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

Larissa Natany Almeida Martins

Imputação de Dados Sintéticos através de Árvores de Classificação

Belo Horizonte
2019

Larissa Natany Almeida Martins

Imputação de Dados Sintéticos através de Árvores de Classificação

Dissertação apresentada, como requisito 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 Galetti

Janeiro
2019

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, que permitiu que tudo isso acontecesse. Obrigada pela força, saúde e esperança que sempre me acompanharam.

À minha mãe Ana Tércia, que sempre foi um exemplo de vida e de perseverança. Obrigada pelas suas palavras de encorajamento e amor.

Aos meus irmãos Wallace e Mateus, pelo companheirismo,carinho, incentivo e apoio.

À toda minha família (avó, tios, tias, primos, primas, etc), pela presença constante em minha vida.

Ao meu querido Alex, pelo amor e por todos os momentos de incentivo, e ajuda ao longo desses anos de estudo.

Aos amigos e companheiros de disciplinas nesses dois anos de mestrado.

À minha orientadora, Thaís Paiva, pela oportunidade, orientação e, principalmente, paciência e dedicação em me ajudar em todo o trabalho.

Aos meus queridos amigos Vicente, Vanderlei e Carla, pelo incentivo e orações ao longo desses anos de estudo.

À CAPES, pelo financiamento desta pesquisa.

À FUMP, pelo apoio financeiro ao longo desses anos de estudo.

Aos Professores da UFMG com que tive mais contato, sejam pelas aulas ou pelo contato em algum projeto.

À esta Universidade, seu corpo docente, funcionários, direção e administração, pela oportunidade de fazer o curso.

Enfim, muito obrigada a todas e todos que me apoiaram em mais esta jornada

Resumo

N. A Martins , Larissa. *Imputação de Dados Sintéticos através de Árvores de Classificação*. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, 2019.

Este trabalho apresenta um estudo sobre a metodologia de geração de dados sintéticos através de árvores de classificação e regressão. Essa metodologia é usada quando existe alguma restrição na divulgação de informações sigilosas por questões éticas ou morais e existe o interesse em divulgar essas informações de maneira segura. Dados sintéticos utilizam a ideia de imputação múltipla, onde os valores originais são imputados por novos valores baseados nas distribuições das variáveis envolvidas no estudo. Várias metodologias podem ser utilizadas para a geração de dados sintéticos. Nesse trabalho utilizamos árvores de classificação e regressão (CART) para a classificação dos grupos envolvidos no estudo, o bootstrap Bayesiano para a estimação da densidade de cada grupo e o método da CDF inversa para a geração final dos dados sintéticos. O objetivo desse trabalho é estender a metodologia utilizada por Reiter e Drechsler (2011) para geração de dados sintéticos utilizando modelos não paramétricos para diferentes distribuições da variável sensível, incluindo o caso de distribuições com caudas pesadas. Iremos também apresentar o cálculo para medida de risco para diferentes hipóteses sobre a informação que um possível intruso possa possuir. Apresentamos a geração dos dados sintéticos para três cenários simulados com distribuições diferentes para verificar a eficiência do modelo. Também foi analisado um banco de dados real. Para os cenários simulados, o cenário 2 apresentou resultados piores do que os cenários 1 e 3, devido a distribuição da variável resposta. Para o banco de dados real os resultados foram considerados satisfatórios.

Palavras Chave: Dados Sintéticos, CART, Divulgação de Dados.

Abstract

This work presents a study on the methodology of synthetic data generation through classification and regression trees. This methodology is used when there is any restriction on disclosure of sensitive information for ethical or moral reasons and there is an interest in disclosing such information. Synthetic data use the idea of multiple imputation, where the original values are imputed by new values based on the distributions of the variables involved in the study. Several methodologies can be used to generate synthetic data. In this work we used classification and regression trees (CART) to classify the groups involved in the study, the Bayesian bootstrap to estimate the density of each group and the inverse CDF method for the final generation of synthetic data. The objective of this work is to extend the methodology used by Reiter and Drechsler (2011) to generate synthetic data using non-parametric models for different distributions of the sensitive variable, including the case of distributions with heavy tails. We will also present the calculation to measure risk for different hypotheses about the information that a possible intruder may have. We present the generation of synthetic data for three simulated scenarios with different distributions to verify the efficiency of the model. We also analyzed a real database. For the simulated scenarios, scenario 2 presented worse results than scenarios 1 and 3, due to the distribution of the response variable. For the real database, the results were considered satisfactory.

Keywords: Synthetic Data, CART, Disclosure data.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A divulgação segura de dados tem sido uma das principais preocupações de várias instituições, como organizações de pesquisas, empresas, órgãos governamentais, entre outras. Essas agências têm o interesse em liberar essas bases de dados, mas em alguns casos podem existir restrições éticas e legais que impedem a divulgação dos dados de pessoas ou até mesmo de empresas. Dessa forma, agências vêm investindo em pesquisas para melhorar a forma de divulgar bases de dados para serem analisadas, mas também mantendo o foco na proteção das informações contidas nas bases de dados (Karr e Reiter, 2013).

O risco de divulgação de informações sigilosas é uma das principais preocupações das agências, visto que a identificação de informações das pessoas podem resultar em danos muito grandes. Usuários mal intencionados podem ser capazes de fazer algum tipo de combinação entre outras bases de dados para conseguir identificar a quem pertencem os dados (Drechsler e Reiter, 2011). Sweeney (2013) apresenta um estudo onde 97% dos registros nas listas de cadastro eleitoral disponíveis em Cambridge, Massachusetts, poderiam ser identificados de maneira exclusiva com data de nascimento e código postal de nove dígitos. Ao combinar as informações nessas listas, ela pôde identificar governadores de Massachusetts em um banco de dados supostamente anônimo.

Além do risco de divulgação é importante medir a utilidade dos dados divulgados. A utilidade dos dados pode ser interpretada como a semelhança entre o banco de dados original e o banco de dados divulgados para uso público. Num cenário ideal, as inferências e análises que são feitas no banco de dados original, também podem ser realizadas nos bancos ou informações divulgadas. Além disso, os resultados das inferências nos dois bancos de dados, real e divulgado, devem ser próximas.

Sendo assim, agências divulgam bancos de dados com informações sigilosas sempre tendo em vista o risco de divulgação e a utilidade dos dados. Alguns métodos já

existentes para a divulgação dessas informações possuem um risco muito baixo, mas em contrapartida sua utilidade é muito ruim. Então, a análise de risco e utilidade é essencial para a divulgação de bancos de dados com informações sigilosas.

Ao divulgar um conjunto de dados para uso público, agências empregam diversas técnicas para limitar a divulgação de informações sigilosas. Dentre as técnicas mais comuns, estão a exclusão e re-codificação das informações mais importantes do banco de dados. O problema é que esses métodos podem distorcer as relações entre as variáveis, fazendo com que as análises dos bancos alterados não sejam confiáveis. Essas análises podem ser desde análises descritivas dos dados, quanto inferências sobre parâmetros, testes de hipóteses e estimação de intervalos de confiança, dentre outras. Logo, o método de divulgação de informações sigilosas impacta diretamente nas inferências desses bancos de dados.

Com a crescente demanda por métodos para a divulgação segura de dados, novos métodos vem sendo criados para tentar sanar os problemas encontrados pelas agências, controlar o risco de divulgação de informações sigilosas, e disponibilizar banco de dados tratáveis para serem analisados.

A técnica utilizada para essa divulgação pode estar ligada à natureza das variáveis envolvidas e até mesmo às características do banco de dados. Karr e Reiter (2013) apresentam uma discussão sobre os problemas que agências tem enfrentado para a divulgação de informações sigilosas. Dependendo da natureza das variáveis envolvidas na divulgação de informação, se são variável contínuas, discretas, binárias ou até mesmo sobre a distribuição dessas variáveis, o método de divulgação pode mudar. Logo, para cada banco de dados a ser divulgado é interessante o estudo sobre a natureza das variáveis e a identificação do melhor modelo de divulgação do banco de dados.

O objetivo desse trabalho é utilizar a abordagem da geração de dados sintéticos através do modelo de árvore de classificação e regressão (CART) e criar novas bases de dados que possam ser utilizadas por pesquisadores, agências, empresas e outros usuários, mantendo a privacidade dos dados originais. Iremos utilizar a metodologia descrita por Drechsler e Reiter (2011) para cenários onde a distribuição da variável que não pode ser divulgada tem cauda pesada. Comparamos 3 cenários com distribuições diferentes para verificar como o modelo se comporta quando uma variável possui maior variação. Faremos a comparação entre os cenários através das árvores geradas pelo CART bem como a comparação entre as medidas de utilidade e risco. Para a medida de risco apresentados alguns cenários de acordo com algumas hipóteses sobre a informação dobre o banco de dados que um possível intruso pode possuir.

A seguir apresentamos nas Seções 1.1, 1.2 e 1.3 as metodologias utilizadas para a divulgação de dados sigilosos, uma introdução sobre dados sintéticos e um modelo não paramétrico utilizado nesse trabalho. No Capítulo 2 apresentamos a metodologia de geração de dados sintéticos. No Capítulo 3, apresentamos as simulações para os três cenários distintos, e Capítulo 4, aplicamos a metodologia para um base de dados real. No capítulo 5, apresentamos uma discussão sobre os resultados da extensão do modelo proposto por Reiter (2003).

1.1 Técnicas para Limitação de Divulgação

Algumas agências utilizam métodos estatísticos para a divulgação segura de dados. Karr e Reiter (2013) apresentam algumas das abordagens mais comuns no tratamento de dados a serem divulgados, entre elas temos:

1. Agregação: várias informações são condensadas em apenas uma. Por exemplo, em um banco de dados onde existem informações ligadas ao endereço de usuários, divulga-se as informações agregadas por bairro ou região. Essa técnica é frequentemente utilizada em dados divulgados pelo IBGE (Instituto Brasileiro de Geografia e Estatística), onde as informações de uma determinada pesquisa são divulgadas por setor censitário (unidade territorial estabelecida para fins de controle cadastral), ou seja, não é possível localizar os usuários daquela região, mas apenas se tem a informação da região como um todo. Esse tipo de abordagem, embora proteja as informações, torna análises mais detalhadas difíceis e muitas vezes impossíveis.

2. Troca de dados: valores de variáveis sensíveis são trocados entre pares de registros semelhantes. Por exemplo, para duas pessoas com a mesma idade, seus salários são trocados para dificultar a identificação. Essa abordagem geralmente destrói as relações verdadeiras entre as variáveis, podendo modificar e até mesmo levar a resultados incorretos na análise dos dados. Além disso, a escolha dos pares de registros a serem trocados pode não ser uma tarefa fácil e pode conter vícios, prejudicando assim as inferências (Drechsler e Reiter, 2011).

3. Ruído Aleatório: para cada variável contínua a ser protegida em um banco de dados, adiciona-se quantidades aleatórias aos valores reais observados. Por exemplo, adicionar um ruído com distribuição Normal com média zero para modificar valores mais sensíveis à identificação. Dessa forma, quanto maior a variância atribuída à distribuição, maior será a proteção dos dados. Entretanto, adicionar ruídos aleatórios distorce as distribuições originais dos dados afetando também a análise dos dados.

4. Supressão: valores de alto risco são excluídos do banco a ser divulgado (Cox,

1980). Por exemplo, em uma base de dados com informações sobre o faturamento de empresas, exclui-se os valores de empresas que destoam das demais, como empresas com grandes faturamentos ou faturamentos muito pequenos, dificultando assim a identificação de empresas. A supressão faz com que a exclusão de valores seja não aleatória, ou seja, os dados faltantes no banco de dados não são aleatórios, comprometendo e dificultando a análise de dados.

5. Dados sintéticos: algumas variáveis da base de dados são substituídas por valores sintéticos gerados através de simulações das distribuições de probabilidade empírica dos dados originais (Reiter e Raghunathan, 2007). Dessa forma, os valores divulgados são simulados, ou seja, nenhuma informação verdadeira do banco de dados original é divulgada, mas as relações originais entre as variáveis são mantidas. Assim, além de proteger os dados originais, possibilitam inferências válidas e melhores conclusões acerca da base de dados original.

Dentre as várias metodologias existentes para a divulgação mais segura de dados, todas apresentam vantagens e desvantagens dependendo do tipo de informação a ser protegida. Por isso, o método utilizado pode impactar diretamente no risco de divulgação e também na utilidade desses dados.

Dessa forma, o método de dados sintéticos foi escolhido para o estudo de divulgação de informações sigilosas nesse trabalho. Esse método é interessante, pois mais de uma técnica pode ser utilizada para sua criação baseando-se na estrutura dos dados originais. O objetivo é preservar ao máximo as relações contidas no banco de dados original, o que possibilita que a inferência com os dados sintéticos divulgados seja parecida com as inferências realizadas no banco de dados original. Além disso, como os dados sintéticos divulgados são valores simulados, é possível controlar o risco de identificação dos valores originais. Apresentamos a seguir a metodologia de criação de dados sintéticos para divulgação de informações sigilosas.

1.2 Metodologia para geração de Dados Sintéticos

A metodologia para a geração de dados sintéticos pode estar associada ao tipo de variável contida no banco de dados. Algumas técnicas se adaptam melhor quando a variável, sensível é contínua ou discreta. A natureza das demais variáveis do banco de dados, aquelas que não possuem risco de divulgação, também é importante para a escolha do modelo para a geração dos dados sintéticos.

Reiter e Dreschsler (2011) apresentam um estudo de comparação de algumas metodologias não paramétricas para a geração dos dados sintéticos. Os métodos não-paramétricos foram utilizados por Reiter e Drechsler (2011) pois esses modelos

estimam melhor as relações não lineares presentes entre as variáveis, o que melhora a estimação de determinadas medidas. Os quatro modelos analisados foram árvores de classificação e regressão (CART) (Breiman, 1984), florestas aleatórias(Breiman, 2001), *bagging* (Breiman, 1996) e *support vector machines*(SVM) (Boser, 1992).

Esses modelos são atraentes, pois não são necessárias hipóteses sobre a distribuição das variáveis envolvidas no estudo. Quando os analistas possuem informações ou fortes hipóteses sobre a distribuição dos dados, outras metodologias paramétricas podem ser empregadas para a geração de dados sintéticos, como modelos de regressão linear, modelo logístico, dentre outros. Ao utilizar modelos paramétricos podemos correr o risco de estar selecionando o modelo incorreto para a geração dos dados sintéticos o que irá impactar diretamente no risco e principalmente na utilidade dos dados.

Dentre as metodologias apresentadas no artigo de Reiter e Drechsler (2011) , o método CART obteve os melhores resultados em relação ao risco de divulgação, além da facilidade de implementação. Embora o método SVM também tenha resultados semelhantes ao CART, sua implementação é mais difícil. Sendo assim, os autores ressaltaram que o CART tem baixo custo computacional e apresentou risco de divulgação aceitável.

Dessa forma a escolha de modelos paramétricos e não paramétricos pode ser feita através do estudo do banco de dados a ser divulgado. Nesse trabalho iremos utilizar o modelo não paramétrico CART, pois como resultado por Reiter e Drechsler (2011), essa metodologia é interessante por ser simples de implementar e sua interpretação ser compreensível.

O CART procura aproximar a distribuição condicional da variável resposta dadas as variáveis preditoras, particionando o espaço de previsão para que subconjuntos da variável resposta sejam relativamente homogêneos. As florestas aleatórias são coleções de CART's e cada árvore é baseada num subconjunto aleatório dos dados. O método *bagging* é fundamentado na geração de amostras *bootstrap* dos dados, permitindo a obtenção de preditores individuais que, posteriormente, são agregados, ou combinados, para formar um preditor final melhor. O SVM é usado para encontrar uma função que separe os dados e prever a partição da variável resposta dados as variáveis explicativas. A seguir na Seção 1.3 é apresentada a metodologia de árvores de regressão e classificação.

1.3 Árvores de Classificação e Regressão - CART

Árvores de Classificação e Regressão (CART) foram inicialmente propostas por Breiman (1984). O algoritmo CART particiona o espaço da variável resposta univariada, a partir de divisões binárias recursivas dos preditores. A série de divisões podem ser efetivamente representadas por uma estrutura de árvore, com folhas correspondentes aos subconjuntos dos dados.

Essas árvores são projetadas para uma determinada variável dependente a partir de outras variáveis. Elas são divididas em duas decisões "sim" e "não". Geralmente são representadas em forma de uma figura de árvore, onde a folha esquerda refere-se à decisão "sim", e a direita à decisão "não".

A Figura 1.1 apresenta o exemplo de árvore de decisão gerada pelo método CART. Temos a árvore de decisão para a variável resposta sexo com duas variáveis explicativas altura e peso. Na primeira folha, temos que a altura foi a variável que melhor dividiu os sexos. Sendo assim, se uma pessoa tem altura maior que 180cm ela é classificada como homem. Se a altura for menor do que 180cm, utiliza-se uma nova decisão, o peso. Então, se o peso for maior que 80kg ela é classificada como homem e caso contrário, é classificada como mulher.

