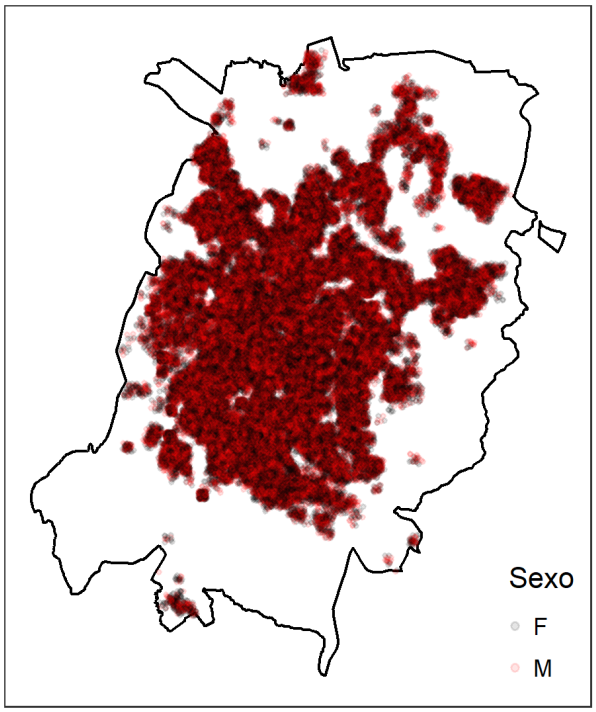
Os mapas com todas as coordenadas originais, com e sem o ruído aleatório, e com as observações mais isoladas que serão omitidas, não foram incluídos nesse texto por questões de confidencialidade. Após aplicar o ruído e omitir as observações mais isoladas, o padrão espacial de distribuição dos dados na área urbana do município pode ser visto na Figura 6.8. Já as distribuições espaciais das variáveis `sexo` e `resultado` são apresentadas nas Figuras 6.9 e 6.10, respectivamente.

Figura 6.8: Área Urbana do Município de Montes Claros com Ruído Aleatório e Omissão nas Coordenadas Geográficas.



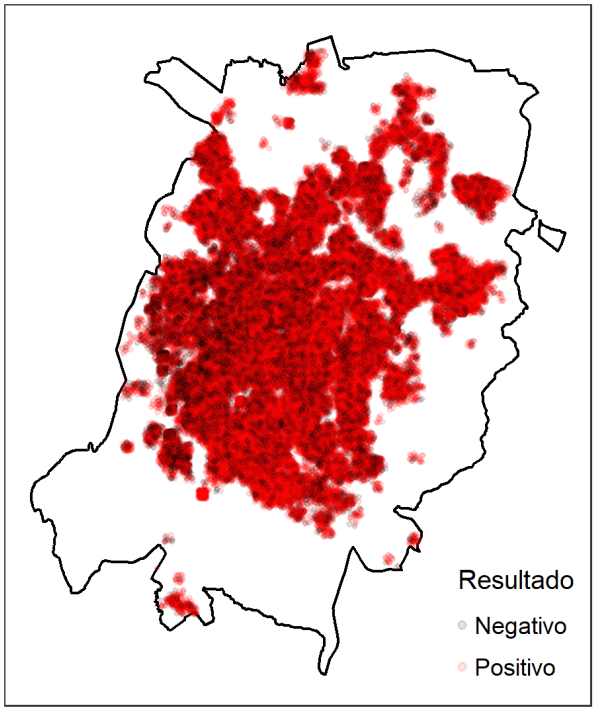
Fonte: Elaborado pela autora.

Figura 6.9: Área Urbana do Município de Montes Claros Dividida pelo Sexo com Ruído Aleatório e Omissão nas Coordenadas Geográficas.



Fonte: Elaborado pela autora.

Figura 6.10: Área Urbana do Município de Montes Claros Dividida pelo Resultado do Teste com Ruído Aleatório e Omissão nas Coordenadas Geográficas.



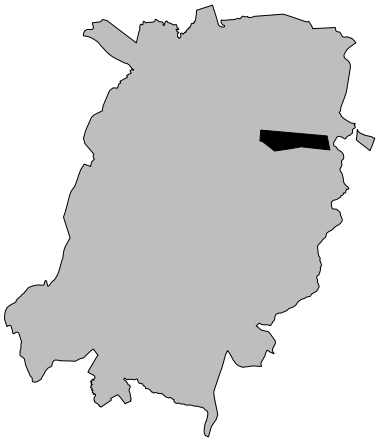
Fonte: Elaborado pela autora.

6.4 Aplicação

Para realizarmos a aplicação deste trabalho ao banco de dados de COVID-19 do município de Montes Claros - MG, primeiramente escolhemos através do *Google Maps*[4] o espaço que denominaríamos de área restrita. Sendo assim optamos por escolher o Aeroporto de Montes Claros - Mário Ribeiro ([15]).

Após essa escolha, retiramos do *Google Maps* as longitudes e latitudes que formam o polígono dessa área, como pode ser verificado na Figura 6.11.

Figura 6.11: Área urbana do município de Montes Claros com destaque na área restrita.



Fonte: Elaborado pela autora.

Como mencionado anteriormente, a utilização de polígonos de áreas pelo `Software R` ([30]) requer que o primeiro par de coordenadas seja igual ao último par de coordenadas de um determinado polígono, uma vez que os mesmos precisam ser fechados para o funcionamento através do *software*.

Após a escolha da área restrita, realizamos a escolha das variáveis do banco de dados de COVID-19 que iríamos considerar para a aplicação. Dessa forma, as variáveis discretas utilizadas foram: Y = `resultado` que corresponde ao resultado do teste diagnóstico para COVID-19 (positivo ou negativo), X_1 = `sexo` feminino ou masculino, e X_2 = `comorbidade` referente à presença de comorbidades (sim ou não); e a variável contínua utilizada foi a Z = `idade` que corresponde à idade do indivíduo, além das coordenadas originais (longitude e latitude).

Devido ao banco de dados possuir 122.774 observações decidimos realizar a aplicação somente em 20.000 observações do banco de dados. Entretanto, para realizar essa mu-

[4] https://www.google.com.br/maps/

dança de tamanho dos dados, utilizamos as proporções de combinações existentes no banco de dados original para amostrar as 20.000 observações que usaríamos na aplicação.

Optamos por reduzir o banco de dados para a aplicação devido ao comprometimento de visualização dos resultados através das figuras dos dados sintéticos.

É importante destacar que, do modo como os algoritmos que verificam as interseções entre as áreas restritas e as células da grade foram implementados, se houver mais de uma área restrita dentro de uma mesma célula da grade, sua proporção de área será retornada considerando todas as áreas de restrições que estão dentro da célula.

Para a aplicação do modelo aos dados, consideramos somente uma área restrita (aeroporto), entretanto ressaltamos que a função comporta a utilização de mais áreas de restrições. Note que, quanto maior for a área restrita, maior será sua cobertura sobre as células da grade, e portanto, maior será o impacto na geração das coordenadas sintéticas.

Por fim, optamos por aplicar dois tamanhos de grade, com $grid = 20 \times 20$ e $grid = 30 \times 30$, sendo assim $G = 400$ e $G = 900$, respectivamente. Observe que, por questões de confidencialidade, todos os gráficos que utilizam as coordenadas geográficas originais possuem tratamentos de ruído aleatório e omissão das observações mais isoladas.

