ISMENIA BLAVATSKY DE MAGALHÃES

AVALIAÇÃO DE REDES BAYESIANAS PARA IMPUTAÇÃO EM VARIÁVEIS QUALITATIVAS E QUANTITATIVAS

São Paulo

2007

ISMENIA BLAVATSKY DE MAGALHÃES

AVALIAÇÃO DE REDES BAYESIANAS PARA IMPUTAÇÃO EM VARIÁVEIS QUALITATIVAS E QUANTITATIVAS

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Engenharia.

Área de concentração: Engenharia de Controle e Automação Mecânica

Orientador: Prof. Dr. Fabio Gagliardi Cozman

São Paulo

2007

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 26 de abril de 2007.

Assinatura do autor: _____

Assinatura do orientador: _____

FICHA CATALOGRÁFICA

Magalhães, Ismênia Blavatsky de
Avaliação de redes bayesianas para imputação em variáveis qualitativas e quantitativas / I.B. de Magalhães. -- ed.rev. -- São Paulo, 2007.
217 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos.

1.Modelagem 2.Controle e decisão 3.Aplicação em redes bayesianas 4.Imputação I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos II.t.

Agradecimentos

"Viver é desenhar sem borracha."

(Millôr Fernandes)

... então só me resta aprender com os meus erros.

Mais que desenvolver esta tese, eu a vivi. Com todos os complementos que possam ter esta frase: sentimentos de alegria, tristeza, euforia, apatia, confiança, desespero e tantos outros que se seguiam em seqüência e inúmeras vezes apareciam em avalanche. Um estado que só quem o vive pode entendê–lo. Certamente todos aqueles que passaram por mim durante este período merecem um agradecimento. Serei injusta por não citar todos os nomes (até porque não lembraria de todos), mas seria muito mais se não registrasse um agradecimento especial àqueles que realmente me acompanharam:

- Deus. Em vários momentos, alguma força inexplicável surgia, principalmente através das palavras que vinham no momento certo. Só posso creditar essa força a Ele;

- ao meu orientador, Fábio Gagliardi Cozman, que confiou no meu trabalho e me guiou, sempre tranqüilo e ponderado em suas colocações;

- ao pessoal do laboratório, em especial ao Daniel Kikuti, que com seu bom humor tornou muito mais agradáveis os dias de trabalho e os cafés da tarde em São Paulo. Nossas discussões e nosso convívio serviram para que eu me tornasse uma pessoa melhor. Ao José Eduardo Ochoa, que com sua companhia animou os dias de trabalho aos sábados e as noites no *lab* durante a semana;

- aos funcionários da Poli–USP, especialmente aos servidores das secretarias e vigias do prédio da Engenharia Mecânica;

- à Sociedade Brasileira de *Coaching* e à Brahma Kumaris, que proporcionaram uma rede de novos amigos e histórias interessantes que motivaram e divertiram, além do

meu treinamento e formação;

- à minha *coach* Vilma Novais, que com sua voz de incentivo, exemplo de luta, perseverança e força de vontade, deu–me força maior para continuar. O "levanta astral" ainda ecoa;

- ao Instituto Brasileiro de Geografia e Estatística (IBGE) que cedeu a base de dados não identificados do Censo Demográfico Brasileiro 2000 na unidade da federação do Rio Grande do Norte, e à pesquisadora Joana Domingues Vargas que cedeu a base de homicídios em Campinas para a realização deste trabalho;

- aos meus revisores e a todos os membros da banca, que em leituras cuidadosas fizeram sugestões e apontaram as correções necessárias;

- aos meus chefes, Vermelho e Sonia, que me incentivaram e compreenderam os momentos de ausência do trabalho por causa da tese;

- aos colegas do IBGE, em especial ao mestre Djalma Pessoa, ao Alexandre Santos, José André Brito e Flávio Montenegro, que com algumas discussões, conversas e sugestões estimularam para o formato que este trabalho assumiu;

- aos médicos Dr. Marco Vianna e Dr. Júlio Cury e sua equipe, que me assistiram com cuidado e carinho;

- aos amigos Ronaldo Figueiró e Rose Utida, que em palavras e atividades terapêuticas surgidas no momento certo, tiraram–me de algumas crises de nervos;

- à minha família em São Paulo por quase cinco meses: Carmen Pedroso, Mércia Silva, Sirley Rodrigues e *Trully* e à minha família em São Paulo por quase três semanas: Bárbara Bicalho, Larissa Batista e Ana Terra;

- e enfim ..., a você Catí, que quase enlouqueceu comigo e mesmo tendo vontade de jogar tudo para o alto, estava ao meu lado em todos os momentos.

Epígrafe

"Homens e mulheres desejam fazer um bom trabalho. Se lhes for dado o ambiente adequado, eles o farão."

William R. Hewlett (1913–2001),
(Co–fundador da Hewlett–Packard)

Resumo

Redes Bayesianas são estruturas que combinam distribuições de probabilidade e grafos. Apesar das redes Bayesianas terem surgido na década de 80 e as primeiras tentativas em solucionar os problemas gerados a partir da não resposta datarem das décadas de 30 e 40, a utilização de estruturas deste tipo especificamente para imputação é bem recente: em 2002 em institutos oficiais de estatística e em 2003 no contexto de mineração de dados. O intuito deste trabalho é o de fornecer alguns resultados da aplicação de redes Bayesianas discretas e mistas para imputação. Para isso é proposto um algoritmo que combina o conhecimento de especialistas e dados experimentais observados de pesquisas anteriores ou parte dos dados coletados. Ao empregar as redes Bayesianas neste contexto, parte–se da hipótese de que uma vez preservadas as variáveis em sua relação original, o método de imputação será eficiente em manter propriedades desejáveis. Neste sentido, foram avaliados três tipos de consistências já existentes na literatura: a consistência da base de dados, a consistência lógica e a consistência estatística, e propôs–se a consistência estrutural, que se define como sendo a capacidade de a rede manter sua estrutura na classe de equivalência da rede original quando construída a partir dos dados após a imputação. É utilizada pela primeira vez uma rede Bayesiana mista para o tratamento da não resposta em variáveis quantitativas. Calcula–se uma medida de consistência estatística para redes mistas usando como recurso a imputação múltipla para a avaliação de parâmetros da rede e de modelos de regressão. Como aplicação foram conduzidos experimentos com base nos dados de domicílios e pessoas do Censo Demográfico 2000 do município de Natal e nos dados de um estudo sobre homicídios em Campinas. Dos resultados afirma–se que as redes Bayesianas para imputação em atributos discretos são promissoras, principalmente se o interesse estiver em manter a consistência estatística e o número de classes da variável for pequeno. Já para outras características, como o coeficiente de contingência entre as variáveis, são afetadas pelo método à medida que se aumenta o percentual de não resposta. Nos atributos contínuos, a mediana apresenta–se mais sensível ao método.

Palavras–chave: Imputação. Redes Bayesianas. Imputação múltipla. Não resposta.

Abstract

Bayesian networks are structures that combine probability distributions with graphs. Although Bayesian networks initially appeared in the 1980s and the first attempts to solve the problems generated from the non–response date back to the 1930s and 1940s, the use of structures of this kind specifically for imputation is rather recent: in 2002 by official statistical institutes, and in 2003 in the context of data mining. The purpose of this work is to present some results on the application of discrete and mixed Bayesian networks for imputation. For that purpose, we present an algorithm combining knowledge obtained from experts with experimental data derived from previous research or part of the collected data. To apply Bayesian networks in this context, it is assumed that once the variables are preserved in their original relation, the imputation method will be effective in maintaining desirable properties. Pursuant to this, three types of consistence which already exist in literature are evaluated: the database consistence, the logical consistence and the statistical consistence. In addition, the structural consistence is proposed, which can be defined as the ability of a network to maintain its structure in the equivalence class of the original network when built from the data after imputation. For the first time a mixed Bayesian network is used for the treatment of the non–response in quantitative variables. The statistical consistence for mixed networks is being developed by using, as a resource, the multiple imputation for evaluating network parameters and regression models. For the purpose of application, some experiences were conducted using simple networks based on data for dwellings and people from the 2000 Demographic Census in the City of Natal and on data from a study on homicides in the City of Campinas. It can be stated from the results that the Bayesian networks for imputation in discrete attributes seem to be promising, particularly if the interest is to maintain the statistical consistence and if the number of classes of the variable is small. Features such as the contingency tables coefficient among variables, on the other hand, are affected by this method as the percentage of non–response increases. The median is more sensitive to this method in continuous attributes.

Keywords: Imputation. Bayesian networks. Multiple imputation. Missing data.