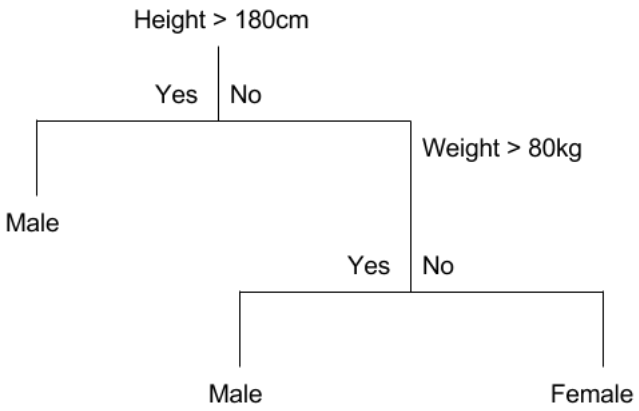

Figura 1.1: Exemplo CART

Essa árvore, embora seja simples, retrata bem o funcionamento do CART. A árvore é criada por particionamento recursivo binário usando a variável resposta na fórmula e escolhendo divisões dado as demais covariáveis. A divisão escolhida é a que maximiza a redução da deviance geral do modelo, que é a soma da deviance de cada folha do modelo final. O conjunto de dados é dividido e o processo é repetido. A divisão continua até que as folhas finais tenham tamanho mínimo de, geralmente

5 (Ripley, 2018).

Com essa divisão, é possível calcular medidas de sensibilidade, que nesse exemplo é a probabilidade de se classificar um indivíduo como homem dado que ele realmente é homem, e a especificidade, que é a probabilidade de se classificar uma pessoa como "não homem", se a pessoa é mulher.

Os valores em cada folha representam a distribuição condicional da variável resposta para os dados que satisfazem o particionamento que define a folha. O particionamento é feito de forma a satisfazer alguns critérios, como o número mínimo de dados em cada folha, variância mínima em cada folha e menor erro de classificação. Após a construção da árvore são realizadas "podas" de acordo com os critérios dos modelos. As podas podem ser utilizadas para diminuir os preditores do modelo. O tamanho final da folha também pode ser usado como critério de "poda". Assim como nesse exemplo, uma covariável pode ser escolhida para a partição do espaço de Y mais de uma vez em pontos diferentes.

Friedman, Hastie e Tibshirani (2001) descrevem o algoritmo do CART para a regressão e para a classificação. Os dados de entrada no modelo consistem em p covariáveis $X_1, ..., X_p$ univariadas e uma resposta Y univariada, para cada uma das N observações, ou seja , temos o vetor $(x_{i1}, x_{i2}, ..., x_{ip}, y_i)$ para $i = 1, 2, ..., N$. O algoritmo precisa decidir a variável explicativa que terá a primeira divisão binária, qual será o ponto que divide melhor o espaço de Y e a construção das demais divisões

Seja o nó da árvore o ponto de divisão de cada divisão da árvore. Suponha que o primeiro nó da árvore tenha M Folhas $F_1, F_2, ..., F_M$, sendo que a variável resposta será uma constante c_m em cada folha. Então seja a função para cada $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})$ definida por:

$$Y_i = f(\mathbf{x_i}) = \sum_{m=1}^{M} c_m I(\mathbf{x_i} \in R_m). \tag{1.1}$$

Suponha, por exemplo que temos um banco de dados com uma variável resposta Y e duas covariáveis X_1 e X_2. A árvore final para esse exemplo resultou em cinco folhas com a média estimada para Y em cada folha. Então para esse exemplo temos F_1, F_2, ..., F_5 folhas distintas, então a fórmula dada em (2.1) para o exemplo será:

$$\hat{Y}_i = \hat{f}(\mathbf{x_i}) = \sum_{m=1}^{5} c_m I\{(x_{i1}, x_{i2}) \in F_m\}.$$

Esse mesmo exemplo é representado na Figura 1.2. Observações que satisfazem a condição em cada ponto de divisão é atribuído ao ramo esquerdo,e os outros para o ramo direito. As folhas correspondem a cada uma das regiões $F_1, ..., F_5$. A Figura 1.2 é um gráfico em perspectiva da superfície de regressão deste modelo, ou seja, a

primeira imagem apresenta a divisão do espaço de Y utilizando as duas covariáveis. A segunda imagem a direita, apresenta a divisão do espaço de Y nos pontos t de divisão de X_1 e X_2. A terceira imagem apresenta a árvore resultando do exemplo, com as divisões em cada covariável e as regiões resultantes F. A última imagem a direita apresenta a representação da partição da segunda imagem em perspectiva do espaço de previsão de Y.

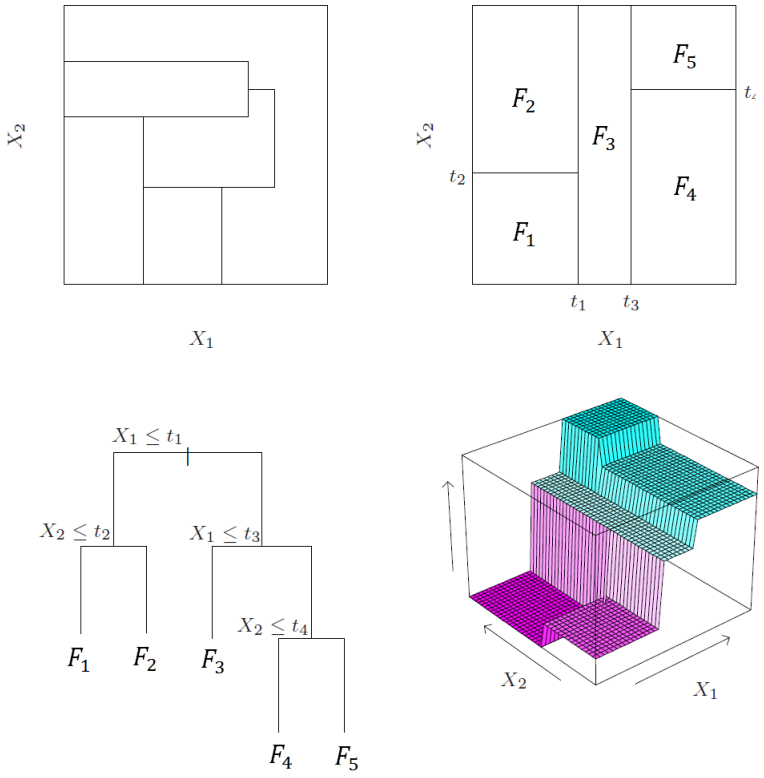

Figura 1.2: Exemplo Partições CART

Se o critério para encontrar a melhor divisão for a minimização da soma de quadrados $\sum(y_i - f(\mathbf{x_i}))^2$, o melhor valor de $\hat{c}_m$ será a média de y_i na região F_m, ou seja,

$$\hat{c}_m = \text{média}(y_i|\mathbf{x_i} \in F_m). \tag{1.2}$$

Encontrar a melhor partição binária em termos de soma mínima de quadrados geralmente é computacionalmente inviável, portanto um algoritmo guloso é criado. Algoritmo guloso é uma solução comum para problemas de otimização onde é realizada a escolha que parece ser a melhor no momento, tal que a mesma acarrete em uma solução de problemas a nível global. Todos os dados são considerados para o modelo.

Considere uma variável de divisão j, um ponto de divisão s e o par de planos F_1 e F_2:

$$F_1(j,s) = \mathbf{x_i}|\mathbf{x_{ij}} \leq s \text{ e } F_2(j,s) = \mathbf{x_i}|\mathbf{x_{ij}} > s. \tag{1.3}$$

Então o algoritmo tenta achar a variável de divisão j e o ponto de divisão s que resolve:

$$\min_{j,s}\left[\min_{c_1}\sum_{\mathbf{x_i}\in F_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{\mathbf{x_i}\in F_2(j,s)}(y_i - c_2)^2\right] \tag{1.4}$$

onde c_1 e c_2 são os valores de x_i que minimizam a soma de quadrados em cada região F_1 e F_2. Para qualquer escolha j e s, a minimização interna é resolvida por

$$\hat{c}_1 = \text{média}(y_i|\mathbf{x_i} \in F_1(j,s)) \text{ e } \hat{c}_2 = \text{média}(y_i|\mathbf{x_i} \in F_2(j,s)). \tag{1.5}$$

Então ao invés de calcular $\sum(y_i - f(\mathbf{x_i}))^2$ para cada região F_m, o algoritmo calcula o mínimo da soma de quadrados para os valores estimados de c_1 e c_2 para cada divisão da árvore, achando assim a soma mínima entre as duas regiões. Sendo assim a divisão final será aquela que é minimizada pelas escolhas da variável j, ponto s,e os respectivos valores de c_1 e c_2.

Para cada variável j de divisão, a determinação do ponto de divisão s pode ser feita através do cálculo de $\hat{c}_1$ e $\hat{c}_2$ de acordo com as equações (2.4) e (2.5). Então a determinação do melhor par (j,s) é viável através desses cálculos e os melhores valores de $\hat{c}_1$ e $\hat{c}_2$ para cada divisão da árvore são encontrados. Tendo encontrado a melhor divisão, particionamos os dados nas duas folhas F_1 e F_2 e repetimos o processo de divisão em cada uma das duas regiões conforme os critérios da construção da árvore. Então esse processo é repetido em todas as regiões resultantes.

Com esse procedimento, árvores muito grandes podem ser construídas. Então é necessário escolher um critério de parada para as divisões das variáveis. Uma estratégia é criar uma grande árvore T_0, parando a divisão quando um tamanho mínimo (usualmente 5 observações, Ripley 2018) é atingido em um determinada folha. Então a árvore grande é diminuída usando a poda de acordo com outros critérios de parada. Esses critérios podem estar ligados ao número de observações em cada folha, variância mínima, dentre outros.

Defina $T \subset T_0$ que será construída a partir da poda de T_0. Seja r o nó que representa a folha F_m e $|T|$ o número de nós em T, sendo assim:

$$\begin{aligned} N_m &= \#\{\mathbf{x_i} \in R_m\} && (1.6)\\ \hat{c}_m &= \frac{1}{N_m} \sum_{\mathbf{x_i} \in F_m} y_i && (1.7)\\ Q_m(T) &= \frac{1}{N_m} \sum_{\mathbf{x_i} \in F_m} (y_i - \hat{c}_m)^2. && (1.8) \end{aligned}$$

Então, o critério de complexidade de custo será dado por:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|. \quad (1.9)$$

O procedimento é , para cada α, encontrar a subárvore $T_\alpha \subseteq T_0$ que minimiza $C_\alpha(T)$, sendo $\alpha \geq 0$. Segundo Friedman, Hastie e Tibshirani (2001) a estimativa de α é obtida por validação cruzada de cinco ou dez vezes para cada banco de dados selecionado: escolhemos o valor α para minimizar a soma dos quadrados em cada validação cruzada. Então a árvore final será T_α. Quando $\alpha = 0$ a solução será a árvore completa T_0.

Para cada α, pode-se mostrar que existe uma subárvore única menor, T_α, que minimiza $C_\alpha(T)$. Para encontrar o T_α usamos a poda de nós mais fracos: sucessivamente recolher o nó interno que produz o menor aumento por nó em $\sum_m N_m Q_m(T)$ e continuar até produzirmos a árvore de nó único.

O resultado será sequência finita de subárvores, e pode-se mostrar que essa sequência deve conter T_α de acordo com Breiman (1984) ou Ripley (1996).

Para a classificação, seja $1, 2, ..., K$ as divisões da árvore, ou seja, por exemplo $K = 1$, representa a primeira divisão da árvore para a variável j no ponto s. Para a regressão, pode-se usar a minimização do erro quadrático e o critério de complexidade de custo descrito em (2.9), mas quando o interesse é classificação podemos utilizar outas medidas para verificar se a árvore escolhida será utilizada.

Em um nó m que representa a região F_m com N_m observações, seja

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\mathbf{x_i} \in F_m} I(y_i \in k)$$

a proporção de observações da classe k na folha m, ou seja, definidas as divisões K, contamos o número de observações em cada classe k e em cada folha m. Usando então os valores calculados $\hat{p}_{mk}$ podemos calcular diferentes medidas $Q_m(T)$:

Erro de classificação $= 1 - \hat{p}_{mk(m)}$, onde $k(m) = \text{argmax}_k \hat{p}_{mk}$.

Índice de Gini $= \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$

Deviance $= -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Por exemplo, para 2 classes, se p é a proporção de observações na segunda divisão, essas três medidas são $1 - max(p, 1-p)$, $2p(1-p)$ e $-p \log p - (1-p)\log(1-p)$, respectivamente. Os três métodos são similares e eficientes na classificação das observações da árvore, mas a deviance e o índice de Gini são diferenciáveis e por isso melhores para a otimização numérica (Friedman, Hastie e Tibshirani, 2001).

Usando a deviance como critério de parada, suponha que escolhemos a primeira variável de divisão e o melhor ponto para a partição do espaço de Y. Após essa escolha é calculado a deviance de acordo com a fórmula apresentada. Para as próximas divisões, calcula-se novamente o valor de deviance, se esse valor for menor do que o valor calculado no passo anterior, a divisão é feito em mais um nó, caso contrário o algoritmo é parado. Então novas folhas da árvore só são estimadas se a deviance diminuir. Com isso iremos obter uma possível melhor árvore para a classificação.

Capítulo 2

Dados Sintéticos

Uma proposta para a divulgação de dados é a utilização de dados sintéticos. Essa proposta foi descrita por Rubin (1993), onde novos dados são simulados com base nos valores verdadeiros a partir de um modelo de imputação múltipla.

A técnica de imputação múltipla foi inicialmente proposta por Rubin(1987) para casos onde existem dados faltantes em bancos de dados. A ideia basicamente é gerar vários valores para cada dado faltante a partir do modelo de imputação, gerando assim mais de um banco de dados completos. Dessa forma, pode-se medir a variação existente em cada banco de dados, e também tendo mais de uma base, medir a variação referente à imputação.

Assim, podemos considerar os dados sintéticos como "ausentes", como na imputação múltipla, e imputar novos valores para esses dados. Os valores verdadeiros são substituídos por novos valores que foram gerados a partir das relações originais presentes no banco de dados. O método de árvores de classificação e regressão (CART), dentre outras metodologias, pode ser utilizado para a geração dos dados sintéticos, pois ele tenta preservar as relações entre as variáveis do banco original. Mais detalhes são descritos na Seção 2.3.

Uma das vantagem desse método é a possibilidade de se obter inferências mais precisas, visto que os bancos de dados sintéticos possuem relações muito parecidas com as relações existentes no banco original. Além disso, esse método permite que os usuários dos dados públicos façam as mesmas análises estatísticas que fariam com os dados originais e sendo assim as inferências serão muito próximas nos dois bancos de dados.

Raghunathan, Reiter e Rubin (2003) apresentam um estudo sobre a criação de dados sintéticos através da imputação múltipla. O método de imputação múltipla descrito por Rubin (1987) é um método utilizado quando um banco de dados possui dados ausentes. Essa técnica utiliza as demais variáveis do banco de dados para

fazer uma estimação dos valores ausentes. Nessa metodologia é possível completar os valores ausentes por estimativas baseadas em outras variáveis do banco de dados.

Embora essa abordagem resolva muitos dos problemas inerentes à imputação de dados ausentes, um problema permanece. Como o valor imputado é uma estimativa , um valor previsto, há incerteza sobre seu valor real. Visto que , toda estatística tem incerteza, medida pelo seu erro padrão, estatísticas calculadas usando dados imputados têm ainda mais incerteza. Essa incerteza está ligada ao fato de que se apenas uma única imputação for feita para um valor ausente, podemos ter algum tipo de viés nessa imputação.

Dessa forma Rubin (1987) apresenta a imputação múltipla, estimando os valores ausentes mais de uma vez e medindo assim o erro inerente a imputação, ou seja, para cada valor ausente são estimados mais de um valor estimado, o valor final imputado será a média desses valores, e o erro da imputação pode ser calculado. A seguir é apresentado os passos para a imputação múltipla:

1. Crie m conjuntos de imputações para os valores ausentes usando um processo de imputação com um componente aleatório.

2. O resultado são m conjuntos de dados completos. Cada conjunto de dados terá valores ligeiramente diferentes para os dados atribuídos, devido ao componente aleatório.

3. Analise cada conjunto de dados concluído. Cada conjunto de estimativas de parâmetros será um pouco diferente porque os dados diferem ligeiramente.

4. Combine os resultados, calculando a variação nas estimativas dos parâmetros.

Dessa forma, o número de imputações suficientes, pode ser de apenas 5 a 10 imputações, embora dependa da porcentagem de dados que estão faltando Rubin (1987). O resultado serão estimativas de parâmetros imparciais e um tamanho de amostra completo. Esse procedimento requer um modelo de imputação muito bom. Criar um bom modelo de imputação requer conhecer muito bem seus dados e ter variáveis que prevejam valores ausentes. Na literatura existem algumas metodologias paramétricas e não paramétricas para a realização dessas imputações. Intuitivamente, os valores ausentes podem ser estimados por exemplo como um modelo de regressão, onde estimamos a equação de previsão do modelo para a variável resposta, que possuem os dados ausentes.