Na aplicação do MCMC, utilizamos a seguinte configuração nas funções:

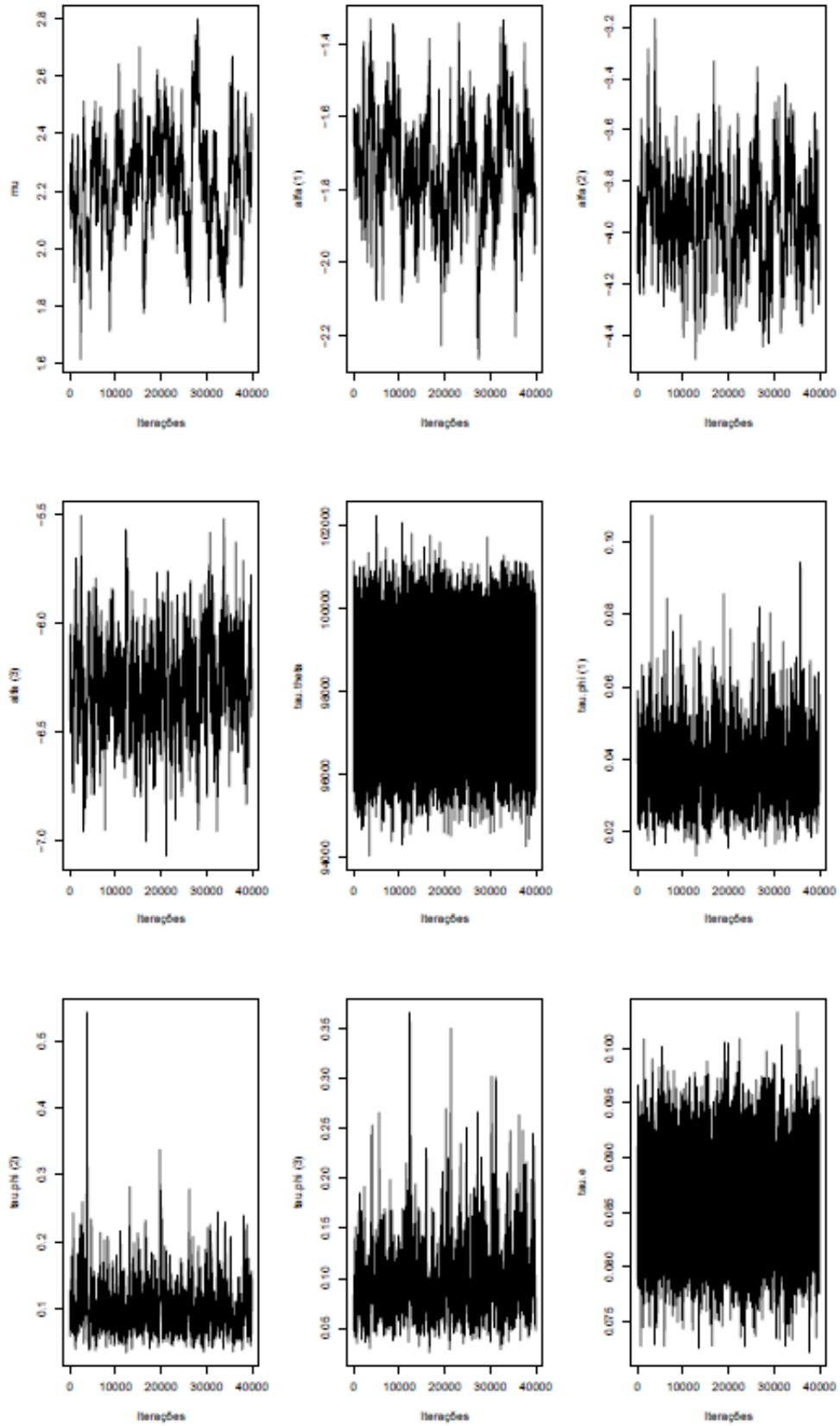
```
syn_mcmc(dataset = dados, coord = coordenadas,
         grid = 20, S = 50000, burn = 10000,
         continuous = 4, spatial_beta = FALSE
         return_parameters = TRUE)

syn_mcmc(dataset = dados, coord = coordenadas,
         grid = 30, S = 50000, burn = 10000,
         continuous = 4, spatial_beta = FALSE
         return_parameters = TRUE)
```

Nas Figuras 6.12 e 6.13 podemos observar a convergência dos parâmetros amostrados através do *MCMC* [16] sem o período de aquecimento. Notamos que existe a convergência para todos os parâmetros amostrados do modelo.

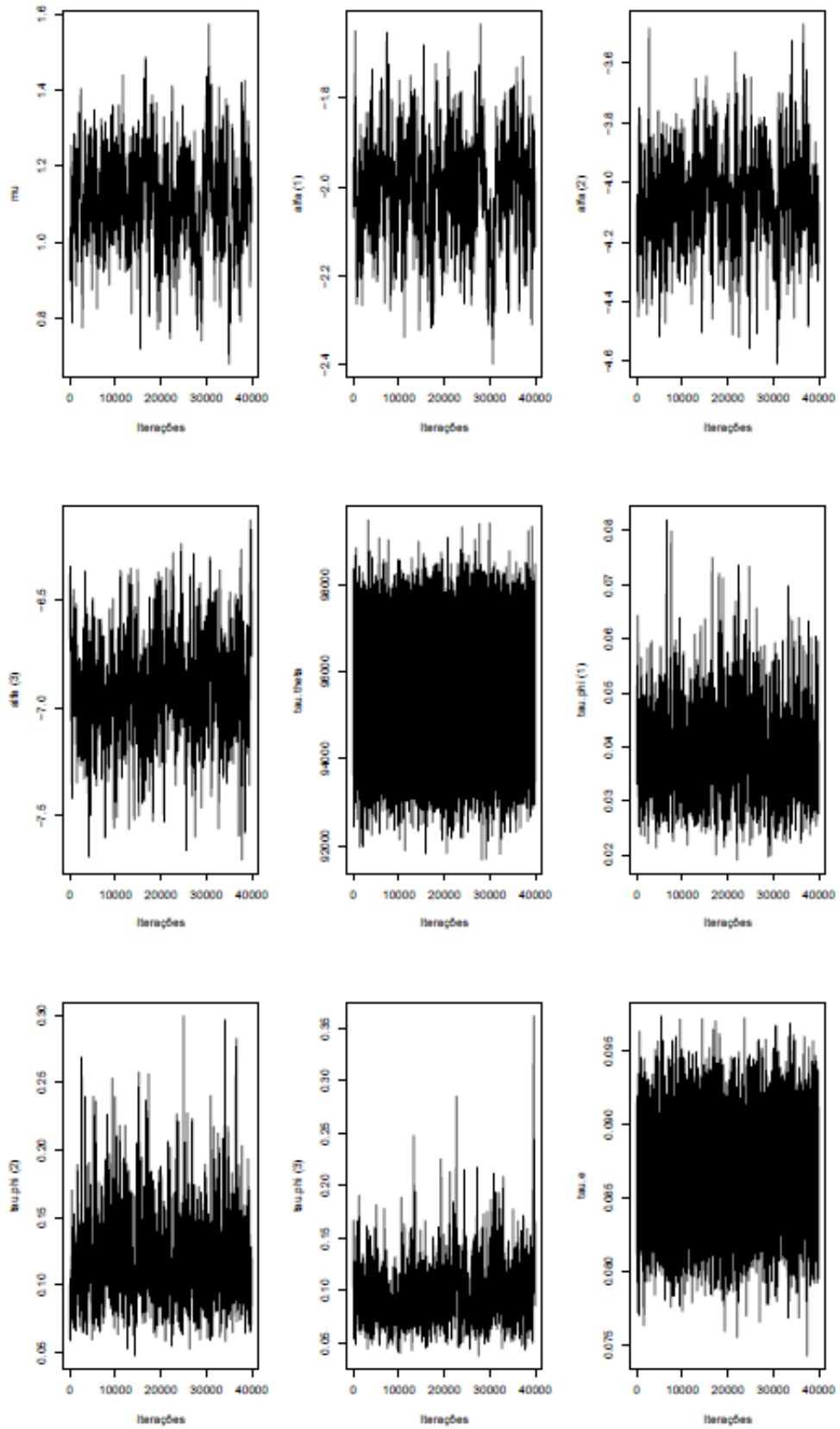
Também analisamos a convergência de λ, entretanto como o parâmetro é estimado pelo modelo em cada célula da grade e em cada uma das combinações, escolhemos a célula da grade número 210 arbitrariamente. Assim, pelas Figuras 6.14 e 6.15 observamos que existe a convergência para o λ em todas as combinações, tanto para a grade $G = 400$ quanto para $G = 900$. Note que o período de aquecimento foi retirado para melhor visualização da convergência.

Figura 6.12: Traceplots dos parâmetros gerados através do modelo com S = 50.000, período de aquecimento de 10.000 e G = 400.



