Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Estatística

Letícia Silva Nunes

Métodos de Simulação de Dados Geográficos Sintéticos Para Bases Confidenciais

Belo Horizonte

2018

Letícia Silva Nunes

Métodos de Simulação de Dados Geográficos Sintéticos Para Bases Confidenciais

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

Orientadora: Profa. Dra. Thaís Paiva Galletti

Belo Horizonte

2018

DEDICATÓRIA

Dedicado a toda a minha família em especial, minha querida irmã Clarice (in memoriam).

AGRADECIMENTO

Agradeço a **Deus** por ter me dado saúde e força par superar as dificuldades.

Agradeço muito à minha orientadora **Thaís** , pelo empenho dedicado à elaboração deste trabalho, pela disponibilidade, pelas conversas que tranquilizaram e por todo o aprendizado recebido.

A **todos professores e funcionários** do Departamento de Estatística da UFMG que contribuíram muito para o meu aprendizado, em especial ao professor **Marcos Prates** por todo o ensinamento, disponibilidade e ajuda neste trabalho.

Aos meus pais **Eloisa e José Duarte**, pelo exemplo de força e persistência, amor, incentivo e apoio incondicional.

Aos meus irmãos **Clarice e José Lucas**, Clarice minha estrela que sei que está sempre me iluminando e José Lucas pelo companheirismo, amizade, carinho, incentivo e palavras sábias nos momentos necessários, me orgulho muito de tê-lo como irmão.

Agradeço **meus familiares** que sempre acreditaram muito no meu trabalho e me ajudaram no que foi preciso.

Aos meus amigos, **Guilherme Oliveira, Ana Cláudia, Guilherme Veloso, Ebert, Juliana, Rafael e Danielle** pela convivência, compreensão e momentos de lazer necessários.

Aos colegas de curso e com certeza futuros/atuais excelentes profissionais, em especial, **Guilherme Oliveira** (novamente) pelas inúmeras dicas e auxílios nos momentos mais complicados, mesmo que para questões simples.

Aos **amigos do intercâmbio** por compartilharem comigo essa incrível experiência.

Aos **colegas de trabalho** da BRAIN que sempre demonstraram apoio imensurável.

Aos **demais amigos** de perto e de longe, não menos importantes, pelo amor e preocupação demonstrados.

Enfim, um **muito obrigada** a todos que me apoiaram em mais esta jornada.

RESUMO

Silva Nunes, Leticia *Métodos de Simulação de Dados Geográficos Sintéticos para Bases Confidenciais*. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, 2018.

Este trabalho apresenta métodos estatísticos para preservação de sigilo de bases de dados confidenciais, em especial, a simulação de dados sintéticos. Este método, simula dados sintéticos que são gerados a partir de distribuições de probabilidade especificadas para reproduzir o máximo de relações possíveis contidas nos dados originais. A simulação de dados sintéticos é atrativa pois permite controlar a preservação da privacidade e utilidade, além de permitir que os pesquisadores utilizem ferramentas convencionais de análise estatística. Além disso, estendemos a metodologia para simulação de coordenadas geográficas sintéticas para dados de área que é proposta em Paiva et al. (2014) ao inserir variáveis contínuas na estimação do modelo.

Palavras-chave: Dados Sintéticos, Confidencialidade, Privacidade, Coordenadas Geográficas.

ABSTRACT

This work presents statistical methods to preserve confidentiality of datasets, in particular simulation of synthetic data. This approach is based on simulate sinthetic data that are generated from from specified probability distributions to reproduce the maximum possible relationships contained in the original data. Simulation of synthetic data is attractive because it allows controlling the disclosure risk and utility. Futhermore, researchers will be able to use conventional statistical analysis tools. In addition, we extended the methodology for simulation of synthetic geographical coordinates for spatial data that is proposed in Paiva et al. (2014) by inserting continuous variables in the estimation of the model.

Keywords: Synthetic Data, Confidenciality, Privacy, Geographical Coordinates.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O objetivo de muitas agências e instituições que coletam dados é divulgá-los de uma forma segura, que sejam informativos e que esses bancos de dados sejam de fácil manipulação para que possamos realizar análises estatísticas. Porém, a divulgação desses dados pode fazer com que indivíduos mal intencionados tenham acesso a informações confidenciais. Segundo Cassa et al. (2006), 87% dos indivíduos de um banco de dados dos EUA que foi divulgado para o público, foram identificados usando o zipcode, a data de nascimento e o gênero. Sendo assim, o desafio para as agências e instituições é divulgar essas informações protegendo a confidencialidade dos dados.

Quando alguma instituição divulga dados confidenciais, ela pode enfrentar vários problemas e sérias consequências. Existem algumas leis e códigos de ética que instituições de pesquisa devem seguir e quando essa política é violada, elas podem enfrentar processos. Além disso, quando coletamos os dados, afirmamos para os indivíduos em estudo que manteremos a privacidade das informações coletadas. Quando isso não é feito, podemos assumir que perderemos a confiança das pessoas. Como consequência, os próximos dados coletados destes entrevistados podem ser duvidosos em razão do medo da divulgação.

Uma solução seria não divulgar os dados, o que é não é uma boa alternativa visto que as informações são importantes em várias áreas de pesquisa e trabalho. Um exemplo são os dados financiados por agências públicas de fomento, como a CNPq, que serão divulgados daqui a alguns anos sob o controle dos coordenadores dos estudos com o intuito de divulgar a pesquisa. Por fim, alguns periódicos exigem que os dados sejam de domínio público para que os revisores possam testar a reprodutibilidade dos resultados. Sendo assim, é interessante encontrar maneiras de divulgar informações com o cuidado de proteger aquelas que sejam confidenciais.

Quando falamos em confidencialidade dos bancos de dados, estamos nos referindo à proteção das informações contidas nestes bancos. Essas informações podem ser endereço de algum indivíduo, número de documento, histórico médico de pacientes, entre outras.

Os métodos de proteção de dados confidenciais vem sendo investigado por vários autores na literatura. Duncan and Lambert (1989) falam sobre alguns métodos para mascarar os dados antes de divulgá-los e seus riscos. Algumas técnicas são divulgar apenas

uma amostra dos dados e trocar as respostas de algumas variáveis para indivíduos selecionados. O problema dessas técnicas é que perderíamos as correlações entre as variáveis prejudicando assim as inferências.

Armstrong et al. (1999) e Sherman and Fetters (2007) falam sobre algumas alternativas de mascarar dados geográficos de saúde protegendo a confidencialidade dos indivíduos e fornecendo boas análises estatísticas. Uma alternativa descrita por eles é adicionar um ruído aleatório às localizações. Porém, essa técnica pode apresentar problemas quando tratamos de áreas pequenas pois prejudicamos a inferência dos dados já que podemos ter uma variância grande nos ruídos adicionados.

Algumas maneiras de proteger os bancos de dados antes de divulgá-los usando metodologias estatísticas, segundo Karr and Reiter (2014), são:

- Agregação: valores que estão em unidade menores, por exemplo, indivíduos em uma classe de aula de uma universidade, são agregados em uma unidade maior, por exemplo, indivíduos de um prédio de uma universidade ou ainda, indivíduos de uma universidade. Porém, este método dificulta a inferência em níveis mais precisos.

- Supressão: os dados que possuem um alto risco de serem identificados são excluídos do banco de dados. Porém, esse método cria dados ausentes que não são aleatórios, o que pode gerar viés de seletividade amostral dificultando as análises. Além disso, mesmo excluindo algumas informações, pessoas má intencionados podem identificar indivíduos a partir de outras variáveis.

- Troca de dados: os valores de algumas varáveis de alguns indivíduos podem ser trocados. O problema desse método é que ele pode destruir a relação entre as variáveis, comprometendo assim as análises.

- Ruído aleatório: são adicionados ruídos aleatórios em algumas variáveis numéricas alterando os seus valores originais e, consequentemente, protegendo os indivíduos da amostra. Geralmente, quanto maior a variância do ruído, maior a proteção dos dados, mas a pertubação das distribuições originais também é maior.

Quando temos dados georreferenciados, como o local de residência do indivíduo, temos um problema na divulgação, já que não podemos divulgar onde a pessoa mora.

Para resolver esse problema, muitas instituições agregam essa informação em unidades espaciais maiores, como por exemplo o bairro, a cidade, ou até mesmo o estado. Outro exemplo de agregação são os setores censitários criados pelo IBGE que são unidades espaciais criadas para gerenciamento de campo e que são usadas na divulgação dos dados coletados nos censos demográficos. Porém, essa técnica prejudica a inferência nos níveis menores de área, apesar de preservar o sigilo dos indivíduos como discutido por Wang and Reiter (2012).

Pensando em proteger a confidencialidade dos dados, Willenborg and De Waal (2001) apresentam metodologias para modificar dados confidenciais de tal forma que possamos fazer boas inferências e tendo pouca perda de informação, controlando os riscos que uma instituição assume ao divulgar os dados. Além disso, ele oferece medidas para controlar a segurança e a perda de informação destes dados.

Segundo Reiter (2004), uma outra maneira de proteger a confidencialidade dos bancos de dados é divulgar dados sintéticos que são simulados e que se assemelham aos dados originais. Os dados sintéticos são gerados a partir de distribuições de probabilidade especificadas para reproduzir o máximo de relações possíveis contidas nos dados originais. Os dados podem ser completamente ou parcialmente sintéticos. Nos dois casos, são gerados várias versões das bases sintéticas para permitir que o usuário possa medir a variabilidade do modelo de simulação ao analisar os dados. Quanto mais bancos de dados forem gerados, mais precisas serão as inferências. Os dados sintéticos permitem então fazer inferências válidas de acordo com o modelo sintético, além de proteger os dados originais substituindo os valores observados por valores simulados.

Usando uma abordagem parecida, Machanavajjhala et al. (2008) nos fornece uma outra maneira de proteger a confidencialidade dos dados usando regressões multinomiais para criar localizações sintéticas de onde a pessoa mora condicionada ao seu local de trabalho e outros atributos. An et al. (2010) propõem proteger os indivíduos utilizando uma abordagem baseada em imputações múltiplas em dados longitudinais. Zhou et al. (2010) apresentam uma maneira de mascarar dados confidenciais de saúde baseada em técnicas de suavização espacial fazendo com que a utilidade dos dados seja mantida. Kounadi and Leitner (2015) quantifica o erro espacial contido nos métodos usados para mascarar os dados confidenciais.

Neste sentido, Paiva et al. (2014) propuseram uma metodologia de simulação de coordenadas sintéticas como uma alternativa para garantir a confidencialidade dos dados e a sua utilidade. Esta metodologia mostra como gerar coordenadas geográficas sintéticas a partir de atributos que são variáveis discretas. Sendo assim, neste trabalho, apresentaremos alguns resultados do modelo proposto, além de uma extensão desse modelo para variáveis contínuas e um estudo de simulação que será feito para a avaliação dos modelos.

Este trabalho será apresentado da seguinte forma: no Capítulo 2, encontra-se a metodologia proposta por Paiva et al. (2014) e a sua extensão, que consiste em incluir variáveis contínuas no modelo e gerar localizações sintéticas usando as informações contidas nestas variáveis contínuas. No Capítulo 3, verificaremos o modelo proposto e comparamos com um modelo mais simplista e que requer menos esforço computacional. No Capítulo 4, apresentaremos os resultados do modelo com a variável contínua aplicando-o a um conjunto de dados simulado. Por fim, no Capítulo 5, concluímos com algumas discussões sobre o que pretendemos fazer no futuro.

2 METODOLOGIA

Para preservar a confidencialidade de dados espaciais, Paiva et al. (2014) propuseram uma metodologia que gera coordenadas sintéticas a partir de variáveis discretas ou atributos de um banco de dados original. Esta metodologia é apresentada a seguir. Apresentamos também o estudo do modelo e o que utilizamos para a implementação do mesmo no "software"livre Rstudio, R Core Team (2014), que tem integrada a versão 3.3.2 do R. Em seguida, apresentamos a extensão do modelo com variável contínua, sua estimação e a geração dos dados sintéticos

2.1 O modelo proposto por Paiva et al. (2014)

Suponha que tenhamos um banco de dados composto pela localização do indivíduo e algumas variáveis discretas que podem ser, por exemplo, idade, gênero e escolaridade. No nosso caso, consideraremos as coordenadas geográficas como sendo a localização do indivíduo. O nosso objetivo é gerar coordenadas geográficas sintéticas a partir do banco de dados original.

Considere que as variáveis discretas ou atributos $(X_1, ..., X_p)$ tenham d_k níveis para $k = 1, ..., p$. Por exemplo, seja X_1 o atributo que representa o gênero do indivíduo. Assim, $X_1 = (1, 2)$, feminino e masculino, e $d_1 = 2$. Logo, o nosso banco de dados pode ser definido como $D = (S, \boldsymbol{X})$ onde S são as nossas coordenadas geográficas e X o vetor de atributos.

Seja $b = 1, ..., B$ o índice para cada combinação diferente dos atributos em X, onde $B \leq \prod_{k=1}^{p} d_k$. Então, $x_k^{(b)}$ é o valor de X_k para a combinação b para cada (b, k).

Para modelar onde os indivíduos com certos padrões demográficos tendem a viver, os autores do artigo sugerem dividir a área de interesse em uma grade regular com G células indexadas por $i = 1, ..., G$. Em cada célula é realizada uma contagem do número de observações naquela célula i com a combinação de atributos b que é chamado de $c_i^{(b)}$.Sendo assim, o modelo a ser estimado, tem a seguinte forma:

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

$$log\lambda_i^{(b)} = \mu + \sum_{k=1}^{p} \alpha_k' \mathbb{1}_{\{x_k^{(b)}\}} + \theta_i + \sum_{k=1}^{p} \phi_{ik}' \mathbb{1}_{\{x_k^{(b)}\}} + \epsilon_i^{(b)} \tag{2.1}$$

onde μ é o intercepto geral, cada $\alpha_k = (\alpha_{k1}, ..., \alpha_{kd_k})$ é um vetor de dimensão d_k x 1 dos efeitos principais para o atributo k; θ_i é o efeito espacial específico para cada célula da grade; e cada $\phi_{ik} = (\phi_{ik1}, ..., \phi_{ikd_k})$ é um vetor de dimensão d_k x 1 dos efeitos espaciais específicos de cada célula e para o atributo k. $\mathbb{1}_{\{x_k^{(b)}\}}$ é o vetor indicador que tem dimensão d_k x 1, sendo um na posição $x_k^{(b)}$ e zero caso contrário, ou seja, ele indica em qual combinação e qual atributo estamos. Os autores sugerem fixar $\alpha_{k1} = 0$ e $\phi_{ik1} = 0$ para garantir a identificabilidade dos parâmetros. Os efeitos espaciais são necessários para que $\lambda_i^{(b)}$ varie pelas células da grade e pelas combinações de atributos. Foi inserido o erro $\epsilon_i^{(b)}$ que adiciona uma flexibilidade extra no modelo. Este modelo implica que as intensidades espaciais são constantes dentro de cada célula da grade.

O modelo utilizado para introduzir a correlação espacial entre células vizinhas, foi o modelo intríseco autoregressivo condicional (ICAR) (Banerjee et al., 2004) utilizado como distribuição *a priori* para $\boldsymbol{\theta} = (\theta_1, ..., \theta_G)$. Logo, para todo i

$$\theta_i | \boldsymbol{\theta_{-i}} \sim N(\overline{\theta}_i, \sigma_\theta^2 / n_i) \tag{2.2}$$

onde $\boldsymbol{\theta_{-i}}$ inclui os valores de θ_j para todo $j \neq i$, e $\overline{\theta}_i$ é a média dos n_i valores de θ_j para as células j que são vizinhas da célula i e σ_θ^2 é a variância comum para todos os valores de θ. Definimos como vizinhos as células que contém vértices comuns, ou pontos comuns entre os lados da célula. Usando uma notação análoga, assumimos para $\{ikj : i = 1, ..., G, k = 1, ..., p, j = 2, ..., d_k\}$,

$$\phi_{ikj} | \boldsymbol{\phi}_{-i,kj} \sim N(\overline{\phi}_{ikj}, \sigma_{\phi_{kj}}^2 / n_i). \tag{2.3}$$

Segundo Banerjee et al. (2004), para garantir a identificabilidade, restringimos os elementos de $\boldsymbol{\theta}$ e $\boldsymbol{\phi_{kj}}$ para que $\sum_{i=1}^{G} \theta_i = 0$ e $\sum_{i=1}^{G} \phi_{ijk} = 0$ para todo (kj).

Como foi falado anteriormente, o erro $\epsilon_i^{(b)}$ foi incluído para adicionar flexibilidade extra no modelo. Qualquer variação residual nas taxas de Poisson que não são explicadas

pelas covariáveis e efeitos espaciais é explicada por este erro. A *priori* assumida é

$$\epsilon_i^{(b)} \sim N(0, \sigma_\epsilon^2). \tag{2.4}$$

Para a distribuição *a priori* dos demais hiperparâmetros, usamos

$$\mu \sim N(0, \upsilon_\mu) \tag{2.5}$$

$$\alpha_{kj} \sim N(0, \upsilon_{\alpha k}) \text{ para todo } (kj) \tag{2.6}$$

$$1/\sigma_\theta^2 \sim Gamma(a_\theta, b_\theta) \tag{2.7}$$

$$1/\sigma_{\phi_{kj}}^2 \sim Gamma(a_{\phi k}, b_{\phi k}) \text{ para todo } (kj) \tag{2.8}$$

$$1/\sigma_\epsilon^2 \sim Gamma(a_\epsilon, b_\epsilon). \tag{2.9}$$

Para que as *prioris* dos hiperparâmetros apresentados acima sejam vagas, os autores recomendam colocar as variâncias $\upsilon.$ altas.

Para calcular as distribuições *a posteriori* dos parâmetros deste modelo, foi usado o algoritmo MCMC (Monte Carlo via Cadeias de Markov).