A geração de dados sintéticos utiliza a ideia de imputação múltipla, pois tratamos dados originais, que não podem ser divulgados, como ausentes. Dessa forma os novos valores imputados são valores baseados nos dados originais. Sendo assim são gerados

mais um banco de dados para estimar o erro associado a imputação desses novos dados.

Para gerar esses dados sintéticos, Raghunathan, Reiter e Rubin (2003), utilizam uma abordagem Bayesiana, e depois é realizada a comparação das análises com os dados originais. Como conclusão, os dados sintéticos obtiveram análises semelhantes às dos dados originais, dependendo do método de imputação aplicado.

Rubin(1993) ressalta que esse método é vantajoso, visto que métodos de imputação múltipla já são largamente utilizados, e parece ser mais fácil de ser aplicado do que algumas técnicas de limitação de divulgação. Neste estudo pretendemos mostrar que essa abordagem pode ser uma boa ferramenta para a divulgação de dados.

2.1 Descrição dos dados sintéticos

O tipo de variável presente no banco de dados, e até mesmo sua importância para a divulgação de dados pode estipular como a geração dos dados sintéticos é realizada. Em alguns casos tem-se o interesse em divulgar dados completamente sintéticos para aumentar a proteção dos dados, mas em algumas bases nem todas as variáveis possuem algum risco de divulgação. Gerar bancos inteiros com dados sintéticos pode não ser uma tarefa simples e muitas vezes nem todas as variáveis são de risco.

Uma proposta apresentada por Reiter (2003a) apresenta um estudo utilizando dados parcialmente sintéticos. Nessa abordagem apenas aqueles valores mais sensíveis à divulgação são imputados. Dessa forma, fórmulas para análise dos dados sintéticos diferem daquelas apresentadas por Raghunathan, Reiter e Rubin (2003).

Dados parcialmente sintéticos são atraentes porque podem manter muitos dos benefícios de dados totalmente sintéticos, protegendo a confidencialidade, e também sua implementação computacional pode se tornar mais simples.

Seja Y_{obs} o vetor de N observações para a variável sensível que desejamos substituir parcialmente por dados sintéticos. Seja X a matriz Nxp de outras variáveis que não apresentam atributos sensíveis. Vamos assumir que os dados $D = (X, Y_{obs})$ são completamente observados.

Os dados parcialmente sintéticos são imputados em duas partes distintas. Primeiro, são selecionados os valores de Y_{obs} que serão substituídos por valores imputados. Depois, novos valores de Y são gerados para substituir os valores selecionados.

Seja $Z_j = 1$ se a unidade j é selecionada para ser substituída por dados sintéticos, e $Z_J = 0$ para as unidades que permanecerão com os valores originais, tal que $Z = (Z_1, ..., Z_N)$. Seja $Y_{rep,i}$ os valores imputados simulados no $i-$ésimo conjunto de

dados sintéticos, e Y_{nrep} os valores originais que não serão substituídas de Y_{obs}. Dessa forma, Y_{nrep} será igual em todos os conjuntos de dados. Assim, cada banco de dados $\tilde{D}_i$ é formado por $(X, Y_{rep,i}, Y_{nrep}, Z)$. As imputações $i = 1, ..., m$ são independentes, produzindo assim m bancos de dados sintéticos diferentes que serão divulgadas para os usuários.

Os valores $Y_{rep,i}$ normalmente serão gerados através do modelo Bayesiano da distribuição preditiva *a posteriori* de $(Y_{rep,i}|X, Y_{obs}, Z)$. Para este estudo, esses valores serão geradora partir dessa distribuição condicional obtida pelo CART.

A inferência para os bancos de dados sintéticos será feita pela combinação dos m bancos completos gerados, para estimadores como média, variância e outras medidas de interesse, visto que os m bancos de dados sintéticos são independentes.

Seja Q uma estatística de interesse, q seu estimador pontual e v a variância desse estimador pontual . Para $i = 1, ..., m$, seja q_i e v_i os respectivos valores de q e v para cada um dos bancos de dados sintéticos. Os valores de q_i e v_i são estimados como se estivessem no banco de dados original. Os seguintes escalares são necessários para estimar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{2.1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2.2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m \tag{2.3}$$

A estatística Q é estimada por $\hat{q}_m$, e a variância é estimada por

$$T_p = b_m/m + \bar{v}_m \tag{2.4}$$

Quando n é suficientemente grande, as inferências para Q podem ser feitas utilizando a distribuição t-student com grau de liberdade $\nu_p = (m-1)(1 + r_m^{-1})^2$, onde $r_m = (m^{-1}b_m/\bar{v}_m)$. Em muitos casos, a distribuição normal fornece uma aproximação adequada para a distribuição t porque r_m é pequeno. Derivações destes métodos são apresentados em Reiter (2003a). Extensões para Q multivariada são apresentadas em Reiter (2003b).

Suponha que em um determinado estudo estamos interessados em estimar a média de uma variável $Y \sim f(\mu, \sigma^2)$, onde μ e σ^2 são a esperança e variância de Y i.i.d (independente e identicamente distribuído), respectivamente. Dessa forma, nosso interesse é Q= μ. Para isso o estimador de $Q = \mu$ será $q = \bar{X}$. A esperança e variância de $\bar{X}$ serão dados por $E(\bar{X}) = \mu$ e $Var(\bar{X}) = \frac{S^2}{n}$, respectivamente, quando

σ^2 é desconhecido.

Como nesse estudo foram gerados $m = 5$ bancos sintéticos, podemos utilizar (2.10)-(2.13) para fazer inferências para μ.

A medida Q pode estar relacionada a diversas medidas de interesse, desde que a média e variância de seu estimador q sejam conhecidas, ou possam ser calculadas.

Sendo assim, utilizando as fórmulas (2.10), (2.11), (2.12), (2.13) é possível, por exemplo, fazer inferências para estatísticas descritivas simples das variáveis dos bancos , bem como estimação de parâmetros de modelos mais complexos como modelos de regressão. Nas Seções 2.3 e 2.4 são apresentados os cálculos para estimação de alguns parâmetros.

Reiter (2005) apresenta um estudo para a geração de dados sintéticos utilizando o CART. Foram realizados estudos de simulação utilizando o algoritmo implementado por Clark e Pregibon (1992).

As simulações sugerem que o CART é eficiente para a geração de dados sintéticos de forma simples. Como desvantagem, o autor sinaliza que as imputações podem introduzir algum tipo de dependência condicional nos dados gerados. Outra questão é a escolha da melhor "poda", para a árvore, pois após a árvore ser gerada, o pesquisador pode diminuir suas divisões antes de gerar os dados sintéticos. Embora a análise da "poda", seja mencionada, o artigo não forneceu um estudo mais aprofundado desse problema.

2.2 Simulação de dados sintéticos

O modelo CART pode ser usado para imputar dados sintéticos, considerando cada folha da árvore ajustada como uma classe diferente de imputação. Os dados são imputados de acordo com a mesma distribuição estimada para os dados originais, ou seja, de acordo com o modelo de imputação para dados ausentes de maneira aleatória (Rubin, 1976).

O primeiro passo para a geração de dados sintéticos usando o CART é gerar a árvore de classificação para $Y_{obs}|X$. A árvore ajustada pode ser podada para algum tamanho mínimo desejado para as folhas. Em seguida, são gerados valores sintéticos para as observações com $Z_j = 1$ a partir da distribuição dos dados em cada folha.

Seja F_k a k-ésima folha gerada para $Y_{obs}|X$, e seja $Y^{(F_k)}$ os n_k valores de Y_{obs} resultantes nessa folha. Dessa forma, para cada folha F_k da árvore são gerados novos conjuntos de valores a partir de $Y^{(F_k)}$ de acordo com o método de *bootstrap* Bayesiano proposto por Rubin (1981).

O *bootstrap* Bayesiano é análogo ao *bootstrap* proposto por Efron(1979). Porém,

em vez de simular a distribuição amostral de uma estatística para estimar um determinado parâmetro, o *bootstrap* Bayesiano simula as distribuições *à posteriori* desse mesmo parâmetro (Rubin, 1981).

O *bootstrap* Bayesiano amostra valores de $Y_{rep,i}$ a partir dos valores $Y^{(F_k)}$ classificados em cada folha F_k. Quando a variável a ser imputada é categórica, os valores amostrados são selecionados a partir das n_k unidades que pertencem à folha F_k. Seja n_s o número de valores sintéticos a serem simulados. O *bootstrap* Bayesiano procede da seguinte maneira:

1. Gere $(n_0 - 1)$ valores de uma distribuição uniforme $(0, 1)$. Organize esse números em ordem crescente, sendo eles de $a_0 = 0, a_1, ..., a_{n_0-1}, a_{n_0} = 1$.

2. Gere n_s valores uniformes $(0, 1)$, sendo $u_1, ..., u_{ns}$. Para cada u_i, selecione $Y_j^{(F_k)}$ quando $a_{j-1} < u \leq a_j$

O *bootstrap* Bayesiano incorpora a incerteza adicional nas distribuições condicionais em cada folha já que temos apenas uma amostra de valores em cada folha.

Para variáveis contínuas, um novo passo é necessário. Para cada folha, é estimada a densidade de probabilidade dos dados $Y^{(F_k)}$ utilizando o método da CDF inversa. Dessa forma, o suporte da densidade se estende do mínimo até o máximo valor de Y em cada folha. Isso garante que a densidade estimada esteja dentro dos limintes da distribuição de cada folha. Os dados sintéticos $Y_{rep,i}$ são então simulados a partir da densidade estimada. Assim, estamos evitando divulgar dados sintéticos iguais aos dados originais.

Esse procedimento é repetido m vezes para cada conjunto de dados sintéticos. Os bancos de dados resultantes $\tilde{D}_i = (X, Y_{rep,i}, Y_{nrep}, Z)$ são então liberados para uso público.

2.3 Avaliação do Risco e Utilidade

Embora algumas técnicas para a divulgação de dados protejam bem os dados, em alguns casos os dados divulgados tornam-se inúteis ou até mesmo enganosos para fornecer análises (Karr e Reiter, 2013). Em contrapartida, alguns métodos também podem ser bons para análises dos dados, mas o risco de identificação dos dados originais sigilosos pode ser grande. Sendo assim, as agências buscam um equilíbrio entre essas duas preocupações quanto à publicação de informações. A divulgação de um banco de dados será mais eficiente quando o risco de identificação é minimizado e a utilidade dos dados é a maior possível.

A divulgação de dados sigilosos pode ser realizada de forma mais eficiente e segura quando essas duas medidas, o risco de identificação e a utilidade dos dados, são analisadas e combinadas da melhor maneira possível. Veja que essas medidas podem ser proporcionalmente inversas. Se o modelo para gerar os dados sintéticos for muito preciso, a utilidade dos novos bancos será grande. Em contrapartida, o risco de divulgação aumenta, pois os novos bancos são muito parecidos com os originais, podendo resultar na identificação de dados originais. Por outro lado, se o modelo de geração de dados sintéticos não for tão preciso, a utilidade pode ser comprometida, mas o rico de identificação diminuirá. Dessa forma, é de interesse medir essas duas condições de maneira a otimizar a uma combinação da utilidade e do risco dos dados a serem divulgados.

2.3.1 Medidas de Risco de Divulgação

Antes da divulgação de dados ser realizada, é possível identificar e mensurar o risco de identificação dos dados divulgados. O risco deve ser analisado de acordo com as características do banco. Por exemplo, se as informações não forem potencialmente perigosas para serem divulgadas, então pode-se assumir um risco de identificação um pouco mais alto. Caso contrário, o risco de identificação aceitável deve ser baixo.

Dreschsler e Reiter (2011) apresentam algumas medidas para o cálculo do risco de divulgação baseado na informação que o invasor possui a respeito do banco de dados.

Suponha que o invasor possua um vetor de informações t, sobre algum indivíduo da população alvo. Por exemplo, $\boldsymbol{t} = (Y^*, X_1^*, ..., X_p^*)$ é o vetor alvo onde Y^* é um dado valor para Y e assim respectivamente. Note que, nesse caso, o alvo do intruso pode não estar no banco sintético $\tilde{D}$. Seja t_0 um identificador único desse alvo, e seja d_{j0} o identificador para o registro j em $\tilde{D}$, onde $j = 1, ..., n$.

Dessa forma, o intruso tem o objetivo de unir a unidade j em $\tilde{D}$ ao alvo, quando $d_{j0} = t_0$ e , da mesma maneira, não unir quando $d_{j0} \neq t_0$ para qualquer $j \in \tilde{D}$.

Seja J a variável aleatória igual a j quando $d_{j0} = t_0$ e $n+1$ quando $d_{j0} \neq t_0$ para algum $j \notin \tilde{D}$. O invasor tenta calcular $P(J = j|t, \tilde{D})$ para $j = 1, ..., n$.Tendo essas probabilidades ele pode decidir sobre a identificação correta baseado nas maiores probabilidades para $j = 1, ..., n$. Mas o intruso não sabe os valores de Y_{obs} para cada banco de dados sintéticos $\tilde{D}_i$, então ele calcula

$$P(J = j|\boldsymbol{t}, \tilde{D}) = \int P(J = j|\boldsymbol{t}, \tilde{D}, Y_{rep}) P(Y_{rep}|\boldsymbol{t}, \tilde{D}) dY_{rep} \tag{2.5}$$

Dessa forma, integrando fora Y_{rep} o intruso obtém as probabilidades de identificação desejadas.

Reiter e Mitra (2009) apresentam alguns cálculos para a probabilidade descrita em (2.14) de acordo com o conhecimento do invasor sobre as informações contidas no banco de dados original e também sobre os modelos utilizados para a geração dos dados sintéticos. Nesse estudo foi assumido que o intruso se aproxima de (2.14) tratando os valores simulados nos conjuntos de dados liberados como empates plausíveis para Y_{rep}. Isso também representa o que os invasores podem fazer sem fortes hipóteses sobre a distribuição condicional de Y_{rep}. A probabilidade correspondente para qualquer registro j será dada por

$$P(J = j|\boldsymbol{t}, \tilde{D}) = \frac{1}{m} \sum_{j} \frac{1}{F_{\boldsymbol{t}}} I((Y_{rep,j}^{(m)}, X_j) = \boldsymbol{t})$$

onde F_t é o número de observações que satisfazem os critérios de correspondência na população. Sendo assim, $I((Y_{rep,j}^{(m)}, X_j) = \boldsymbol{t})$ é uma indicadora se a observação j do banco sintético m é similar ao alvo, isto é, $(X_{1j} = X_1^*, ..., X_{pj} = X_p^*)$ e $Y_{rep,j}^{(m)} \in [Y^* - l; Y^* + l]$, onde l é um valor escolhido pelo intruso quando Y é uma variável contínua.

Essa indicadora será igual a 1 quando os valores de correspondência para registro j são iguais aos valores correspondentes em $\boldsymbol{t}$, e zero, caso contrário. Como $F_{\boldsymbol{t}}$ pode não ser facilmente obtido, nesses casos, $F_{\boldsymbol{t}}$ pode ser substituído por $N_{i\boldsymbol{t}}$ que representa o número de registros em D_i que satisfazem os critérios de correspondência. Isso explica o fato de que os dados originais representam apenas uma amostra da população, de modo que o intruso geralmente não sabe se o alvo está incluído em $\tilde{D}$.

Segundo Reiter (2005) e Dreschsler e Reiter (2008) a avaliação do risco pode ser verificada utilizando algumas medidas de resumo baseadas em probabilidades. É razoável assumir que o invasor irá selecionar o registro j quando $P(J = j|\boldsymbol{t}, \tilde{D})$ for a maior possível. Dessa maneira os autores consideram três medidas para avaliar o risco de divulgação:

1. *Expected match risk* (EMR): risco esperado.

2. *True match risk*(TMR): risco de correspondência verdadeira.

3. *False match rate* (FMR): taxa de correspondência falsa.

Seja c_j o número de registros com a maior probabilidade $P(J = j|\boldsymbol{t}, \tilde{D})$ no banco de dados $\tilde{D}_i$ para o alvo t_j para $j = 1, ...n$. Seja $I_j = 1$ se o verdadeiro alvo está em

c_j, e 0 caso contrário. Seja $K_j = 1$ quando $c_j I_j = 1$, ou seja, se a maior probabilidade for única e correspondente ao valor verdadeiro. Seja $Q_j = 1$ se $c_j(1 - I_j) = 1$ e 0 caso contrário. Seja s o número de registros com $c_j = 1$. Dessa forma as três medidas de resumo serão dadas por:

$$EMR = \sum_j \frac{1}{c_j} I_j \tag{2.6}$$

$$TMR = \sum_j \frac{K_j}{n} \tag{2.7}$$

$$FMR = \sum \frac{Q_j}{s} \tag{2.8}$$

Quando $I_j = 1$ e $c_j > 1$, isto é, o verdadeiro alvo está entre as maiores probabilidades de correspondência, mas há mas de uma observação, a contribuição da unidade j para EMR mostra que o intruso adivinha aleatoriamente o alvo para as unidades de c_j. TMR se refere a probabilidade de correspondência correta para a unidade j, ou seja, a maior probabilidade percente a unidade j que é alvo do intruso. Já a probabilidade FMR é associada a correspondências incorretas que o intruso pode fazer, ou seja, a maior probabilidade de correspondência existe é única, mas não pertence ao registro correto.