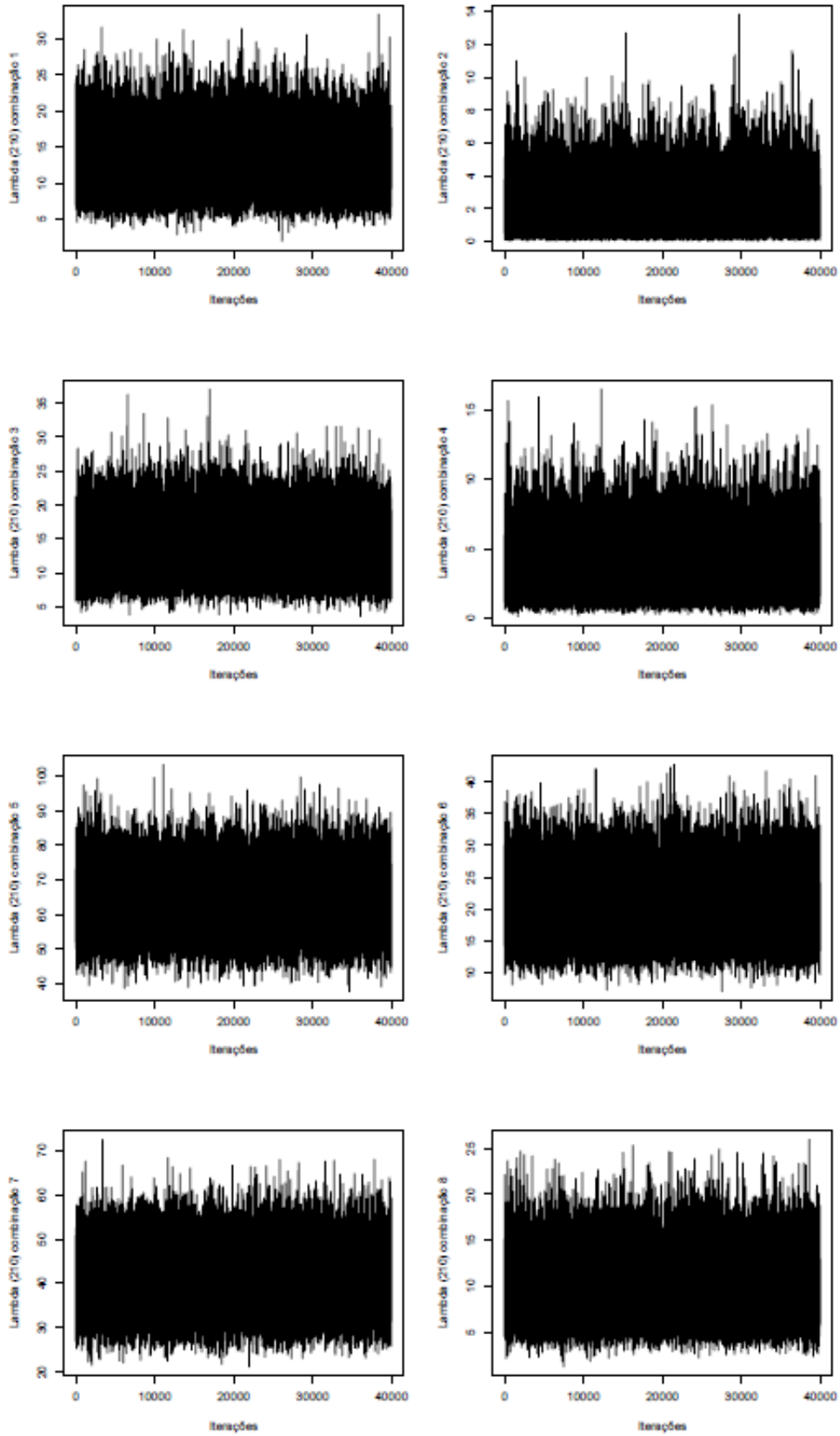
Fonte: Elaborado pela autora.

Figura 6.13: Traceplots dos parâmetros gerados através do modelo com S = 50.000, período de aquecimento de 10.000 e G = 900.



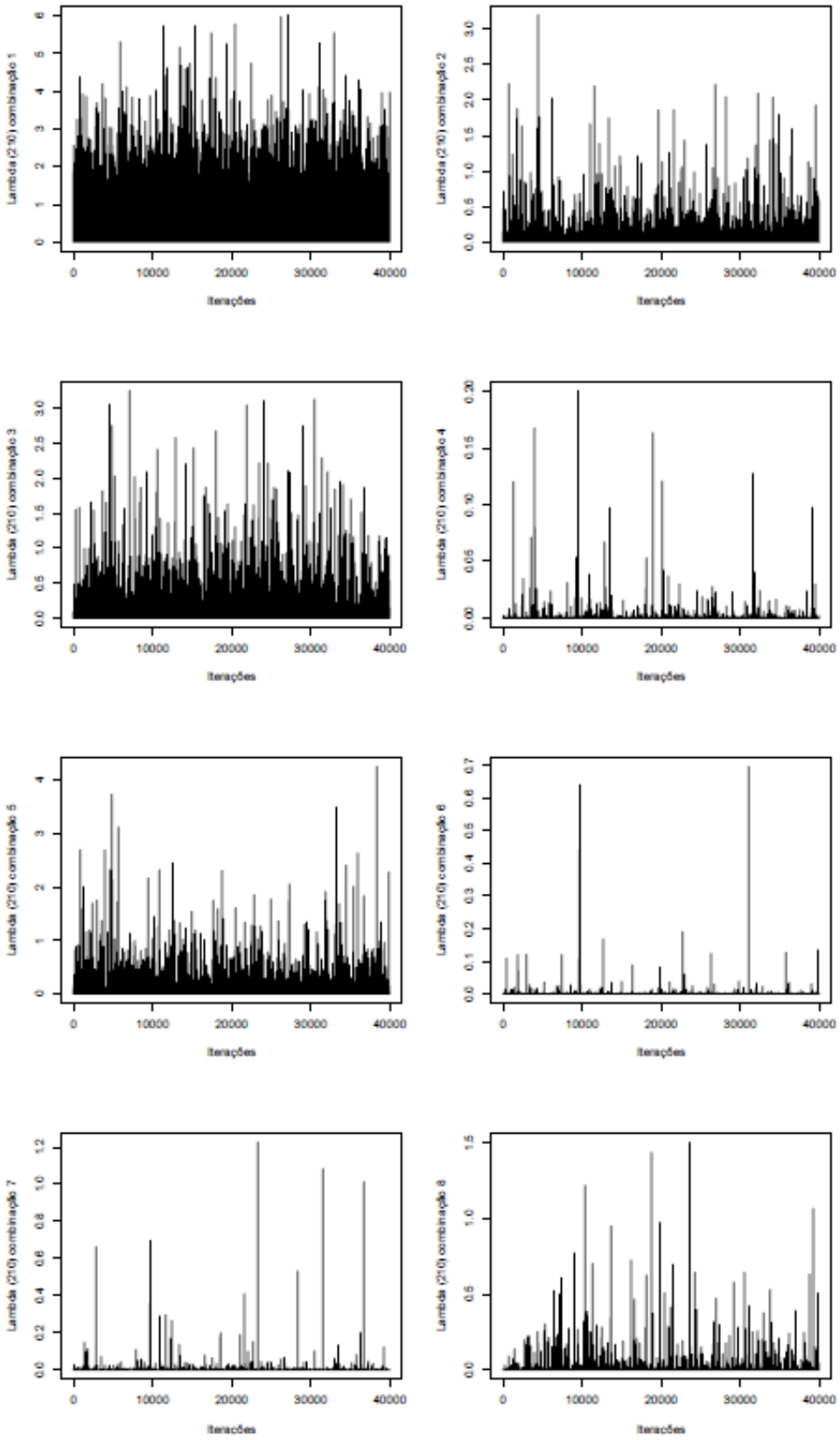
Fonte: Elaborado pela autora.

Figura 6.14: Traceplots do parâmetro λ das combinações gerado através do modelo com S = 50.000, desconsiderando o período de aquecimento de 10.000 e G = 400.



Fonte: Elaborado pela autora.

Figura 6.15: Traceplots do parâmetro λ das combinações gerado através do modelo com S = 50.000, desconsiderando o período de aquecimento de 10.000 e G = 900.



Fonte: Elaborado pela autora.