2.1.1 Estimação

A seguir apresentamos as condicionais completas para cada um dos parâmetros utilizados no algoritmo MCMC.

$$p(\mu|\alpha, \theta, \phi, \epsilon) \propto \exp\left\{\mu \sum_{i=1}^{G}\sum_{b=1}^{B} c_i^{(b)} - \frac{1}{2\upsilon_\mu}\mu^2 - e^\mu \sum_{i=1}^{G}\sum_{b=1}^{B} e^{\sum_{k=1}^{p}\alpha'_k \mathbb{1}_{\{x_k^{(b)}\}} + \theta_i + \sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}} + \epsilon_i^{(b)}}\right\} \tag{2.10}$$

$$p(\alpha_{k1}|\mu,\theta,\phi,\epsilon) \propto \exp\left\{\alpha_{k1}\sum_{i=1}^{G}\sum_{b}c_i^{(b)} - \frac{1}{2\upsilon_{\alpha k}}\alpha_{k1}^2 - e^{\alpha k1}\sum_{i=1}^{G}\sum_{b}e^{\mu+\sum_{j\neq 1}\alpha'_{kj}\mathbb{1}_{\{x_k^{(b)}\}}+\theta_i+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\epsilon_i^{(b)}}\right\} \tag{2.11}$$

$$p(\theta_i|\mu,\alpha,\phi,\epsilon) \propto \exp\left\{\theta_i\sum_{b=1}^{B}c_i^{(b)} - e_i^{\theta}\sum_{b=1}^{B}e^{\mu+\sum_{k=1}^{p}\alpha'_k\mathbb{1}_{\{x_k^{(b)}\}}+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\epsilon_i^{(b)}} - \frac{n_i\tau_\theta}{2}(\theta_i-\overline{\theta_i})^2\right\} \tag{2.12}$$

$$p(\phi_{i1}|\mu,\alpha,\theta,\epsilon) \propto \exp\left\{\phi_{i1}\sum_{b}c_i^{(b)} - e_{i1}^{\phi}\sum_{b}e^{\mu+\sum_{k=1}^{p}\alpha'_k\mathbb{1}_{\{x_k^{(b)}\}}+\theta_i+\sum_{k}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\epsilon_i^{(b)}} - \frac{n_i\tau_{\phi 1}}{2}(\phi_{i1}-\overline{\phi_i})^2\right\} \tag{2.13}$$

$$p(\epsilon_i^{(b)}|\mu,\alpha,\theta,\phi) \propto \exp\left\{\epsilon_i^{(b)}c_i^{(b)} - e^{\epsilon_i^{(b)}+\mu+\sum_{k=1}^{p}\alpha'_k\mathbb{1}_{\{x_k^{(b)}\}}+\theta_i+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}} - \frac{\tau_{\sigma^2}}{2}{\epsilon_i^{(b)}}^2\right\}. \tag{2.14}$$

Os hiperparâmetros de precisão apresentaram *posterioris* que seguem uma distribuição Gama.

$$p(\tau_\theta|\mu,\alpha,\theta,\phi,\epsilon) \propto {\tau_\theta}^{a_\theta+G/2-1}\exp\left\{-\tau_\theta\left[\frac{\theta^T(D_W-W)\theta}{2}+b_\theta\right]\right\} \tag{2.15}$$

$$p(\tau_\phi|\mu,\alpha,\theta,\phi,\epsilon) \propto {\tau_\phi}^{a_\phi+G/2-1}\exp\left\{-\tau_\phi\left[\frac{\phi^T(D_W-W)\phi}{2}+b_\phi\right]\right\} \tag{2.16}$$

onde W é uma matriz de vizinhos de tamanho $G\times G$ cujo elemento W_{ij} é igual a um se as células i e j são vizinhas, e zero caso contrário, e D_W é uma matriz diagonal também de W com tamanho $G\times G$ cuja as entradas indicam o número de vizinhos de cada célula da grade.

$$p(\tau_{\sigma^2}|\mu,\alpha,\theta,\phi,\epsilon) \propto {\tau_{\sigma^2}}^{a_\sigma^2+\frac{GxB}{2}-1}\exp\left\{-\tau_{\sigma^2}\left[b_{\sigma^2}+\frac{\sum_i\sum_b{\epsilon_i^{(b)}}^2}{2}\right]\right\}. \tag{2.17}$$

Como podemos ver, algumas condicionais completas não apresentaram forma fe-

chada. Para elas, usamos o algoritmo Adaptive Rejection Sampling (ARS) (Gilks and Wild, 1992).

O algoritmo Adaptive Rejection Sampling foi escrito por Gilks and Wild (1992) e foi implementado no R dentro do pacote MfuSampler (Alireza and Mansour, 2017). Seu objetivo é oferecer um maneira eficiente de amostrar de distribuições que são algebricamente complexas, mas que pertençam à classe de densidades log-côncava.

Sendo assim, podemos usar o ARS para estimar os parâmetros das distribuições de 2.10 a 2.14 já que eles não apresentaram forma fechada e são log-concâvas.

Para utilizarmos essa função, precisamos encontrar o logaritmo da condicional completa e sua derivada em relação ao parâmetro desejado.

Essa função foi usada dentro do algoritmo Gibbs Sampler via MCMC (Monte carlo via Cadeias de Markov) para calcular as amostras da distribuição *a posteriori* dos parâmetros deste modelo: μ, α, θ, ϕ, ϵ.

2.1.2 Dados sintéticos

Depois de estimar as distribuições *a posteriori* de $\boldsymbol{\lambda} = \{\lambda_i^{(b)}\}$, da equação 2.1, geramos as localizações sintéticas para os n indivíduos. Primeiramente, amostramos um único valor de $\boldsymbol{\lambda}$, por exemplo $\lambda^{(l)}$, da sua distribuição *a posteriori*. Para todo (i, b), fazemos

$$p_i^{(lb)} = \lambda_i^{(lb)} / \sum_{i=1}^{G} \lambda_i^{(lb)} \tag{2.18}$$

Então, amostramos aleatoriamente e independentemente a célula da grade de cada indivíduo com a combinação de atributos b com probabilidade $(p_1^{(lb)}, ..., p_G^{(lb)})$. A localização sintética para cada indivíduo pode ser a célula amostrada. Por outro lado, se queremos saber as coordenadas geográficas do indivíduo, amostramos as coordenadas uniformemente dentro da célula já amostrada. O resultado é um conjunto de localizações sintéticas, $\tilde{S}^{(l)} = (\tilde{s}_1^{(l)}, ..., \tilde{s}_n^{(l)})$, que quando combinado às covariáveis X, obtemos um banco de dados parcialmente sintético, $\tilde{D}^{(l)} = (\tilde{S}^{(l)}, X)$. Para reduzir o vício nas estimativas do modelo, geramos m conjuntos de localizações sintéticas independentes, $\tilde{S} = (\tilde{S}^{(1)}, ..., \tilde{S}^{(m)})$, e seus respectivos bancos de dados $\tilde{D} = (\tilde{D}^{(1)}, ..., \tilde{D}^{(m)})$, que serão

divulgados para o público.

O grande desafio aqui é encontrar o tamanho da célula da grade adequado. Se temos um tamanho de célula muito grande, estamos protegendo o indivíduo, já que as coordenadas sintéticas serão geradas em uma área grande, porém, os dados sintéticos retratarão pouco as relações contidas nos dados originais. Por outro lado, se o tamanho da célula é muito pequeno, retratamos bem as relações contidas nos dados originais, porém, pode ser mais fácil para uma pessoa mal intencionada identificar a localização verdadeira dos indivíduos pesquisados, prejudicando assim, a confidencialidade dos dados.

2.2 Extensão do modelo com variável contínua

O modelo proposto em 2.1 assume que os atributos $(X_1, ..., X_p)$ sejam variáveis discretas. O nosso objetivo aqui é inserir nesse modelo, variáveis contínuas e para isso, é necessário incluir uma função flexível na estimação de $\lambda_i^{(b)}$ para capturar as relações entre a intensidade de ocorrências dos eventos e as variáveis contínuas.

Considere os mesmos dados descritos na seção anterior onde temos $(X_1, ..., X_p)$ como as variáveis discretas ou atributos, S sendo as nossas coordenadas geográficas e incluiremos agora Z que é uma variável contínua. Podemos assumir, por exemplo, que Z represente a idade do indivíduo. Sendo assim, o nosso banco de dados passa a ser definido como $D = (S, X, Z)$.

Considere também que continuamos a ter $b = 1, ..., B$ combinações de atributo em X e dividimos a área de interesse em uma grade com G células indexadas por $i = 1, ..., G$. Como estamos modelando a contagem de indivíduos em cada célula e para cada combinação, e cada indivíduo tem um valor de Z, sendo esta uma variável contínua, precisamos obter um resumo dessa variável que tenha a mesma dimensão da nossa contagem. Para isso, propomos $\overline{Z}_i^{(b)}$ que é a média de Z em cada célula i e para cada combinação de atributo b.

Escolhemos calcular a média $\overline{Z}_i^{(b)}$, pois como estamos olhando para indivíduos que pertençam a mesma combinação e estão localizados na mesma célula da grade, assumimos que eles são suficientemente parecidos. Logo, faz sentido resumir a informação da variável contínua Z calculando a sua média para cada i, b.

Além disso, incluímos no modelo o efeito associado à variável contínua para cada célula da grade, mas comum entre as combinações. Podemos então, incluir uma nova função na modelagem descrita na equação 2.1 da seguinte maneira:

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

$$log\lambda_i^{(b)} = \mu + \sum_{k=1}^{p} \alpha'_k \mathbb{1}_{\{x_k^{(b)}\}} + \theta_i + \sum_{k=1}^{p} \phi'_{ik} \mathbb{1}_{\{x_k^{(b)}\}} + \beta_i \overline{Z}_i^{(b)} + \epsilon_i^{(b)} \tag{2.19}$$

As distribuições *a priori* assumidas para μ, α, θ, ϕ e ϵ são as mesmas da seção 2.1 assim como as suas interpretações. O parâmetro β_i é o efeito da variável contínua para cada célula da grade.

Primeiramente, assumimos uma distribuição Normal para a distribuição *a priori* de β_i, já que a princípio, este parâmetro tem uma interpretação parecida com a interpretação do parâmetro α. Logo,

$$\beta_i \sim N(0, \upsilon_\beta). \tag{2.20}$$

Assim, como nos demais hiperparâmetros recomendamos colocar a variância de υ grande para que *a priori* seja vaga.

Porém, podemos pensar também que os indivíduos possam estar correlacionados espacialmente de acordo com essa variável contínua. Por exemplo, podemos assumir que em determinado bairro tenham mais estudantes por causa da proximidade à universidade e se estamos considerando que a variável contínua seja a idade, os morados deste bairro teriam entre 18 a 30 anos. Assim, também utilizamos o modelo ICAR como distribuição *a priori* para $\boldsymbol{\beta} = (\beta_1, ..., \beta_G)$. Logo, para todo i

$$\beta_i | \boldsymbol{\beta_{-i}} \sim N(\overline{\beta}_i, \sigma_\beta^2 / n_i) \tag{2.21}$$

onde $\boldsymbol{\beta_{-i}}$ inclui os valores de β_j para todo $j \neq i$, e $\overline{\beta}_i$ é a média dos n_i valores de β_j para as células j que são vizinhas da célula i. Para garantir a identificabilidade, restringimos os elementos of $\boldsymbol{\beta}$ para que $\sum_{i=1}^{G} \beta_i = 0$.

2.2.1 Estimação

A seguir apresentamos as condicionais completas para β_i usando as duas *prioris* especificadas em 2.20 e 2.21.

Usando a *priori* Normal:

$$p(\beta_i|\mu,\alpha,\theta,\phi,\epsilon) \propto \exp\left\{\beta_i \sum_{b=1}^{B} c_i^{(b)}\overline{Z}_i^{(b)} - \sum_{b=1}^{B} e^{\mu+\sum_{k=1}^{p}\alpha'_k \mathbb{1}_{\{x_k^{(b)}\}}+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\beta_i\overline{Z}_i^{(b)}+\epsilon_i^{(b)}} - \frac{1}{2\upsilon_\beta}\beta_i^2\right\} \tag{2.22}$$

Usando a *priori* ICAR:

$$p(\beta_i|\mu,\alpha,\theta,\phi,\epsilon) \propto \exp\left\{\beta_i \sum_{b=1}^{B} c_i^{(b)}\overline{Z}_i^{(b)} - \sum_{b=1}^{B} e^{\mu+\sum_{k=1}^{p}\alpha'_k \mathbb{1}_{\{x_k^{(b)}\}}+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\beta_i\overline{Z}_i^{(b)}+\epsilon_i^{(b)}} - \frac{n_i\tau_\beta}{2}(\beta_i-\overline{\beta_i})^2\right\} \tag{2.23}$$

A condicional completa de τ_β, onde $\tau_\beta = 1/\sigma_\beta^2$ também segue uma distribuição Gama como pode ser visto na equação 2.24.

$$p(\tau_\beta|\mu,\alpha,\theta,\phi,\beta\epsilon) \propto {\tau_\beta}^{a_\beta+G/2-1}\exp\left\{-\tau_\beta\left[\frac{\beta^T(D_W-W)\beta}{2}+b_\beta\right]\right\}. \tag{2.24}$$

Para gerar amostras destes parâmetros, não foi possível usar o ARS, pois temos um formato diferente das distribuições anteriores. Sendo assim, usamos o algoritmo Slice Sampling (Neal, 2003) que permite a especificação da função de densidade em outro formato.

O algoritmo Slice Sampling foi escrito por Neal (2003) e também está implementado no R no pacote MfuSampler. O Slice Sampling é um algoritmo MCMC e seu objetivo é mostrar que para amostrar de variáveis aleatórias, podemos simplesmente amostrar uniformemente da região sob o gráfico da sua função de densidade.

Além disso, padronizamos a média $\overline{Z}_i^{(b)}$ para cada combinação b afim de termos valores em torno de zero e desvio-padrão igual a um. Também fizemos uma transformação no β da seguinte forma:

$$\beta^* = \frac{e^\beta}{1+e^\beta}$$

Logo,

$$\beta = \log\left(\frac{\beta^*}{1-\beta^*}\right)$$

Assim, as condicionais completas descritas em 2.22 e 2.23 são reescritas da seguinte maneira:

$$p(\beta_i|\mu,\alpha,\theta,\phi,\epsilon) \propto \exp\{\beta_i \sum_{b=1}^{B} c_i^{(b)}\overline{Z}_i^{(b)} - \sum_{b=1}^{B} e^{\mu+\sum_{k=1}^{p}\alpha'_k \mathbb{1}_{\{x_k^{(b)}\}}+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\beta_i\overline{Z}_i^{(b)}+\epsilon_i^{(b)}} - \frac{1}{2\upsilon_\beta}\beta_i^2 - \log(\beta^*) - \log(1-\beta^*)\} \tag{2.25}$$

$$p(\beta_i|\mu,\alpha,\theta,\phi,\epsilon) \propto \exp\{\beta_i \sum_{b=1}^{B} c_i^{(b)}\overline{Z}_i^{(b)} - \sum_{b=1}^{B} e^{\mu+\sum_{k=1}^{p}\alpha'_k \mathbb{1}_{\{x_k^{(b)}\}}+\sum_{k=1}^{p}\phi'_{ik}\mathbb{1}_{\{x_k^{(b)}\}}+\beta_i\overline{Z}_i^{(b)}+\epsilon_i^{(b)}} - \frac{n_i\tau_\beta}{2}(\beta_i - \overline{\beta_i})^2 - \log(\beta^*) - \log(1-\beta^*)\} \tag{2.26}$$

Essas transformações foram necessárias para termos estabilidade computacional. Além disso, $\beta \in \mathbb{R}$ não sendo possível usar o algoritmo Slice Sampling, já que para usá-lo, precisamos que o suporte da função de densidade seja especificado. Fazendo a transformação logit descrita nas equações 2.25 e 2.26 conseguimos um suporte de $0 \leq \beta \leq 1$.

2.2.2 Dados sintéticos

Anteriormente, para gerar as coordenadas sintéticas, calculamos a probabilidade de sortear uma célula da grade, e dentro dessa célula amostramos as coordenadas uniformemente. Essa probabilidade foi calculada a partir dos valores de λ *a posteriori* e ela dependia da combinação do atributo e da célula da grade de cada indivíduo do banco de dados. Agora, temos uma variável contínua no modelo de estimação de λ. Então, além de gerar as coordenadas sintéticas da mesma forma como foi descrito na Seção 2.1.2, será necessário gerar também valores sintéticos para a variável contínua Z.

Em 2.1.2, ao gerarmos as localizações sintéticas, conseguiríamos saber quais seriam

os atributos associados a estas coordenadas, visto que amostramos a célula da grade de cada individuo com a combinação de atributo b. Assim, os valores das covariáveis X são dados de acordo com a combinação b. Agora, com a inclusão da variável contínua, cada indivíduo possui um valor de Z diferente na base original. Como não há uma correspondência entre os dados originais e os dados sintéticos é necessário criar valores sintéticos para Z.

Para gerar os valores sintéticos de Z, olhamos qual a combinação de atributos b e qual a célula da grade i para cada individuo j com $j = 1, ..., n$. Primeiro, amostramos Z_j^* de uma Normal com a média de $\overline{Z}_i^{(b)}$, que foi padronizado, tendo variância unitária. Depois, voltamos o valor de Z para sua escala original. Seja $\tilde{Z}_j$ o valor sintético que queremos. Então,

$$\tilde{Z}_j = Z_j^* \times Var(\overline{Z}_i^{(b)}) + E(\overline{Z}_i^{(b)}) \tag{2.27}$$

onde $E(\overline{Z}_i^{(b)})$ e $Var(\overline{Z}_i^{(b)})$ são a média e a variância de $\overline{Z}_i^{(b)}$, respectivamente, calculadas em relação a cada combinação.

Lembrando que usamos $\overline{Z}_i^{(b)}$ pois essa média é uma medida razoável para resumir a informação que temos sobre $Z_i^{(b)}$ e é a maneira como incluímos a variável contínua no modelo.

O resultado será um conjunto de valores para $\tilde{Z}$ que, quando combinados às covariáveis X e as coordenadas sintéticas $\tilde{S}$, resultará em um banco de dados parcialmente sintético, $\tilde{D} = (\tilde{S}, X, \tilde{Z})$. Assim, como na Seção 2.1.2, para reduzir o vício nas estimativas do modelo, geramos m valores para $\tilde{Z} = (\tilde{Z}^{(1)}, ..., \tilde{Z}^{(m)})$ e de localizações sintéticas independentes $\tilde{S} = (\tilde{S}^{(1)}, ..., \tilde{S}^{(m)})$, formando seus respectivos bancos de dados $\tilde{D} = (\tilde{D}^{(1)}, ..., \tilde{D}^{(m)})$, que serão divulgados para o público.