As medidas apresentadas são úteis para resumir o risco geral de divulgação dos dados sintéticos. Essas medidas são interessantes, pois as probabilidades são calculadas para cada observação divulgada em cada banco de dados.

Para avaliar a medida de risco geral, vamos considerar todo o vetor $\boldsymbol{t}_i = (Y_{orig,i}, X_{1i}, ..., X_{pi})$, para $i = 1, ..., n$, isto é, avaliar a probabilidade de identificar cada uma das observações originais Reiter (2005). Para isso vamos calcular a matriz:

$$\boldsymbol{\theta}_{i,j}^{(m)} = \frac{I\left\{(Y_{rep,j}^{(m)} \in [Y_{orig,i} - l; Y_{orig,i} + l]) \bigcap (\boldsymbol{X_j} = \boldsymbol{X_i})\right\}}{F_t}$$

Dessa maneira é possível calcular o risco geral no pior cenário, onde o intruso sabe todos os valores verdadeiros exceto um, para a combinação entre os m bancos de dados sintéticos. Também foi considerado que o intruso não precisa saber o valor exato do alvo, mas sim um intervalo de tamanho l ao redor do verdadeiro valor de Y_{orig} para correspondência.

Sejam $X_1, ..., X_p$ as covariáveis não sintéticas. O algoritmo para o cálculo das três medidas de resumo é dado por

1. Escolha $Y_{obs,j}$ e um valor l de intervalo.

2. Para cada $\tilde{D}_i$, seja $a_i = 1$ se Y_{rep} pertence ao intervalo $[Y_{jobs} - l; Y_{jobs} + l]$ e possui as mesmas covariáveis, e 0 c.c;

3. Calcule a probabilidade para cada j, sendo $1/\sum a_i$, quando $a = 1$;

4. Crie um banco de dados combinando as probabilidades p para cada j sendo $p = \frac{1/\sum a_i}{m}$;

Agora com as probabilidades de identificação de cada indivíduo calculados, é possível obter as medidas c_j, I_j, K_j e Q_j. Então as equações (2.15), (2.16) e (2.17) são calculadas para medir o risco de divulgação.

Os valores de l afetam diretamente nas medidas de risco, pois assumindo intervalos maiores, a probabilidade de identificação do alvo é menor, pois podem existir muitos valores no intervalo l e com as covariáveis iguais. Na Seções 3.3.1 é apresentado o estudo dessas medidas para alguns valores de l com o objetivo de verificar comportamento do risco em função das hipóteses sobre l.

2.3.2 Medida de Utilidade

A utilidade dos dados pode ser medida comparando-se as análises feitas com os dados originais e os dados a serem divulgados. Pode-se gerar novos dados utilizando qualquer uma das técnicas de limitação de divulgação, depois realiza-se análises estatísticas nos dois bancos de dados e é verificado se as conclusões são parecidas.

Para a estimação de estatísticas descritivas e coeficientes de regressão, vamos utilizar a metodologia apresentada na Seção 2.2. Note que, para verificar a utilidade dos dados, basta fazer a comparação entre os valores obtidos no banco de dados original e os valores obtidos com as fórmulas para a combinação dos dados sintéticos. Os bancos de dados sintéticos terão medida de utilidade grande quando esses valores estiverem próximos aos reais.

Woo et. al. (2009) apresentam um estudo sobre medida de utilidades gerais, ou seja, medidas de utilidade dos dados sintéticos que podem ser utilizadas para qualquer tipo de dados sintéticos para comparar os dados originais e os dados sintéticos. Algumas medidas de utilidade visam captar as diferenças existentes entre os bancos de dados. As medidas apresentadas são: o escore de propensão para avaliar a diferença nas distribuições, análise de cluster que avalia se os dados possuem valores semelhantes, e medidas que utilizam estatísticas de Kolmogorov-Smirnov para avaliar a diferença entre as distribuições empíricas de cada banco de dados. Nesse trabalho utilizamos o escore de propensão para medir a utilidade dos dados.

Nenhuma das medidas descritas estão associadas à natureza dos dados sintéticos. Isso permite calcular a utilidade na mesma escala para qualquer metodologia de geração de dados sintéticos, o que facilita as comparações entre métodos de geração de bancos de dados sintéticos.

O escore de propensão é largamente utilizado em estudos observacionais. Ele é a probabilidade atribuída ao tratamento dado os valores das covariáveis do estudo. Atribuições de tratamento e covariáveis são condicionalmente independentes, dado o escore de propensão segundo Rosenbaum e Rubin (1983). Assim, quando dois grandes grupos têm as mesmas distribuições dos escores de propensão, os grupos devem ter distribuições similares de covariáveis.

Esta teoria sugere uma abordagem para medir a utilidade de dados sintéticos. Primeiro, nós fundimos o banco de dados original com um banco de dados sintético D_i e adicionamos uma variável $U = 1$ para todos os registros do conjunto de dados sintéticos e $U = 0$ para todos os registros do conjunto de dados original. Segundo, para cada registro nos dados originais e sintéticos, calculamos a probabilidade de estar no conjunto de dados sintéticos, ou seja, o escore de propensão.

Em terceiro lugar, comparamos as distribuições dos valores de propensão nos dados originais e sintéticos. Quando essas distribuições são semelhantes, as distribuições dos dados originais e sintéticos são semelhantes, portanto, a utilidade dos dados deve ser relativamente alta, ou seja, o escore de propensão é próximo a zero.

Os escores de propensão geralmente são estimados por meio da regressão logística da variável sintética dado as covariáveis. Os escores de propensão são as probabilidades previstas no modelo de regressão. Seja p_j o escore de propensão calculado através do modelo logístico. A similaridade dos escores de propensão para as observações sintéticas e originais pode ser avaliados de várias maneiras, por exemplo, comparando seus percentis em cada grupo. Dessa forma, podemos calcular o escore de propensão G da seguinte forma:

$$G = \frac{1}{m}\frac{1}{2n}\sum_{j=1}^{2n}(\hat{p}_j - c)^2$$

Nesse estudo foi considerado que os dados sintéticos e originais possuem o mesmo tamanho. Os escores de propensão devem ser próximos a c para que a soma seja minimizada. Usualmente $c = 0.5$, ou seja, a probabilidade do indivíduo estar em um dos bancos de dados é aleatória, então os dados sintéticos e originais são parecidos.

Dessa forma o valor final de G deve ser o mínimo possível, indicando que as probabilidades de classificação são próximos a meio, ou seja, a classificação entre banco de dados sintético e original é aleatória. Então os dois bancos de dados são

similares. Com o valor de G pequeno, podemos inferir que a utilidade dos dados sintéticos é satisfatória.

Capítulo 3

Dados Simulados

Para os dados simulados, foi considerado o caso mais simples onde todos os valores de Y_{obs} serão substituídos por valores sintéticos. Dessa forma $Z = 1$ para todas as unidades. Três cenários diferentes são comparados para verificar as características do modelo proposto. O cenário 1 apresenta um modelo de regressão com distribuição Normal para os erros. Já o segundo cenário a distribuição do erro tem a cauda mais pesada, sendo uma distribuição T-student com 2 graus de liberdade. E o cenário 3 apresenta um modelo com erro com distribuição T-student com 10 graus de liberdade. Assim, com esses três cenário queremos estudar as diferenças na geração de dados sintéticos para modelos mais e menos dispersos.

3.1 Descrição dos dados simulados

Para verificar o funcionamento do modelo proposto e suas características, foram realizadas simulações de três cenários distintos de tamanho $n = 1.000$ observações cada. Os três cenários possuem os mesmos valores em cada covariável com os mesmos valores dos coeficientes de regressão. A diferença entre cada um dos cenários é a distribuição assumida para os erros. A seguir é apresentada a geração dos três bancos de dados simulados.

Primeiramente geramos a variável X_1 com as probabilidades descritas na Tabela 3.1. As probabilidades foram escolhidas de forma arbitrária para criar dependência entre as variáveis.

Dados os valores da Tabela 3.1, usamos a distribuição de probabilidade condicional para gerarmos $X_2|X_1$ com as probabilidades apresentadas na Tabela 3.2.

Agora, utilizando as probabilidades apresentadas nas Tabelas 3.1 e 3.2 , geramos então $X_3|X_2, X_1$, como é mostrado na Tabela 3.3.

Dessa forma, foram geradas 3 covariáveis categóricas com as probabilidades apre-

Tabela 3.1: Distribuição de Probabilidade de X_1

x	$P(X_1 = x)$
0	$0,35$
1	$0,65$

Tabela 3.2: Distribuição de Probabilidade de $X_2|X_1$

x	$P(X_2 = x\|X_1 = 0)$	$P(X_2 = x\|X_1 = 1)$
0	$0,38$	$0,55$
1	$0,62$	$0,45$

sentadas nas Tabelas 3.1, 3.2 e 3.3 para que as covariáveis tivessem relações entre si, fazendo com que o modelo final pudesse ser avaliado.

Por fim, geramos os valores de $Y|X_1, X_2, X_3$ de acordo com cada cenário simulado. Os valores dos preditores lineares foram escolhidos arbitrariamente de forma a dividir melhor as covariáveis, sendo:

$$Y_i = 23 + 1,5X_1 - 5,4X_2 + 2,5X_3 + e_\alpha \tag{3.1}$$

onde e_α, para $\alpha = 1, 2$ e 3, tem as seguintes distribuições conforme os cenários:

1. Cenário 1: $e_1 \sim N(0, 1)$
2. Cenário 2: $e_2 \sim$ T-student(2)
3. Cenário 3: $e_3 \sim$ T-student(10)

Considerando o modelo de Y_i, para cada resultado de X_1, X_2 e X_3, podemos obter a distribuição teórica de todas as oito combinações possíveis. As médias teóricas para cada combinação das covariáveis foram obtidas através da $E(Y_i)$ onde Y_i é dado pela equação (3.1), considerando os valores de X_1, X_2, e X_3. As probabilidades de cada combinação podem ser calculadas considerando $P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1)P(X_3 = x_3|X_1 = x_1, X_2 = x_2)$. As médias teóricas e as probabilidades π_i de cada combinação são apresentadas na Tabela 3.4.

Tabela 3.3: Distribuição de Probabilidade de $X_3|X_2, X_1$

x	$P(X_3 = x\|X_1 = 0, X_2 = 0)$	$P(X_3 = x\|X_1 = 1, X_2 = 0)$	$P(X_3 = x\|X_1 = 0, X_2 = 1)$	$P(X_3 = x\|X_1 = 1, X_2 = 1)$
0	0,70	0,40	0,25	0,54
1	0,30	0,60	0,75	0,46

Tabela 3.4: Tabela de Probabilidades

X_1	X_2	X_3	Média Teórica	π
0	1	0	17,6	0,039
1	1	0	19,1	0,162
0	1	1	20,1	0,214
1	1	1	21,6	0,134
0	0	0	23,0	0,093
1	0	0	24,5	0,054
0	0	1	25,5	0,143
1	0	1	27,0	0,157

Utilizando as probabilidades da Tabela 3.4 são apresentadas a seguir as densidades teóricas de cada combinação entre X_1, X_2 e X_3 para os três cenários. Na legenda de cada figura são apresentadas as combinações entre as covariáveis e suas respectivas médias.

A Figura 3.1 apresenta o histograma de Y_i para o cenário 1 e as respectivas densidades teóricas de cada combinação.

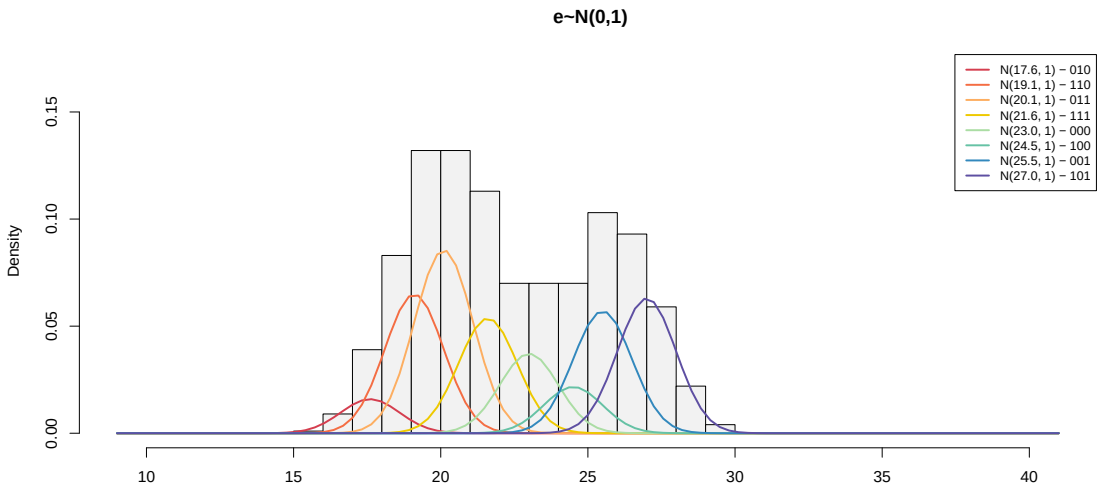

Figura 3.1: Cenário 1 - Histograma e Densidades Teóricas

A Figura 3.2 mostra a densidade teórica de cada combinação de valores entre X_1, X_2 e X_3 para o cenário 2.

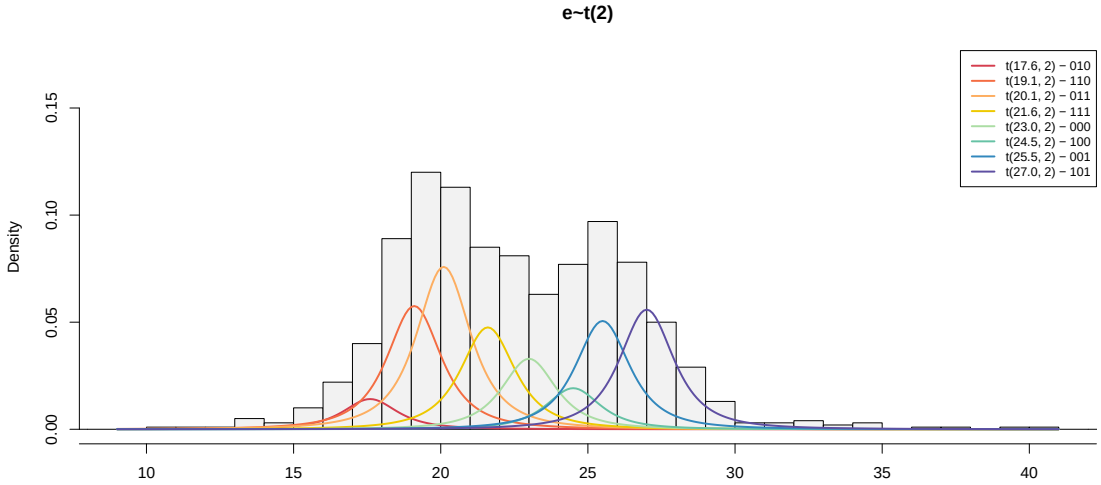

Figura 3.2: Cenário 2 - Histograma e Densidades Teóricas

A Figura 3.3 mostra a densidade teórica de cada combinação de valores entre X_1, X_2 e X_3 para o cenário 3.

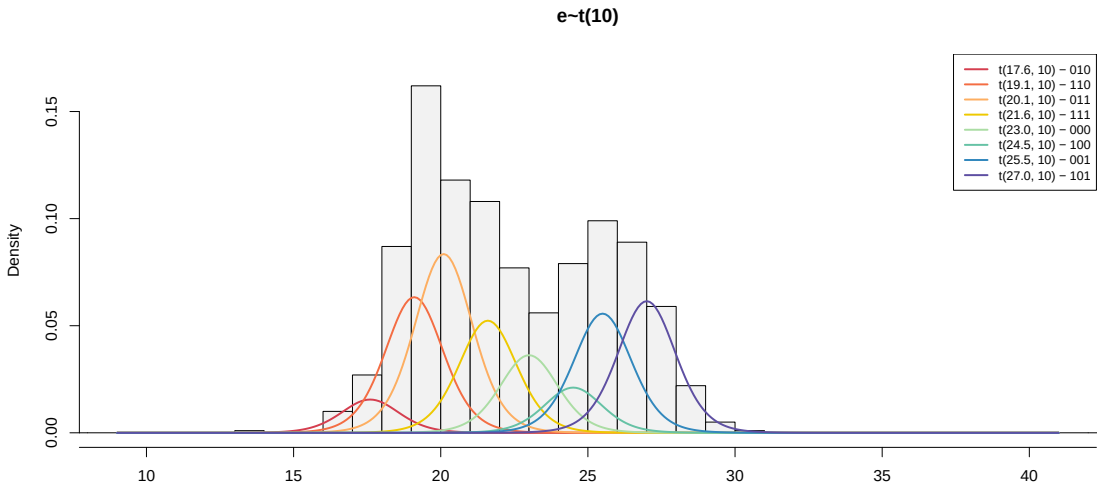

Figura 3.3: Cenário 3 - Histograma e Densidades Teóricas

Nos três cenários apresentados é possível notar que a distribuição de Y_i possui duas modas principais. Analisando as densidades e seus respectivos pesos, verificamos que existem densidades que se sobrepõe, ou seja, espera-se que no modelo apresentado pelo CART algumas distribuições podem se unir. Apresentamos a seguir o ajuste feito pelo CART em cada cenário.