Sumário

1 Introdução

É difícil medir o tamanho do crescimento na demanda de dados e das considerações estatísticas acerca destes desde o surgimento dos computadores e seu aumento em capacidade e velocidade no processamento de informações. A transmissão de conhecimento, que nas décadas de 20, 30 e 40 estava restrita a poucos, hoje já está, em parte, massificada pela possibilidade de se estabelecer contatos diretos e pelo seu acesso facilitado.

Vive–se hoje na era da informação e do acesso interativo a repositórios de informações (SILVA-FILHO, 2001). E nesta realidade deparamo–nos com uma carga de informações cada vez maior. Pode–se acompanhar simultaneamente projeções econômicas e movimentações financeiras de diversos países e em diversas metodologias diferentes. Sabe–se de fenômenos meteorológicos, catástrofes e tragédias quase que no mesmo momento de suas ocorrências. Consegue–se prever a preferência de consumo de um ou mais segmentos de clientes a partir da avaliação de dados captados, por exemplo, por uma companhia telefônica e associados a dados bancários, de administradoras de cartões de crédito, de sistemas de saúde.

A convergência de diversos tipos de tecnologias possibilitou as mais variadas conexões entre unidades de processamento e acelerou o desenvolvimento dos sistemas de informação e das redes de comunicação (PORCARO, 2003).

Durante a captação de dados, quer seja automaticamente por algum sistema específico, quer pela entrada manual, existe a possibilidade de perda de informações ou inserção de erro nestes procedimentos.

Na preocupação de se manter a qualidade nos dados e análises, institutos de pesquisa, gerenciadores dos processos de captação de dados, pesquisadores e responsáveis pelas informações estudam metodologias e desenvolvem técnicas para identificar e minimizar problemas decorrentes destas perdas de dados ou inserção de erros.

Fazem parte deste cenário os processos de crítica e imputação que garantem aos usuários a qualidade da informação utilizada e permitem a aplicação de procedimentos de análise sem o seu comprometimento. Segundo Dias e Albieri (1992), métodos ou procedimentos para imputação são aqueles que se preocupam em substituir os valores ausentes, de uma unidade ou item, por estimativas dos mesmos. A aplicação de redes Bayesianas para imputação é bem recente: surgiu em 2002 no contexto de estatísticas oficiais e em 2003 em mineração de dados. Faz–se ainda necessário o aprimoramento da sua teoria e uma discussão dos resultados até então obtidos. Nas próximas duas seções contextualizam–se as duas áreas de aplicações em que surgiram as redes Bayesianas para imputação. É a partir dos resultados existentes nestas duas áreas que se baseia este trabalho.

1.1 Estatísticas oficiais

Bivar (2005) afirma que com o aumento na disponibilidade de informações, aumenta–se também a responsabilidade dos órgãos coordenadores dos sistemas estatísticos nacionais (informação verbal)[1]. A estes coube, ao longo dos últimos anos, além de manter a coleta e o processamento de dados de pesquisas tradicionais (como o Censo Demográfico, por exemplo), ampliar seus domínios metodológicos para atender a demandas do governo, de órgãos de classe, universidades, pesquisadores e da sociedade como um todo, na busca constante e crescente de dados nas mais variadas temáticas e para os mais diversos objetivos.

Além disso, os órgãos responsáveis pelas estatísticas oficiais passaram a trabalhar em conjunto com instituições internacionais diante da busca pela comparabilidade das estatísticas e do acompanhamento da mudança no relacionamento diplomático surgido com as áreas de livre comércio entre os países e a economia globalizada. As realidades dos países inserem–se agora em um contexto macro e isso demanda o conhecimento de muitos fatores antes não avaliados ou avaliados com pouca ênfase.

No Brasil, o órgão responsável por esta atividade é o Instituto Brasileiro de Geografia e Estatística (IBGE), fundado em 1936 e cuja missão institucional é: *Retratar o Brasil com informações necessárias ao conhecimento da sua realidade e ao exercício da cidadania*[2]. Dentre as suas principais funções estão a produção, análise, coordenação e consolidação das informações estatísticas oficiais do País.

[1]Informação fornecida por Wasmália Bivar no Seminário Internacional de Crítica e Imputação no Rio de Janeiro em 29 de novembro de 2005.

[2]Ver em: www.ibge.gov.br\home\disseminacao\eventos\missao.

Da responsabilidade de coletar, processar e disseminar os dados, uma etapa fundamental que compõe este sistema é a fase da crítica. Nesta etapa, previnem–se grandes inconsistências nas análises decorrentes de erros no fornecimento da informação, de captura do dado ou da fase de codificação.

Devido às políticas para a qualidade praticadas pelos diversos institutos de estatísticas oficiais pelo mundo, fica evidente a necessidade de se preocupar com as estimativas e com os usuários dos microdados e das bases de dados disponíveis a partir das pesquisas realizadas. Diante deste fato, alguns métodos de imputação de dados vêm sendo amplamente discutidos na comunidade estatística internacional para se obter as melhores metodologias para cada tipo de variável ou pesquisa.

1.2 Mineração de dados

A mineração de dados (ou *data mining*) é parte de um processo conhecido como "busca de conhecimentos em bancos de dados" (ou *Knowledge Discovery in Databases* (KDD)) que inclui diversos passos para identificar e utilizar informações contidas em grandes bases de dados. Existem diversas definições e aplicações variadas para mineração de dados, mas a mais abrangente e escolhida para este trabalho é a encontrada no *Clementine User Guide*[3] (2000, apud DINIZ; LOUZADA–NETO, 2000):

> Mineração de dados refere–se ao uso de uma variedade de técnicas para identificar informações úteis em bancos de dados e a extração dessas informações de tal maneira que elas possam ser usadas em áreas tais como teoria de decisão, estimação, predição e previsão. Os bancos de dados são geralmente volumosos, e na forma que se encontram nenhum uso direto pode ser feito deles; as informações escondidas nos dados é que são realmente úteis.

Apesar de ter–se iniciado a partir de ferramentas e por profissionais da área de tecnologia e engenharia de computação, a multidisciplinaridade em mineração de dados está presente por concatenar técnicas estatísticas de análises, procedimentos de inteligência artificial (aprendizado de máquina, reconhecimento de padrões, etc.), tratamento em grandes bancos de dados e otimização (DINIZ; LOUZADA–NETO, 2000). A Figura 1.1 a seguir mostra as etapas constituintes do KDD, das quais se forma a mineração de dados: seleção, pré processamento, transformação, extração de informações e assimilação dos resultados.

[3]Na internet em www.spss.com\clementine\.

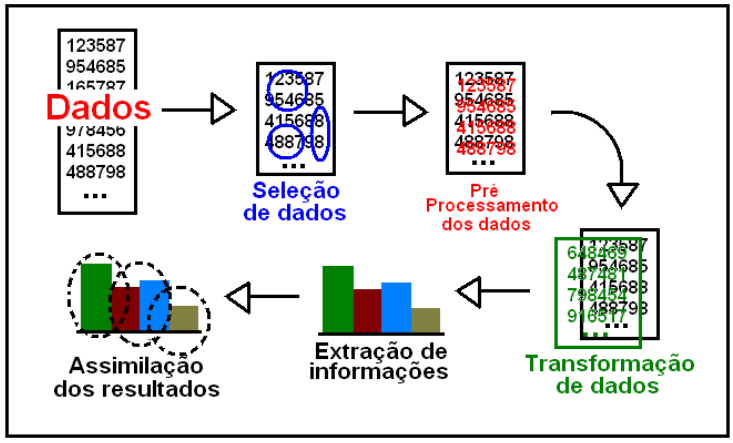


Figura 1.1: Etapas do processo KDD

Na fase de pré processamento é feita uma verificação da qualidade dos dados, revisando sua estrutura e obtendo algumas medidas que retratem a confiança que é possível depositar neles. Isso pode ser feito de diversas maneiras, combinando métodos estatísticos, técnicas de visualização de dados e conhecimento de especialistas. Dois problemas que naturalmente são resolvidos nesta etapa são a identificação de erros, valores atípicos (ou *outliers*) e valores faltantes. Mesmo sendo grande o volume de dados total, se a quantidade de casos a serem corrigidos abranger percentuais consideráveis, não é prudente excluí–los da análise.

Em determinadas condições, por exemplo quando os objetivos se concentram em detecção de fraude ou acompanhamento de lotes para controle de qualidade, mesmo se forem poucas as ocorrências de erros ou valores perdidos, pode acontecer a perda de informação (DINIZ; LOUZADA–NETO, 2000). Dessa maneira, a decisão de eliminar observações pode trazer consequências que nem sempre são fáceis de serem avaliadas.

Assim, também em mineração de dados existe a preocupação com erros e valores faltantes, sendo as técnicas de imputação e modelos de predição, sugestões para compor a etapa de pré processamento dos dados. Para maiores detalhes a respeito destas técnicas aplicadas ao KDD, ver por exemplo (HAIR et al., 1998) e (HRUSCHKA–JR.; HRUSCHKA; EBECKEN, 2005).