O próximo passo é avaliar o que foi feito pelo modelo proposto por Paiva et al. (2014) e depois aplicarmos a extensão do modelo com a variável contínua em um banco de dados simulado. Logo, no Capítulo 3, avaliaremos o modelo apresentado na Seção 2.1 e o compararemos com um modelo mais simplista analisando os seus riscos de divulgação e utilidade ao fazermos inferências.

3 AVALIAÇÃO DO MODELO

Neste capítulo, avaliaremos o modelo apresentado no Capítulo 2, na seção 2.1, verificando a sua utilidade e risco e comparando-o com um modelo simplista, que chamaremos de Naive. Para isso, realizaremos algumas simulações em cenários cujas intensidades espaciais são previamente conhecidas.

3.1 Dados Simulados

O primeiro passo para avaliar o nosso modelo é gerar algumas intensidades espaciais para depois verificarmos se, com o modelo proposto, conseguimos encontrar aproximadamente os mesmos valores. Assim, considere um banco de dados formado por $s = (s_1, s_2)$, a localização do indivíduo, e algumas covariáveis $x_1 \in \{1, 2\}$, $x_2 \in \{1, 2, 3\}$ e $y \in \{0, 1\}$. O nosso objetivo é gerar s a partir das informações contidas no conjunto de covariáveis (x_1, x_2, y).

Geramos uma amostra de 500 observações. Primeiramente, geramos x_1 com probabilidade dada pela Tabela 1.

x	$\mathrm{p}(\mathrm{X}_1 = x)$
1	0,6
2	0,4

Tabela 1 : Distribuição de probabilidade de x_1

Então, usamos a distribuição de probabilidade condicional para gerarmos $x_2|x_1$ de acordo com as probabilidades mostradas na Tabela 2.

x	$\mathrm{p}(\mathrm{X}_2 = x\|X_1 = 1)$	$\mathrm{p}(\mathrm{X}_2 = x\|X_1 = 2)$
1	0,333	0,6
2	0,333	0,1
3	0,333	0,3

Tabela 2 : Distribuição de probabilidade de $x_2|x_1$

As probabilidades de y são determinadas pela regressão logística a seguir:

$$\text{logit } (p(y = 1)) = \beta_0 + \beta_1 \mathbb{1}_{(x_1=2)} + \beta_2 \mathbb{1}_{(x_2=2)} + \beta_3 \mathbb{1}_{(x_2=3)} \tag{3.1}$$

onde $\beta_0 = -1$, $\beta_1 = 1,5$, $\beta_2 = -0,5$ e $\beta_3 = 0,5$.

Para gerar as localizações espaciais (s_1, s_2) usamos uma mistura de normais bivariadas para cada combinação possível da seguinte forma:

Tabela 3 Distribuições dos dados simulados

y	x_1	x_2	**Distribuição de** $\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix}$
0	1	1	$\mathrm{N}_2\left[\begin{pmatrix}3\\7\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$ & $\mathrm{N}_2\left[\begin{pmatrix}7\\3\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$
		2	$\mathrm{N}_2\left[\begin{pmatrix}3\\3\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$ & $\mathrm{N}_2\left[\begin{pmatrix}7\\7\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$
		3	$\mathrm{N}_2\left[\begin{pmatrix}5\\5\end{pmatrix};\begin{pmatrix}0,3&0\\0&4,5\end{pmatrix}\right]$
	2	1	$\mathrm{N}_2\left[\begin{pmatrix}3\\5\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$ & $\mathrm{N}_2\left[\begin{pmatrix}7\\5\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$
		2	$\mathrm{N}_2\left[\begin{pmatrix}5\\3\end{pmatrix};\begin{pmatrix}0,8&0\\0&0,8\end{pmatrix}\right]$ & $\mathrm{N}_2\left[\begin{pmatrix}5\\7\end{pmatrix};\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$
		3	$\mathrm{N}_2\left[\begin{pmatrix}7\\3\end{pmatrix};\begin{pmatrix}1&0\\0&2\end{pmatrix}\right]$
1	1	1	$\mathrm{N}_2\left[\begin{pmatrix}1{,}5\\5\end{pmatrix};\begin{pmatrix}0,3&0\\0&4,5\end{pmatrix}\right]$
		2	$\mathrm{N}_2\left[\begin{pmatrix}5\\5\end{pmatrix};\begin{pmatrix}1,5&0{,}9*\sqrt{(1,5*2)}\\0{,}9*\sqrt{(1,5*2)}&2\end{pmatrix}\right]$
		3	$\mathrm{N}_2\left[\begin{pmatrix}8{,}5\\5\end{pmatrix};\begin{pmatrix}0,3&0\\0&4,5\end{pmatrix}\right]$
	2	1	$\mathrm{N}_2\left[\begin{pmatrix}5\\2\end{pmatrix};\begin{pmatrix}2,5&0\\0&0,7\end{pmatrix}\right]$
		2	N(5,5); $\mathrm{N}_2\left[\begin{pmatrix}5\\5\end{pmatrix};\begin{pmatrix}1,5&\text{-}0{,}9*\sqrt{(1,5*1,5)}\\\text{-}0{,}9*\sqrt{(1,5*1,5)}&2\end{pmatrix}\right]$
		3	N(5;7,5); $\mathrm{N}_2\left[\begin{pmatrix}5\\7{,}5\end{pmatrix};\begin{pmatrix}2,5&0\\0&0,7\end{pmatrix}\right]$

As distribuições normais bivariadas apresentadas na Tabela 3 foram escolhidas para apresentar diferentes padrões espacias para cada combinação.

Usando essas intensidades, podemos agora gerar as localizações. Seja n_b o número de pontos a ser gerado para cada combinação. Esse número já está estabelecido uma vez que já geramos os valores de y, x_1 e x_2. Sendo assim, para cada combinação b, geramos n_b localizações da seguinte forma:

- Selecionamos uma célula do grid de $i = 1, ..., G$ com probabilidade igual a $\frac{\lambda_i^{(b)}}{\sum_i \lambda_i^{(b)}}$,

onde $\lambda_i^{(b)}$ é a intensidade fixada, ou seja, o número de indivíduos para cada célula da grade e cada combinação de atributo.

- Geramos a localização (s_1, s_2) uniformemente dentro da grade da célula selecionada.

Devemos lembrar que as intensidades bivariadas definidas pelas misturas na Tabela 3 foram calculadas nos centros das células da grade, pois dessa forma é possível calcular a distância entre os vizinhos mais próximos, e que podemos usar diferentes tamanhos para as células.

Na Figura 1, temos uma amostra de coordenadas originais geradas para cada combinação de atributo e as intensidades verdadeiras fixadas pelas distribuições definidas na Tabela 3.

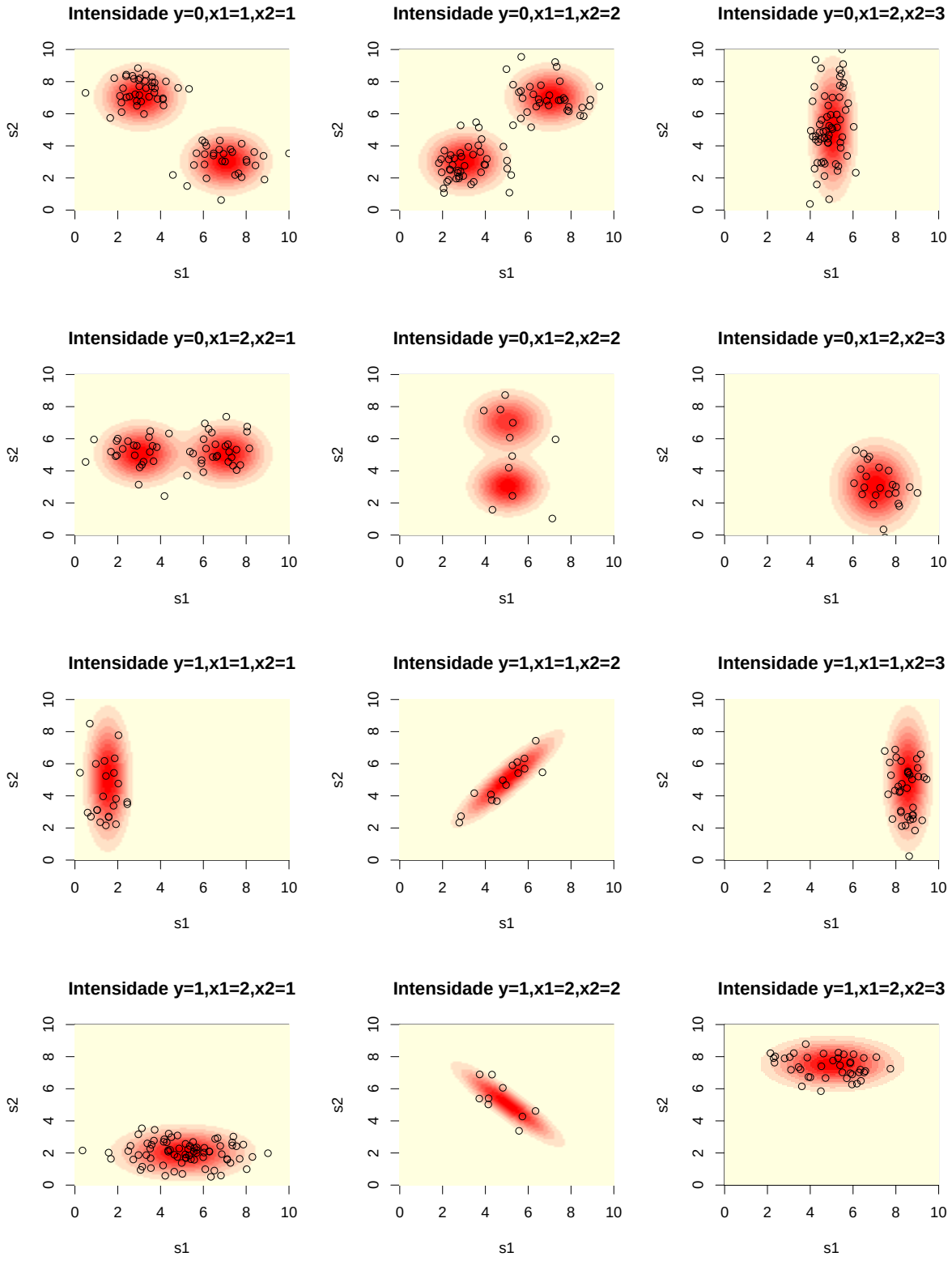

Figura 1 : Intensidades fixadas e coordenadas originais simuladas para cada combinação

Depois de gerar as coordenadas originais, usamos os valores de $D = (y, x_1, x_2, s_1, s_2)$

para estimar as intensidades e gerar as localizações sintéticas. Os valores estimados das intensidades e as coordenadas sintéticas serão, então, comparados aos valores originais plotados na Figura 1.

Assim, estimamos as intensidades como descrito no Capítulo 2. Geramos 15001 observações das condicionais completas apresentadas em (2.10) a (2.14). Na Figura 2, apresentamos os gráficos de série para avaliar as convergências dos parâmetros. Plotamos os valores de μ, alguns valores de α, já que para garantir a identificabilidade fixamos $\alpha_{k1} = 0$, e os valores das precisões τ_θ, τ_ϕ e τ_ϵ. Os valores de θ_i e ϕ_{ik} não foram plotados, pois teríamos um gráfico para cada i, onde $i = 1, ..., G$. Descartamos 5000 observações do período de aquecimento e utilizamos uma grade regular de 10 x 10, com $G = 100$ células no total.

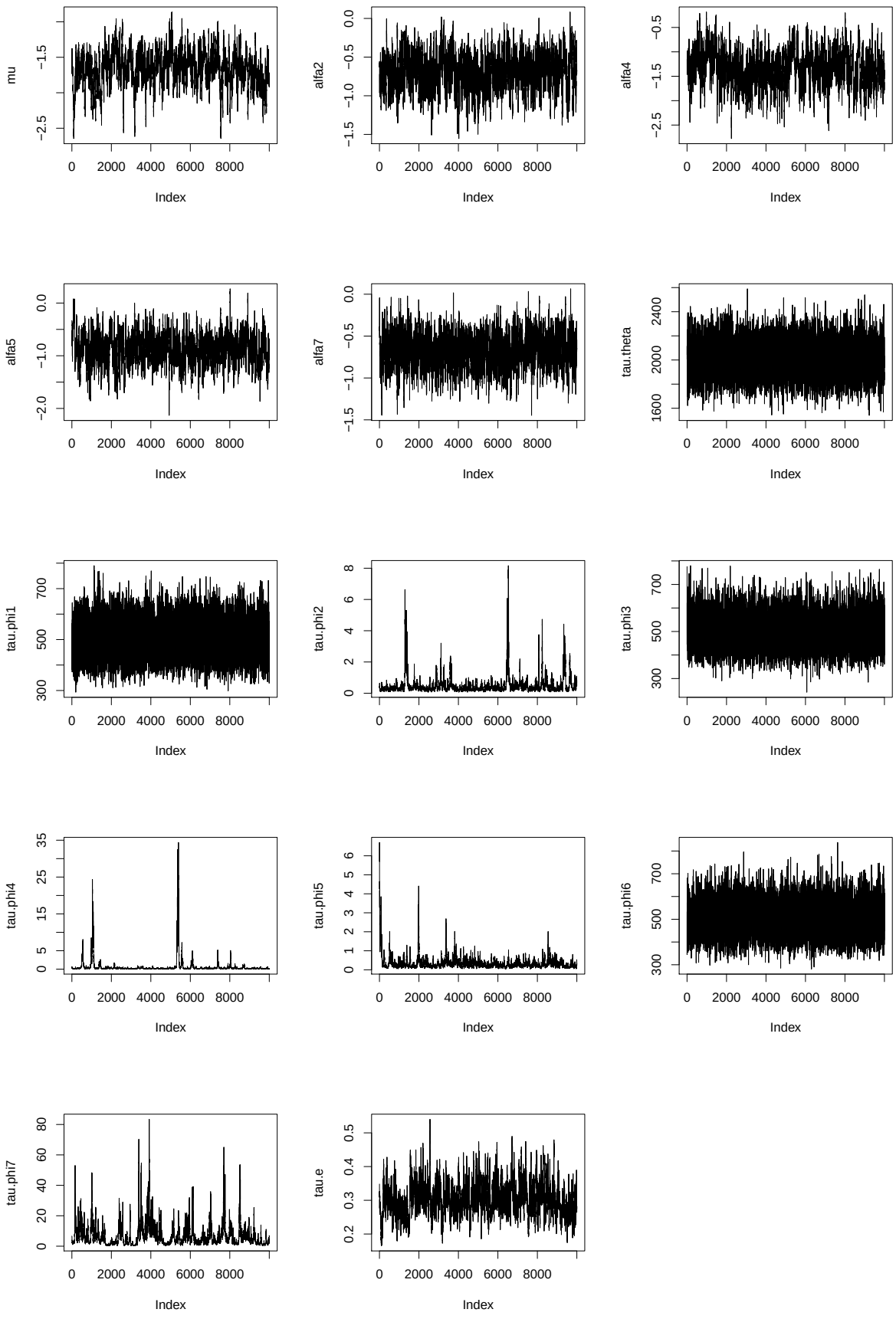

Figura 2 : Gráfico de série para os parâmetros

Podemos perceber que todos os parâmetros convergiram. Sendo assim, podemos fazer as comparações entre as intensidades estimadas e originais. Na Figura 3, apresentamos as intensidades estimadas pelo modelo.

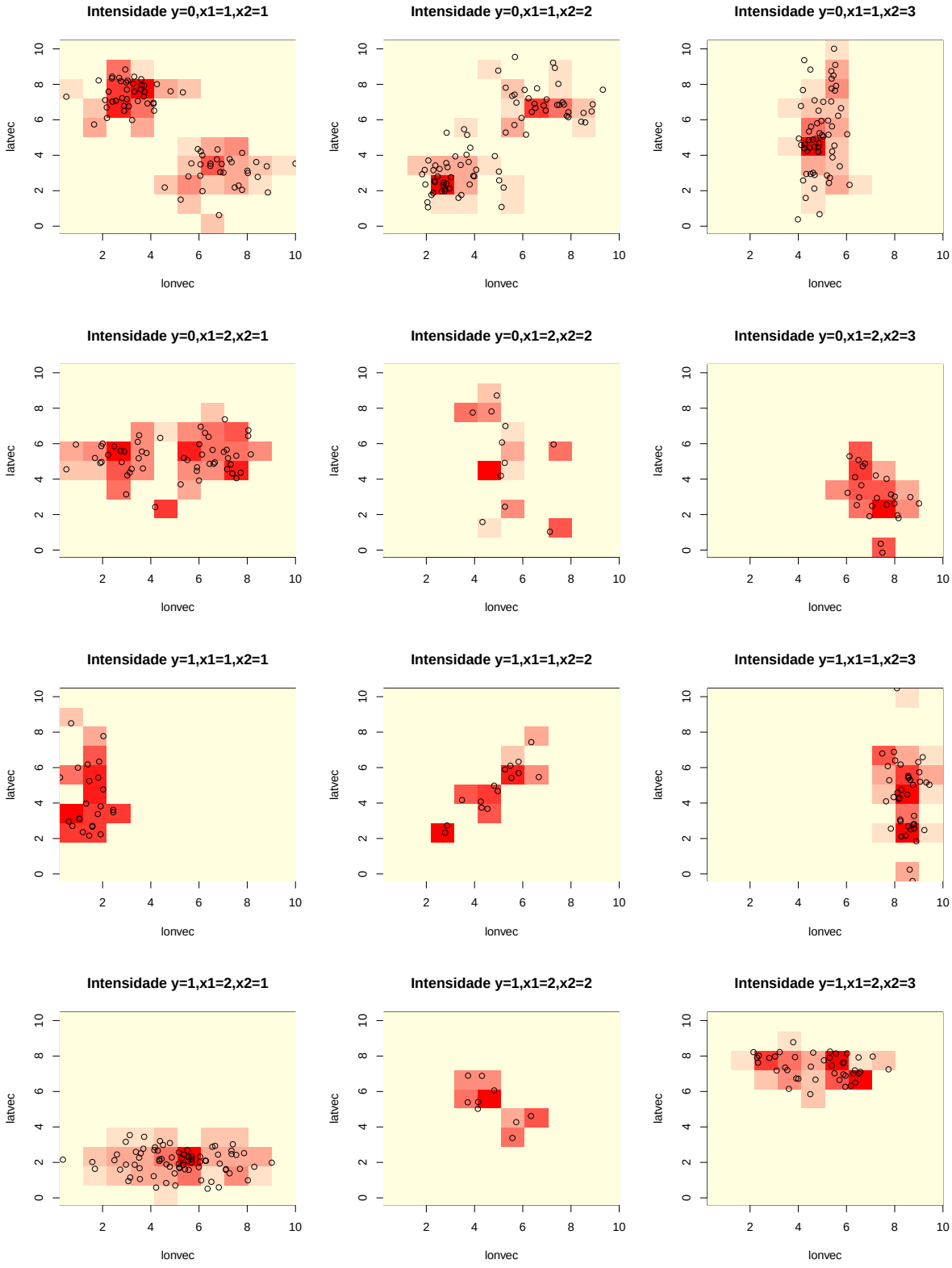

Figura 3 : Intensidades estimadas

Nos gráficos da Figura 3, podemos perceber que as intensidades geradas pelo

modelo seguem os pontos, que indicam as localizações originais.