3.2 Ajuste das Árvores

O primeiro passo para a geração dos dados sintéticos é o ajuste da árvore de classificação para os dados simulados. A árvore foi ajustada utilizando o *software* R, usando a função `tree` do pacote `Tree` (Ripley, 2018).

A Figura 3.4 apresenta a árvore estimada para o cenário 1 e as densidades teóricas das folhas. O gráfico à esquerda representa a árvore construída para o cenário 1. A primeira partição foi feita de acordo com o valor $X_2 = 0$ ou 1, depois cada galho foi dividido com relação a X_3, e por fim, alguns galhos divididos por X_1. As folhas resultantes apresentam um grupo em que a distribuição condicional da variável resposta Y é mais homogênea. Os valores mostrados em cada folha são as médias de Y para as observações em cada grupo.

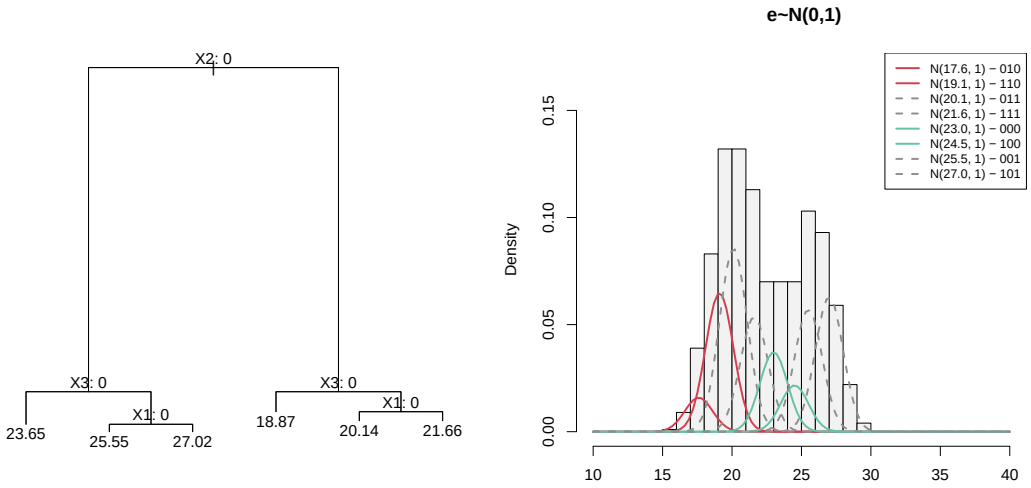

Figura 3.4: Cenário 1 - Árvore estimada e densidades para o cenário com erros gerados com distribuição normal padrão

O gráfico à direita representa o histograma de Y para o cenário 1 e as densidades estimadas de cada uma das oito combinações. As linhas traçadas representam as densidade das combinações que foram captadas pela árvore, As curvas com linha contínua representam as combinações que foram unidas pela árvore, Por exemplo as densidades das combinações (010) e (110) foram unidas em uma única folha pela árvore. No gráfico das densidades vemos que uma curva de sobrepõe a outra. O mesmo ocorreu com as combinações (000) e (100).

Cada folha obtida no modelo final foi enumerada de 1 a 6 em sequência da esquerda para a direita de acordo com a árvore da Figura 5. Sendo assim, por exemplo, a folha 1 (F_1) diz respeito à folha com $X_2 = 0$ e $X_3 = 0$, sendo a média

observada igual à $23,65$. Agora a folha 5 (F_5) corresponde a folha com $X_2 = 1$, $X_3 = 1$ e $X_1 = 0$, com média observada igual à $20,14$.

A Tabela 3.5 mostra a média teórica, a média estimada e a variância observada de cada folha. As médias teóricas foram obtidas a partir da equação (3.1). As médias teóricas são as calculadas com a equação (3.1) usando os valores de X de cada combinação. Mas para as duas folhas em que houve a junção de duas curvas foi calculada a média ponderada usando probabilidades proporcionais aos pesos π de cada combinação.

Tabela 3.5: Cenário 1 - Estatísticas descritivas CART

	F_1	F_2	F_3	F_4	F_5	F_6
n	152	146	153	194	216	139
$\mu_{\text{teór.}}$	$23,55$	$25,50$	$27,00$	$18,81$	$20,10$	$21,60$
μ_{obs}	$23,65$	$25,55$	$27,02$	$18,87$	$20,14$	$21,66$
σ^2_{obs}	$(1,45)$	$(1,05)$	$(0,87)$	$(1,03)$	$(1,21)$	$(0,96)$

Após a estimação da árvore, geramos dados sintéticos utilizando o *bootstrap* Bayesiano (Rubin, 1981) descrito na Seção 2,3. Quando a variável resposta é contínua, os valores de Y são amostrados aleatoriamente a partir da densidade estimada usando o método da função de distribuição acumulada (CDF) inversa (Stehfest, 1970).

Então, utiliza-se o *bootstrap* Bayesiano para gerar valores sintéticos a partir das densidades estimadas com as observações classificadas em cada folha.. O número de pontos utilizados para estimar a densidade da distribuição através do *bootstrap* foi definido como 300 observações, após ter sido feito estudo com os valores 250, 300, 500 e 1000. Foi notado que a distribuição da densidade foi muito parecida para 300, 500 e 1000. O tamanho dessa amostra deve ser maior que o valores de observações de cada folha. Essas observações são utilizadas para a geração da densidade teórica em cada folha. A densidade da amostra foi estimada utilizando Kernel Gaussiano (Wegman, 1972), e o suporte de cada densidade estimada deve estar restrito ao mínimo e máximo de cada amostra.

Tendo as densidades estimadas, utiliza-se o método da CDF inversa para gerar os valores sintéticos. Seja W uma variável aleatória com função de distribuição F. Definimos F^{-1} como a função $F^{-1}(y) = inf\{w : F(w) \geq y\}, 0 \leq y \leq 1$. Se $U \sim U(0,1)$, a CDF transformada inversa de $F^{-1}(U)$ é dada por $P(F^{-1}(U) \leq x) = P(U \leq F(W)) = F(w)$.

Portanto, para gerar a variável W, dado uma variável aleatória $U \sim U(0,1)$ e sua CDF inversa F^{-1}, aplicamos o seguinte algoritmo:

1. Gere $U \sim U(0,1)$.

2. Retorne $W = F^{-1}(U)$

Então esse procedimento é repetido para cada folha da árvore estimada, e os dados sintéticos $Y_{rep,i}$ são gerados de acordo com o tamanho e densidade de cada folha. Esse procedimento é repetido $m = 5$ vezes, resultando nos bancos sintéticos $(\tilde{D}_1, ..., \tilde{D}_5)$.

O mesmo procedimento descrito foi aplicado para todos os cenários. As Figuras 3.5 e 3.6 apresentam as árvores estimadas e as densidades teóricas de cada combinação para os outros dois cenários, respectivamente.

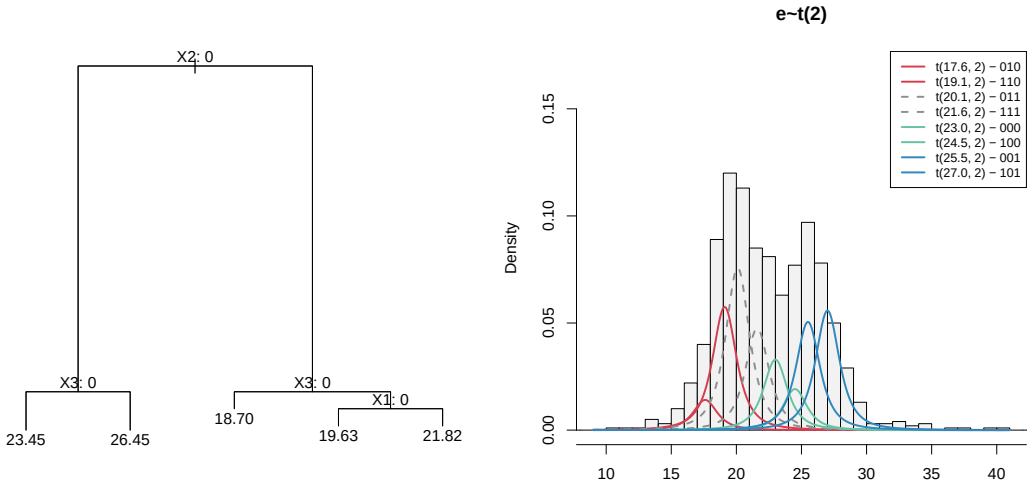

Figura 3.5: Cenário 2 - Árvore Estimada e Densidades para o cenário com erros gerados com distribuição t com 2 graus de liberdade

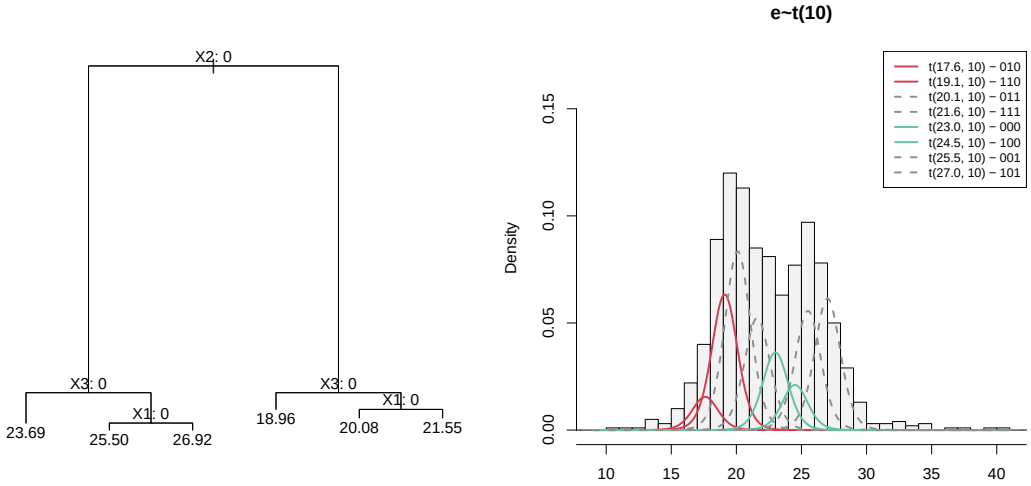

Figura 3.6: Cenário 3 - Árvore Estimada e Densidades para o cenário com erros gerados com distribuição t com 10 graus de liberdade

De acordo com as Figuras 3.4, 3.5 e 3.6 vemos que as árvores estimadas para os três cenários resultou em diferentes partições de Y. Para o cenário 2, apenas cinco folhas finais foram estimadas, o que pode estar associado ao fato da distribuição de Y possuir uma cauda mais pesada. Para os cenários 1 e 3, vemos que as partições foram idênticas, a diferença foi apenas nas médias de cada folha, embora os valores foram bem parecidos. Na sessão a seguir apresentamos os resultados para a geração dos dados sintéticos.

A Tabela 3.6 apresenta as médias encontradas para o cenário 2 para cada folha e a Tabela 3.7 apresenta as médias encontradas para o cenário 3.

Tabela 3.6: Cenário 2 - Estatísticas descritivas CART

	F_1	F_2	F_3	F_4	F_5
n	152	299	194	216	139
$\mu_{\text{teor.}}$	$23,55$	$26,28$	$18,81$	$20,10$	$21,60$
μ_{obs}	$23,45$	$26,45$	$18,70$	$19,63$	$21,82$
σ^2_{obs}	$(1,03)$	$(1,10)$	$(1,05)$	$(1,22)$	$(1,04)$

Tabela 3.7: Cenário 3 - Estatísticas descritivas CART

	F_1	F_2	F_3	F_4	F_5	F_6
n	152	146	153	194	216	139
$\mu_{\text{teor.}}$	$23,55$	$25,50$	$27,00$	$18,81$	$20,10$	$21,60$
μ_{obs}	$23,69$	$25,50$	$26,92$	$18,96$	$20,06$	$21,55$
σ^2_{obs}	$(1,33)$	$(1,06)$	$(1,00)$	$(1,33)$	$(1,02)$	$(1,39)$

3.3 Análises dos Dados Sintéticos

Após gerar os dados sintéticos, obtemos as estimativas da média em cada folha pra cada base $\tilde{D}_i$, e combinamos os resultados com os métodos de combinação de dados parcialmente sintéticos descritos na Seção 2.3. A Tabela 3.8 mostra os valores da média (μ_{rep}) e variância (σ^2_{rep}) dos dados sintéticos para o cenário 1.

Tabela 3.8: Cenário 1 - Média e variância dados sintéticos para o cenário com erros gerados com distribuição normal padrão

	F_1	F_2	F_3	F_4	F_5	F_6
n	152	146	153	194	216	139
μ_{obs}	$23,650$	$25,551$	$27,023$	$18,873$	$20,140$	$21,661$
μ_{rep}	$23,260$	$25,394$	$27,209$	$18,040$	$20,142$	$23,612$
σ^2_{rep}	$(1,722)$	$(1,752)$	$(1,631)$	$(1,162)$	$(1,304)$	$(1,631)$

A Figura 3.7 apresenta o intervalo de confiança de 95% para a média de cada folha para os dados sintéticos e para os dados originais.

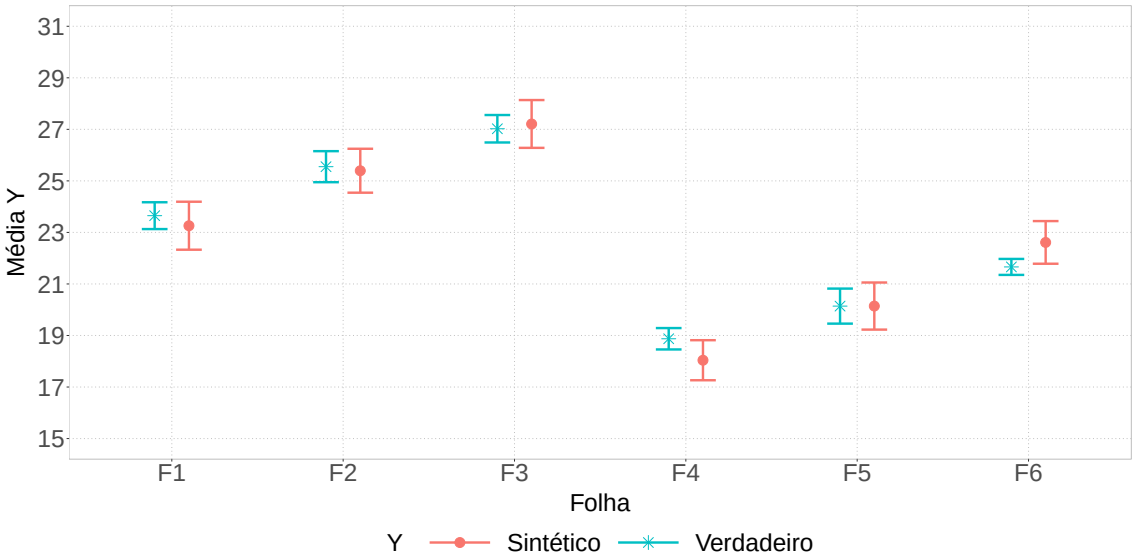

Figura 3.7: Cenário 1 - Intervalo de confiança da média por folhas obtidos com os dados originais e com os dados sintéticos

Para medir a utilidade dos dados, um modelo de regressão foi ajustado para cada banco de dados e seus resultados combinados, conforme o procedimento de combinação descrito na Seção 2.3. A Tabela 3.9 mostra a comparação entre os $\beta's$ estimados para os dados originais e para os dados sintéticos. Os resultados dos dados sintéticos foram similares aos calculados para os dados originais. Isso indica que, tanto para a estimação da média quanto para a regressão linear, os resultados obtidos com os dados sintéticos gerados a partir do CART se mostraram satisfatórios.

Tabela 3.9: Estimação pontual dos coeficientes de regressão estimados com os dados originais e com os dados sintéticos para o cenário 1

	β_{orig}	β_{sin}	Erro Relativo
β_0	23.000	23.203	0.009
β_1	1.500	1.508	0.0004
β_2	-5.400	-5.519	0.0220
β_3	2.500	2.441	-0.0236

A Figura 3.8 apresenta os intervalos de confiança para os betas estimados através dos dados originais e das combinações dos dados sintéticos. Vemos que a amplitude dos intervalos dos betas sintéticos são um pouco maiores que as dos betas originais.