1.3 Motivação e justificativa

Identificada uma inconsistência, um item que não esteja de acordo com a distribuição da variável ou simplesmente um campo vazio na base de dados, como proceder? Responder a esta pergunta não é tão simples quanto parece. Ou pelo menos não sem antes

avaliar o impacto de qualquer procedimento sobre os dados e estatísticas derivadas. Uma das formas de se tratar os dados faltantes, ou a não resposta, utiliza–se da imputação, que consiste de técnicas para substituir valores ausentes de uma unidade por estimativas destes (DIAS; ALBIERI, 1992).

Existem diversos métodos de imputação na literatura e tantos outros a serem estudados, cada um deles com vantagens e limitações. A proposta deste trabalho é acrescentar, a estes vários procedimentos, o conhecimento sobre um com propriedades ainda não estudadas. As redes Bayesianas vêm aparecendo na literatura como uma opção de grandes vantagens e aplicabilidades. Embora estas apresentem teoria bem consolidada nas questões de aprendizado, classificação e inferência, para imputação os resultados encontram–se dispersos e faz–se necessária uma investigação de suas propriedades, bem como uma formalização e avaliação da teoria neste contexto.

O objetivo é avaliar o método de imputação a partir de redes Bayesianas para variáveis discretas, unificando conceitos e verificando características inerentes à sua aplicação. Trata–se mais precisamente do estudo do comportamento dos estimadores de quantidades básicas, observando a variabilidade associada ao método de imputação a partir de imputação múltipla (RUBIN, 1987). É verificada a necessidade de adaptação dos algoritmos propostos na literatura para a aplicação desta técnica em variáveis discretas. Uma contribuição neste contexto está na consolidação das consistências propostas por Di Zio et al. (2004) e a proposição da consistência estrutural, que é a propriedade de se manter a estrutura da rede Bayesiana ajustada com a base de dados após a imputação.

Uma segunda contribuição deste trabalho está na extensão do método para imputação em variáveis aleatórias discretas e contínuas a partir de redes Bayesianas mistas. Avaliam–se as mesmas consistências já existentes, bem como a consistência estrutural proposta para a rede discreta.

A terceira contribuição está na aplicação da imputação múltipla para obter a variabilidade associada ao método de imputação que se utiliza de redes Bayesianas mistas, para as modelagens mais comumente realizadas pelos usuários de bases de dados. Neste contexto estão os modelos lineares, modelos de riscos proporcionais e regressão logística. Esta avaliação é conduzida sob a suposição de que estes modelos descrevem de fato o comportamento real das variáveis de interesse.

Antes de descrever este trabalho, a seção seguinte padroniza a notação que será usada ao longo do texto, ressaltando que este é o produto de três artigos a serem submetidos para revistas técnicas que tratam de aplicações na área de estatística, imputação

de dados e estatísticas oficiais.

1.4 Padronizando a notação

Nesta seção, padroniza–se a notação que será utilizada ao longo do trabalho para evitar inconsistências.

Considere uma base de dados com n observações (ou unidades) e $X_1, X_2, \ldots, X_k$ variáveis aleatórias, que podem ser discretas ou contínuas. As variáveis aleatórias são denotadas por letras maiúsculas e os seus valores observados são representados por letras minúsculas. Dessa maneira, $X_{11} = x_{11}$ significa que o valor da variável aleatória X_1 para a primeira unidade da base é representado por x_{11}. $X_{ij} = x_{ij}$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$, é chamado de item. Na maioria dos bancos de dados, as variáveis referem–se a características que são captadas, coletadas ou observadas e as unidades ou observações referem–se a representantes diferentes de alguma população de interesse.

Associada a cada x_{ij} está uma variável indicadora $M_{ij} = m_{ij}$ da não observação deste dado, ou seja, $m_{ij} = 1$, se estiver ausente e $m_{ij} = 0$, se estiver presente.

A Figura 1.2 mostra o formato de uma base de dados completa, típica das aplicações que seguem neste texto. Outras representações existem, mas quaisquer delas podem ser convertidas para a representação em questão.

		Variáveis aleatórias			
	Controle	X_1	X_2	...	X_k
Observações ou unidades	**1**	x_{11}	x_{12}	...	x_{1k}
	2	x_{21}	x_{22}	...	x_{2k}
	3	x_{31}	x_{32}	...	x_{3k}
	⋮	⋮	⋮	⋱	⋮
	$\boldsymbol{n-2}$	$x_{n\text{-}2,1}$	$x_{n\text{-}2,2}$	...	$x_{n\text{-}2,k}$
	$\boldsymbol{n-1}$	$x_{n\text{-}1,1}$	$x_{n\text{-}1,2}$	...	$x_{n\text{-}1,k}$
	$\boldsymbol{n}$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,k}$

Figura 1.2: Representação de uma base de dados de trabalho

Um vetor $\boldsymbol{X_1}$ é aquele que contém $\{x_{11}, x_{21}, \ldots, x_{n1}\}$ ocorrências da variável aleatória X_1. A matriz $\mathbf{X}$ representa as ocorrências das variáveis aleatórias $X_1, X_2, \ldots, X_k$. Após ordenar as observações de cada uma das variáveis, identificam–se estas com $X_{(j)}$, $j = 1, 2, \ldots, k$, onde $X_{(j)}$ é a j–ésima variável aleatória após o vetor ordenado.

Se uma variável é imputada, então um *til* acima de sua identificação representará

que esta foi imputada, por exemplo, $\widetilde{X}_i$. O mesmo ocorre para os valores em cada variável; $\widetilde{x}_{ij}$, por exemplo, representa o valor imputado da j–ésima variável para a i–ésima unidade da base.

As quantidades de interesse são representadas por letras gregas μ, σ, ρ, etc. Se estimadas a partir da base completa, as estimativas passam a ser $\widehat{\mu}, \widehat{\sigma}, \widehat{\rho}$. Se estimadas a partir de uma base de dados imputada, estas quantidades são representadas por $\widetilde{\mu}, \widetilde{\sigma}, \widetilde{\rho}$. Por exemplo, os estimadores de um modelo de regressão após a imputação são identificados por $\widetilde{\beta}$.

Os grafos direcionados acíclicos, que são componentes das redes Bayesianas serão representados por letras caligráficas tais como $\mathcal{A}$, $\mathcal{B}$, $\mathcal{G}$, etc. Estes são compostos por um conjunto de vértices e de arestas onde, para cada vértice, associa–se uma variável aleatória da base de dados. Dessa maneira, o conjunto de vértices $V = \{X_1, X_2, \ldots, X_k\}$ está definido da mesma maneira como estão definidas as variáveis aleatórias. O conjunto de arestas A, ou os relacionamentos entre as variáveis, são determinados com arcos direcionados na rede e $X_i \rightarrow X_j$ implica que X_i exerce uma relação de causa sobre X_j, ou seja, é pai de X_j. Mais adiante, $X_i - X_j$ representará um relacionamento qualquer entre X_i e X_j, apenas a título de composição das classes de equivalência de Markov para uma dada rede.

Se uma rede Bayesiana é construída a partir de uma base de dados imputada, então os grafos $\widetilde{\mathcal{A}}$, $\widetilde{\mathcal{B}}$, $\widetilde{\mathcal{G}}$ representarão estas redes.

Outros operadores e notações específicas são descritos quando utilizados em suas aplicações ao longo do texto.

1.5 Descrição do texto

Este texto é composto por sete capítulos, três dos quais são resultados de textos voltados à publicação em revistas técnicas sobre o assunto. Dessa maneira, pode existir algum conceito que se encontre repetido no texto. O Capítulo 2 trata de imputação e imputação múltipla, desde o histórico no tratamento da não resposta e tipos de não resposta, passando pelos tipos de erros e mecanismos de não resposta, a revisão de alguns métodos de imputação existentes na literatura, os cinco critérios para se avaliar um método de imputação e uma seção que descreve resumidamente a técnica de imputação múltipla (RUBIN, 1987).

O Capítulo 3, que trata sobre redes Bayesianas discretas, foi submetido para publicação à Revista Brasileira de Estatística, cujo público está voltado para alunos e pesquisadores que fazem o uso mais aplicado da estatística. Neste capítulo, encontram–se a descrição das redes Bayesianas discretas e sua utilização como ferramenta para imputação, onde é proposto um novo algoritmo. Ainda se avalia a rede discreta para imputação com a proposta da consistência estrutural e se aplica este método às bases de dados do Censo Demográfico Brasileiro e a uma base de dados de homicídios ocorridos no município de Campinas.