6.5 Resultados

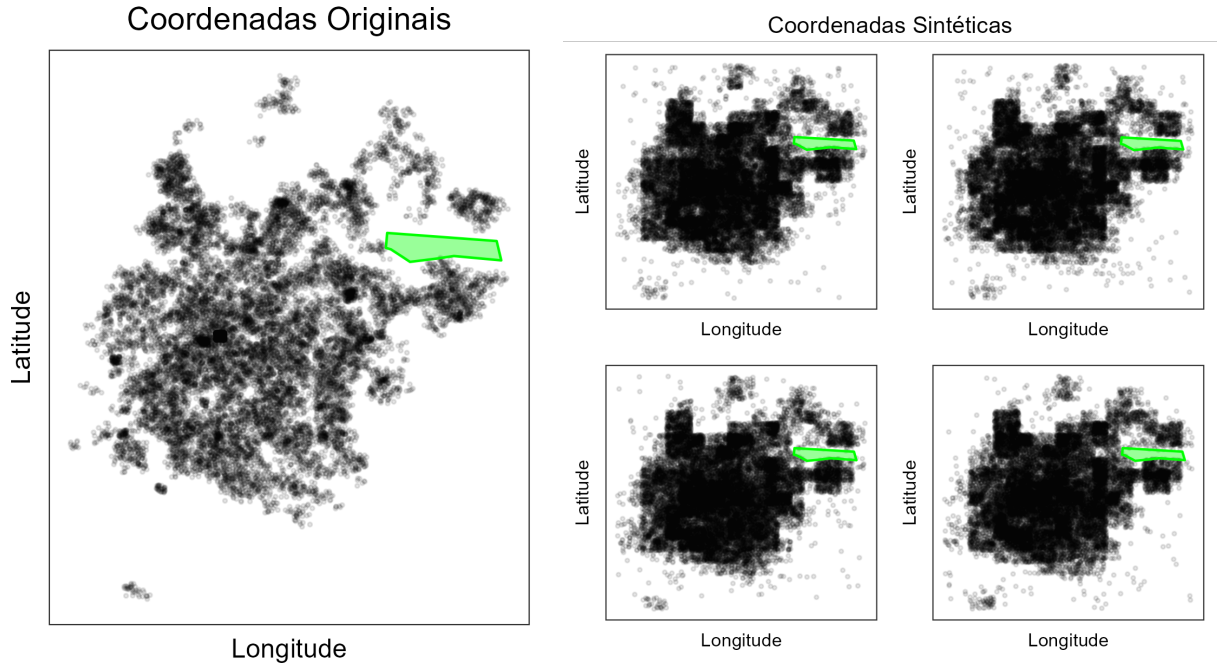
Nessa seção, iniciaremos mostrando os resultados dos dados sintéticos para $G = 400$, utilizamos a função com a seguinte configuração:

```
syncoordinates(dataset = dados, coord = coordenadas,
               grid = 20, continuous = 4,
               restricted_area = TRUE,
               coord_restricted_area = coordenadas_area_restrita,
               list_mcmc = mcmc_dados, n.syn = 4)
```

Na Figura 6.16, podemos observar que o padrão de distribuição das coordenadas originais se manteve nos quatro banco de dados sintéticos para as coordenadas sintéticas. E que, dentro da área restrita, não existe a geração de coordenadas geográficas sintéticas.

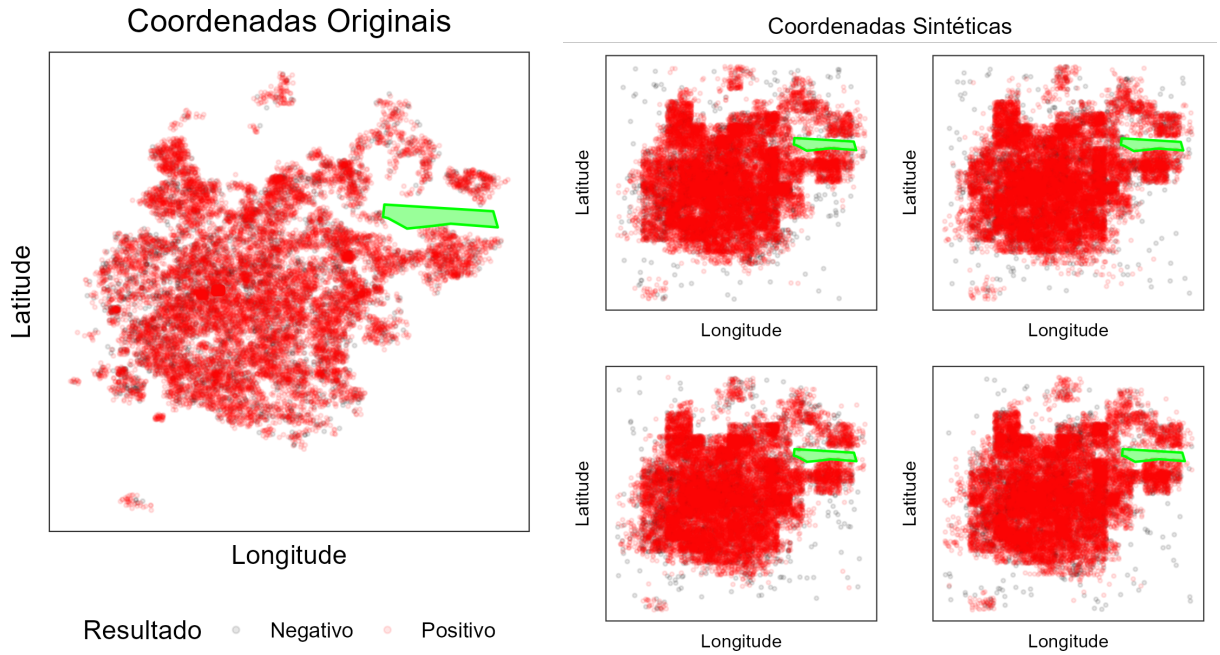
Quando analisamos por cada uma das variáveis, percebemos também que o padrão das coordenadas originais das variáveis originais se repete nos valores para as coordenadas sintéticas.

Figura 6.16: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG com a inclusão de uma área de restrição ($G = 400$).



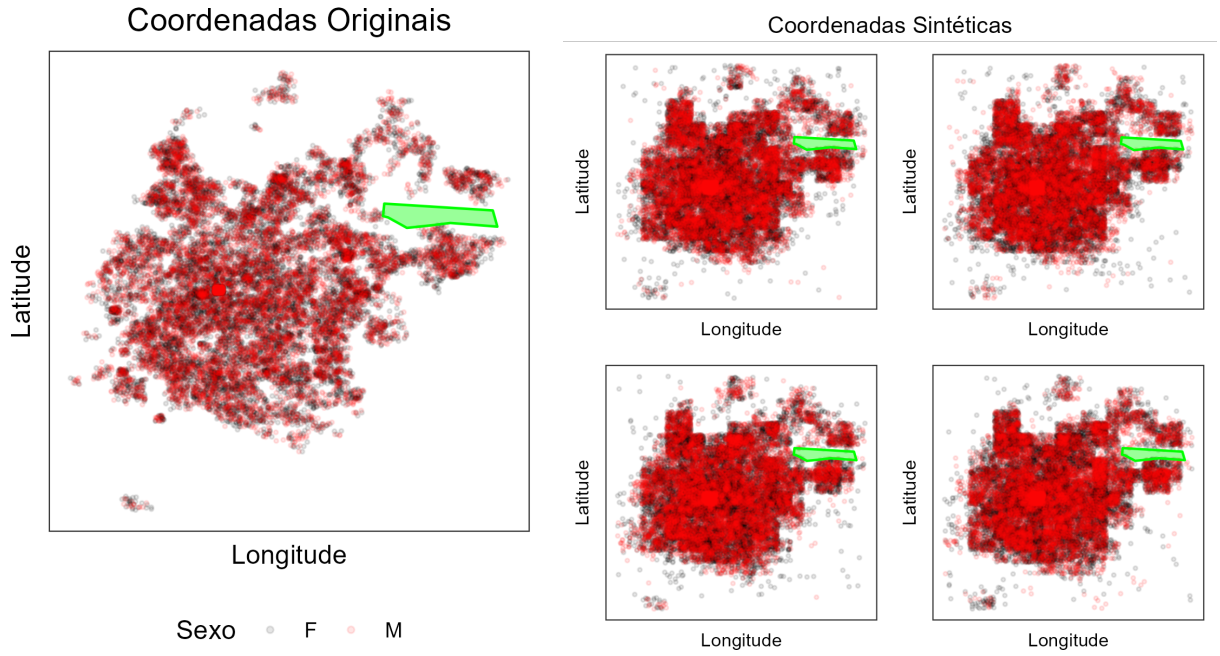
Fonte: Elaborado pela autora.

Figura 6.17: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `resultado` com a inclusão de uma área de restrição ($G = 400$).