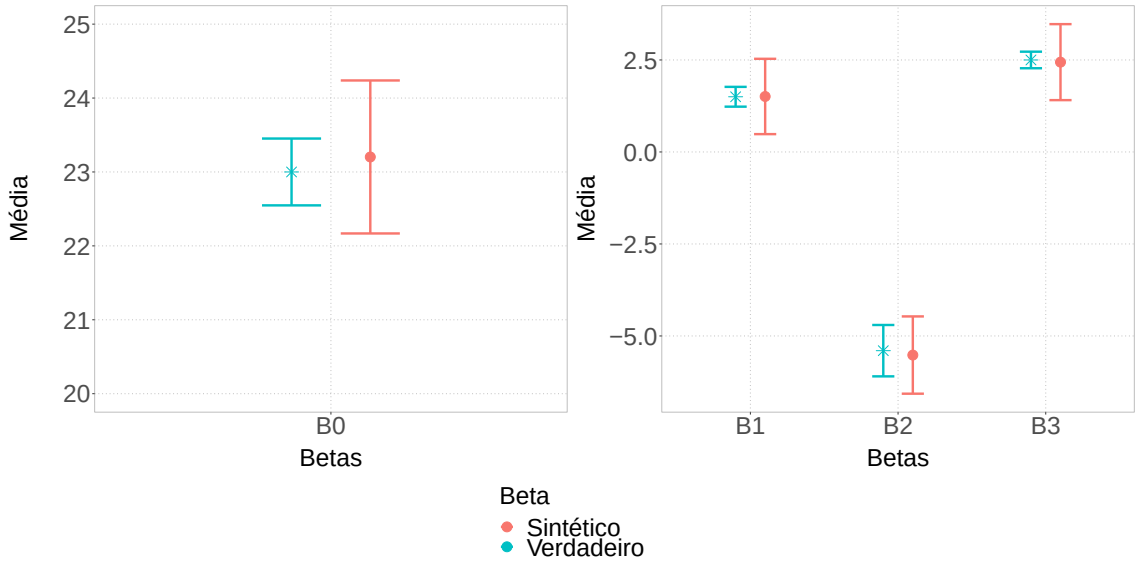

Figura 3.8: Cenário 1 - Intervalo de Confiança de 95% coeficientes de regressão estimados com os dados originais e dados sintéticos

As Tabelas 3.10 e 3.11 apresentam os resultados de estimação das médias para os cenários 2 e 3 respectivamente.

Tabela 3.10: Média e variância dados sintéticos - Cenário 2

	F_1	F_2	F_3	F_4	F_5
n	152	299	194	216	139
μ_{obs}	$23,452$	$26,451$	$18,703$	$19,630$	$21,821$
μ_{rep}	$24,031$	$27,312$	$19,051$	$20,361$	$22,394$
σ^2_{rep}	$(2,591)$	$(2,952)$	$(3,013)$	$(2,210)$	$(2,551)$

Tabela 3.11: Média e variância dados sintéticos - Cenário 3

	F_1	F_2	F_3	F_4	F_5	F_6
n	152	146	153	194	216	139
μ_{obs}	$23,691$	$25,512$	$26,923$	$18,960$	$20,080$	$21,552$
μ_{rep}	$24,553$	$26,019$	$28,235$	$18,321$	$21,122$	$20,522$
σ^2_{rep}	$(1,991)$	$(1,852)$	$(1,965)$	$(1,621)$	$(1,556)$	$(2,122)$

O cenário 3 apresentou resultados muito semelhantes ao cenário 1. O mesmo número de folhas foi estimado e também as combinações das covariáveis foram as mesmas. Já o cenário 2 também captou algumas combinações, mas nesse caso a folha F_2 foi construída com um grupo maior de observações.

A Figura 3.9 apresenta o intervalo de confiança de 95% para a média de cada folha dos dados originais e sintéticos para o cenário 2. Comparando os dois intervalos das médias das folhas dos dados sintéticos e originais, vemos que no geral as estimativas são próximas.

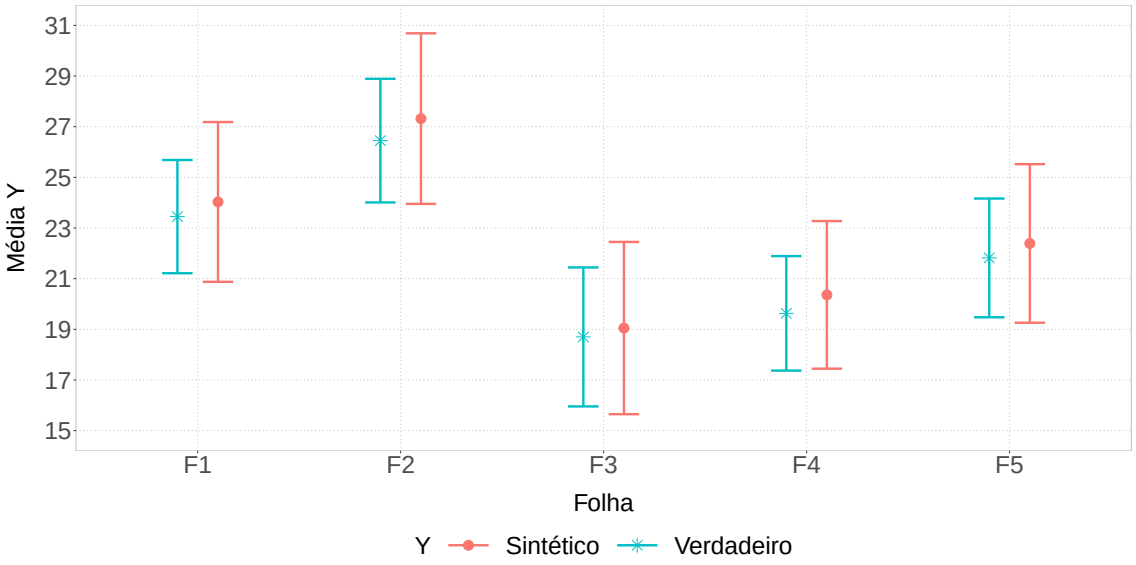

Figura 3.9: Cenário 2 - Intervalo de confiança da média por folhas obtidos com os dados originais e com os dados sintéticos

A Figura 3.10 apresenta o intervalo de confiança de 95% para a média de cada folha entre dos dados originais e sintéticos para o cenário 3. Vemos que os intervalos para as médias obtidas com os dados originais e sintéticos foram próximos Comparando esse cenário com o cenário 3, podemos afirmar que os resultados foram próximos. Isso se deve a distribuição dos dois cenários serem semelhantes.

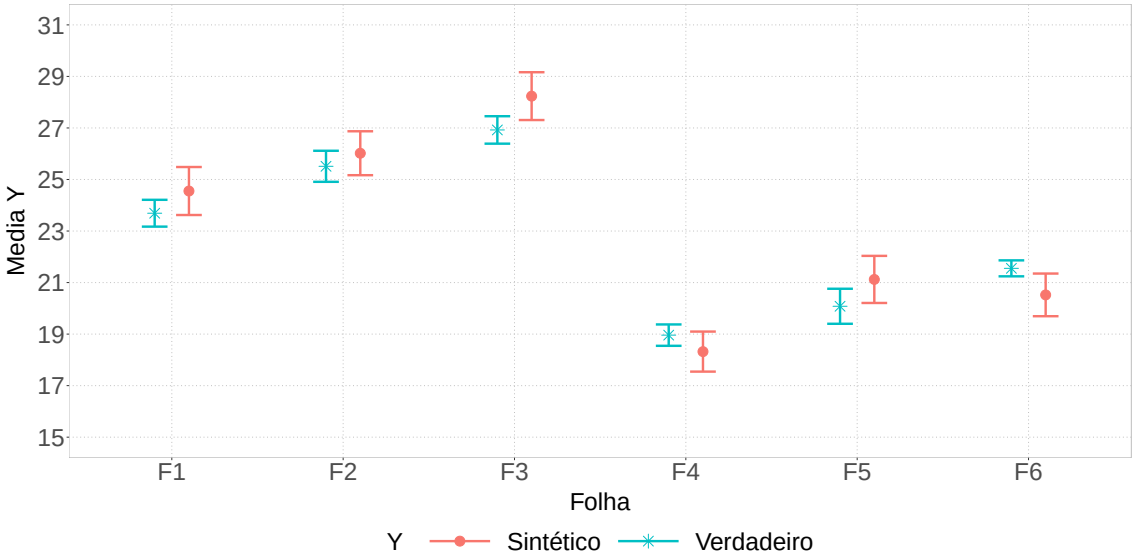

Figura 3.10: Cenário 3 - Intervalo de confiança da média por folhas obtidos com os dados originais e com os dados sintéticos

Entre os três cenários, o cenário que obteve resultados mais divergentes foi o cenário 2, o que era esperado devido à distribuição assumida por Y. Vemos na construção da árvore no cenário 2 que apenas 5 folhas foram estimadas. Esse resultado teve impacto na geração dos dados sintéticos, visto que mais de uma densidade teórica foi unida pela estimação da árvore. Nos resultados pontuais, embora os intervalos nem sempre se sobrepõem, os valores das médias originais e sintéticas ficaram próximos.

Da mesma forma como visto no cenário 1, foi ajustado um modelo de regressão linear normal para os dados sintéticos dos cenários 2 e 3. A Tabela 3.12 apresenta a estimação pontual dos betas da regressão linear para os cenários 2 e 3 em comparação com os betas originais e os respectivos erros relativos (E.R) de cada modelo.

Tabela 3.12: Estimação pontual dos coeficientes de regressão estimados com os dados originais e com os dados sintéticos para os cenários 2 e 3

		Cenário 2	Cenário 3		
	β_{orig}	β_{sin}	β_{sin}	E.R Cenário 2	E.R Cenário 3
β_0	$23,000$	$28,241$	23.546	0.097	0.0237
β_1	$1,500$	$2,653$	1.781	0.0536	0.0130
β_2	$-5,400$	$-1,556$	-5.010	-0.7118	-0.0722
β_3	$2,500$	$3,489$	2.785	0.3956	0.1140

As Figuras 3.11 e 3.12 apresentam o intervalo de confiança de 95% para os coeficientes de regressão dos cenários 2 e 3 respectivamente.

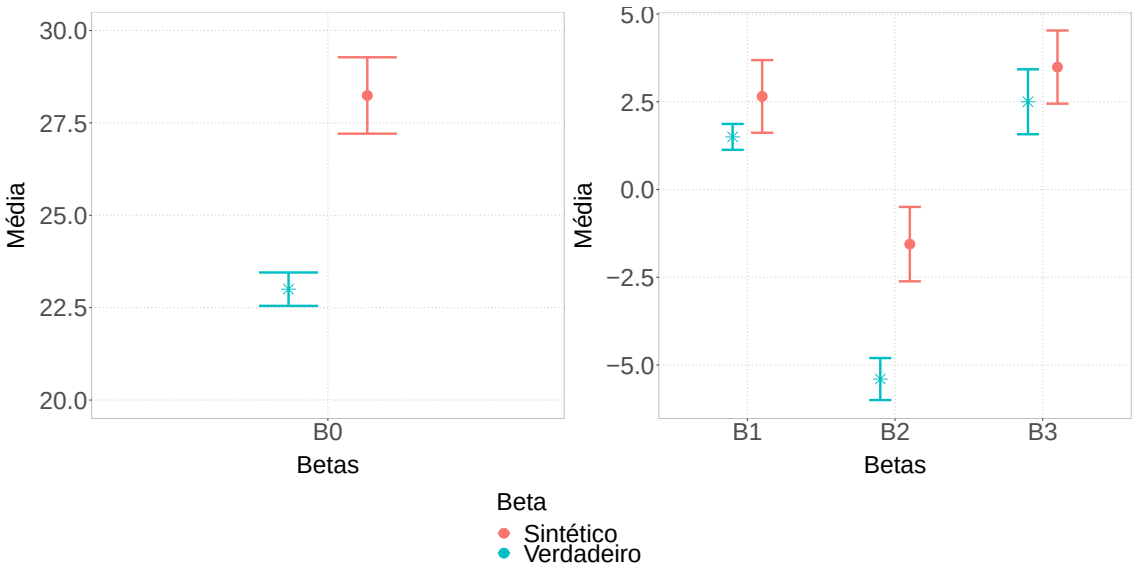

Figura 3.11: Cenário 2 - Intervalo de Confiança de 95% coeficientes de regressão estimados com os dados originais e dados sintéticos

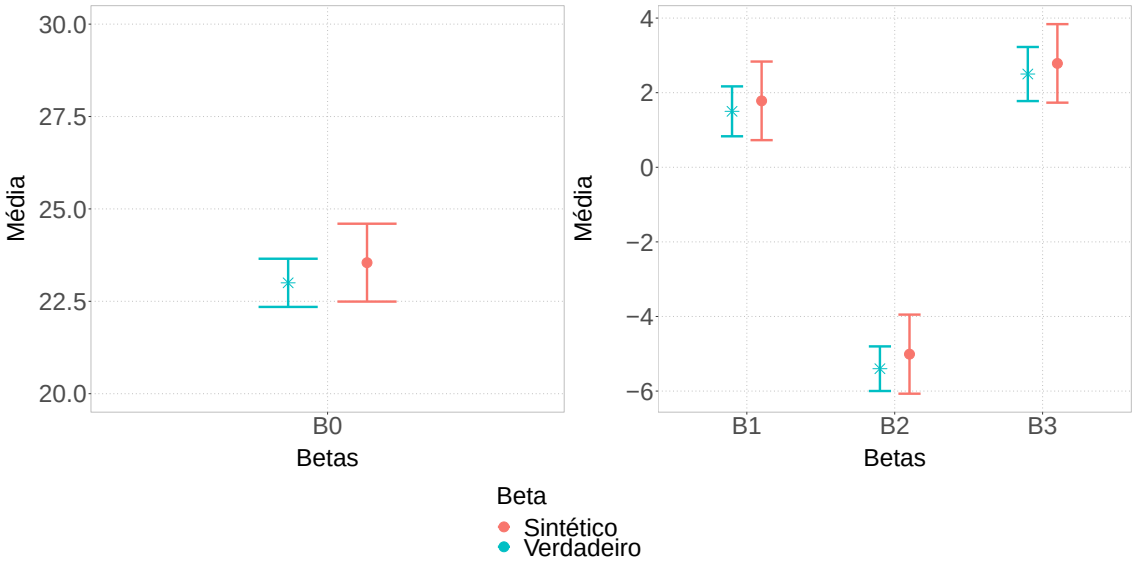

Figura 3.12: Cenário 3 - Intervalo de Confiança de 95% coeficientes de regressão estimados com os dados originais e dados sintéticos

Vemos que, como esperado, o cenário 2 obteve resultados com os dados sintéticos mais distantes dos valores originais. Portanto, o cenário 2 apresentou piores resultados para as estimações pontuais das médias de cada folha e dos coeficientes da regressão linear. Embora medidas pontuais sejam importantes para a comparação entre dados originais e sintéticos, é preciso também fazer uma análise sobre a utilidade geral e risco de divulgação, pois essas duas medidas combinadas mostram melhor como os dados sintéticos podem ser utilizados para fazer inferências. Nas Seções 3.3.1 e 3.3.2 são apresentados esses cálculos para os três cenários.

3.3.1 Medida de Risco e Utilidade

Utilizando as fórmulas descritas na Seção 2.3.1 foram calculadas as medidas de resumo de risco para cada cenário e para diferentes valores de l. Foram testados alguns valores próximos a variância. Esses valores são escolhidos pelo intruso de acordo com a sua confiança em relação a informação t que ele possui. Para ilustrar o impacto no risco de divulgação usamos os valores $l = 0,3; 0,5; 0,7; 1,0$. A Tabela 3.13 apresenta as medidas de risco para $l = 1,0$. Os demais valores são apresentados a seguir.

Tabela 3.13: Medidas de Risco

	Cenário 1	Cenário 2	Cenário 3
EMR/n	0,259	0,361	0,272
TMR	0,145	0,120	0,148
FMR	0,000	0,001	0,000

Os cenários 1 e 3 apresentaram medidas de risco parecidas. O cenário 2 apresenta o maior valor de EMR/n, o que indica que é esperado que o intruso consiga identificar algum alvo mais vezes do que nos outros 2 cenários. Para os 3 cenários o valor de FMR permaneceu baixo, ou seja, a identificação falsa de um alvo é muito baixa nos três casos. Essa medida pode ser usada para medir o quanto podemos "enganar" o intruso com os dados sintéticos, ou seja, qual a probabilidade do invasor escolher um alvo incorretamente.

A medida de TMR, que representa a probabilidade do intruso identificar corretamente o alvo, pode ser considerada baixa nos três casos, embora no cenário 2 essa medida seja um pouco mais baixa. Então para os 3 cenário gerados, os respectivos bancos de dados sintéticos para serem liberados ao público possuem baixa probabilidade de identificação correta.

A Figura 3.13 apresenta os gráficos com as medidas de risco para os 3 cenários e diferentes valores de l. Esses valores podem representar o quão confiante o intruso está quanto a informação que ele possui sobre o banco de dados original. Assumindo intervalos maiores, podemos inferir que o intruso não sabe ao certo o possível valor real para o alvo, então ele escolhe um valor de l maior. A escolha de l pode estar ligada tanto a informação que o invasor detém, quanto a escala original dos dados. Por exemplo, em um estudo relacionado à renda, dependendo da magnitude dos dados, o valor de l estará ligado a quanto o intruso considera uma margem razoável pra identificar a renda de alguém. Na Seção 4 discutimos a escolha dos valores de l para um caso aplicado.