Até aqui, reproduzimos o que foi indicado pelo modelo apresentado em 2.1 e verificamos que ele é capaz de nos fornecer boas estimações para as intensidades originais fixadas. Porém, quão útil e qual o risco dos bancos de dados sintéticos gerados por esse modelo? A seguir mostraremos o modelo Naive e discutiremos sobre o Risco e a Utilidade dos bancos de dados sintéticos gerados pelos dois métodos.

3.2 **Método Naive**

O modelo apresentado em 2.1 é complexo por ter muitos parâmetros para estimar, tendo assim um grande custo computacional. Pensando nisso, propomos um método simplista que consiste em gerar de forma aleatória as coordenadas sintéticas dentro da célula da grade considerando a informação da célula em que está o indivíduo e a sua combinação de atributo. Por exemplo, para os indivíduos que estão na mesma célula e possuem a mesma combinação de atributos, geramos coordenadas sintéticas uniformemente dentro da mesma célula. Assim como no método de Paiva et al. (2014), fazemos isso m vezes e obtemos no final m bases sintéticas. Aqui, queremos saber qual o ganho temos ao usar o modelo apresentado anteriormente comparado com o modelo simplista. Para isso, avaliaremos os riscos de divulgação e a utilidade das bases sintéticas geradas pelos dois métodos.

3.3 **Risco e Utilidade**

Como discutido no Seção 1, queremos gerar bases que sejam úteis e que protejam o indivíduo em estudo. Sendo assim, é necessário verificar quais riscos assumimos ao divulgarmos essas bases sintéticas, assim como o quão útil para fazermos inferências elas são.

Existem algumas medidas para avaliar os riscos de divulgação que são descritos em Paiva et al. (2014). Estas medidas assumem que a pessoa mal intencionada esteja em posse de qualquer informação sobre os dados sintéticos, incluindo os m banco de dados gerados e o modelo que gerou estes bancos de dados. Além disso, os riscos propostos no artigo são baseados na probabilidade da pessoa mal intencionada acertar corretamente

qual a célula da grade aquele indivíduo pertence. Como no método Naive os valores são gerados dentro da mesma célula em que os valores originais estão, estas medidas não seriam adequadas, pois a pessoa mal intencionada já teria a informação da localização daquele indivíduo violando assim as regras existentes de avaliação do risco.

Sendo assim, para avaliarmos os riscos de divulgação das bases sintéticas, olharemos para as distâncias euclidianas de cada coordenada sintética com a coordenada original mais próxima. Para isso, usaremos um algoritmo que nos retornará os valores das distâncias mínimas entre as coordenadas sintéticas e originais. No algoritmo precisamos de uma matriz que contenha todas as distâncias entre as coordenadas sintéticas e originais das observações que tenham a mesma combinação de atributos b. Esse algoritmo usado para criar as medidas de distâncias é descrito a seguir

Algoritmo 1: Função para criar as medidas de distâncias

Entrada: *matriz*

Saída: Distância entre a coordenada original e a sintética

```
1  início
2      nb = número de linhas de matriz
3      d = valor inicial para a distância entre a coordenada original e a sintética
       para cada i ∈ nb faça
4          pos = encontre a distância mínima entre uma coordenada sintética e
             uma original na matriz
5          d[i] = matriz[pos]
6          se i == nb então
7              fim
8          senão
9              Elimine a coluna e a linha que contém o valor mínimo da matriz
10         fim
11     fim
12     retorna d
13 fim
```

O valores d retornados pelo Algoritmo 1 representam as distâncias mínimas entre cada observação original e uma observação sintética de mesma combinação. Esses valores

são guardados em um vetor de tamanho n, e calculamos algumas estatísticas descritivas para essas distâncias mínimas para avaliar o risco de identificação das bases sintéticas.

Para verificarmos a utilidade das bases sintéticas geradas, fazemos inferências usando uma análise frequentista e uma Bayesiana.

Na análise frequentista, fizemos um intervalo de confiança de 95% da estimativa de y para o banco de dados original e os bancos de dados sintéticos gerados pelos dois métodos. Primeiramente, selecionamos dois quadrantes da nossa área da grade. Pensamos em selecionar duas células da grade, porém o intervalo de confiança do método Naive e original seriam os mesmos já que o número de observações geradas pelo método Naive é o mesmo numero de observações do banco de dados original. Por esta razão, resolvemos usar quadrantes que não coincidam com as células como mostrado na Figura 4.

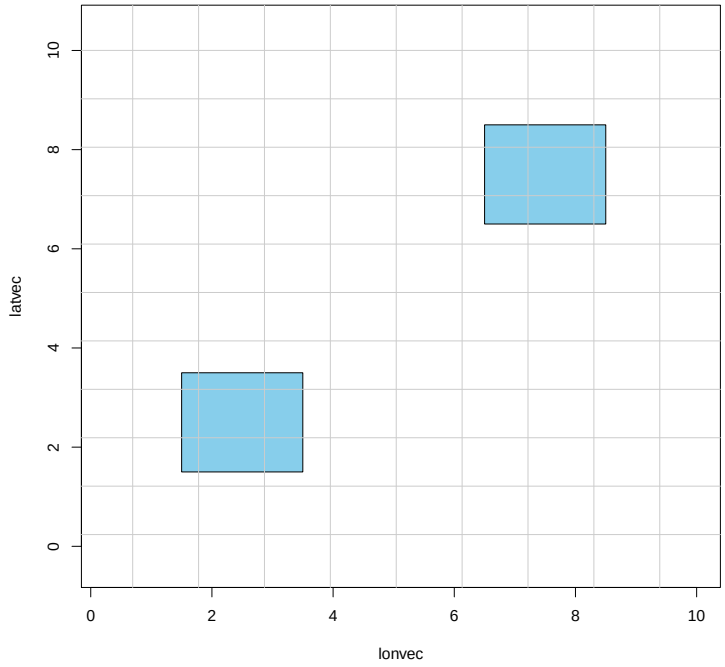

Figura 4 : Quadrantes selecionados

Dentro destes quadrantes, estimamos a proporção de $\hat{y}$ e então fazemos o intervalo de confiança para a proporção de y e comparamos com o intervalo de confiança original.

Seguindo uma mesma lógica, geramos 1000 quadrantes e estimamos a proporção de y em cada um deles. Fazemos então, 1000 intervalos de confiança para cada método e

comparamos com os intervalos originais ao verificar as taxas de cobertura. Para calcular essa taxa, primeiro, encontramos a proporção do intervalo gerado com os dados original que está contido no intervalo gerado por um dos métodos. Depois, fazemos ao contrário, ou seja, encontramos a proporção do intervalo gerado por um dos métodos que está contido no intervalo gerado com os dados originais. Então, fazemos uma média geométrica entre as duas taxas encontradas. Logo, teremos 1000 taxas de cobertura encontradas para cada um dos métodos. Para comparar as taxas, fazemos então, um média entre entre as 1000 taxas encontradas.

Na análise Bayesiana, fizemos uma regressão Bayesiana espacial e estimamos os seus parâmetros. A regressão Bayesiana espacial é da seguinte forma:

$$Y(s) = \mathbf{x}'\boldsymbol{\beta} + \delta(s) \tag{3.2}$$

onde $\mathbf{x}'$ é o vetor de covariáveis, $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão e $\delta(s)$ é o erro do processo Gaussiano. Para que esse erro seja estacionário, ou seja, sua média e variância não muda (Banerjee et al., 2004), devemos assumir uma função de covariância para δ. A função que assumimos é a exponencial da seguinte forma:

$$C(s) = \begin{cases} \sigma^2 \exp(-\phi s), & \text{se} \quad s > 0, \\ \tau^2 + \sigma^2, & \text{se} \quad s = 0. \end{cases}$$

onde τ^2 é a variância do efeito não espacial, σ^2 é a variância do efeito espacial e ϕ é o parâmetro de decaimento.

Estamos interessados em estimar os valores de $\boldsymbol{\beta}$, σ^2 e ϕ. Então, fazemos uma regressão espacial com abordagem Bayesiana e estimamos valores para cada parâmetro. Com estes valores, fazemos então, um intervalo de credibilidade de 95% para cada parâmetro dos dados originais, bases geradas pelo método Paiva et al. (2014) e pelo método Naive.

Para a análise frequentista, utilizamos as regras de combinação de estimativas sob imputação múltipla descritas por Raghunathan et al. (2003). Considere Q como o parâmetro que desejamos estimar, q um estimador pontual para Q e u a variância estimada. Para cada $l = 1, ..., m$, seja $q^{(l)}$ e $u^{(l)}$ os estimadores de q e u para cada $D^{(l)}$.

Para fazer inferências sobre Q, precisamos das seguintes quantidades:

$$\overline{q}_m = \sum_{l=1}^{m} q^{(l)}/m \tag{3.3}$$

$$b_m = \sum_{l=1}^{m} (q^{(l)} - \overline{q}_m)^2/(m-1) \tag{3.4}$$

$$\overline{u}_m = \sum_{l=1}^{m} u^{(l)}/m \tag{3.5}$$

Então, usamos $\overline{q}_m$ como estimador pontual de Q com variância $T_m = b_m/m + \overline{u}_m$. As inferências são baseadas na distribuição t, $(Q - \overline{q}_m) \sim t_\nu(0, T_m)$ com $\nu_m = (m-1)(1 + m\overline{u}_m/b_m)^2$ graus de liberdade.

Essas regras de combinação para inferência sob imputação múltipla não são válidas para análises Bayesianas. De acordo com Zhou and Reiter (2010), o modelo deve ser ajustado para cada base sintética, e as amostras *a posteriori* de cada parâmetro devem ser empilhadas. Os intervalos de credibilidade são então calculados considerando essa amostra empilhada.

Na Seção 3.4 apresentamos os resultados das análises de risco e utilidade para os métodos de Paiva et al. (2014) e Naive.

3.4 **Resultados**

Nesta seção, avaliaremos os riscos e utilidade do modelo proposto por Paiva et al. (2014) e Naive. Primeiramente, escolhemos a combinação de número 2 que tem $y = 0$, $x_1 = 1$ e $x_2 = 2$, pois essa é a combinação que mais possui observações. Então, compararemos as bases sintéticas e originais geradas pelos dois métodos. Na Figura 5 e 6 temos as localizações originais e as geradas cada uma das bases sintéticas $m = 1, ..., 5$, para os métodos de Paiva et al. (2014) e Naive, respectivamente, plotados na superfície dos valores médios de λ obtidos pelo método de Paiva et al. (2014).

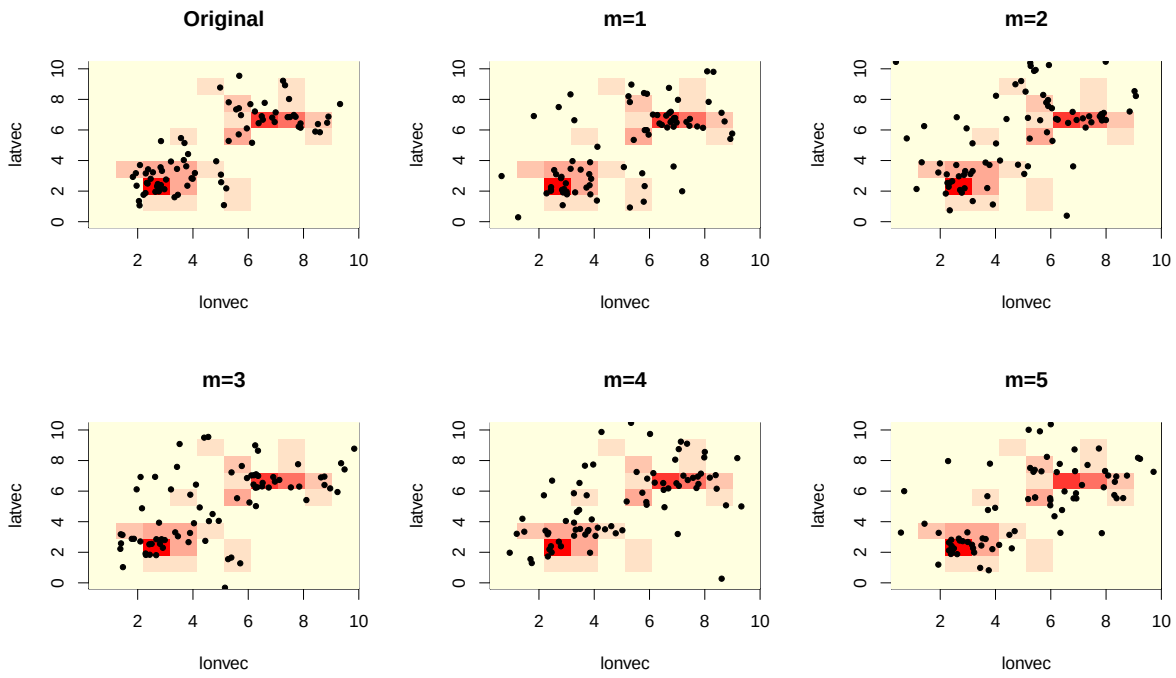

Figura 5 : Localizações originais e sintéticas geradas pelo método de Paiva et al. (2014)

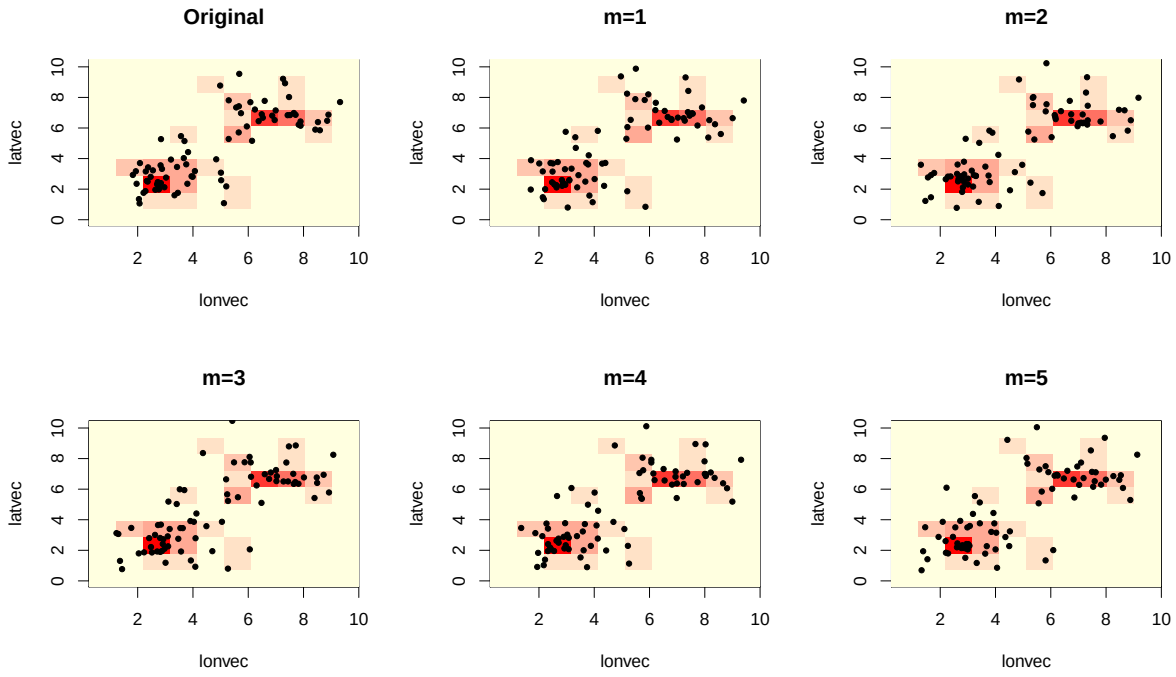

Figura 6 : Localizações originais e sintéticas geradas pelo método Naive

Como era de se esperar, as localizações sintéticas geradas pelo método Naive se

parecem mais com as localizações originais, visto que as observações neste método são geradas dentro da mesma célula da grade em que se encontra a localização original e não sorteados como em Paiva et al. (2014).

Agora, avaliaremos o risco de divulgação destas bases sintéticas simuladas pelos dois métodos. Na Tabela 4, temos algumas estatísticas descritivas feitas com o vetor retornado pelo Algoritmo 1 que contém as distâncias mínimas entre cada observação original e uma observação sintética de mesma combinação.

Tabela 4 Comparação das distâncias das coordenadas sintéticas geradas

	Método de Paiva et al		**Método Naive**	
B	**Mínimo**	**Média**	**Mínimo**	**Média**
1	0.0343	1.2996	0.0617	1.4056
2	0.0975	2.4576	0.2108	2.9278
3	0.0334	1.0290	0.0357	1.1320
4	0.1510	1.8115	0.2000	1.7052
5	0.0333	1.0390	0.0521	2.0284
6	0.0533	1.7041	0.0863	2.9856
7	0.0719	1.7755	0.0862	2.3301
8	0.0323	1.2704	0.0377	1.7766
9	0.4473	2.1181	0.3776	2.0892
10	0.4439	2.7035	0.2952	1.9161
11	0.1377	2.1867	0.4423	3.6277
12	0.0462	1.3958	0.0640	2.0882
Média	**0.1318**	**1.7325**	**0.1625**	**2.1677**

Na Tabela 4, vemos que os valores mínimos e médios das distâncias no método Naive e no métodos de Paiva et al. (2014) não são tão diferentes. Porém, as distâncias são um pouco maiores no método Naive. Com isso, o risco de uma pessoa mal intencionada encontrar aquele indivíduo é menor. Contudo, devemos lembrar que neste método podemos gerar observações somente dentro da mesma célula em que se encontram as observações originais. Já no método de Paiva et al. (2014), as coordenadas não são neces-

sariamente geradas dentro da mesma célula da grade em que elas se encontram no banco de dados original.