No Capítulo 4 imputam–se dados quantitativos a partir de redes Bayesianas mistas, onde se estendem os conceitos apresentados no Capítulo 3 para a aplicação em dados discretos e contínuos. O texto deste capítulo será submetido à publicação no *The Imputation Bulletin* do *Statistics Canada*, o instituto oficial de estatísticas do Canadá, e também conta com aplicações aos dados do Censo Demográfico e aos dados de homicídios do município de Campinas. No Capítulo 5, avalia–se de forma preliminar a consistência estatística em redes Bayesianas a partir da imputação múltipla. A partir deste procedimento pode–se quantificar a variabilidade associada à imputação quando esta é conduzida com base em redes Bayesianas. Por se tratar de um estudo inicial sobre o assunto, os resultados deste capítulo, em combinação com um maior número de aplicações e maior detalhamento teórico, serão submetidos ao *Journal of Statistical Computations and Simulation*. No Capítulo 6 faz–se uma síntese dos resultados obtidos ao longo do texto e no Capítulo 7 destacam–se os pontos em aberto e futuros direcionamentos de pesquisa para a continuidade do trabalho neste tópico.

2 *Imputação e imputação múltipla*

É comum ver observações não classificadas ou dados faltantes em grandes bancos capturados automaticamente. Em bases de dados resultantes de pesquisas (por amostra ou censo), além dos dados faltantes originados da recusa do entrevistado, domicílio fechado, má fé de entrevistadores, etc., ainda podem ser encontradas inconsistências no registro ou transcrição das informações (como, por exemplo, um morador de 10 anos ou menos de idade que tenha filhos, ou que este seja responsável pelo domicílio).

As análises conduzidas com dados faltantes ou inconsistentes geram estimativas viciadas e indicam conclusões errôneas acerca dos dados. Nesse contexto, os mantenedores de grandes bancos de dados e gestores de informações estatísticas oficiais devem se preocupar com as análises divulgadas a partir destes dados e com as bases disponibilizadas para os usuários.

Este capítulo trata especificamente sobre imputação, uma das várias formas descritas na literatura para o tratamento de dados ausentes, desde o histórico, na Seção 2.1, passando pelos tipos de não resposta na Seção 2.2. Na Seção 2.3 tem–se a revisão de alguns métodos de imputação enquanto que a Seção 2.4 traz alguns critérios para se avaliá–los. A Seção 2.5 apresenta os principais aspectos de imputação múltipla, que serão usados posteriormente para avaliar a influência do método proposto para imputação em quantidades de interesse.

2.1 Histórico do tratamento da não resposta

Como citado no capítulo anterior, a não resposta, quer gerada pela impossibilidade de se coletar uma característica ou por identificação de um valor não coerente, deve ser

tratada de forma a minimizar erros que possam surgir a partir da manipulação e análise de tais dados. De acordo com Cochran (1977) a consequência mais óbvia e direta da não resposta é dispor de uma amostra menor do que a planejada para a análise.

Segundo Dias e Albieri (1992), métodos ou procedimentos para imputação são aqueles que se preocupam em substituir os valores ausentes, de uma unidade ou de um item, por estimativas dos mesmos. Além da imputação, outros dois métodos são descritos na literatura como formas de tratamento de dados ausentes. Dentre esses métodos estão os que se baseiam no uso das unidades que apresentam todas as informações. Estes são os mais comuns de serem encontrados em pacotes computacionais. Consistem em desconsiderar as unidades com algum valor faltante e conduzir análises apenas com as observações completas. Em alguns casos, estes métodos podem ser aplicados sem que a qualidade das estimativas seja muito afetada, especificamente quando exista um baixo percentual de não resposta e quando a ocorrência dos dados ausentes seja completamente aleatória (o que não ocorre, por exemplo, com variáveis do tipo rendimento, pois existe uma tendência à não resposta nas classes de rendas mais altas (PESSOA; SANTOS, 2004)(PESSOA; MOREIRA; SANTOS, 2004). Uma outra classe é a dos procedimentos de ponderação, que são utilizados para modificar os pesos do desenho amostral nas unidades respondentes, de forma a considerar a não resposta. Platek e Gray (1983) descrevem o método de ponderação para estimativas de razão, usando uma compensação para a não resposta.

A terceira classe compreende os métodos de imputação. A seguir, são apresentados alguns fatos históricos relevantes sobre imputação de acordo com a sua cronologia. O primeiro registro do uso de imputação deve–se a Allan e Wishart (1930) ao estimar resultados para um valor perdido em experimentos agropecuários. Ainda na mesma área, Yates (1933) mostrou previsões para várias observações faltantes a partir da minimização da soma de quadrados dos resíduos. Desde então, vários pesquisadores têm–se dedicado ao tema. Politz e Simmons (1949) foram os primeiros a utilizar probabilidades estimadas de resposta. Até o início dos anos 70 tinha–se ênfase apenas em mecanismos determinísticos de resposta e Cochran (1977) descreve algumas de suas limitações.

As primeiras experiências com imputação em Censos Demográficos foram dos Estados Unidos em 1960 e do Canadá em 1961 (BEAUMONT, 2005). Fellegi (1975) e Fellegi e Holt (1976) aparecem como marco estabelecendo princípios para edição e imputação e estes ainda são usados em vários institutos oficiais de estatística.

Kalton e Kasprzyk (1982) descrevem diversos métodos de imputação avaliados posteriormente por Albieri (1989) em relação aos estimadores de média, onde apresenta

seus vícios e variâncias. Silva (1989) implementou em seu trabalho rotinas computacionais para a metodologia de crítica e imputação de dados quantitativos propostos por Little e Smith (1987). Dias (1990) apresenta alternativas para se trabalhar com dados ausentes em análise fatorial com aplicação na construção de indicadores de saúde. Barroso (1995) estuda a imputação de dados em painéis para populações finitas onde obtém um previsor linear não viesado de erro quadrático médio mínimo para efeitos fixos e aleatórios sob o modelo linear misto e Chambers (2000) estabelece alguns critérios para avaliação de métodos de edição e imputação.

Atualmente, esforços para padronizar métodos de imputação têm sido despendidos por diversos organismos, a exemplo do projeto EUREDIT e seminários nacionais e internacionais que ocorrem com esta ênfase (a exemplo, o Seminário Internacional de Crítica e Imputação (SICI) realizado no Rio de Janeiro em 2005).

Antes de se estudar os efeitos da não resposta nos resultados de uma pesquisa, ou de se definir métodos aplicáveis para minimizar o efeito da não resposta nas estimativas, faz–se necessário distinguir os tipos de não resposta e suas possíveis causas. Isto porque existem diferenças conceituais entre a não resposta por impossibilidade de se atingir uma unidade selecionada e a não resposta por recusa no fornecimento da informação.

2.2 Tipos de erros e mecanismos de não resposta

Existem basicamente dois tipos de erros comuns em pesquisas: os erros amostrais e os erros não amostrais (SÄRNDAL; SWENSSON; WRETMAN, 1992). Os primeiros devem–se ao fato de que somente uma parte da população está sendo investigada e os erros não amostrais podem ser devido a diversas causas, como por exemplo, questão mal formulada no questionário, falhas de planejamento, falta de treinamento, respostas falsas por parte dos respondentes, erros de transcrição ou digitação, entre outros.

Esta seção descreve os tipos de erros que podem ocorrer em uma pesquisa e que, em geral, afetam as bases de dados disponíveis para análise. Além disso, nesta seção mostram–se também os mecanismos de não resposta existentes.

2.2.1 Erros amostrais e não amostrais

2.2.1.1 Erros amostrais

Suponha que exista o interesse em se obter alguma quantidade a partir de um conjunto de observações, por exemplo o total, a média, a mediana ou outra medida qualquer de uma variável. Denominemos esta variável de X, que pode representar o rendimento, a altura, a idade, etc. Sabemos que o interesse está em obter um resumo para uma população de N elementos, mas só dispomos de uma amostra de n ($n \leq N$) elementos deste conjunto. Dizemos que o erro amostral é aquele que se carateriza pela diferença entre o verdadeiro valor do parâmetro populacional de interesse e a estimativa deste calculada com base na amostra (COCHRAN, 1977). Em geral, os erros amostrais podem ser controlados no planejamento e desenvolvimento dos planos amostrais, que tentam combinar, da melhor maneira possível, as limitações da pesquisa (custo, tempo, operacional) com a perda máxima aceitável de qualidade das estimativas.

2.2.1.2 Erros não amostrais

Dentre os erros não amostrais, podem ser citados principalmente os erros de cobertura, a não resposta e os erros de medida, que são descritos a seguir.