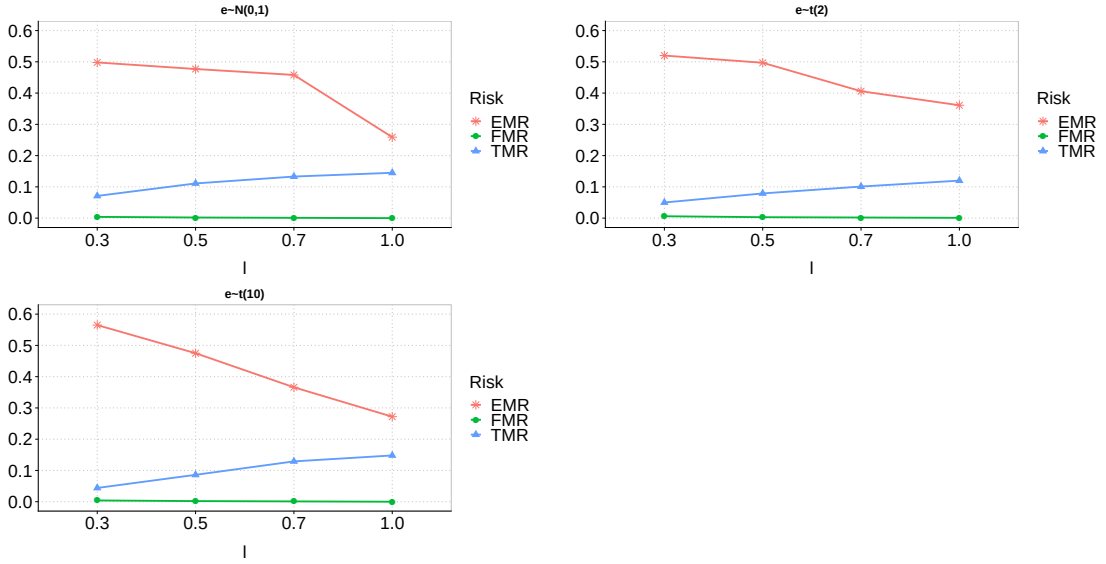

Figura 3.13: Medidas de Risco para os três cenários e diferentes valores de l

Para o cenário 1, aumentando os valores de l, o valor de EMR tende a diminuir. Isso ocorre pois, sendo o intervalo de valores maior, mais observações dos bancos de dados sintéticos podem estar dentro desse critério. Então, conforme a equação (2.15) mais valores de máximos podem ser achados, e a probabilidade de cada máximo é diluída em cada observação. Dessa forma, espera-se que tendo mais valores de comparação a correspondência esperada seja menor.

Em relação à medida TMR, conforme os valores de l aumentam, essa medida diminui. A diminuição também se deve ao fato de que tendo um intervalo maior de valores, mais máximos de probabilidade de correspondência podem ser encontrados, então a probabilidade de achar um único máximo, e esse máximo ser a verdadeira correspondência, tende a ser pequena.

As medidas de risco para os cenários 1 e 3 foram parecidas. Os valores de FMR em todos os cenários foram pequenos indicando que, mesmo aumentando os valores de l em cada cenário, o intruso faz poucas combinações falsas. Os valores de TMR são considerados bons de acordo com a decisão dos responsáveis pelo banco de dados. Nos casos simulados foi considerado que esses valores para qualquer l são baixos, ou seja, a probabilidade do invasor acertar um determinado alvo é pequena.

Sendo o risco considerado satisfatório para a divulgação dos dados sintéticos, é preciso avaliar se os dados divulgados podem fornecer inferências válidas para análises. Embora na Seção 3.3 apresentamos as comparações entre as medidas pontuais de cada cenário, é preciso verificar se no geral, dados sintéticos e dados originais são semelhantes. Para isso usamos a medida de utilidade apresentada na Seção 2.3.2, onde é feito o cálculo do escore de propensão entre o banco original e os dados

sintéticos para cada cenário.

A Tabela 3.15 apresenta os valores finais do escore de propensão para os 3 cenários estudados.Para cada cenário foi calculado o escore de propensão entre os dados originais e cada banco de dados sintético. Depois a média dos escores calculados para cada cenário é calculado, tendo assim a medida de utilidade final em cada cenário. Segundo Dreschsler e Reiter (2011) a medida de utilidade, embora não possua um valor padrão de utilidade considerada boa, deve ser próxima a zero. Vemos que para os cenário 1 e 3 essas medidas foram mais próximas a zero. O cenário 2 apresentou um valor um pouco maior, o que pode estar relacionado a distribuição dos dados ter cauda um pouco mais pesada, mas ainda assim pode ser considerada baixa.

Tabela 3.14: Medidas de Utilidade para os três cenários

Cenário 1	Cenário 2	Cenário 3
0,0632	0,1172	0,0592

A Figura 3.14 apresenta o gráfico com as medidas de utilidade para cada um dos cenários e para cada banco de dados sintéticos, com as medidas destacadas em vermelho.

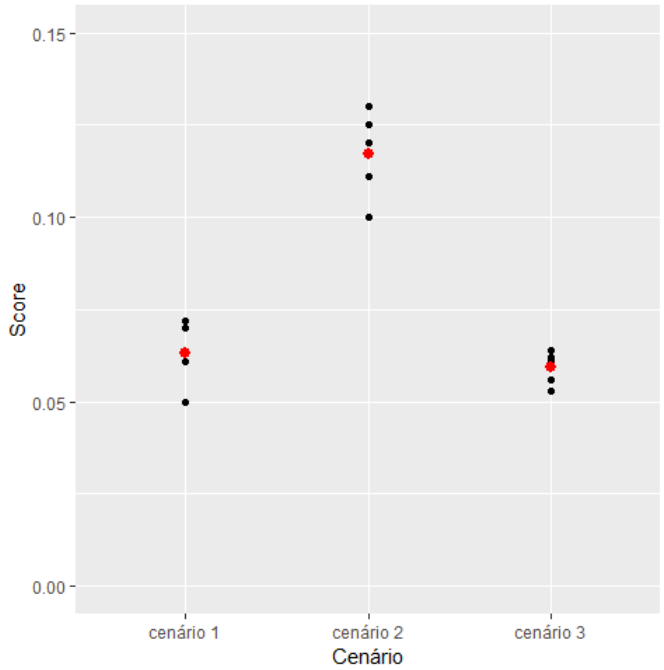

Figura 3.14: Medidas de Utilidade para os três cenários

Analisando o cenário 2, vemos que os valores de escore são maiores que os outros dois cenários, mas ainda assim os valores são próximos a zero. Então podemos concluir que para os três cenários os bancos de dados sintéticos, apenas os cenário 1 e 3 apresentaram resultados satisfatórios. O cenário 2, que possui uma cauda mais pesada para a distribuição do erro, não obteve resultados satisfatórios nas estimações pontuais, embora o risco esteja controlado, a utilidade ficou comprometida. Então, agências podem utilizar os bancos de dados sintéticos combinados para fazerem inferências muito próximas aquelas feitas nos bancos originais. Como vimos na Seção 3.3, as medias pontuais e os coeficientes das regressões realizadas em cada cenário foram muito semelhantes entre dados verdadeiros e sintéticos. Isso indica que as relações presentes nos bancos de dados originais foram captadas pelos bancos de dados sintéticos. Caso os analistas queiram fazer outras inferências com os bancos sintéticos, basta utilizar as equações apresentadas na Seção 2.2.1 para a combinação dos bancos sintéticos e estimação das medidas de interesse.

Analisando todas as etapas do processo de geração de dados sintéticos e avaliações das medidas de risco e utilidade, podemos notar que o CART apresentou resultados satisfatórios para o modelo final. A construção das árvores em todos os cenários conseguiu captar as relações existentes entre as variáveis presentes nos bancos de dados observados. A seguir é apresentada a análise da geração de dados sintéticos para um banco de dados real.

Capítulo 4

Aplicação em Dados Reais

A seguir analisa-se um banco de dados real. O conjunto de dados apresenta informações sobre a renda de famílias americanas no ano de 2004. O banco de dados possui informações de residências onde apenas um morador é entrevistado. Dentre as principais perguntas estão a idade, estado civil, anos de educação dentre outras. Existem também informações sobre a renda geral da família, bem como algumas informações sobre o cônjuge do entrevistado, se esse existir.

O banco de dados possui algumas informações financeiras sobre o entrevistado como o valor da apólice de seguro, renda, dentre outras. Uma determinada agência pode estar interessada em divulgar esse banco de dados, mas por questões éticas, não pode divulgar informações sensíveis como renda. Portanto, a divulgação do banco de dados através da metodologia de geração de dados sintéticos pode ser útil.

Dessa forma, suponha que o intruso tem interesse em descobrir informações sobre um alvo nesse banco de dados. Ele pode possuir alguma pré-informação e sabe que o alvo está nesse banco de dados. Visto que o invasor sabe as covariáveis do banco original, ele precisa apenas identificar a renda verdadeira da pessoa de interesse através dos m bancos de dados sintéticos. A seguir é apresentada a análise descritiva desse banco de dados real, bem como os resultados para a geração dos dados sintéticos.

4.1 Descrição

O conjunto de dados reais utilizados para aplicação da geração de dados sintéticos utilizando o CART é proveniente da pesquisa sobre finanças do consumidor (SCF) realizada em 2004 nos EUA. Esse banco apresenta uma amostra nacionalmente representativa que contém informações abrangentes sobre ativos, passivos, receita e características demográficas daqueles amostrados. O banco utilizado contém uma

amostra aleatória de 500 domicílios com rendimentos positivos que foram entrevistados na pesquisa de 2004. Esses dados são largamente utilizados para o cálculo de seguros de vida baseadas nas variáveis presentes no banco de dados, segundo Frees (2011).

O banco de dados apresenta 500 observações e 18 variáveis. Apresentamos abaixo a descrição das variáveis.

1. **INCOME**: Renda anual da família;
2. **TOTINCOME**: Total da renda;
3. **GENDER**: 0= Homem, 1= Mulher;
4. **AGE**: Idade do entrevistado;
5. **MARSTAT** : 1=Casado, 2= mora junto, 0 caso contrário;
6. **EDUCATION** : Número de anos de estudo do entrevistado;
7. **ETHNICITY** :1=Branco, 2=Negro, 3=Latino, 7= Outro
8. **SMARSTAT**: Estado civil do cônjuge do entrevistado;
9. **SGENDER**: Gênero do cônjuge;
10. **SAGE** : Idade do cônjuge;
11. **SEDUCATION** : Anos de estudo do cônjuge;
12. **NUMHH** : Número de membros da família;
13. **CHARITY**: contribuições para a caridade;
14. **FACE** : Valor de seguro pago pela seguradora;
15. **FACECVLIFEPOLICIES**: Valor nominal da apólice de seguro de vida com valor em dinheiro;
16. **CASHCVLIFEPOLICIES**: Valor em dinheiro da apólice de seguro de vida;
17. **BORROWCVLIFEPOL**: Montante emprestado na apólice de seguro de vida com um valor em dinheiro;
18. **NETVALUE**: Valor líquido em risco na apólice de seguro de vida com valor em dinheiro;

Suponha que um intruso tem como alvo o banco de dados apresentado e o interesse é descobrir a renda anual familiar de um determinado alvo. Nesse caso a agência detentora dessas informações pode ter o interesse de divulgar esse banco de dados, mas por questões éticas não poderá fazer a divulgação dos dados originais. Esse é um exemplo onde a aplicação de dados sintéticos pode ser utilizada, pois as informações mais sensíveis do banco de dados não serão divulgadas.

Assim, iremos considerar a variável INCOME como a variável mais sensível do banco de dados. As covariáveis escolhidas para o estudo foram aquelas associadas ao entrevistado que estão relacionados à renda do indivíduo. Sendo assim, as covariáveis selecionadas para o estudo foram: INCOME, GENDER, AGE, MARSTAT, EDUCATION, ETHINICITY, NUMHH e CHARITY.

4.2 Análise Descritiva

O banco de dados possui $82,6\%$ de mulheres e $17,4\%$ de homens. As idades variam entre 20 e 85 anos. Dos entrevistados, 27% são solteiros, 66% são casados ou amigados, e 7% divorciados. Considerando a etnia, 73% são brancos, 14% negros, 8% latinos, e 5% de outras etnias.

A Tabela 4.1 apresenta algumas medidas descritivas para a variável INCOME. Essa variável possui uma dispersão muito grande, o que pode dificultar na geração de dados sintéticos, por isso foi utilizada a transformação log nos dados. Dessa forma, os dados sintéticos foram gerados com a transformação, e depois retornando à escala dos dados originais. A Tabela 4.2 apresenta as estatísticas descritivas para a variável log(INCOME).

Tabela 4.1: Estatísticas descritivas INCOME

Min.	1º Qu.	Mediana	Média	3º Qu.	Max.	Desvio Padrão
260	28.000	54.000	321.022	106.000	75.000.000	3.410.936

Tabela 4.2: Estatísticas descritivas log(INCOME)

Min.	1º Qu.	Mediana	Média	3º Qu.	Max.	Desvio Padrão
5,561	10,240	10,897	10,925	11,571	18,133	1,394

A Figura 4.1 apresenta o histograma da variável renda (INCOME) e também para o log(INCOME). A dispersão presente no banco de dados é diminuída com a transformação log. Dessa forma utilizaremos essa transformação para as análises a seguir.

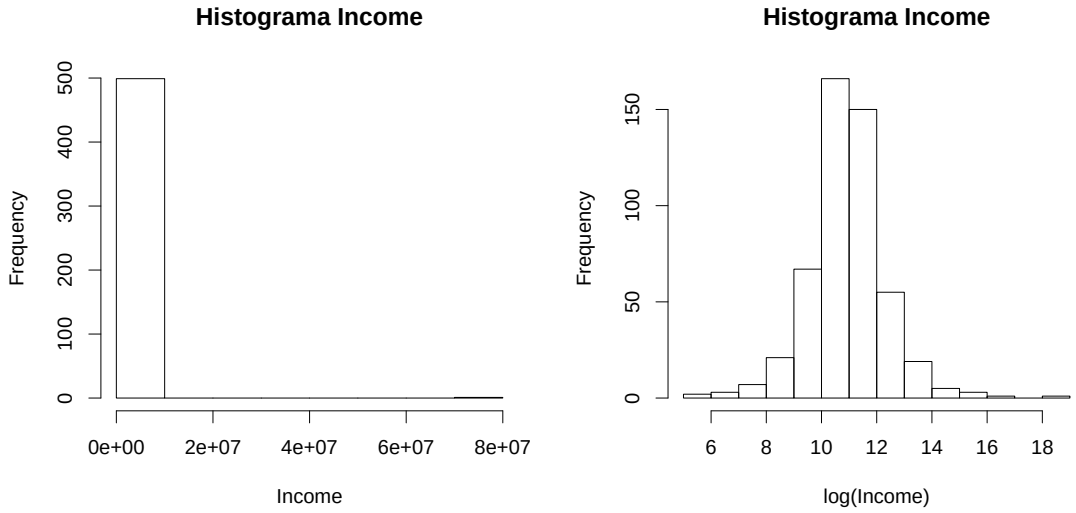

Figura 4.1: Histograma INCOME e Log(INCOME)

A variável AGE foi recategorizada de acordo com as faixas etárias mais comumente utilizadas, sendo:

- 1, AGE entre 16 e 29 anos - (12%);
- 2, AGE entre 30 e 59 - (68%);
- 3 AGE $\geq$ 60 - (20%);

A variável EDUCATION também foi recategorizada conforme os anos de estudo sendo:

- 0, se EDUCATION $\leq$ 9 - (6%);
- 1, EDUCATION entre 10 e 12 anos - (25%);
- 2, EDUCATION $\geq$ 13 - (69%);

A Tabela 4.3 apresenta os testes de comparação entre médias da variável sensível INCOME para as diferentes categorias das covariáveis. Existe apenas uma covariável contínua, CHARITY, e sua correlação com INCOME foi de $0,072$ com p-valor de 0,110, indicando não correlação entre os dados.

Tabela 4.3: Testes de comparação entre covariáveis e INCOME

Covariável	P-valor
GENDER	<0,001 *
EDUCATION	<0,001**
ETHNICITY	<0,001**
GENDER	<0,001**
MARSTAT	<0,001**
NUMHH	0,071**

* Teste T-student

* Teste Anova

4.3 Geração e Análise dos Dados sintéticos

Para a geração dos dados sintéticos, primeiramente foi gerada a árvore de classificação e regressão para as variáveis associadas ao entrevistado conforme descrito na Seção 4.1. Foram excluídos 2 registros com renda superior a $US\$1.000.000$, pois além do ajuste ser melhor sem eles, o risco de divulgação é maior por serem valores únicos muito altos. A Figura 4.2 apresenta a árvore gerada. A variável que melhor divide os valores de Renda foi o estado civil. Apenas 3 variáveis foram significativas para o modelo, MARSTAT, EDUCATION e AGE.