Para avaliarmos a utilidade dos bancos de dados, fizemos um intervalo frequentista de 95% de confiança para y. Calculamos a proporção de y em cada um dos quadrantes da Figura 4 na base original e em cada uma das bases sintéticas geradas pelos dois métodos. Na base original, fizemos um intervalo de confiança para a proporção de y, e para as bases sintéticas, calculamos os intervalos de confiança de acordo com as regras de combinação descritas nas equações 3.3 a 3.5. Os intervalos são apresentados na Figura 7.

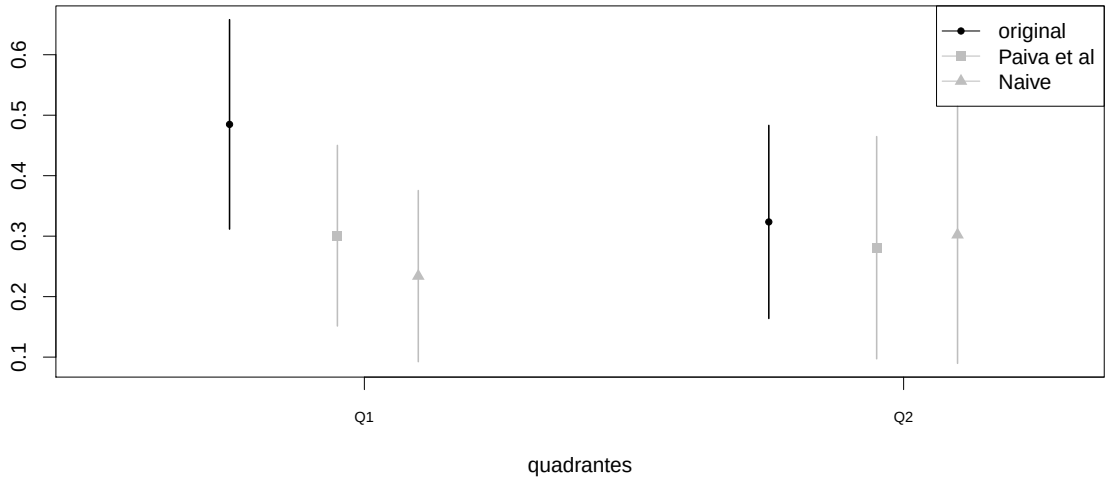

Figura 7 Intervalo de 95% de confiança para y

Podemos perceber que os intervalos de confiança para as bases sintéticas geradas pelo método de Paiva et al. (2014) se parecem mais com os intervalos de confiança feitos para as bases originais. A taxa de cobertura gerada pelo método de Paiva et al. (2014) foi de 0,725 enquanto a do método Naive foi 0,697.

Na Figura 8, apresentamos os intervalos de 95% de credibilidade para os coeficientes da regressão Bayesiana feita para as bases sintéticas geradas pelos dois métodos. Geramos 100000 valores para cada parâmetro e excluímos 90000 no período de aquecimento. Os intervalos para os parâmetros da regressão feitos para o método de Paiva et al. (2014) também se assemelham mais aos parâmetros da regressão feita com os dados originais.

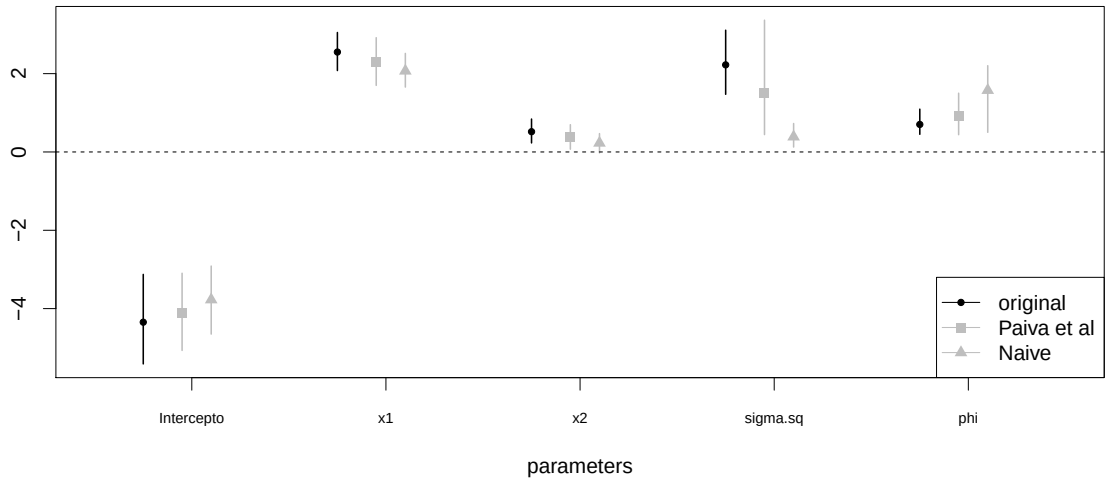

Figura 8 Intervalo de 95% de credibilidade

Com isso, concluímos que os bancos de dados sintéticos gerados pelo método de Paiva et al. (2014) e pelo método Naive têm os riscos de divulgação parecidos, apesar das distâncias do método Naive serem um pouco maiores. Porém, a utilidade das bases geradas pelo método de Paiva et al. (2014) parece um pouco melhor do que a utilidade das bases sintéticas geradas pelo método Naive.

No Capítulo 4, mostraremos os resultados da extensão do modelo de Paiva et al. (2014) com a variável contínua.

4 RESULTADOS

Neste Capítulo, mostraremos os resultados das simulações feitas usando a metodologia apresentada em 2.2 com distribuições a *priori* Normal e ICAR. Para as duas *prioris*, usaremos o mesmo banco de dados simulado na seção 3.1, acrescentando uma variável contínua Z.

A variável contínua Z foi gerada da seguinte forma:

$$\overline{Z}^{(b)} = \beta_0 + \beta_1 \mathbb{1}_{(y=1)} + \beta_2 \mathbb{1}_{(x_1=2)} + \beta_3 \mathbb{1}_{(x_2=2)} + \beta_4 \mathbb{1}_{(x_2=3)} \tag{4.1}$$

onde $\beta_0 = 50$, $\beta_1 = -3,5$, $\beta_2 = 1$, $\beta_3 = -1,5$ e $\beta_4 = 2$. Assim, para cada combinação temos uma média de Z e, com essa média, geramos um valor de Z para cada indivíduo de uma distribuição Normal($\overline{Z}^{(b)}$,10). Na Figura 9, temos os boxplots dos valores de Z para cada combinação de atributo.

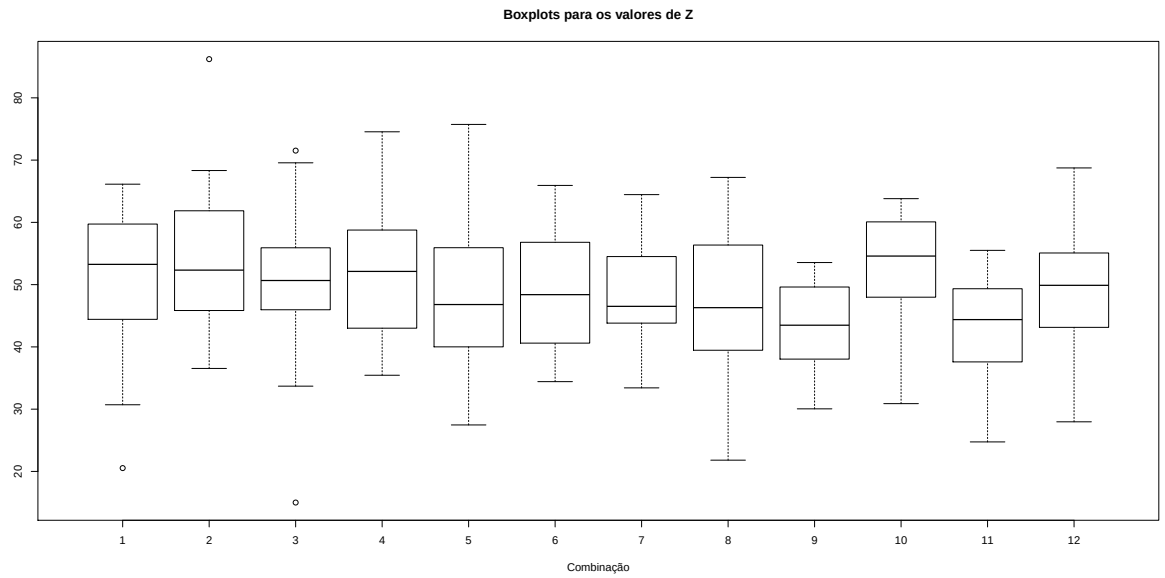

Figura 9 : Boxplot para os valores de Z

Na Figura 10 temos as coordenadas s_1 e s_2 plotadas de acordo com os valores de Z.

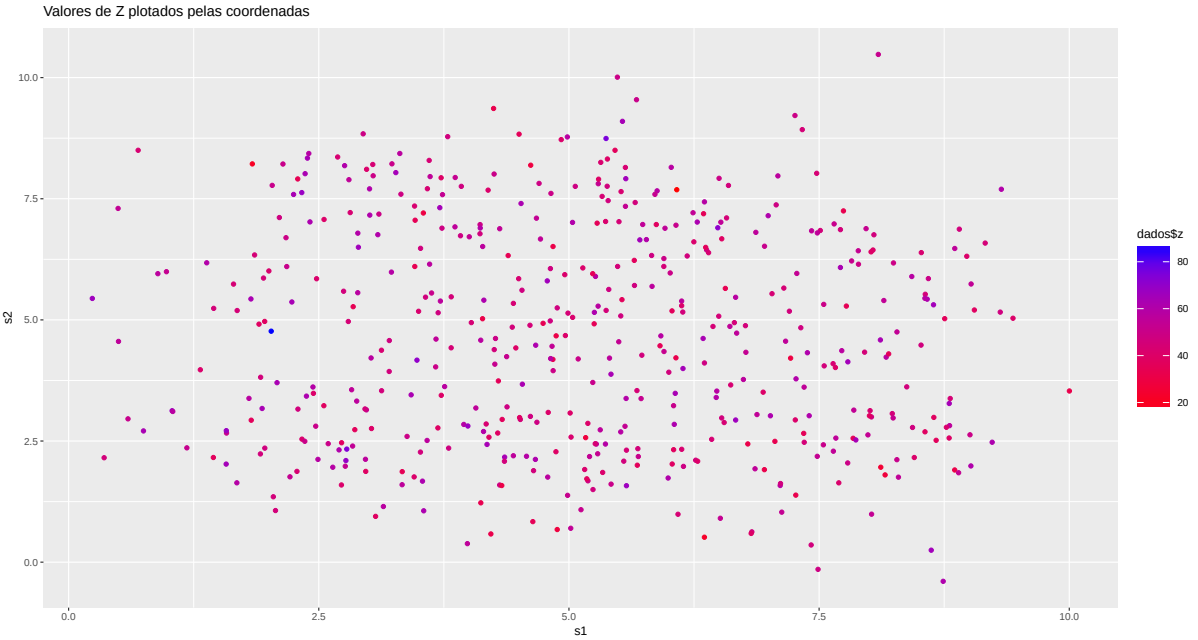

Figura 10 : Valores de Z plotados pelas coordenadas s_1 e s_2

4.1 ***Priori* Normal**

Para gerar as intensidades $\lambda_i^{(b)}$, usamos o modelo apresentado na equação 2.19 e usando como distribuição *a priori* para o β uma distribuição Normal. Assim como anteriormente, geramos uma amostra de 15001 e descartamos as 5001 primeiras observações. Como sabemos que o tamanho da célula da grade é importante para verificarmos o quão útil é o banco de dados gerado e qual o risco ao divulgá-lo, dividimos a região em três tamanhos: 5 x 5, 10 x 10 e 20 x 20. Em todas as estimações para λ, os parâmetros convergiram. As intensidades para as grades de tamanho 5 x 5 são apresentadas na Figura 11, 10 x 10 na Figura 12 e 20 x 20 na Figura 13, os pontos são as localizações originais e os seus tamanhos variam conforme o valor de Z correspondente aquela observação.

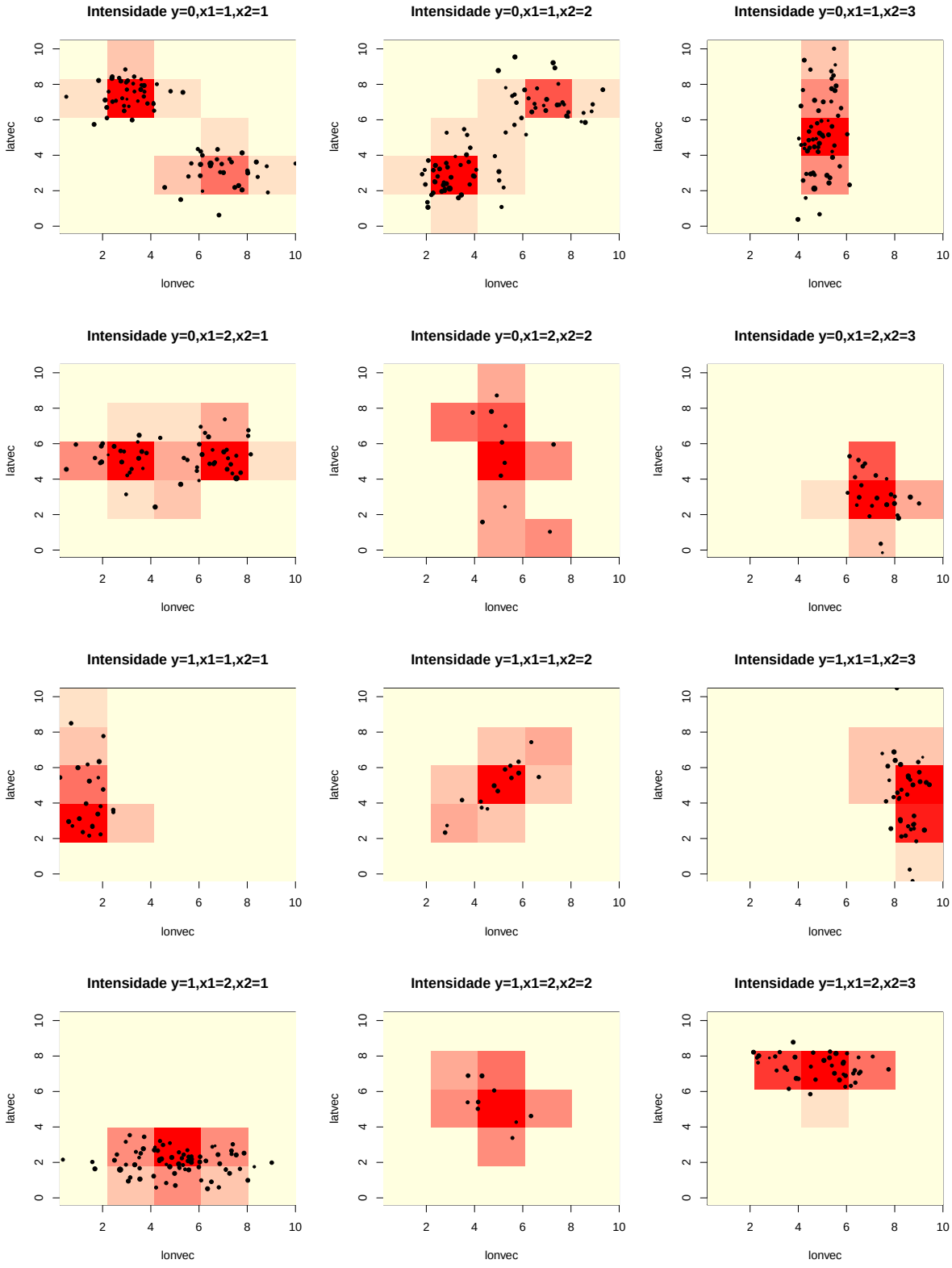

Figura 11 : Intensidades estimadas usando a *priori* Normal com grade de tamanho 5 x 5