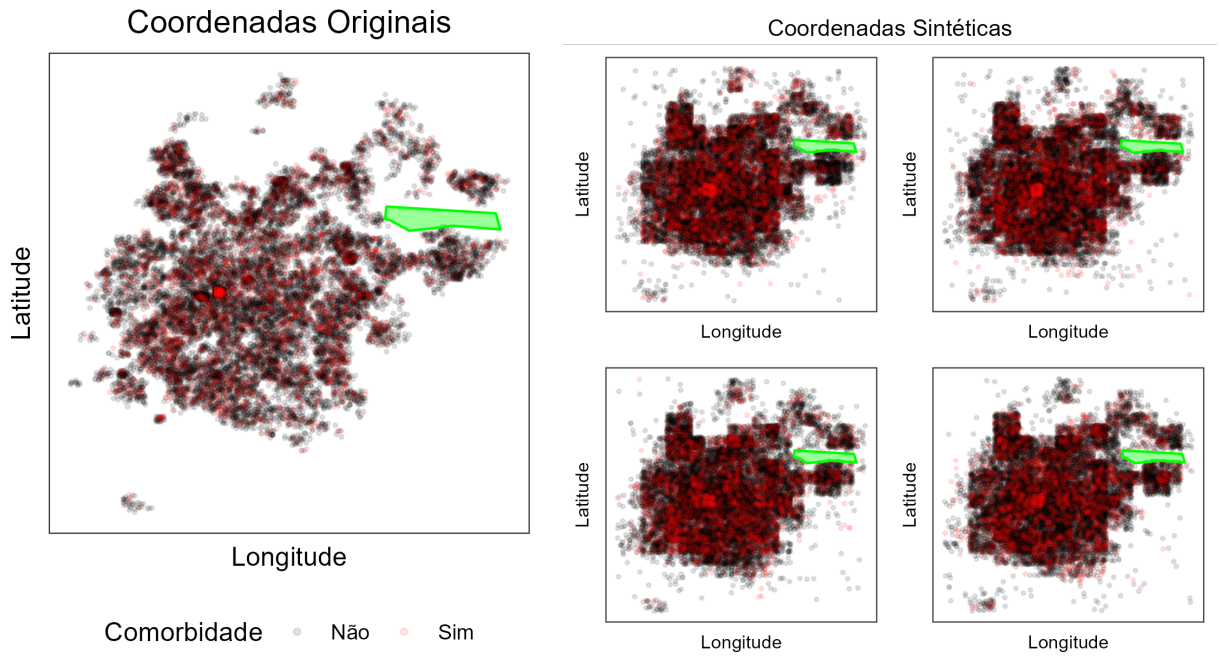
Fonte: Elaborado pela autora.

Figura 6.18: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `sexo` com a inclusão de uma área de restrição ($G = 400$).



Fonte: Elaborado pela autora.

Figura 6.19: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `comorbidade` com a inclusão de uma área de restrição ($G = 400$).



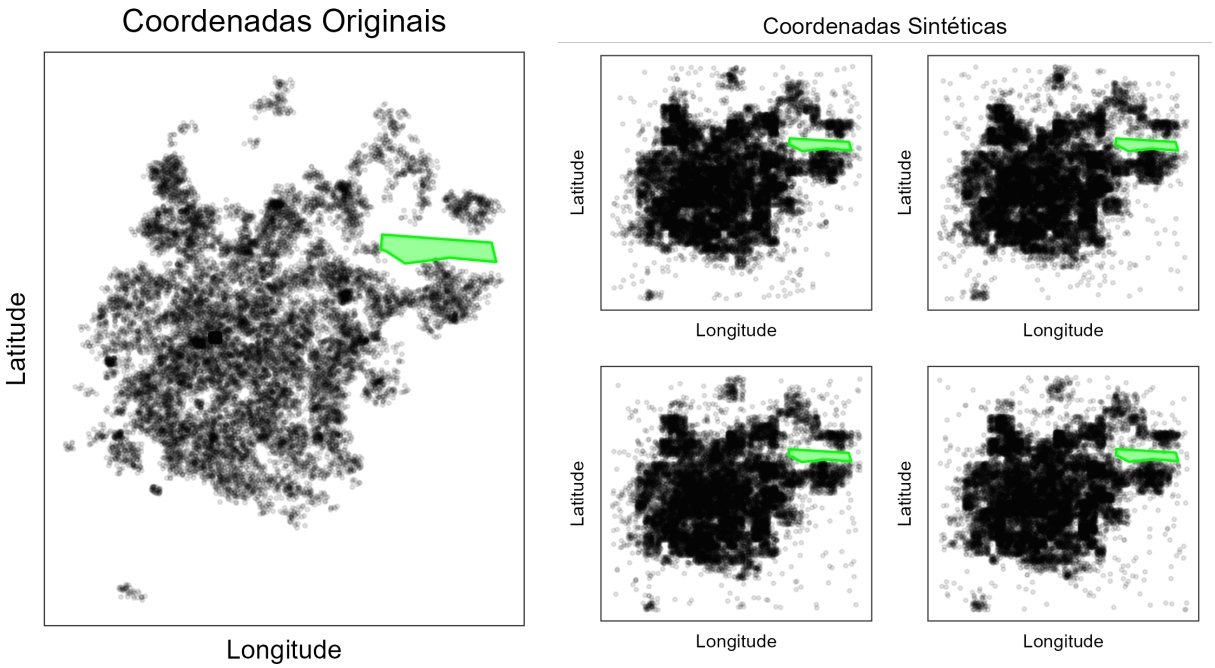
Fonte: Elaborado pela autora.

A seguir, apresentaremos os resultados para a aplicação com $G = 900$, ou seja, $grid = 30 \times 30$, em que utilizamos a função com a seguinte configuração:

```
syncoordinates(dataset = dados, coord = coordenadas,
               grid = 30, continuous = 4,
               restricted_area = TRUE,
               coord_restricted_area = coordenadas_area_restrita,
               list_mcmc = mcmc_dados, n.syn = 4)
```

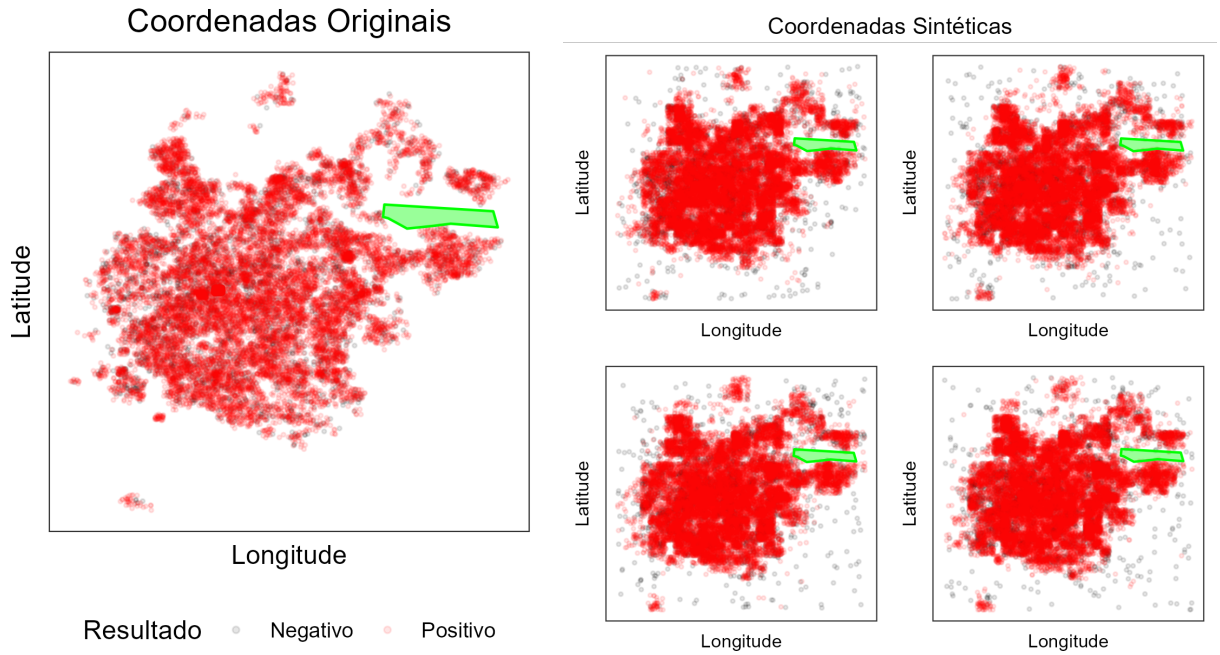
Nas Figuras 6.20, 6.21, 6.22 e 6.23 observamos que o padrão das coordenadas originais são preservados pelas coordenadas sintéticas, também nas divisões pelas variáveis discretas utilizadas.