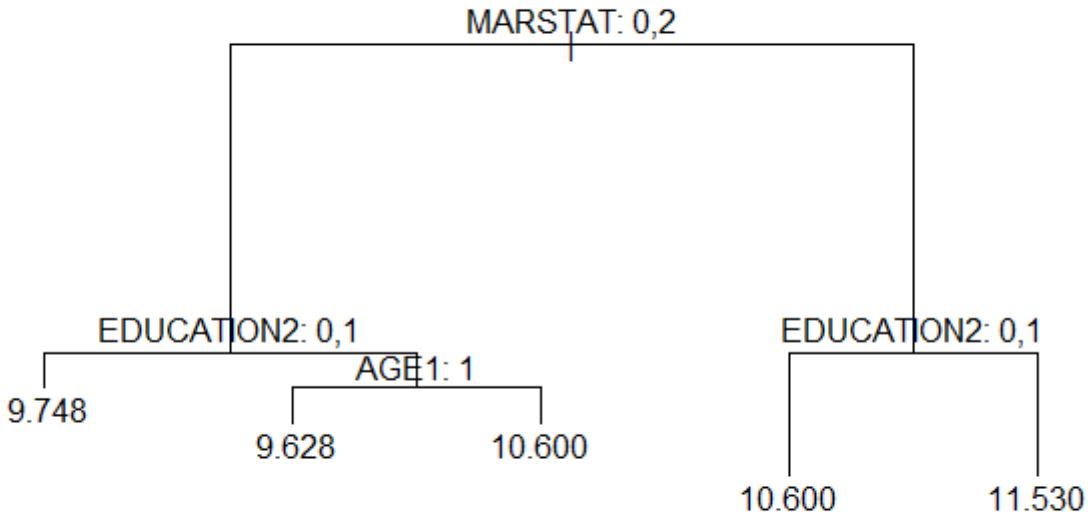

Figura 4.2: Árvore INCOME

Todas as covariáveis finais do modelo possuem mais de um fator. Vemos que a primeira variável escolhida para iniciar a divisão do modelo foi o estado civil. As categorias 0 e 2 foram unidas, permanecendo apenas a categoria 1 de estado civil (casado ou amigado) sozinha para a divisão ao lado direito. Tanto para o lado esquerdo da primeira divisão, quanto para o direito, a segunda variável de divisão foi educação, sendo as categorias 0 e 1 (menos de 12 anos de estudo) unidas nos dois casos. E por fim, na última divisão a variável Idade na categoria 1 (entre 16 e 29 anos).

Usando a mesma nomenclatura das folhas apresentada na Seção 3.2, temos que a folha F_1 representa o grupo de pessoas solteiras ou divorciadas, com educação menor que 12 anos. Esse grupo possui média do logaritmo da renda igual a $9,748$. Já a folha F_2 representa as pessoas solteiras ou divorciadas, com mais de 12 anos de estudo e idade entre 16 e 29 anos, e a média do logaritmo da renda é $9,628$. A folha F_3 representa o grupo de pessoas solteiras ou divorciadas, com mais de 12 anos de estudo e idade superior a 30 anos e possuem a média do logaritmo da renda igual a $10,600$. A folha F_4 representa pessoas que não são solteiras, com até 12 anos de educação e média do logaritmo da renda igual a $10,600$. A última folha F_5 representa o grupo de pessoas casadas ou amigadas, com educação superior a 12

anos e com média do logaritmo da renda igual a $11,53$.

Assim, verificamos que a renda anual familiar pode ser explicada pelo estado civil, os anos de escolaridade e a idade do entrevistado. A seguir repetimos o mesmo procedimento descrito na Seção 3 para a geração dos dados sintéticos.

Foram gerados $m = 5$ banco de dados sintéticos. A Tabela 4.4 apresenta as medidas pontuais para a média do logaritmo da renda em cada folha da árvore.

Tabela 4.4: Medidas Estimadas – Combinação Bancos Sintéticos

	F_1	F_2	F_3	F_4	F_5
n	63	22	81	90	242
μ_{rep}	$9,812$	$9,752$	$10,502$	$10,589$	$11,789$
σ^2_{rep}	$1,387$	$2,066$	$2,079$	$1,931$	$1,612$

Dado os dados sintéticos de renda, suponha que um intruso deseja descobrir um alvo no banco de dados original. Ele acha plausível estipular uma faixa de renda para a qual ele aceite fazer uma correspondência com o valor original. Apresentamos a seguir alguns valores de l, que representam quais as faixas de renda o alvo pode ter. A Tabela 4.5 apresenta as medidas de risco de identificação para as diferentes faixas de renda para correspondência.

Tabela 4.5: Medidas de Risco para diferentes faixas de renda

l	8,517	9,210	9,903	10,308
$exp(l)$	5.000	10.000	20.000	30.000
EMR/n	0,402	0,385	0,306	0,245
TMR	0,172	0,186	0,199	0,206
FMR	0,100	0,093	0,080	0,075

A Figura 4.3 apresenta o gráfico de comparação das medidas de risco para diferentes valores de l.

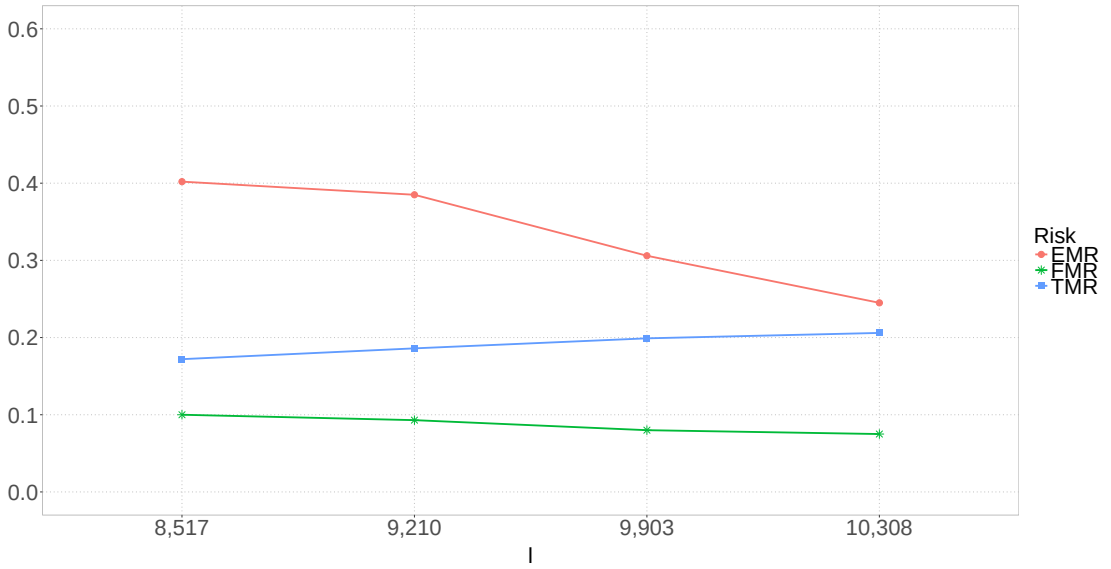

Figura 4.3: Medidas de Risco de identificação da renda

O valor de EMR/n representa o valor esperado de correspondência correta feita pelas informações do alvo. Esse valor tende a diminuir com o aumento do intervalo de correspondência para a renda, pois mais valores sintéticos podem estar dentro do intervalo de interesse. A medida TMR aumenta conforme os valores de l aumentam, ou seja, o intruso terá maiores chances de identificar corretamente seu alvo. As medidas de FMR podem ser consideradas baixas para todos os casos, logo o invasor não é facilmente enganado pelos dados sintéticos, ou seja, baixa correspondência falsa. Apesar disso, os valores podem ser considerados bons, visto que aqui consideramos o pior caso, onde o invasor sabe que o alvo em questão está nos bancos de dados sintéticos. No caso em que o intruso não sabe se o alvo está presente no banco de dados, os valores de EMR/n, TMR e FMR podem ser significativamente diminuídos.

Com relação às medidas de utilidade,a média dos valores do escore de propensão para cada banco de dados foi igual a $0,012$. Isso indica que os valores sintéticos gerados tem distribuições semelhantes aos do banco de dados originais.

Valores de renda presentes em um banco de dados usualmente não são divulgados por questões legais e éticas. A metodologia da geração de dados sintéticos pode ser uma melhor alternativa para a divulgação segura de dados do que aquelas apresentadas na Seção 1.1. Para essa aplicação no banco de dados SCF, o CART apresentou bons resultados tanto para as medidas pontuais, quanto para as medidas de risco e utilidade.

Capítulo 5

Discussão

A divulgação segura de informações tem sido cada vez mais de interesse de agências. Dentre as várias metodologias para a divulgação de dados sigilosos, a metodologia de dados sintéticos para divulgação de informações tem sido muito usada para esse fim. Nesse trabalho apresentamos a metodologia de divulgação de dados confidenciais através da geração de dados sintéticos. O interessante dessa técnica é a possibilidade maior de divulgação de informações mais parecidas com aquelas presentes no banco de dados original.

Dentre as várias técnicas disponíveis na literatura para a geração de dados sintéticos, escolhemos CART para gerar um modelo de classificação para dividir o espaço da variável escolhida como mais sensível. O CART é uma metodologia não paramétrica interessante,pois não é preciso fazer suposições sobre a distribuição dos dados envolvidos no estudo e sua utilização já é largamente desenvolvida em *softwares* estatísticos.

Nosso objetivo neste trabalho foi apresentar a técnica de geração de dados sintéticos para casos onde existe apenas uma variável com restrição de divulgação, utilizando o CART juntamente com o *bootstrap* Bayesiano. Por fim,utilizamos o método da CDF inversa para a geração dos dados sintéticos.

De acordo com as medidas de risco calculadas, os bancos de dados sintéticos gerados podem ser divulgados seguramente para a análise de dados. Vemos que o risco está sempre ligado ao nível de informação que o intruso já possui sobre o banco de dados. Nesse estudo foi considerado o caso mais grave, onde o intruso tem uma informação sobre parte do registro de algum alvo e o interesse é tentar descobrir o valor da variável sensível através dos bancos de dados sintéticos.

Para a utilidade, vemos que as medidas pontuais da média e dos coeficientes de regressão para os cenário 1 e 2 são bem similares entre o banco de dados sintético e original. Apenas para o cenário 2 esses resultados não foram considerados satis-

fatórios. Utilizando o escore de propensão para medir a similaridade entre os dados originais e sintéticos, vemos que em todos os cenários essa estimativa foi próxima a zero, mas o cenário 2 teve o maior escore, indicando não adequação dos dados sintéticos. Para os cenário 1 e 3 as distribuições da variável são próximas, ou seja, as inferências feitas com a combinação dos bancos de dados sintéticos são similares às realizadas no banco de dados original.

Podemos concluir que para distribuições mais dispersas a metodologia de geração de dados sintéticos através de árvores de classificação pode não resultar em resultados confiáveis. Embora a medida de risco foi considerada baixa para o cenário 2, e a utilidade ser relativamente próxima a zero, vemos que as medidas pontuais da média e dos coeficientes de regressão não foram próximas. Logo, nesse caso onde a distribuição dos dados é mais dispersa é interessante o uso de outras metodologidas de geração de dados sintéticos.

Nos estudos de simulação percebemos que, para os três cenários apresentados, os bancos de dados sintéticos conseguiram captar as relações existentes no banco de dados original. Percebemos que o cenário 2 obteve resultados um pouco piores que os outros três cenários em relação as medidas de risco e utilidade. Isso pode estar ligado ao fato da distribuição do cenário 2 ter uma cauda mais pesada.

O banco de dados real apresentou conclusões muito próximas daquelas mencionadas para o estudo simulado. As medidas pontuais foram próximas entre o banco de dados original e banco de dados sintéticos. No estudo sobre o risco, para os valores de l mencionados, as medidas de risco foram consideradas satisfatórias para esse banco. Embora não exista na literatura valores estipulados para uma boa medida de risco, podemos inferir que as agências devem estipular qual o valor mais aceitável para cada caso.

Para a escolha do melhor modelo para a geração dos dados sintéticos é preciso verificar o tipo de variável envolvida no estudo. O CART apresentou resultados satisfatórios para a geração dos dados sintéticos. Esse método pode ser uma estratégia simples e eficaz para a divulgação de dados sintéticos quanto temos o intuito de divulgar informações sigilosas.

Embora o CART tenha apresentado resultados satisfatórios para a geração de dados sintéticos, em estudos onde existam muitas variáveis envolvidas, a construção da árvore pode não ser simples. Quando as covariáveis possuem mais de uma categoria, a divisão da árvore acaba juntando categorias que parecem ser mais similares. Tanto nas simulações quanto no banco de dados real, as covariáveis escolhidas eram categóricas.

Em geral, podemos concluir que a metodologia de geração de dados sintéticos

é uma ótima ferramenta para a divulgação segura de dados. A importância desse método se deve ao fato de podermos controlar os dois principais focos da divulgação segura de dados, o risco existente na divulgação e se as inferências feitas nos bancos divulgados são próximas daquelas feitas no banco original.

Para nossos trabalhos futuros, temos o interesse em estudar a metodologia apresentada para outros tipos de variáveis ou até mesmo a combinação delas. Além disso, avaliar outras metodologias para a separação do espaço da variável sensível, como florestas aleatórias e modelos Bayesianos.

Capítulo 6

Bibliografia

1. Boser, B., Guyon, I., Vapnik, V., (1992). *A training algorithm for optimal margin classifiers.* In: Proceedings of the Fifth ACM Workshop on Computation Learning Theory. COLT. ACM Press, New York, pp. 144–152.

2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., (1984). *Classification and Regression Trees.* Wadsworh, Inc., Belmont, CA.

3. Breiman, L., (1996). *Bagging predictors.* Machine Learning 24, 123–140.

4. Breiman, L., (2001). *Random forests.* Machine Learning 45, 5–32.

5. Clark, L. and Pregibon, D. (1992). *Tree-based models.* In J. Chambers and T. Hastie, eds., Statistical Modelsin S. Belmont, CA: Wadsworth, Inc.

6. Cox, L. H. (1980). *Suppression methodology and statistical disclosure control.* Journal of the American Statistical Association, 75:377–385.

7. Drechsler, J. and Reiter, J. P. (2011). *An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets.* Computational Statistics and Data Analysis 55, 3232–3243.

8. Efron, D. (1979a). *The bootstrap methods: another look at the jakknife.* Ann Statistic. 7 1-26.

9. Ein-Dor, P. and Feldmesser, J. (1987). *Attributes of the performance of central processing units: a relative performance prediction model.* Comm. ACM. 30, 308–317.

10. Frees, E.W. (2011). *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press.

11. Friedman, J., Hastie, T., Tibshirani, R. (2001). *The elements of statistical learning.* Springer series in statistics. (Vol. 1, No. 10)

12. Karr A.F.. Reiter J. P. (2013) . *Using Statistics to Protect Privacy.* In Privacy, Big Data, and the Public Good: Frameworks for Engagement, 276 – 95 .

13. Raghunatha, T.T, Reiter, P.P. e Rubin, D.B.(2003). *Multiple imputation for statistical disclosure limitation.* Jornal of Official Statistics, 19,1-16.

14. Ripley, B. (2018). *Tree: Classification and Regression Trees.* R package version 1.0-39. Disponível em https://CRAN.R-project.org/package=tree.

15. Reiter, J. P. (2003a). *Inference for partially synthetic, public use microdata sets.* Survey Methodology forthcoming.

16. Reiter, J. P. (2003b). *Significance tests for multi-component estimands from multiply-imputed, synthetic microdata.* Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

17. Reiter, J.P., (2005). *Using CART to generate partially synthetic, public use microdata.* Journal of Official Statistics 21, 441–462.

18. Reiter, J. P. and Mitra, R. (2009). *Estimating Risks of Identication Disclosure in Partially Synthetic Data.* JThe Journal of Privacy and Condentiality, 99-110.

19. Reiter, J. P. and Raghunathan, T. E. (2007). *The multiple adaptations of multiple imputation.* Journal of the American Statistical Association, 102:1462–1471.

20. Rosenbaum, P. R. and Rubin, D. B. (1983). *The Central Role of the propensity score in observational studies for Causal Efects.* Biometrika, 70: 4155.

21. Rubin, D.B. (1981). *The Bayiesian boostratp. The Annals of Statistics.* Vol. 9, pp. 130-134.

22. Rubin, D.B.(1987). *Multiple Imputation for Nonresponse in Surveys.* Jonh Wiley & Sons, New York.

23. Rubin, D.B., (1993). *Discussion: Statistical disclosure limitation.* Journal of Official Statistics 9, 462–468.

24. Stehfest, H. (1970). *Algorithm 368: Numerical inversion of Laplace transforms [D5].* Communications of the ACM, 13(1), 47-49.

25. Sweeney L. (2013). *Known Patients to Health Records in Washington State Data.* Data Privacy Lab, IQSS, Harvard University

26. Wegman, E. J. (1972). *Nonparametric probability density estimation.* Technometrics 14, 533-546.

27. Woo, M. J., Reiter, J. P., Oganian, A., Karr, A. F. (2009). *Global measures of data utility for microdata masked for disclosure limitation.* Journal of Privacy and Confidentiality.