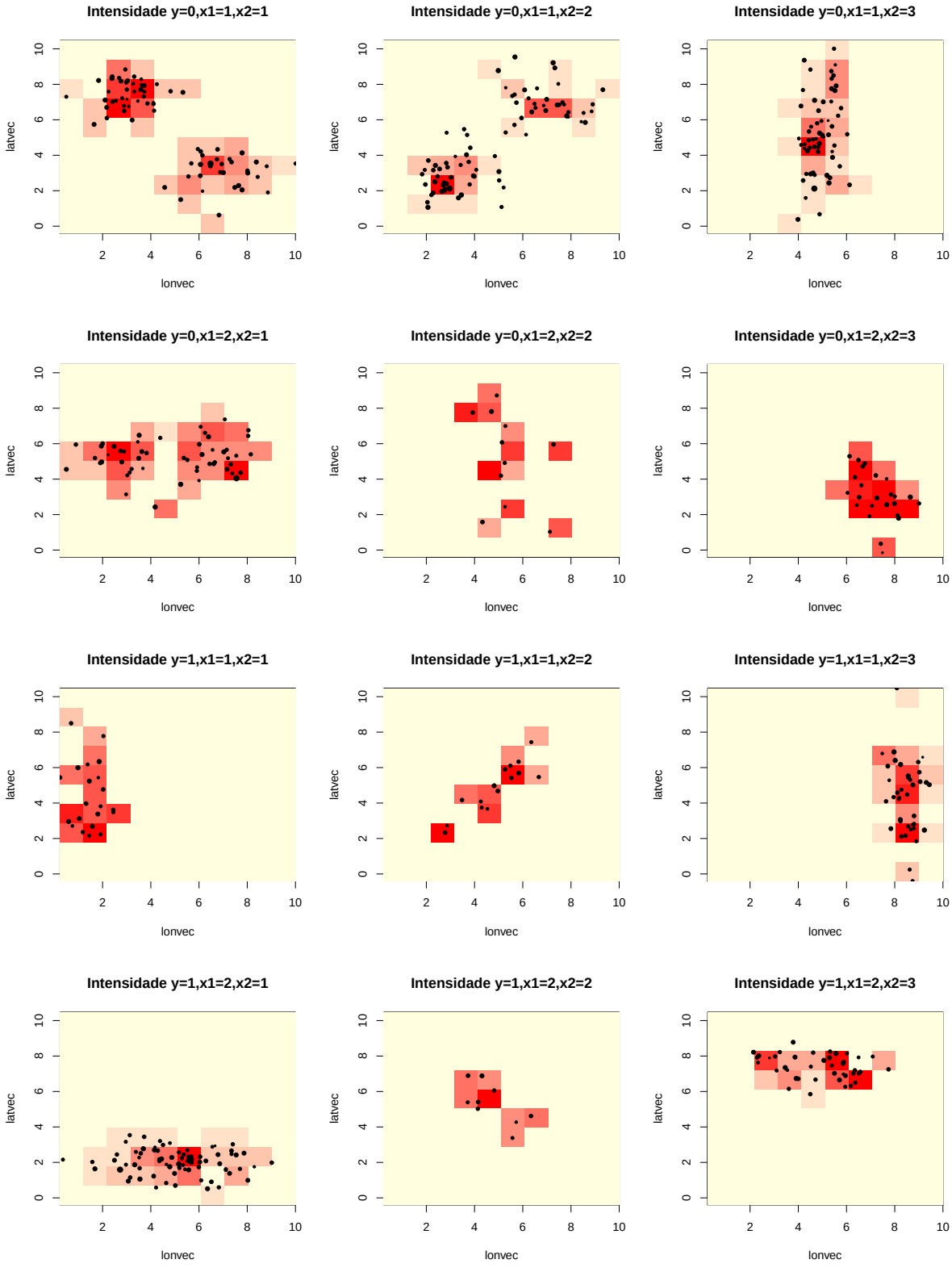

Figura 12 : Intensidades estimadas usando a *priori* Normal com grade de tamanho 10 x 10

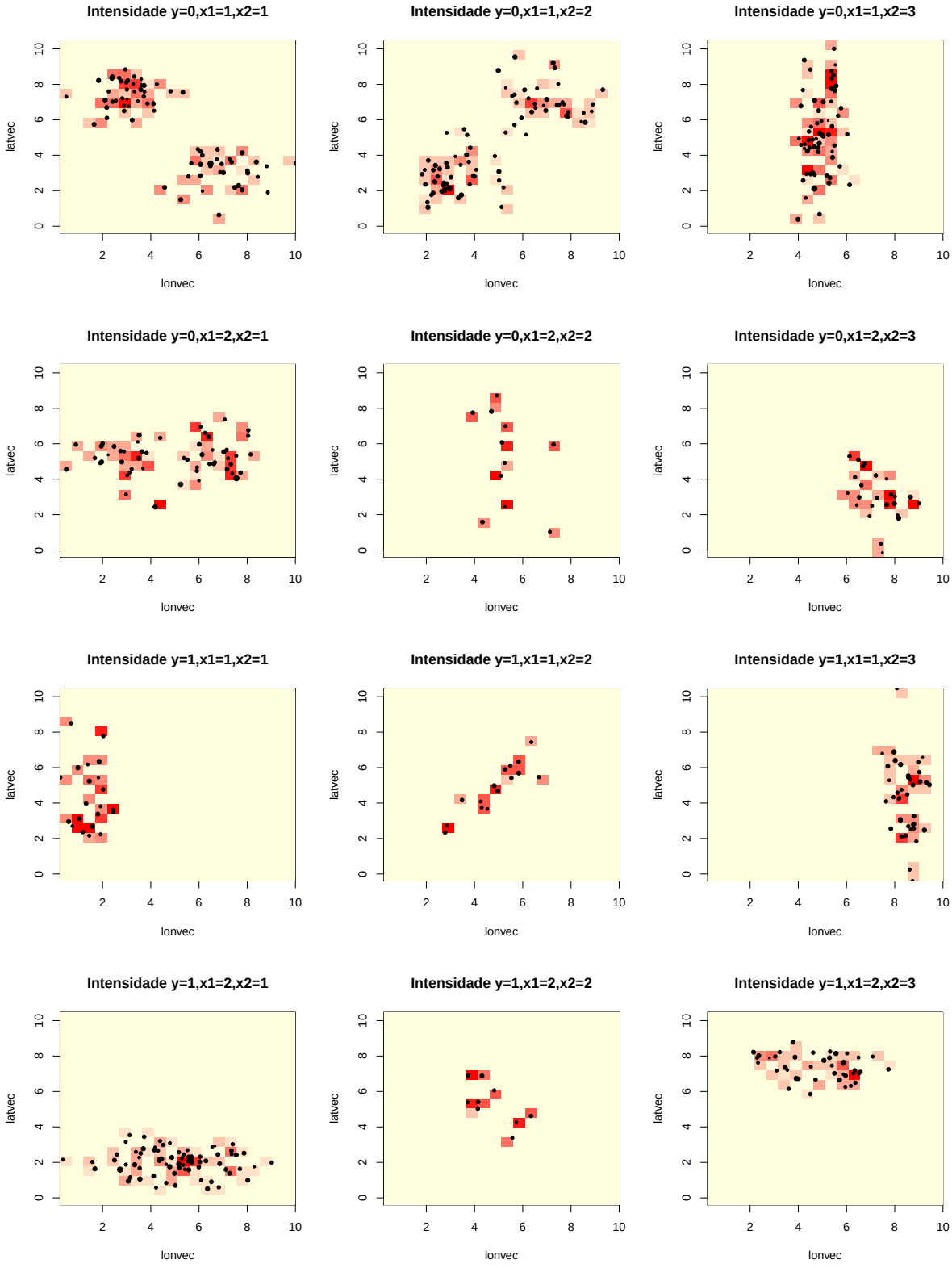

Figura 13 : Intensidades estimadas usando a *priori* Normal com grade de tamanho 20 x 20

Podemos ver que quanto mais fina é a célula da grade, melhor é a estimação de λ, porém maior é o risco de identificação das localizações originais. Verificaremos então, qual o risco e utilidade para cada tamanho de grade fazendo os mesmos procedimentos descritos nas Seções 3.1 e 3.2.

Apresentamos também, as localizações originais e as geradas cada uma das bases sintéticas $m = 1, ..., 5$, para os tamanhos de grade 5 x 5, 10 x 10 e 20 x 20, nas Figuras 14, 15 e 16, respectivamente, para a combinação de atributos $y = 0$, $x_1 = 1$ e $x_2 = 2$, que é a que possui maior número de observações plotados na superfície dos valores médios de λ obtidos por cada um dos métodos.

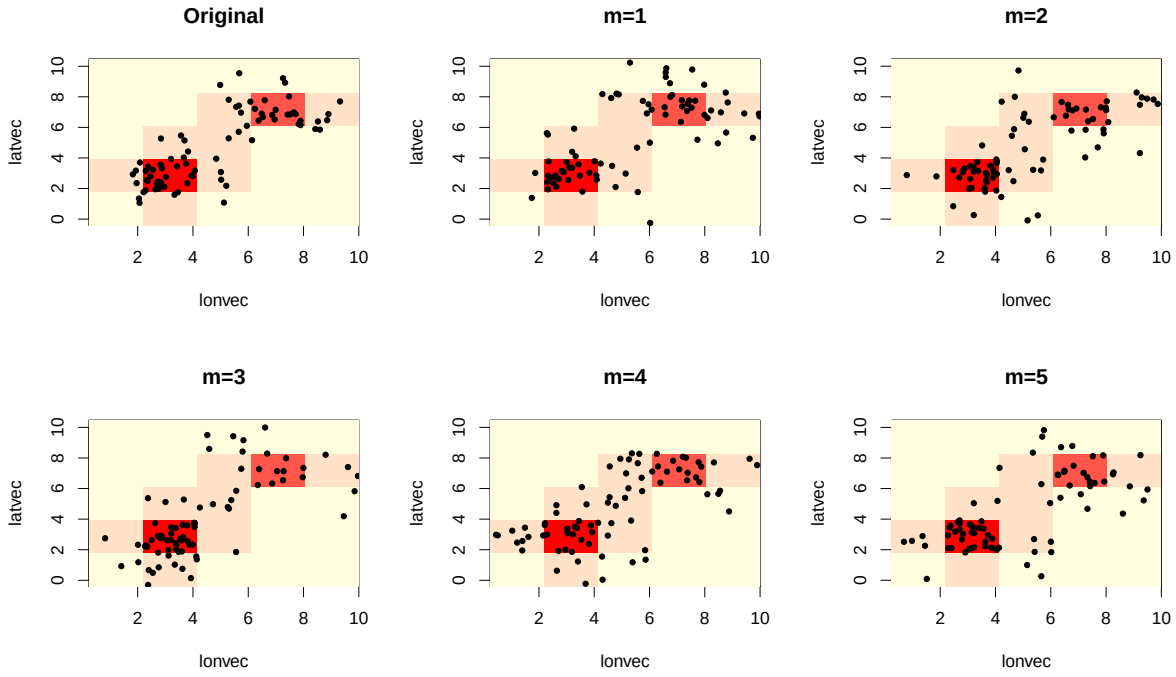

Figura 14 : Localizações originais e sintéticas geradas para o tamanho de grade 5x5 com a *priori* Normal

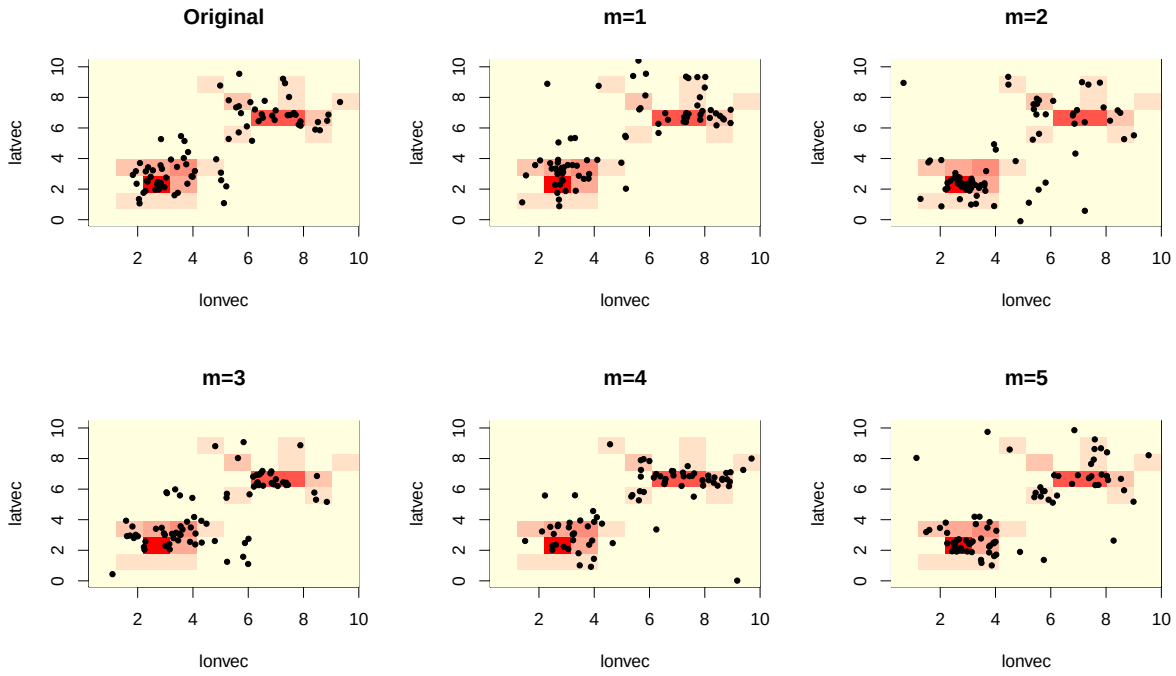

Figura 15 : Localizações originais e sintéticas geradas para o tamanho de grade 10x10 com a *priori* Normal

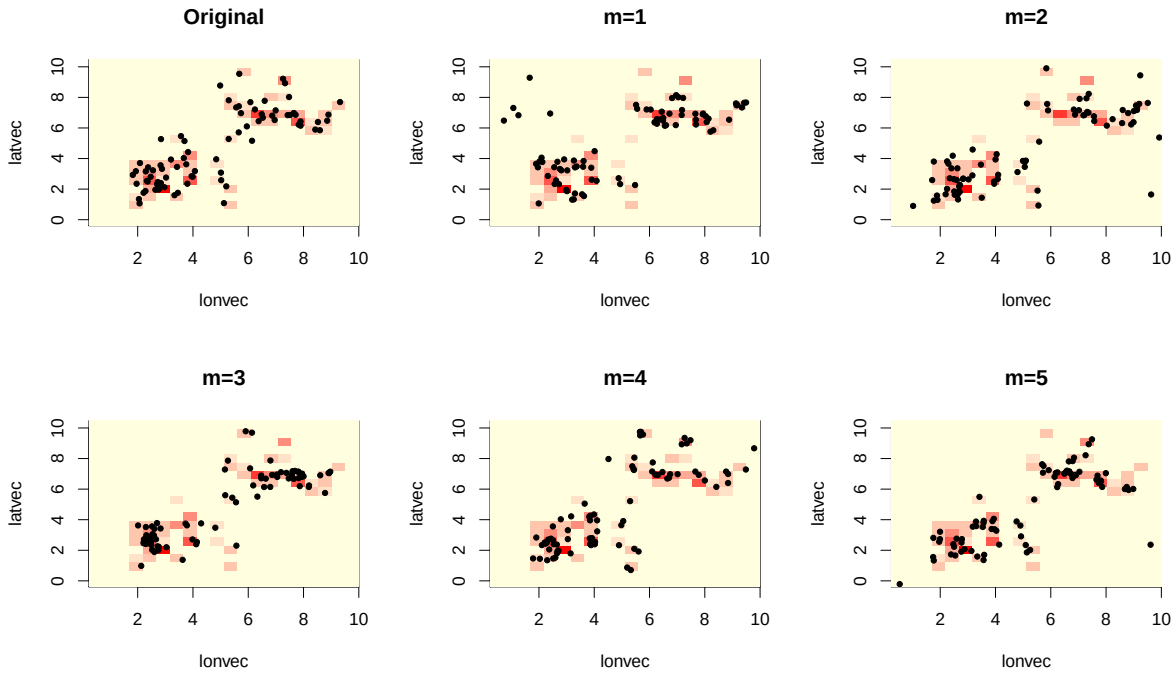

Figura 16 : Localizações originais e sintéticas geradas para o tamanho de grade 20x20 com a *priori* Normal

Podemos ver que quando mais fina é a grade, mais próximas da base original são as observações geradas pelas bases sintéticas. Os gráficos de $m = 1, ..., 5$ da Figura 16, por exemplo, estão bem parecidos com o gráfico original.

O próximo passo é avaliar o risco e, para isso, apresentamos na Tabela 5, as distâncias mínimas e médias para cada tamanho de grade.

Podemos ver na Tabela 5 que, quanto menor o tamanho da célula, menor a distância e consequentemente, maior o risco da divulgação. Isso comprova o que havíamos falado anteriormente, quanto mais fina é a grade, menos protegido está o indivíduo.

Fizemos também os intervalos de confiança de y para o banco de dados original e os bancos de dados sintéticos gerados para cada tamanho de célula usando a metodologia frequentista, e para os parâmetros da regressão usando a metodologia Bayesiana como descrito na Seção 3.2. Na Figura 17 apresentamos o intervalo de confiança de 95% de $\hat{y}$ para o método frequentista para os dois quadrantes mostrados na Figura 4.

Tabela 5 Distâncias mínima e média para cada tamanho de grade com *priori* Normal

	Grade 5x5		**Grade 10x10**		**Grade 20x20**	
B	**Mínimo**	**Média**	**Mínimo**	**Média**	**Mínimo**	**Média**
1	0.0413	1.0981	0.0261	0.9980	0.0296	0.9732
2	0.1135	1.2555	0.0618	1.3934	0.0759	1.5691
3	0.0428	1.0215	0.0359	1.0302	0.0236	0.8542
4	0.1420	1.7441	0.1173	1.0426	0.0608	0.8975
5	0.0263	0.7947	0.0323	0.9575	0.0141	0.8403
6	0.0658	1.0237	0.0498	0.9893	0.0238	0.7435
7	0.0280	0.9700	0.0460	1.1765	0.0169	0.8915
8	0.0292	1.0361	0.0235	0.6916	0.0129	0.6868
9	0.2260	1.4684	0.1839	1.4904	0.0712	1.7970
10	0.3171	1.6341	0.1110	0.9453	0.0739	0.9956
11	0.0767	1.0924	0.0689	1.0569	0.0210	0.9662
12	0.0589	0.9286	0.0294	0.7578	0.0246	0.7247
Média	**0.0973**	**1.1723**	**0.0655**	**1.0441**	**0.0373**	**0.9950**

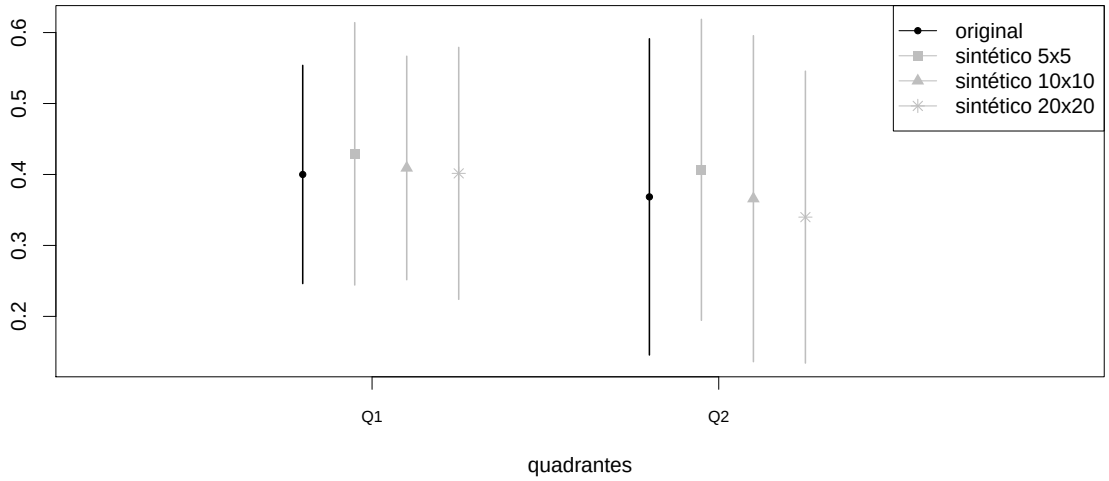

Figura 17 : Intervalo de confiança de 95% para $\hat{y}$ usando o método frequentista

Calculamos também as taxas de cobertura dos intervalos de confiança para y feitos em cada um dos 1000 quadrantes gerados. Para a grade de 5x5, a taxa de cobertura foi de 0,730, para a grade de 10x10 foi 0,676 e para a grade de 20x20 foi de 0,737. Assim, podemos ver que, em média, os intervalos gerados quando o tamanho da grade é 20x20 se aproximaram mais dos intervalos de confiança originais.