Figura 6.20: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG com a inclusão de uma área de restrição ($G = 900$).



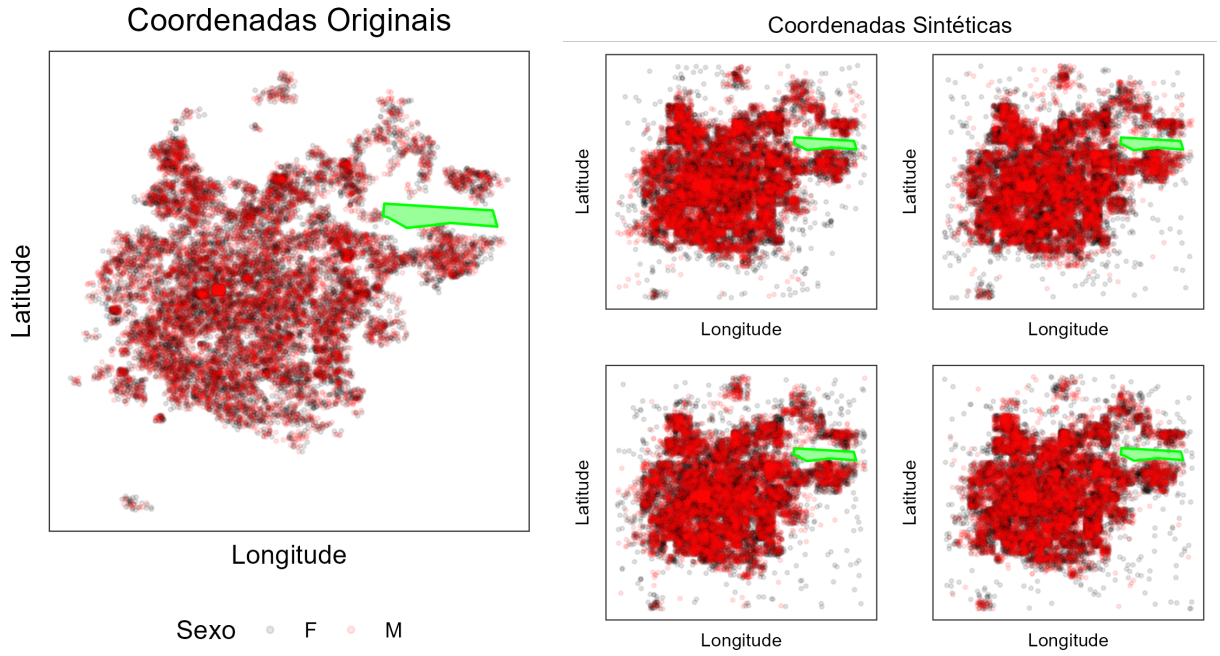
Fonte: Elaborado pela autora.

Figura 6.21: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `resultado` com a inclusão de uma área de restrição ($G = 900$).



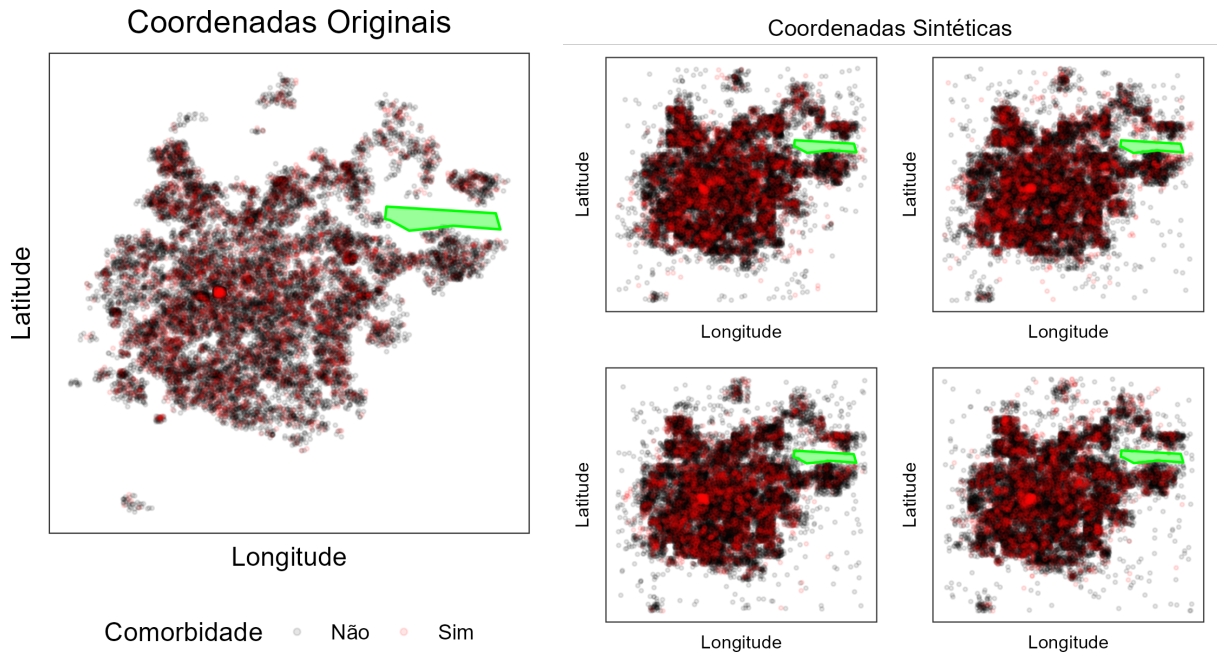
Fonte: Elaborado pela autora.

Figura 6.22: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `sexo` com a inclusão de uma área de restrição ($G = 900$).



Fonte: Elaborado pela autora.

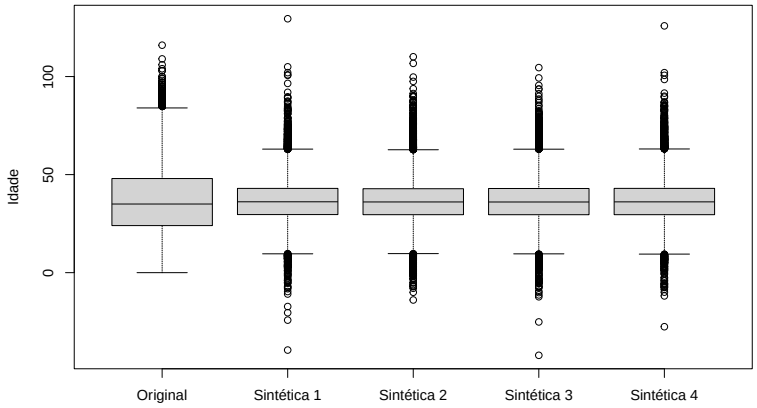
Figura 6.23: Coordenadas originais e coordenadas sintéticas geradas para o banco de dados do município de Montes Claros - MG para a variável `comorbidade` com a inclusão de uma área de restrição ($G = 900$).



Fonte: Elaborado pela autora.

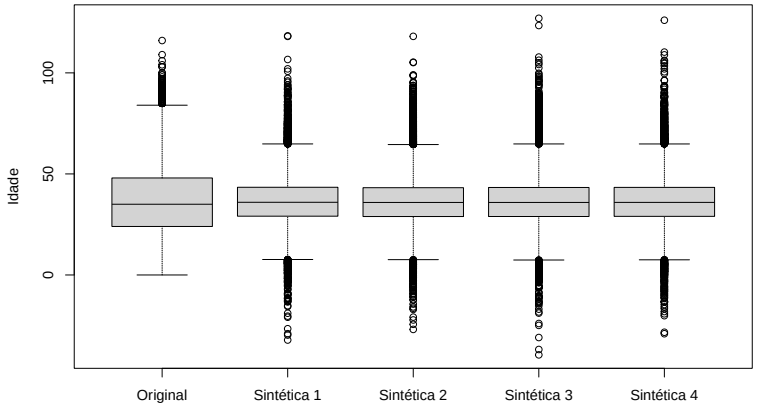
Por fim, para analisarmos a variável contínua da aplicação, podemos observar pelas Figuras 6.24 e 6.25 a distribuição dos valores originais e os valores sintéticos, para as aplicações com $G = 400$ e $G = 900$, para uma área restrita. Para a `idade` notamos que a média é aproximadamente 37 anos, a idade mínima e máxima são 0 ano e 116 anos, respectivamente.