- **Erros de cobertura**

  Os erros de cobertura ocorrem devido a problemas no cadastro (ou lista) para seleção em pesquisas feitas por amostragem e podem ser por subcobertura ou supercobertura (OLKIN, 1983). A subcobertura acontece se algumas unidades que deveriam estar no cadastro não se encontram no mesmo. Pode ainda ocorrer se algumas unidades foram incorretamente classificadas como inelegíveis[1] para a pesquisa, ou se as unidades forem omitidas ou desconsideradas da amostra pelo entrevistador. A supercobertura pode causar erro pelo oposto da subcobertura, ou seja, pode haver duplicação de unidades no cadastro. Como exemplo podem ser citadas as pesquisas em que os cadastros são obtidos a partir de listas telefônicas. Nesse caso, um mesmo domicílio pode conter mais de uma linha de telefone, ou uma mesma pessoa pode ter diversos telefones em seu nome. Nesta mesma pesquisa poderíamos ter um problema de subcobertura definido por aqueles que não possuem telefone.

[1] O conceito de elegibilidade depende muito da definição da população a ser pesquisada (COCHRAN, 1977).

No caso de erros de cobertura, se uma unidade é erroneamente classificada como elegível, a informação adicional obtida na entrevista pode ser usada para corrigir a pesquisa posteriormente. Já se uma unidade é erroneamente classificada como inelegível, então o erro introduzido na pesquisa não tem como ser corrigido. Muitas vezes a solução para este tipo de problema não é facilmente obtida (OLKIN, 1983). O que diversos institutos de pesquisa fazem é obter algum tipo de estimativa de sub ou supercobertura para o cadastro que utilizam e depois efetuar a correção das estimativas a partir de defasagens para uma melhor precisão.

- **Erros de não resposta**

A não resposta pode ser na unidade ou no item (DIAS; ALBIERI, 1992). Se uma unidade está elegível para a pesquisa e é selecionada para a amostra mas nenhuma resposta é obtida para a mesma (ou então é obtida uma resposta mas ela é descartada), tem–se a chamada não resposta na unidade. As principais razões para as não respostas na unidade em pesquisas domiciliares são: a inexistência de algum morador no domicílio selecionado, a dificuldade de comunicação entre o estrevistador e o entrevistado, a recusa total do selecionado a responder o questionário ou uma quebra na entrevista que leve a desconsiderar o que já foi respondido. Além das respostas que são classificadas como inconsistentes e descartadas da pesquisa.

A não resposta no item pode acontecer se o entrevistado não tem ou se nega a fornecer a informação para uma ou mais perguntas, dando resposta apenas para algumas questões da pesquisa (OLKIN, 1983). Nestes casos é importante conhecer o mecanismo gerador destas perdas para possibilitar um tratamento adequado às mesmas.

- **Erros de resposta**

Outra possibilidade é a identificação de erros em alguma parte das respostas nas fases de entrevista, transcrição, captura, codificação ou edição. Por exemplo, um entrevistador pode registrar um ano de nascimento de 1677 quando na verdade deveria ser 1977, gerando uma idade de mais de 300 anos. Este é o caso também quando o informante responde algo inconsistente ou fornece um dado errado. Se identificada, esta informação pode gerar a não resposta no item.

Quando todos os erros são detectados, o mais natural é que se efetuem as correções necessárias para que as estimativas não se tornem viciadas levando a conclusões errôneas a respeito do estudo que se está desenvolvendo. Na impossibilidade de se corrigí–los, é comum descartá–los, o que gera lacunas e conseqüentemente, erros de não resposta nos dados. Ainda pode ocorrer a recusa da resposta em alguma questão específica ou a impossibilidade de se obter os dados por algum motivo. Para se definir o melhor método de imputação a ser utilizado é importante conhecer o mecanismo de não resposta, pois a partir dele pode–se ter idéia a respeito do relacionamento entre a perda da informação e os valores das variáveis presentes na matriz de dados.

2.2.2 Mecanismo de não resposta

Existe uma forma mais genérica de se classificar a não resposta, que pode ser de acordo com a característica da informação perdida. Esta forma diz respeito à probabilidade da não reposta e, dependendo da sua classificação, pode–se melhor definir a estratégia de imputação de modo a não comprometer as análises (PLATEK; GRAY, 1983).

O mecanismo de não resposta pode ser inicialmente classificado em ignorável ou não ignorável[2]. O mecanismo ignorável indica que não é necessário especificar um modelo de não resposta. Nesse tipo de mecanismo, podem ser aplicados de forma satisfatória os métodos que se baseiam na imputação a partir de doadores. Já o mecanismo não ignorável limita o recurso de uso dos doadores por haver dependência entre os dados ausentes e a variável de interesse.

Considere um conjunto de variáveis, por exemplo, sócio–demográficas em uma pesquisa de rendimentos, definidas por $X_1, X_2, \ldots, X_k$. Estas características são coletadas para n pessoas, o que determina uma matriz de dados de dimensões $n \times k$, onde x_{ij} é o valor da variável X_j observado na unidade i, $i = 1, 2, \ldots, n$. Se for associada a cada x_{ij} uma variável indicadora $M_{ij} = m_{ij}$ da não observação deste dado (ou seja, $m_{ij} = 1$, se estiver ausente e $m_{ij} = 0$, se estiver presente), ter–se–á uma matriz $\mathbf{M}$ que define o padrão dos dados faltantes. Na Figura 2.1 (a) mostra–se a matriz $\mathbf{X}$ dos dados e sua correspondente matriz $\mathbf{M}$, indicadora de dados ausentes. A Figura 2.1 (b) mostra um exemplo hipotético.

Little e Rubin (2002) definem o mecanismo de não resposta usando as matrizes $\mathbf{X}$ e $\mathbf{M}$ em relações de dependência entre elas. Se a distribuição dos dados faltantes não

[2]Barroso (1995) salienta que é o processo de não resposta que pode ser ignorado e não o dado ou a unidade ausente.

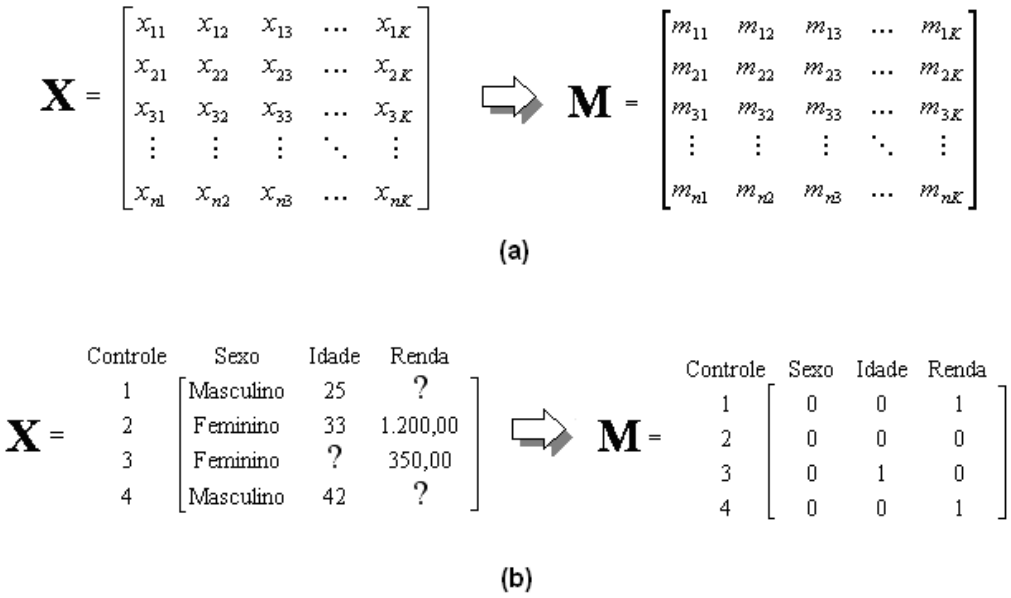


Figura 2.1: (a) Associação da matriz de dados à matriz de indicadores de não resposta (b) Exemplo hipotético de associação

depende dos valores dos dados observados ou perdidos tem–se o chamado mecanismo de perda completamente ao acaso – *Missing Completely at Random* (MCAR). Ou seja,

$$f(\mathbf{M}|\mathbf{X},\theta) = f(\mathbf{M}|\theta) \qquad \forall \mathbf{X},\theta, \tag{2.2.1}$$

onde θ é o conjunto de parâmetros. Nesse caso a não resposta é dita MCAR se os dados são perdidos por um processo totalmente aleatório (LITTLE, 1976)(LAIRD, 1988)(BARROSO, 1995). Por exemplo, se a variável de interesse for a idade na população masculina e existirem valores ausentes, essas perdas podem não depender de seus valores, ou seja, não existe alguma idade específica em que os homens tencionem não respondê–la comparando–a com outras. Então diz–se que este é um processo de perda completamente aleatório.