Na Figura 18, apresentamos o intervalo de credibilidade de 95% para os parâmetros

da regressão que foi feita usando a metodologia Bayesiana. Geramos 100000 valores para cada parâmetro e excluímos 90000 no período de aquecimento.

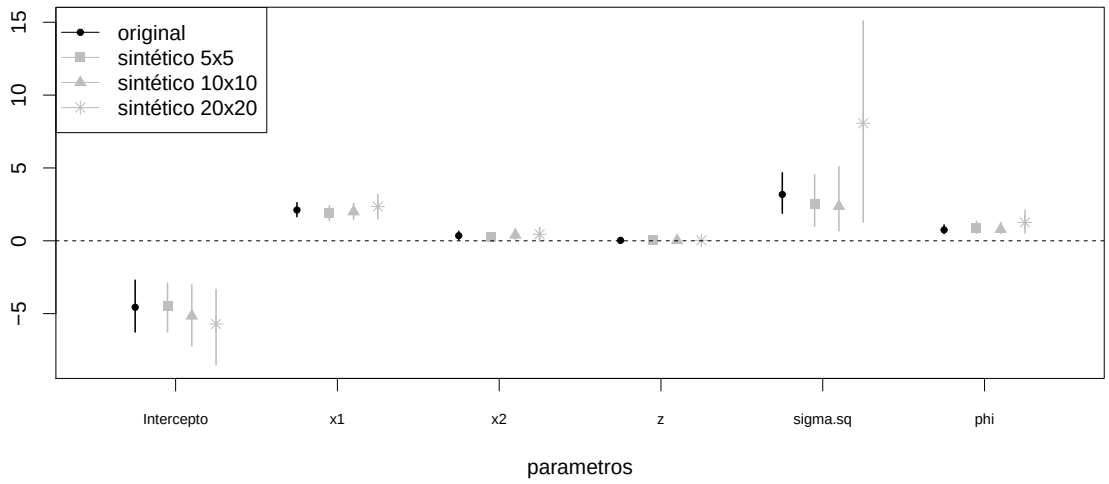

Figura 18 : Intervalo de confiança de 95% usando o método Bayesiano

Na Figura 18, podemos ver que os intervalos de credibilidade feito com as base sintéticas quando o tamanho da grade é 10 x 10 se aproximou mais ao intervalo de credibilidade feito com a base original apesar dos intervalos para todos os parâmetros e bases estarem parecidos. Os intervalos de credibilidade para σ^2 estão um pouco grandes para a grade de 20 x 20 devido a alguma dificuldade de estimação do modelo para essa base. No entanto, estamos somente avaliando se os intervalos dos dados sintéticos se assemelham ao intervalo dos dados originais.

4.2 *Priori* ICAR

Como podemos ter uma correlação espacial para o β, assumimos o modelo ICAR como distribuição *a priori* para este parâmetro. Novamente, para estimar o λ, geramos uma amostra de 15001 da equação 2.19 e descartamos as 5001 primeiras observações. Também consideramos três tamanhos de grade: 5 x 5, 10 x 10 e 20 x 20. Em todas as estimações para λ, os parâmetros convergiram. As intensidades para a grade de tamanho 5 x 5 são apresentadas na Figura 19, 10 x 10 na Figura 20 e 20 x 20 na Figura 21. Os

pontos são as localizações originais e os seus tamanhos variam conforme o valor de Z da observação.

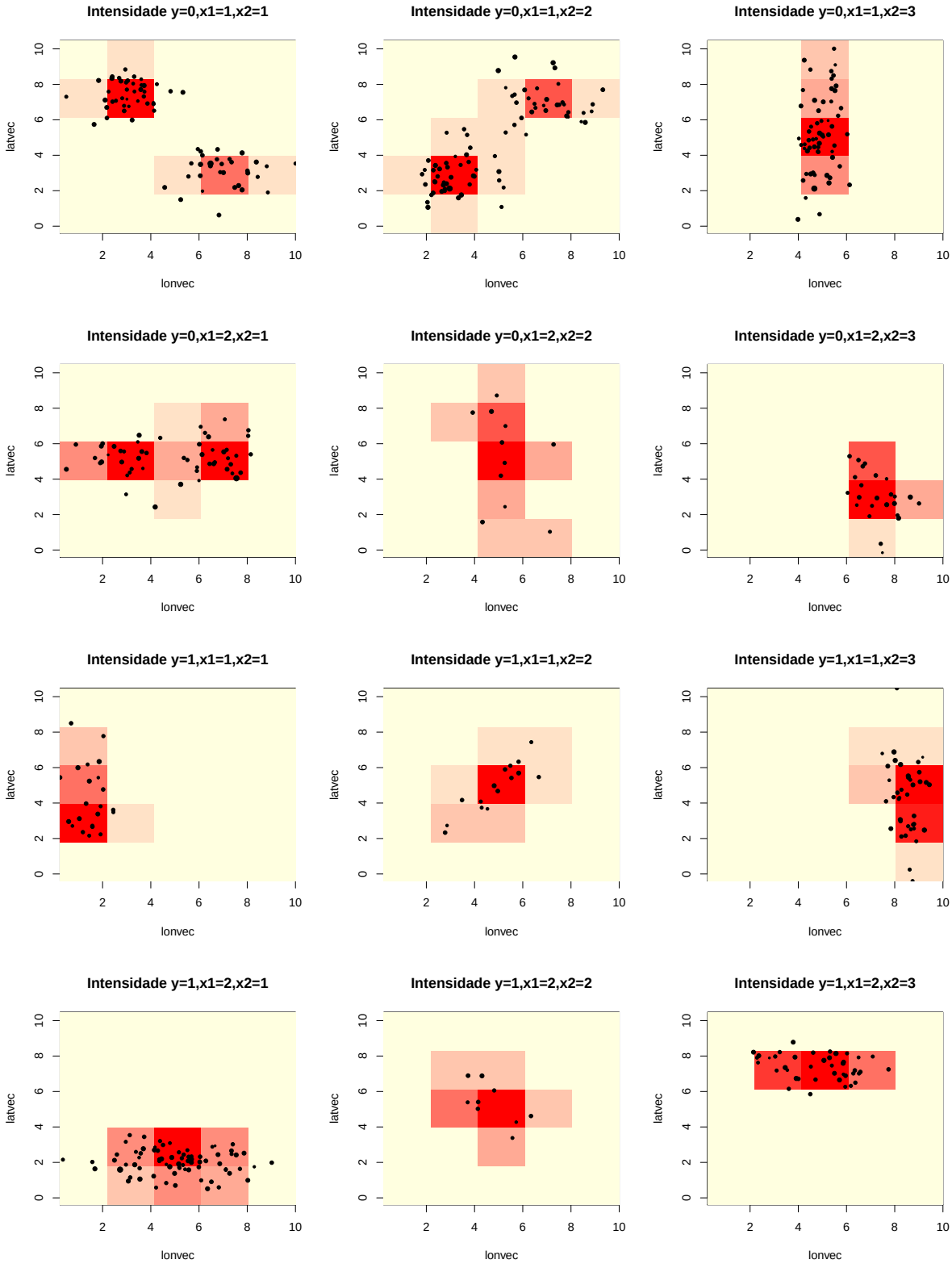

Figura 19 : Intensidades estimadas usando a *priori* ICAR em células de tamanho 5 x 5

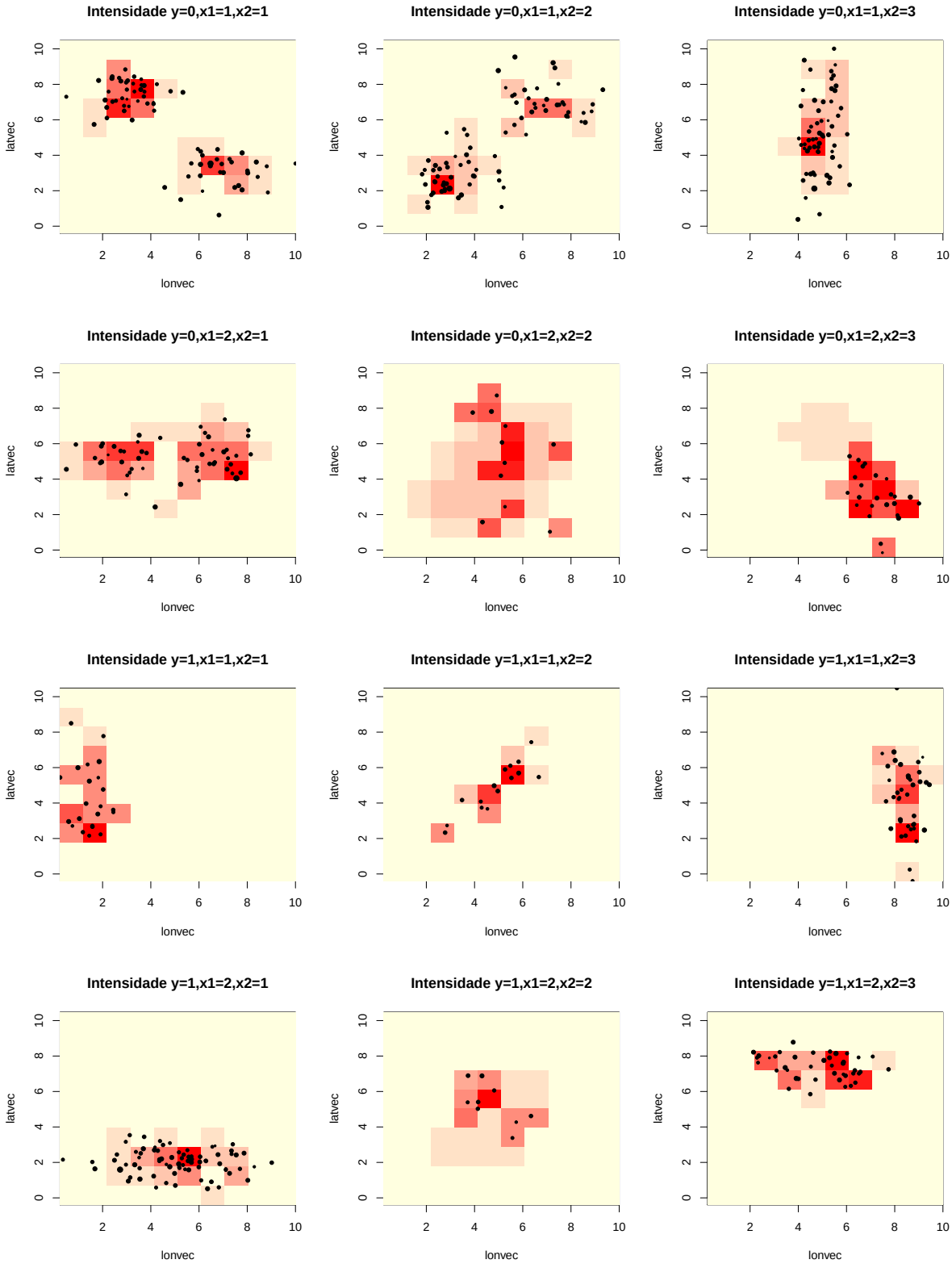

Figura 20 : Intensidades estimadas usando a *priori* ICAR em células de tamanho 10 x 10

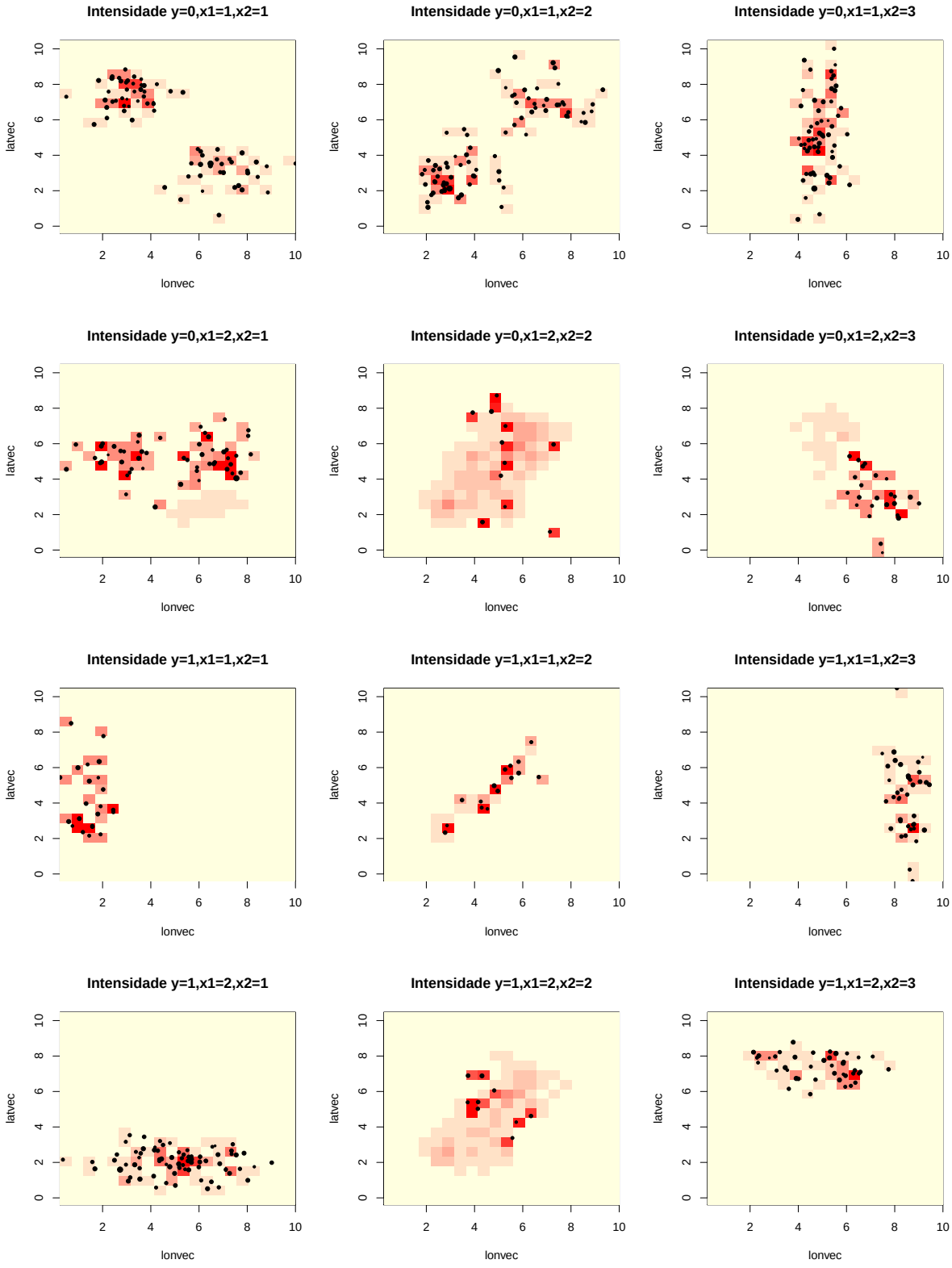

Figura 21 : Intensidades estimadas usando a *priori* ICAR em células de tamanho 20 x 20

Novamente, quanto mais fina é a célula da grade, melhor é a estimação de λ, porém maior é o risco de divulgação. Analisaremos então, qual o risco e utilidade para cada tamanho de grade fazendo os mesmos procedimentos descritos nas Seções 3.1 e 3.2.

Apresentamos também, as localizações originais e as geradas cada uma das bases sintéticas $m = 1, ..., 5$, para os tamanhos de grade 5 x 5, 10 x 10 e 20 x 20, nas Figuras 22, 23 e 24, respectivamente, para a combinação de atributos $y = 0$, $x_1 = 1$ e $x_2 = 2$, que é a que possui maior número de observações plotados na superfície dos valores médios de λ obtidos por cada um dos métodos.