Figura 6.24: Box-plots da variável contínua `idade` dos dados originais e dos dados sintéticos com uma área restrita ($G = 400$).



Fonte: Elaborado pela autora.

Figura 6.25: Box-plots da variável contínua `idade` dos dados originais e dos dados sintéticos com uma área restrita ($G = 900$).



Fonte: Elaborado pela autora.

Capítulo 7

Discussões

Neste trabalho, partimos dos modelos desenvolvidos por [26] e [25], e incluímos a possibilidade de delimitar regiões onde não existem habitações de indivíduos. Tais regiões foram denominadas áreas restritas, como parques, praças, aeroportos, lagoas ou áreas de preservação ambiental. Com essa inclusão, observamos que as coordenadas sintéticas se aproximam cada vez mais da reprodução das coordenadas originais, entretanto sem a divulgação das coordenadas originais.

Exemplificamos os impactos na metodologia após essas alterações, aplicando o método em um banco de dados simulado e avaliando seus resultados. Nesta exemplificação, verificamos tanto o modelo proposto por [26], quanto a atualização do modelo proposto por [25]. Conseguimos comprovar que a metodologia não realizava um tratamento das áreas inabitáveis, dessa forma percebemos a importância da inclusão das áreas de restrições espaciais na geração de coordenadas sintéticas.

Realizamos a implementação do modelo no banco de dados de COVID-19 no município de Montes Claros - MG. Observamos nessa aplicação, a importância da inclusão de áreas de restrição para geração de coordenadas sintéticas. Ainda trabalhamos com tamanhos diferentes da grade, $G = 400$ e $G = 900$, e embora tenhamos refinado o tamanho da célula da grade devido à grande quantidade de observações, não perdemos o padrão entre as coordenadas originais e as sintéticas.

Por fim, para trabalhos futuros, gostaríamos de inserir no modelo a possibilidade de escolha de tamanhos de células da grade irregulares, porque atualmente o modelo assume o mesmo tamanho para todas as células da grade.

Referências

[1] Sham Kakadet Alina Beygelzimer, Sunil Arya John Langford (cover tree library), and Shengqiao Li David Mount (ANN library 1.1.2 for the kd-tree approach). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2019. Version 1.1.3.1.

[2] Sudipto Banerjee, Bradley Carlin, and Alan Gelfand. Hierarchical modeling and analysis of spatial data. *New York,NY: Chapman and Hall/CRC*, 101, 01 2004.

[3] Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregression. *Biometrika, Oxford University Press, Biometrika Trust*, 82(4):733–746, 1995.

[4] Ming-Hui Chen, Qi-Man Shao, and Joseph G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag, 2000.

[5] Lawrence Cox, Alan Karr, and Satkartar Kinney. Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act [with discussions]. *International Statistical Review / Revue Internationale de Statistique*, 79:160–199, 08 2011.

[6] Perry de Valpine, Christopher Paciorek, Daniel Turek, Nick Michaud, Cliff Anderson-Bergman, Fritz Obermeyer, Claudia Wehrhahn Cortes, Abel Rodrìguez, Duncan Temple Lang, and Sally Paganin. *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*, 2023. R package version 1.0.1.

[7] Perry de Valpine, Christopher Paciorek, Daniel Turek, Nick Michaud, Cliff Anderson-Bergman, Fritz Obermeyer, Claudia Wehrhahn Cortes, Abel Rodrìguez, Duncan Temple Lang, and Sally Paganin. *NIMBLE User Manual*, 2023. R package manual version 1.0.1.

[8] Perry de Valpine, Daniel Turek, Christopher Paciorek, Cliff Anderson-Bergman, Duncan Temple Lang, and Ras Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413, 2017.

[9] George Duncan, Sallie Keller-McNulty, and Lynne Stokes. Disclosure risk vs. data utility: The r-u confidentiality map. *Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico*, 01 2001.

[10] D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Boca Raton, FL: Chapman and Hall/CRC Press, 1997.

[11] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, (410):398–409, 06 1990.

[12] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[13] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics), Wiley*, 41(2):337–348, 1992.

[14] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall., 1996.

[15] Google Maps. Aeroporto de Montes Claros - Mário Ribeiro. *Montes Claros, MG. [acessado em 25-Junho-2023]*, 2023.

[16] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[17] HHS, U.S. Department of Health and Human Services. Summary of the health insurance portability and accountability act privacy rule. *[acessado em 18-Junho-2023]*, 2000.

[18] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul Wolf. *Wiley Series in Survey Methodology*, pages 287–288. 07 2012.

[19] David Kahle and Hadley Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[20] Alan F. Karr and Jerome P. Reiter. Using statistics to protect privacy. *Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press*, pages 276–295, 2014.

[21] Satkartar Kinney, Alan Karr, and Joe Gonzalez, Jr. Data confidentiality: The next five years summary and guide to papers. *Journal of Privacy and Confidentiality*, 1:125–134, 01 2010.

[22] Roderick J. A. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9(2):407–426, 1993.

[23] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[24] R Neal. Slice sampling. *The Annals of Statistics*, 31:705–767, 01 2003.

[25] Letícia S. Nunes. Métodos de simulação de dados geográficos sintéticos para bases confidenciais. Dissertação de mestrado, Departamento de Estatística - Universidade Federal de Minas Gerais, 2018.

[26] Thais Paiva, Avishek Chakraborty, Jerry Reiter, and Alan Gelfand. Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine*, 33(11):1928–1945, 2014.

[27] Carlos Daniel Paulino, M. Antónia Amaral Turkman, Bento Murteira, and Giovani L. Silva. *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa, 2a edition, 2018.

[28] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018.

[29] Edzer Pebesma and Roger Bivand. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://r-spatial.org/book/, 2023.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[31] J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189, 2003.

[32] Jerome P. Reiter. Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly*, 76(1):163–181, 02 2012.

[33] Donald B. Rubin. Multiple imputation for nonresponse in surveys. *Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley*, 1987.

[34] Donald B. Rubin. Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

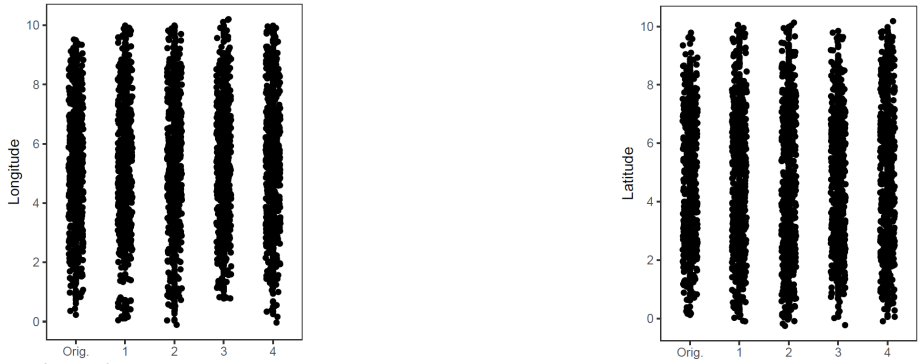
[35] Camila A. de S. P. Teixeira. Implementação de métodos estatísticos para preservação de sigilo de bases de dados confidenciais via nimble. Monografia, Departamento de Estatística - Universidade Federal de Minas Gerais, 2023.

[36] Hao Wang and Jerome P. Reiter. Multiple imputation for sharing precise geographies in public use data. *The Annals of Applied Statistics*, 6(1):229–252, 03 2012.

[37] Leon Willenborg and Ton de Waal. *Disclosure Risks for Microdata*, pages 39–70. Elements of Statistical Disclosure Control. Springer New York. New York, NY, 2001.