Ainda de acordo com Little e Rubin (2002) uma suposição menos restritiva que o mecanismo MCAR é aquela em que a perda depende somente dos dados observados em $\mathbf{X}$ (que chamaremos $\mathbf{X}_{obs}$) e não das informações faltantes. Este mecanismo é dito perda ao acaso – *Missing at Random* (MAR) e pode ser delimitado por:

$$f(\mathbf{M}|\mathbf{X},\theta) = f(\mathbf{M}|\mathbf{X}_{obs},\theta) \qquad \forall \mathbf{X}_{per},\theta, \tag{2.2.2}$$

onde $\mathbf{X}_{per}$ identifica os componentes perdidos da matriz $\mathbf{X}$. Aqui diz–se que os dados são perdidos por um processo aleatório, quando a probabilidade de não resposta depende dos dados presentes mas não dos ausentes (LITTLE; RUBIN, 1983). Os dados MAR e MCAR

fazem parte dos mecanismos ignoráveis de não resposta.

O mecanismo de não resposta chamado de perda não ao acaso – *Not Missing at Random* (NMAR) é aquele em que a distribuição de **M** depende dos valores faltantes da matriz **X**. Neste caso, a não resposta é não ignorável e o seu exemplo mais claro está na variável de rendimento, onde sabe–se que a perda de informações é maior em classes de maiores rendas. Pessoa e Santos (2004) e Pessoa, Moreira e Santos (2004) procedem a imputação de variáveis de renda no Censo Demográfico Brasileiro utilizando–se do recurso de transformar as não respostas do tipo NMAR em não respostas ignoráveis dentro de grupos homogêneos para busca de doadores.

Dependendo do tipo de mecanismo de não resposta é mais indicado um ou outro método de imputação e algumas técnicas têm eficiência condicionada no tipo de não resposta. A próxima seção descreve alguns métodos de imputação existentes na literatura.

2.3 Alguns métodos de imputação

Os métodos de imputação podem ser classificados em duas dimensões: a partir da modelagem dos dados e observação dos resíduos (ou métodos baseados no modelo) ou a partir do uso de doadores. Kalton e Kasprzyk (1982) e Albieri (1989) descrevem alguns métodos com estudos simulados sobre o efeito de cada um deles em estimadores de média e total populacional. Durrant (2005) faz uma revisão metodológica de alguns métodos para tratamento da não resposta no item em pesquisas da área social. A seguir, percorrem–se os principais métodos de imputação para tratamento da não resposta citados nestes textos.

- **Imputação dedutiva** – Este método de imputação depende da existência de alguma informação complementar nos dados, de tal forma que o dado perdido possa ser recuperado com base nas variáveis informadas. Baseia–se fundamentalmente em regras de imputação definidas por um raciocínio lógico, e a informação imputada em uma variável é exatamente obtida das demais disponíveis, ou seja, $\widetilde{x}_{ij} = f(x_{i1}, \ldots, x_{ij-1}, x_{ij+1}, \ldots, x_{ik}); i = 1, \ldots, n$. Bethlehem e Hofman (1998) destacam que esta é a forma ideal de imputação. Como exemplo, se a informação faltante for o sexo do respondente e na mesma unidade consta já haver cometido o aborto, percebe–se claramente que a variável sexo a ser imputada só pode ser *feminino*. Um outro exemplo é o número total de moradores de um domicílio se já estiverem contabilizados o número total de homens e de mulheres do mesmo.

- **Imputação pela média geral** – Este método substitui as informações faltantes pela média geral da variável em questão. Portanto, além de não fazer uso dos dados disponíveis em outras variáveis, este método pode alterar a distribuição da variável imputada e o relacionamento desta com as demais (DURRANT, 2005). Então

$$\widetilde{x}_{ij} = \frac{\sum_{l=1}^{n_{obs}} x_{lj}}{n_{obs}} = \beta_j, \qquad i = 1, \ldots, n_{per},$$

onde n_{obs} é o número de unidades respondentes na variável X_j e n_{per} é o número de informações perdidas, ou com informação faltante. É a forma determinística do método da regressão e usada apenas para variáveis contínuas. Como exemplo, pode–se citar como aplicação a variável tempo até o término de inquérito policial em avaliações de sistemas de justiça.

- **Imputação geral aleatória** – Este método designa aleatoriamente para cada não resposta, uma informação existente na base de dados dentro da mesma variável. Neste caso, são identificados possíveis doadores e posteriormente é testada a manutenção da distribuição da variável. Segundo Kalton e Kasprzyk (1982) este é um método que funciona bem quando a variável apresenta mecanismo de não resposta do tipo MCAR. Albieri (1989) cita que este método é a forma estocástica da função linear sem variáveis auxiliares, ou seja,

$$\widetilde{x}_{ij} = \beta_j + e_{ij},$$

onde $e_{ij} = x_{kj} - \beta_j$, o que reduz a $\widetilde{x}_{ij} = x_{kj}$, $i = 1, \ldots, n_{per}$ e $k \in X_{obs}$.

- **Imputação pela média dentro de classes** – Este método divide os dados em classes, de acordo com uma ou mais variáveis categorizadas auxiliares e imputa os itens faltantes de uma classe pela média das suas unidades respondentes, ou seja:

$$\widetilde{x}_{ij|h} = \frac{\sum_{l=1}^{n_{obs}} x_{lj|h}}{n_{obs}} = \beta_{j|h}, \qquad i = 1, \ldots, n_{per},$$

e h é a classe de imputação definida. É aplicado apenas a variáveis quantitativas e exerce menos efeito sobre a distribuição da variável a imputar se comparado com a imputação pela média geral (ALBIERI, 1989).

- **Imputação aleatória dentro de classes** – Aqui, seleciona–se também um doador para a informação faltante, mas dentro de classes de semelhança que previamente são definidas a partir de uma ou mais variáveis categorizadas auxiliares. Este método

é o equivalente estocástico do método de imputação pela média da classe, onde

$$\widetilde{x}_{ij|h} = x_{kj|h}, \qquad i = 1, \ldots, n_{per}; k \in X_{obs},$$

para alguma classe h de imputação. Funciona bem quando o mecanismo de não resposta dentro das classes é do tipo MCAR (DURRANT, 2005).

- **Imputação hot–deck** – Existem várias formas de se executar a imputação hot–deck. Este método baseia–se na especificação de um registro da base de dados para ser o "doador" da informação para um item faltante em uma dada variável X_j, $j = 1, \ldots, k$. Em geral, classes de imputação são construídas a partir de variáveis aleatórias categorizadas auxiliares e, para cada uma delas é feita uma seleção aleatória do doador. Kalton e Kasprzyk (1982) descrevem este método e citam que em cada classe de imputação é definido um ponto inicial que armazena um registro doador para ser imputado em X_j. Este valor vai sendo atualizado à medida que se encontram outros itens faltantes nessa mesma variável, e essa atualização pode ser feita de diversas maneiras a partir do estudo da distribuição conjunta. Bethlehem e Hofman (1998) citam que este é uma implementação especial da imputação aleatória dentro de grupos e que os registros devem ser processados sequencialmente: se o valor é observado em $X_{ij} = x_{ij}$ $(m_{ij} = 0)$, então x_{ij} é armazenado como doador; se $m_{ij} = 1$, então $X_{ij} = \widetilde{x}_{ij}$ receberá o valor do doador corrente. A Figura 2.2 apresenta um esquema gráfico do uso do método hot–deck.

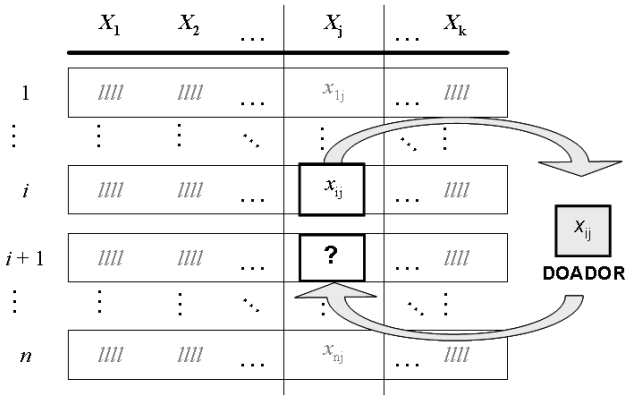


Figura 2.2: Esquema gráfico da realização do método hot–deck de imputação

Durrant(2005) enfatiza que uma vantagem deste método está em imputar valores que são realmente observados na pesquisa e que é adequado para tratar não resposta em variáveis categorizadas. Além disso, sob a imputação hot–deck a distribuição dos valores imputados terá a mesma forma da distribuição dos dados

observados (RUBIN, 1987).