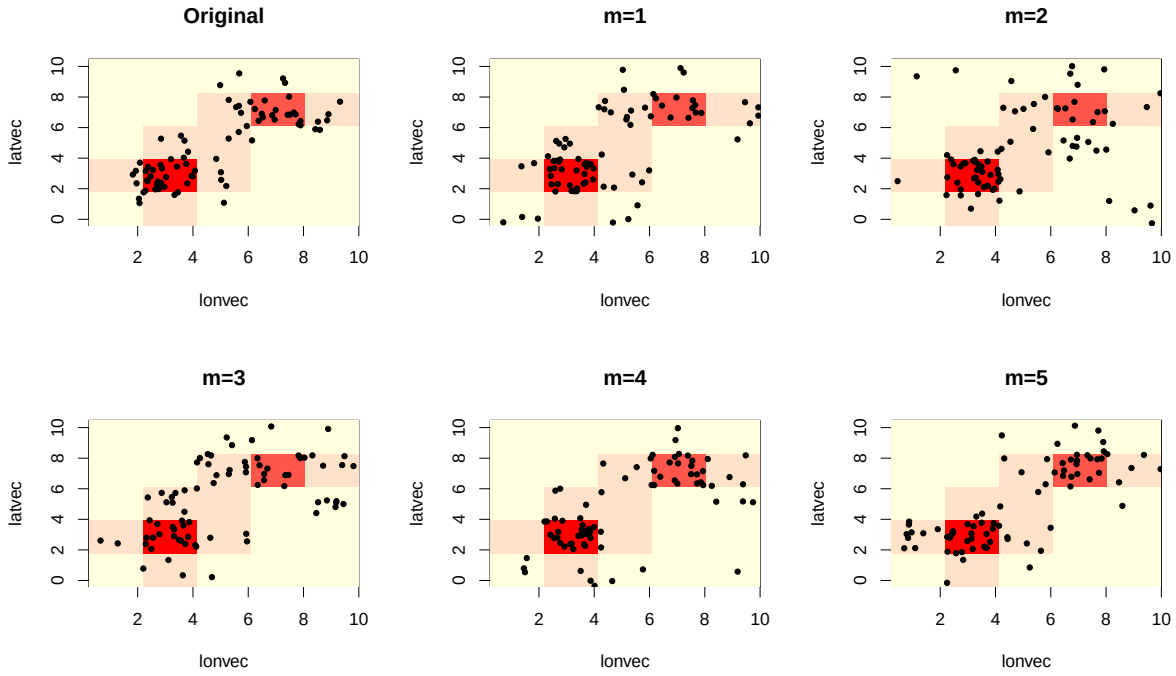

Figura 22 : Localizações originais e sintéticas geradas para o tamanho de grade 5x5 com a *priori* ICAR

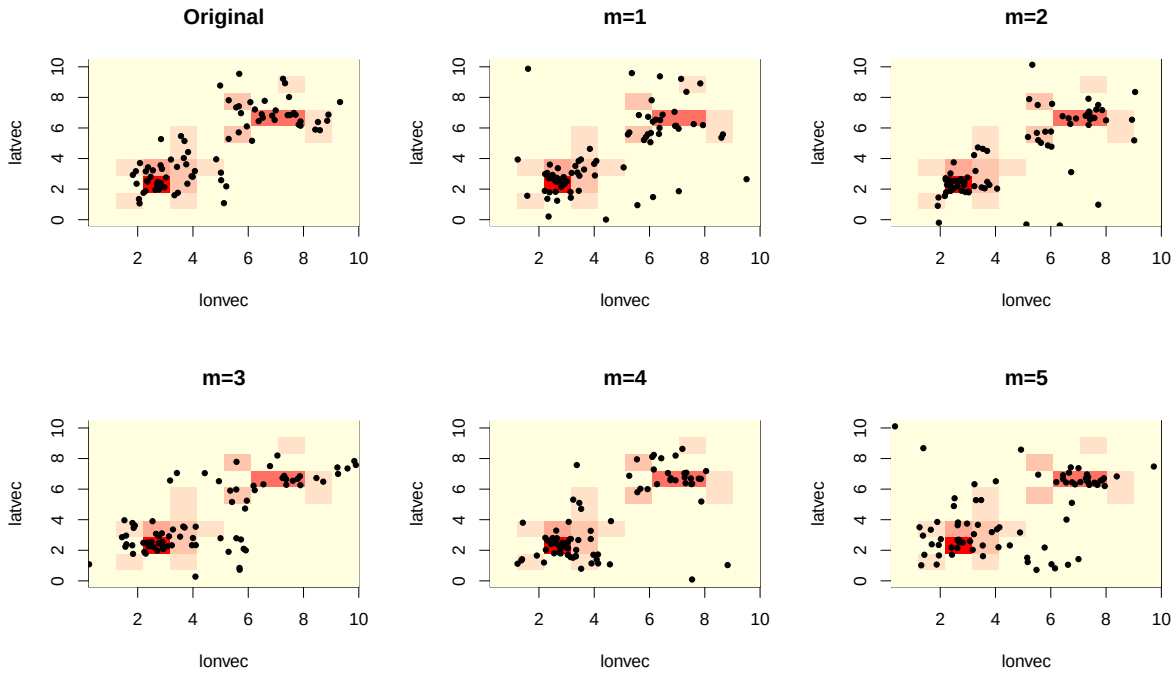

Figura 23 : Localizações originais e sintéticas geradas para o tamanho de grade 10x10 com a *priori* ICAR

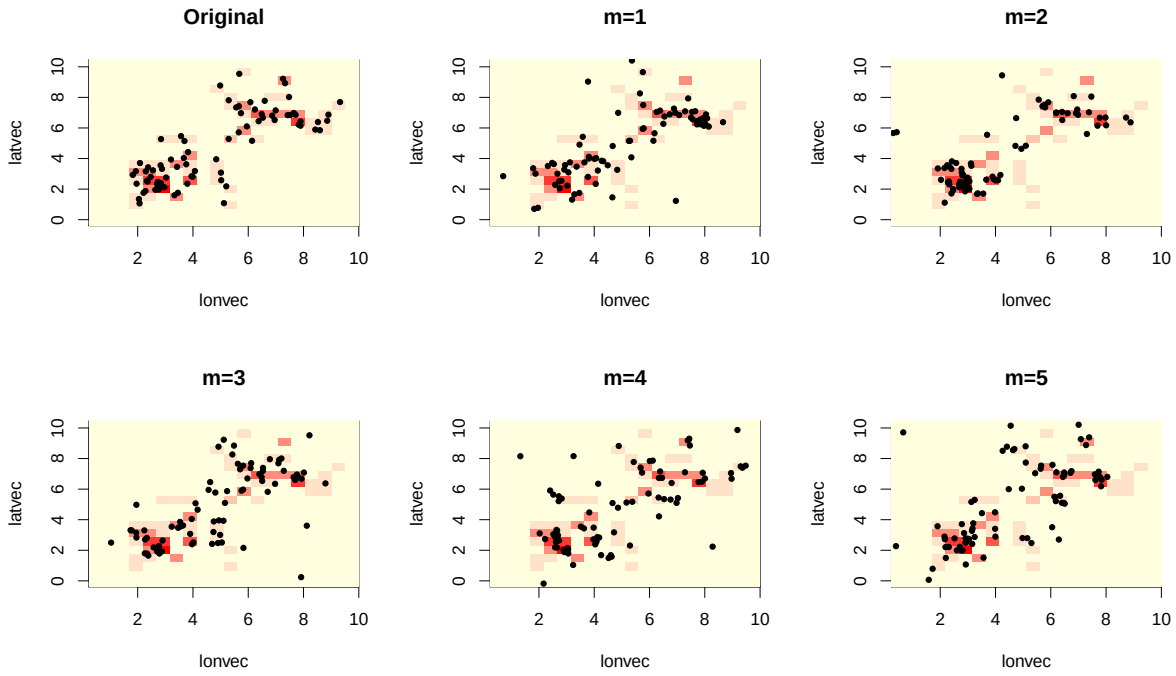

Figura 24 : Localizações originais e sintéticas geradas para o tamanho de grade 20x20 com a *priori* ICAR

Nas Figuras 22 a 24 vemos que as localizações sintéticas geradas para os três tamanhos de grades se aproximam muito das localizações originais.

Avaliaremos agora, o risco que uma agência assumirá ao divulgar os dados. Para isso, apresentamos na Tabela 6, as distâncias mínimas e médias para cada tamanho de grade.

Na Tabela 6, vemos que as distâncias mínimas e médias diminuem conforme a grade fique mais fina, indicando que o risco de uma pessoa mal intencionada identificar um indivíduo é maior. Porém, as distâncias são maiores quando usamos a *priori* ICAR do que quando usamos a *priori* Normal.

Na Figura 25 apresentamos o intervalo de confiança de 95% para $\hat{y}$ usando o método frequentista. Observamos que no quadrante 1 o melhor intervalo de confiança foi o feito com as bases sintéticas geradas com o tamanho da célula da grade 20 x 20, já que ele se assemelha mais ao intervalo de confiança para $\hat{y}$ feito na base original. Já no quadrante 2, o melhor intervalo de confiança foi feito com as bases sintéticas geradas com o tamanho da célula da grade 10 x 10.

Tabela 6 Distâncias mínima e média para cada tamanho de grade com *priori* ICAR

	Grade 5x5		**Grade 10x10**		**Grade 20x20**	
B	**Mínimo**	**Média**	**Mínimo**	**Média**	**Mínimo**	**Média**
1	0.0405	1.2290	0.0467	1.1059	0.0209	1.3203
2	0.0923	1.6371	0.1069	2.5584	0.1047	3.1110
3	0.0341	1.0212	0.0185	1.1129	0.0173	0.9065
4	0.0768	1.6733	0.1669	1.8520	0.0849	1.4662
5	0.0356	0.8407	0.0172	1.1114	0.0356	1.3999
6	0.0676	1.4947	0.0381	1.5874	0.0309	1.5389
7	0.0556	1.0140	0.0457	1.2040	0.0401	1.4927
8	0.0526	0.9275	0.0308	0.9742	0.0188	1.2176
9	0.2747	2.0850	0.3353	2.4162	0.2243	2.0362
10	0.4035	1.7733	0.3772	2.2918	0.2522	2.2800
11	0.0772	1.8050	0.0975	2.0689	0.0736	1.9174
12	0.0434	0.9712	0.0486	1.2187	0.0208	1.6042
Média	**0.1045**	**1.3727**	**0.1108**	**1.6251**	**0.0770**	**1.6909**

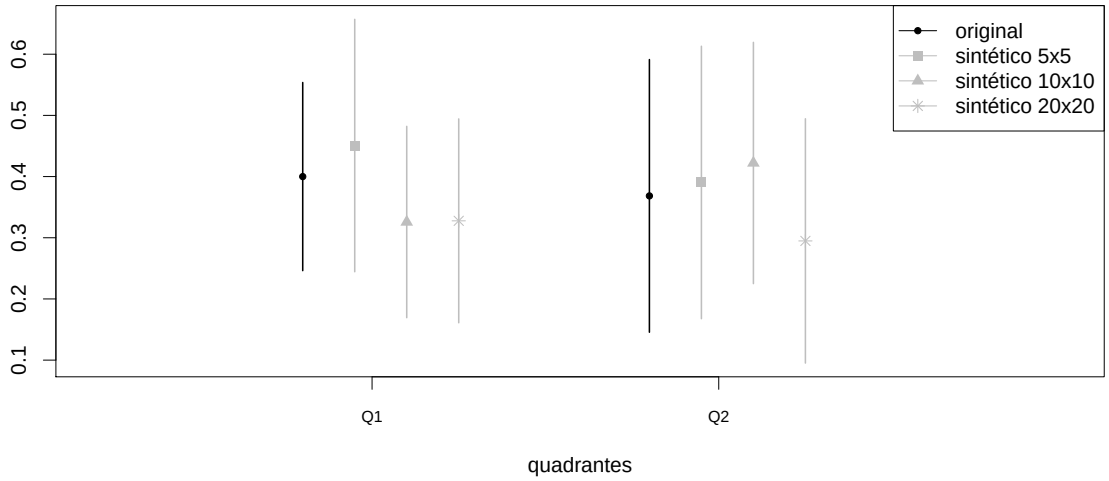

Figura 25 : Intervalo de confiança de 95% para y usando o método frequentista

Calculamos também as taxas de cobertura dos intervalos de confiança para y feitos em cada um dos 1000 quadrantes gerados. Para a grade de 5x5, a taxa de cobertura foi de 0,665, para a grade de 10x10 foi 0,746 e para a grade de 20x20 foi de 0,668. Assim, podemos ver que os intervalos de gerados quando o tamanho da grade é 10x10 se aproximaram mais dos intervalos originais.

Na Figura 26, apresentamos o intervalo de credibilidade de 95% para os parâmetros

da regressão que foi feito usando a metodologia Bayesiana quando usamos o modelo ICAR como distribuição *a priori* para β. Geramos 100000 valores para cada parâmetro e excluímos 90000 no período de aquecimento.

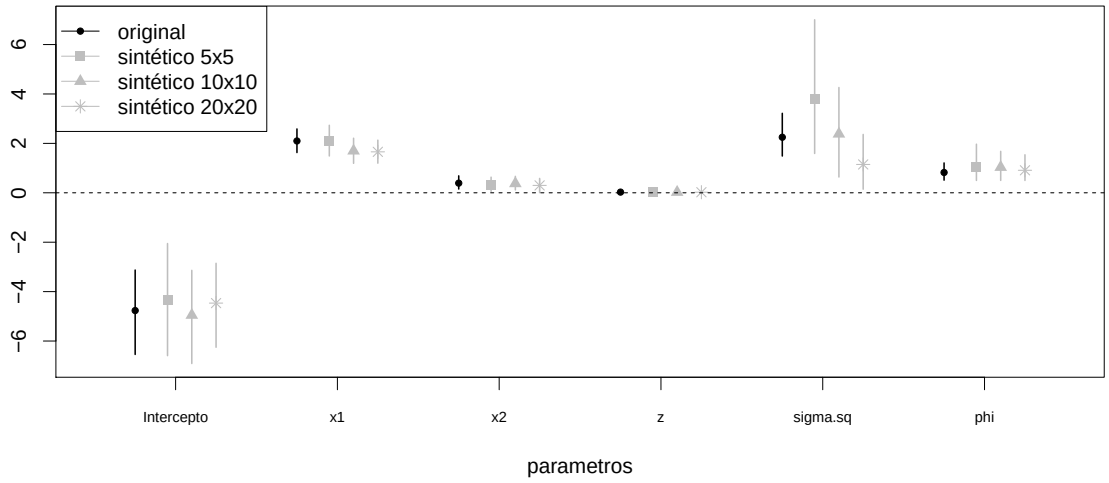

Figura 26 : Intervalo de credibilidade de 95% usando o método Bayesiano

Podemos ver na Figura 26 que a base de dados sintética que apresentou uma melhor utilidade é a base sintética gerada com o tamanho da grade 20 x 20.

No Capítulo 5 discutiremos os modelos até aqui vistos e concluiremos apontando o melhor modelo. Além disso, falaremos sobre as propostas de trabalho futuro.

5 DISCUSSÕES

Neste trabalho, apresentamos a proposta de Paiva et al. (2014), avaliamos este modelo e estendemos ao colocarmos uma variável contínua na estimação de $\lambda_i^{(b)}$. Além disso, propomos duas distribuições *a priori* para o parâmetro β que acompanha a variável contínua. Podemos estender este modelo quando temos mais de uma variável continua sem perda de generalidade.

Na avaliação do modelo, percebemos que o modelo proposto por Paiva et al. (2014) é melhor quando avaliamos o seu risco em conjunto com a utilidade quando comparado com o modelo Naive, que é simplista. Logo, poderíamos estendê-lo colocando a variável contínua na estimação do $\lambda_i^{(b)}$.

Então, propusemos estendê-lo colocando uma função flexível na estimação de $\lambda_i^{(b)}$, a qual dependia da variável contínua como apresentado em (2.19). Para isso, era necessário assumir uma distribuição *a priori* para β. Assumimos dois modelos: distribuição Normal e modelo ICAR. Além disso, sabemos que o tamanho da célula da grade é importante ao avaliarmos o risco e a utilidade das bases sintéticas geradas por esses modelos. Sendo assim, geramos estas bases sintéticas assumindo três tamanhos de grade: 5 x 5, 10 x 10 e 20 x 20.

Entre estes modelos, podemos ver que quanto mais fina é a grade, maior é o risco de uma pessoa mal intencionada identificar as localizações originais e melhor é a sua utilidade. Por outro lado, quanto maior o tamanho da célula, menor é este risco e menor é a sua utilidade. Cabe à agência que coletou os dados avaliar qual o risco está disposta a aceitar de acordo com cada nível de utilidade.

Sendo assim, como trabalho futuro, propomos encontrar o tamanho de grade ótimo através de simulações, incluindo um parâmetro adicional no modelo que corresponde ao tamanho ideal da grade. Então, o risco de encontrarem as localizações dos indivíduos estudados e a utilidade dos bancos de dados gerados serão adequadas.

Além disso, suponhamos que, ao simular as localizações sintéticas, podemos estar gerando coordenadas sintéticas em regiões como parques, lagos, etc., onde não é possível residir. Sendo assim, uma proposta de trabalho futuro é ajustar o modelo de tal forma que as localizações simuladas possuam restrições.

Por fim, queremos avaliar a possibilidade de aplicar esse método a uma base de dados real que contenha as informações geográficas necessárias para aplicarmos os modelos propostos neste trabalho.

REFERÊNCIAS

Alireza, S. M. and T. Mansour (2017). *Multivariate-from-Univariate (MfU) MCMC Sampler*.

An, D., R. Little, and J. McNally (2010). A multiple imputation approach to disclosure limitation for high-age individuals in longitudinal studies. *Statistics in Medicine 29*, 1769–1778.

Armstrong, M., G. Rushton, and D. Zimmerman (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine 18*, 497–525.

Banerjee, S., A. Gelfand, and B. Carlin (2004). *Hierarchical Modeling and Analysis for Spatial Data*. New York,NY: Chapman and Hall/CRC.

Cassa, C., S. Grannis, M. Overhage, and K. Mandl (2006). A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *13*, 160–165.

Duncan, G. and D. Lambert (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics 7*, 207–217.

Gilks, W. and P. Wild (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society 41*, 337–348.

Karr, A. and J. Reiter (2014). Using statistics to protect privacy, in privacy, big data, and the public good: Frameworks for engagement. *Cambridge University Press*, 276–295.

Kounadi, O. and M. Leitner (2015). Defining a threshold value for maximun spatial information loss of masked geo-data. *4*, 572–590.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: Theory meets practice on the map. Cancun, Mexico, pp. 277–286.

Neal, R. (2003). Slice sampling. *The Annals of Statistics 31*, 705–767.

Paiva, T., A. Chakraborty, J. Reiter, and A. Gelfand (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine 33*, 1928–1945.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raghunathan, T., J. Reiter, and D. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Oficial Statistics 19*, 1–16.

Reiter, J. (2004). New approaches to data dissemination: A glimpse into the future (?). *Chance 17*, 12–16.

Sherman, J. E. and T. Fetters (2007). Confidentiality concerns with mapping surbey data in reproductive health research. *Studies in Family Planning 38*, 309–321.

Wang, H. and J. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics 6*, 229–252.

Willenborg, L. and T. De Waal (2001). *Elements of Statistics Disclosure Control*. New York: Springer-Verlag.

Zhou, X. and J. Reiter (2010). A note on bayesian inference after multiple imputation. *The American Statistician 64*, 159–163.

Zhou, Y., F. Dominici, and T. Louis (2010). A smoothing approach for masking spatial data. *Annals of Applied Statistics 4*, 1451–1475.