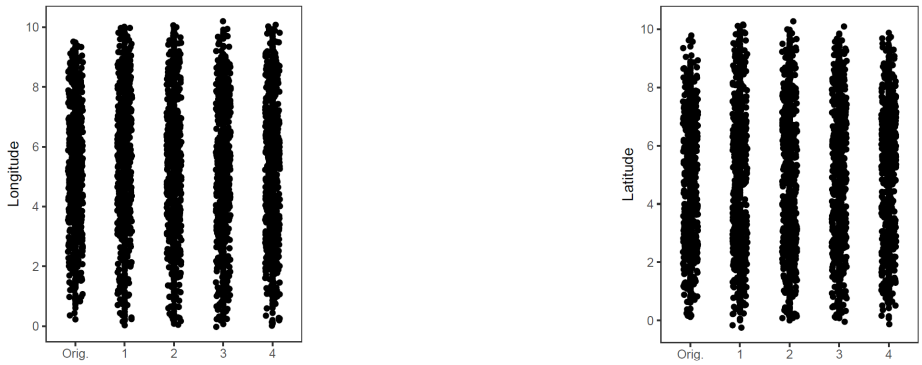
Apêndice A

Figura A.1: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 1 ($grid = 10 \times 10$, com área restrita, com variável contínua).



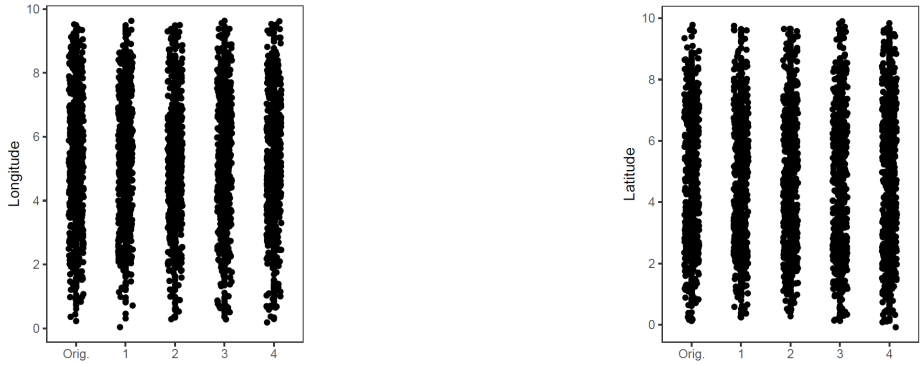
Fonte: Elaborado pela autora.

Figura A.2: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 2 ($grid = 10 \times 10$, com área restrita, sem variável contínua).



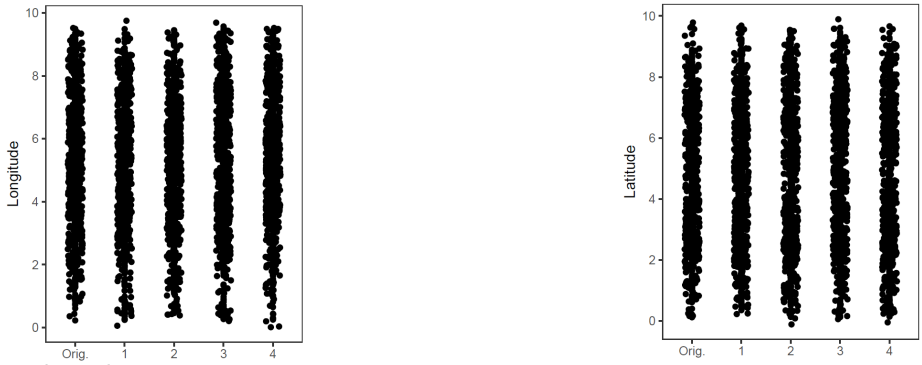
Fonte: Elaborado pela autora.

Figura A.3: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 3 ($grid = 10 \times 10$, sem área restrita, com variável contínua).



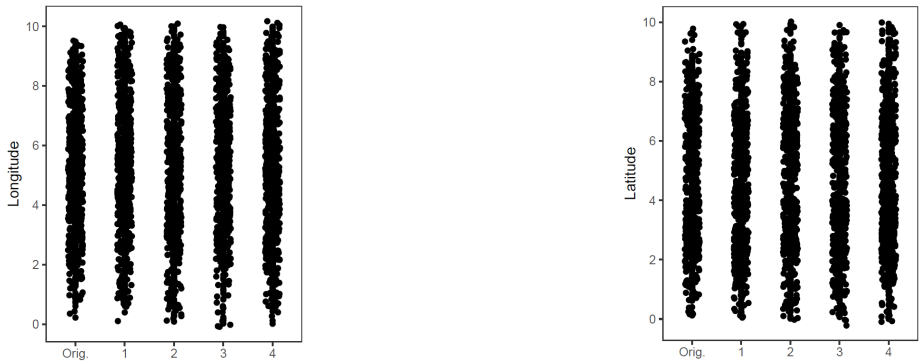
Fonte: Elaborado pela autora.

Figura A.4: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 4 ($grid = 10 \times 10$, sem área restrita, sem variável contínua).



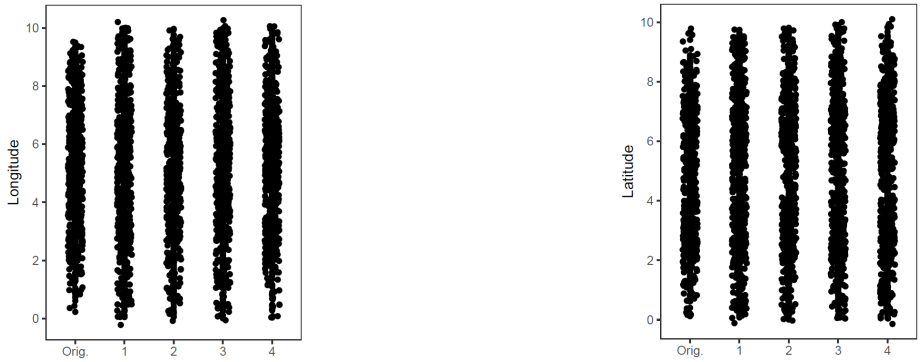
Fonte: Elaborado pela autora.

Figura A.5: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 5 ($grid = 20 \times 20$, com área restrita, com variável contínua).



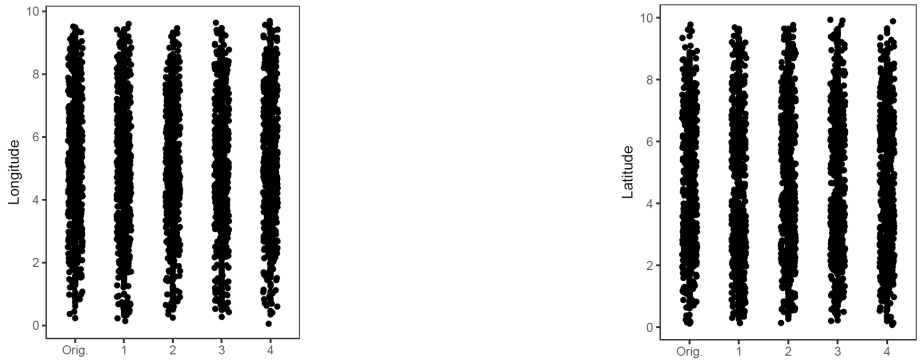
Fonte: Elaborado pela autora.

Figura A.6: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 6 ($grid = 20 \times 20$, com área restrita, sem variável contínua).



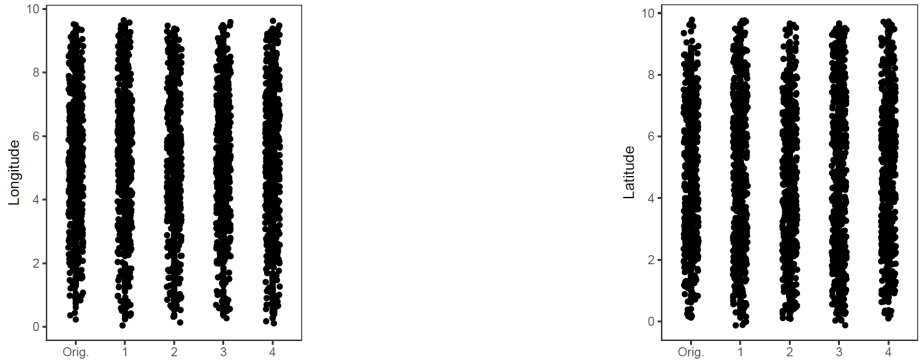
Fonte: Elaborado pela autora.

Figura A.7: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 7 ($grid = 20 \times 20$, sem área restrita, com variável contínua).



Fonte: Elaborado pela autora.

Figura A.8: Gráfico de pontos das longitudes e latitudes das coordenadas originais e sintéticas para os dados simulados - Caso 8 ($grid = 20 \times 20$, sem área restrita, sem variável contínua).



Fonte: Elaborado pela autora.