- **Imputação por emparelhamento flexível** – Este método ordena respondentes e não respondentes dentro de classes de imputação, construídas a partir da combinação das categorias de um conjunto de variáveis discretas auxiliares. Define–se uma hierarquia entre estas variáveis de tal forma que, no menor nível faz–se um emparelhamento entre as unidades respondentes e não respondentes. O doador para uma ou mais observações será aquela unidade mais semelhante dentro de uma mesma classe de imputação. Se não existir um doador viável em algum nível, agregam–se as classes de imputação para um nível imediatamente acima iniciando–se novamente o emparelhamento. Um ponto a ser destacado é o fato de que os resultados dependem da ordem hierárquica entre as variáveis auxiliares estabelecida antes do início do processo (KALTON; KASPRZYK, 1982)(ALBIERI, 1989).

- **Imputação por regressão preditiva** – Aqui utilizam–se as informações disponíveis dos respondentes para ajustar uma regressão, onde a variável dependente é aquela a ser imputada e as independentes são as variáveis auxiliares que podem ser do tipo discreto ou contínuo. Dessa forma:

$$\widetilde{x}_{ij} = \beta_0 + \sum_{l \neq j} \beta_l x_{il}, \qquad i = 1, \ldots, n_{per}; j = 1, \ldots, k.$$

  Se a variável X_j a ser imputada é do tipo qualitativo então os modelos logístico ou log–linear são utilizados. Em geral, para variáveis quantitativas é ajustado o modelo linear normal e faz–se $e_{ij} = 0$ para se obter $\widetilde{x}_{ij}$ por este método. Por isso, neste método a imputação é determinística pois só são considerados os efeitos fixos dos parâmetros estimados a partir dos dados, não levando em conta o erro aleatório na estimativa do valor a ser imputado.

  Durrant (2005) cita que a maior desvantagem em métodos de imputação por regressão é distorcer a forma da distribuição da variável X_j a ser imputada e a correlação desta com as variáveis auxiliares que não são usadas na regressão.

- **Imputação por regressão aleatória** – O mesmo método anterior de forma que o erro aleatório da regressão é considerado na imputação, tornando não determinístico o valor a ser imputado, ou seja:

$$\widetilde{x}_{ij} = \beta_0 + \sum_{l \neq j} \beta_l x_{il} + e_{ij}, \qquad i = 1, \ldots, n_{per}; j = 1, \ldots, k.$$

Existem diversos estudos para determinar a melhor distribuição associada ao erro para os métodos de imputação (BETHLEHEM; HOFMAN, 1998).

- **Imputação pela função distância** – Este é um método que utiliza como doador a unidade que estiver mais próxima daquela que contém um ou mais itens faltantes. Por este motivo é também chamado de imputação por vizinho mais próximo, e envolve definir uma medida de distância adequada que será função de variáveis auxiliares. A maneira mais simples de se pensar neste método está em considerar uma variável contínua auxiliar, por exemplo, X_1, em que:

$$D_{hi} = |x_{i1} - x_{h1}|,$$

onde x_{i1} é o valor da variável auxiliar na unidade que contém o item faltante na base de dados. O doador para a variável faltante na unidade i será aquele onde

$$D_{hi} = min_h|x_{i1} - x_{h1}|.$$

Uma vantagem deste método está na possibilidade de se incorporar efeitos geográficos se variáveis deste tipo forem consideradas no cálculo das distâncias entre as unidades. Durrant (2005) observa que a variância de estimativas de interesse com o uso deste método pode ser inflacionada se uma mesma unidade for usada mais freqüentemente como doadora do que outras.

Em todos estes métodos, podem–se descrever pontos positivos e limitantes de suas aplicações. As referências citadas (e algumas que estas próprias referenciam) também trazem resultados comparativos entre eles, sob várias situações e aspectos. Além dos descritos aqui, ainda podem ser citados alguns métodos de imputação em estudo, como aqueles que se utilizam de redes neurais, *support vector machines* (SVM's), redes Bayesianas (DI ZIO et al., 2004)(HRUSCHKA–JR., 2003), etc. A próxima seção descreve alguns critérios para se avaliar um método de imputação.

2.4 Critérios para avaliar um método de imputação

Tendo como objetivos principais desenvolver novas técnicas para edição e imputação de dados, e avaliar estes métodos de tal maneira a se estabelecer as "melhores práticas" para diferentes aplicações, realizou–se entre os anos de 2000 e 2003 o projeto

EUREDIT[3], que reuniu pesquisadores de doze instituições, institutos oficiais de estatística e universidades da Inglaterra, Holanda, Finlândia, Suíça, Alemanha, Itália e Dinamarca, sob suporte financeiro do Programa IST da União Européia.

O projeto EUREDIT destacou–se por abranger diversos métodos, tradicionais e os mais modernos (ou computacionalmente intensivos), com o objetivo de compará–los e avaliá–los, estabelecendo–se os melhores a serem aplicados para os diferentes tipos de dados, além da disseminação de tais métodos via pacotes computacionais e publicações. Uma das principais motivações para o desenvolvimento deste projeto foi a existência de grande quantidade de estudos e experiências em crítica e imputação, mas dispersos e sem um formato comparável dos resultados.

As investigações dividiram–se em planos de trabalho que contemplaram redes neurais, *support vector machines* (SVM's), redes *multi–layer perceptron*, correção por métodos robustos e métodos baseados em doador ou modelo. Outros métodos surgiram no curso das atividades do EUREDIT, como é o caso das redes Bayesianas para imputação, cujos resultados preliminares não puderam ser comparados aos demais pelas características dos estudos conduzidos. Este fato gerou lacunas em que cabem estudos mais aprofundados tanto na parte teórica quanto nas simulações e aplicações.

Em seu relatório técnico[4], Chambers (2000) padroniza alguns critérios para avaliação dos métodos de edição e imputação que foram estudados no projeto EUREDIT, de forma a manter a comparabilidade entre eles. Cabe observar, segundo o próprio autor salienta, que estes critérios são apropriados para se avaliar o desempenho dos métodos de imputação quando os verdadeiros valores, além dos dados incorretos e/ou faltantes, são conhecidos. Ainda cabe observar que estes critérios não são necessariamente apropriados para obter resultados sob os métodos de imputação quando não se conhecem os verdadeiros valores dos parâmetros associados à base de dados ou o padrão da não resposta, o que ocorre na prática. Mas estes podem fornecer um bom guia sob a aplicação a ser considerada.

Assumindo que toda a perda é identificável[5], Chambers (2000) enumera cinco propriedades que se deseja satisfazer com o uso de um método de imputação[6]:

[3]Ver em www.cs.york.ac.uk\euredit\.

[4]Também disponível como relatório de número 28 do *National Statistics Methodology Series*.

[5]Como perda identificável considera–se aquela onde se sabe da sua ocorrência, como a não resposta ao item ou a recusa de uma unidade selecionada.

[6]Na prática, é quase impossível obter excelentes resultados para todas as propriedades, sendo sugerido buscar aquela (ou aquelas) que melhor se identifica com o objetivo a atingir.

1. Precisão preditiva – o método de imputação deve preservar ao máximo os valores observados na base de dados;

2. Exatidão da ordem – deve manter a ordenação das variáveis contida nos dados reais;

3. Precisão distribucional – deve preservar as distribuições (marginal e de maiores ordens) associadas aos dados reais;

4. Precisão da estimação – deve reproduzir os momentos de menor ordem da distribuição dos dados reais;

5. Plausibilidade da imputação – o método de imputação deve conduzir a valores imputados que sejam coerentes com os procedimentos de crítica e edição.

Deve–se observar que a propriedade (2) está voltada para variáveis aleatórias categorizadas do tipo ordinal e que a propriedade (4) pode ser avaliada sob um contexto mais amplo que apenas para os momentos de baixa ordem das distribuições, como por exemplo, parâmetros de modelos e percentis. Chambers (2000) divide as medidas de desempenho para as três classes de variáveis: nominal, ordinal e contínua. As seções a seguir apresentam os critérios técnicos para a avaliação de um método de imputação segundo Chambers (2000).

2.4.1 Precisão preditiva, distribucional e da estimação

Para avaliar estas propriedades é necessário construir uma matriz que compara as ocorrências entre os registros reais e os imputados nas c classes da variável X_j. As propriedades de precisão preditiva, distribucional e da estimação podem ser avaliadas a partir de uma medida D definida pela proporção de valores fora da diagonal principal dessa matriz, ou equivalentemente,

$$D = 1 - \frac{\sum_{i=1}^{n^*} I(X_{ij} = \widetilde{X}_{ij})}{n^*}, \tag{2.4.3}$$

onde n^* é o número de itens imputados na variável X_j e $I(X_{ij} = \widetilde{X}_{ij})$ é a função indicadora da igualdade entre o valor observado e o valor imputado. Se os valores reais são preservados após a imputação, então espera–se que o valor de D seja próximo de zero, sendo interessante testar se este valor é significantemente maior que uma constante.

Uma maneira de se verificar o quanto o valor de D está próximo de uma quantidade de interesse, considera o máximo de imputações incorretas (ε) em que se permite

flutuar em cada caso. Dessa forma, avaliar a precisão preditiva é o equivalente a verificar a inequação

$$D > \varepsilon + 2\sqrt{\widehat{Var}(D)},$$

onde

$$\widehat{Var}(D) = \frac{1}{n^*} - \frac{1}{(n^*)^2}\mathbf{1}^t\Big\{diag(\mathbf{R}+\mathbf{S}) - \mathbf{T} - diag(\mathbf{T})\Big\}\mathbf{1}. \tag{2.4.4}$$

Em (2.4.4), $\mathbf{1}$ é o vetor unitário de dimensão $c-1$, $\mathbf{R}$ é o vetor das $c-1$ contagens dos valores imputados nas classes da variável X_j, $\mathbf{S}$ é o vetor das contagens dos correspondentes valores reais observados nas $c-1$ classes da variável X_j, $\mathbf{T}$ é a matriz quadrada de ordem $c-1$ que corresponde à classificação cruzada nas n^* unidades imputadas *versus* as reais para as $c-1$ categorias consideradas, $diag(.)$ equivale à diagonal da matriz e as operações $(+)$ e $(-)$ representam a soma e a subtração direta dos elementos. Chambers (2000) sugere que ε seja definido como:

$$\varepsilon = max\left(0, D - 2\sqrt{\widehat{Var}(D)}\right). \tag{2.4.5}$$

Quanto menor este valor, melhor é o método de imputação em preservar os verdadeiros valores dos dados. A precisão distribucional pode ser verificada calculando uma estatística do tipo Wald para a distribuição marginal da variável imputada, dada por:

$$W = (\mathbf{R}-\mathbf{S})^t\Big[diag(\mathbf{R}+\mathbf{S}) - \mathbf{T} - \mathbf{T}^t\Big]^{-1}(\mathbf{R}-\mathbf{S}), \tag{2.4.6}$$

onde $\mathbf{R}$, $\mathbf{S}$ e $\mathbf{T}$ são as mesmas matrizes definidas anteriormente. Este resultado é uma extensão da estatística de McNemar (STUART, 1955) sem a correção de continuidade para uma tabela de contingência.

Em variáveis do tipo escalar ou contínuas, pode–se utilizar do recurso da categorização para se avaliar a imputação, mas deve–se considerar que este processo pode gerar arbitrariedade na construção das classes de avaliação. No sentido de evitar essa arbitrariedade, definiram–se algumas medidas que podem indicar o desempenho do método de imputação sem recorrer–se às transformações. No que diz respeito à precisão preditiva, uma quantidade que fornece uma indicação da proximidade entre X_j e $\widetilde{X}_j$ é o coeficiente de correlação calculado para os n^* valores imputados:

$$\rho(X_j, \widetilde{X}_j) = \frac{\sum_{i=1}^{n^*}\Big[X_{ij} - E(X_j)\Big]\Big[\widetilde{X}_{ij} - E(\widetilde{X}_j)\Big]}{\sqrt{\sum_{i=1}^{n^*}\Big[X_{ij} - E(X_j)\Big]^2\sum_{i=1}^{n^*}\Big[\widetilde{X}_{ij} - E(\widetilde{X}_j)\Big]^2}}, \tag{2.4.7}$$

onde, quanto mais próximo $\rho(X_j, \widetilde{X}_j)$ estiver da unidade, mais próximos dos valores reais estão os valores imputados.

Chambers (2000) recomenda focar a avaliação em estimativas da regressão de X_j em $\widetilde{X}_j$, onde se conduz o teste com respeito ao parâmetro β do modelo ($\beta = 1$) em $X_j = \beta\widetilde{X}_j + e_j$, auxiliado pelo erro quadrático médio da regressão,

$$\widehat{\sigma}^2 = \frac{1}{n^* - 1}\sum_{i=1}^{n^*}\left(X_{ij} - \widehat{\beta}\widetilde{X}_{ij}\right)^2.$$

Quanto menor o valor de $\widehat{\sigma}^2$ associado à não significância do teste em $\beta = 1$, melhor o método de imputação para X_j no sentido da precisão preditiva.

Uma outra classe de avaliadores da imputação em variáveis quantitativas está na observação de uma medida de distância entre o valor real e o valor imputado. No mesmo documento, Chambers (2000) cita diversas funções para o caso ponderado (aplicado a dados amostrais), sendo escolhida para este trabalho a distância:

$$d(X_j, \widetilde{X}_j) = \sqrt{\frac{\sum_{i=1}^{n^*}(X_{ij} - \widetilde{X}_{ij})^2}{n^*}},$$

ressaltando que esta foi uma escolha arbitrária, encorajando–se maior detalhamento posterior em outras funções.

Em variáveis ordinais a precisão preditiva é avaliada no item a seguir.

2.4.2 Exatidão da ordem

No caso da variável categorizada ordinal, o método de imputação deve ser capaz, se ocorrer um erro de imputação, de manter–se não muito afastado da classe original. Isso porque se há uma ordenação lógica entre as classes, por exemplo $X_j = 1, 2, 3$, como é o caso de variáveis do tipo faixa etária ou faixa de renda, uma imputação na classe $X_j = 3$ é um erro mais *grave* que uma imputação na classe $X_j = 2$, se a classe real for $X_j = 1$.

Para tratar deste caso, Chambers (2000) sugere que sejam contabilizados os erros de imputação considerando uma ponderação pela distância entre a classe real e a classe imputada. Na verdade, o cálculo para avaliar um método de imputação em uma variável X_j passaria a ser:

$$D_j = \frac{1}{n^*}\sum_{i=1}^{n^*} d\left(X_{ij} - \widetilde{X}_{ij}\right),$$

onde $d(c_1, c_2)$ é a função distância entre as categorias c_1 e c_2 da variável aleatória X_j

ordinal.

Essa função distância pode ser definida de várias maneiras dando pesos diferentes de acordo com a importância da variável ou das classes que ela contém. Por exemplo, se $X_{ij} = c$ e $\widetilde{X}_{ij} = c + k$, então uma possibilidade pode ser $d_1(X_{ij}, \widetilde{X}_{ij}) = k$, onde c representa uma classe qualquer da variável X_j.

Uma outra distância pode ser definida da forma $d_2(X_{ij}, \widetilde{X}_{ij}) = k^2$, que eleva a contribuição dos maiores erros no cálculo de D_j. O projeto EUREDIT conduziu as avaliações dos métodos de imputação estudados utilizando como distância entre as classes c_1 e c_2 a medida (CHAMBERS, 2000):

$$d(c_1, c_2) = \frac{1}{2}\left[\frac{|c_1 - c_2|}{max(c) - min(c)} + I(c_1 \neq c_2)\right],$$

onde $I(c_1 \neq c_2)$ é a função indicadora da ocorrência de erro de imputação no dado imputado, $min(c)$ e $max(c)$ correspondem ao valor mínimo e máximo respectivamente de uma classe qualquer da variável X_j.

2.4.3 Plausibilidade da imputação

Segundo Chambers (2000), a avaliação da plausibilidade da imputação está relacionada com a não inserção de *outliers* ou erros (no caso de uma variável contínua) ou então de uma classe que retorne falha a alguma regra de edição (no caso de variáveis discretas). Nesse cenário, cada variável bem como sua interação com as demais, apresenta seus critérios para o levantamento dessa medida de desempenho. Por exemplo, a variável faixa etária, que pode seguir a regra de edição condicionada às variáveis estado civil e anos de estudo, pode ter sua coerência avaliada de forma semelhante à propriedade de exatidão da ordem citada na seção anterior.

Já uma variável contínua depende da definição de seus limites mínimo e máximo aceitáveis, além da necessidade de verificação em subgrupos definidos pelas classes das variáveis categorizadas que possam influenciá–las. Um exemplo estaria na variável renda a partir da variável anos de estudo. Uma medida da avaliação nesse caso poderia ser um indicador de imputações incoerentes dado um determinado conjunto de regras pré–estabelecidas para cada variável, que dependeria de cada aplicação e de cada método de imputação. Mais adiante, vê–se que Di Zio et al. (2004) estabelecem um critério para imputação a partir de redes Bayesianas onde a própria estrutura permite criar facilmente indicadores de desempenho do método.

2.5 Imputação múltipla

Ao se imputar um único valor para cada não resposta no item, corre–se o risco de se tratar posteriormente dos dados como se estes fossem os realmente observados. Dessa maneira, muitas das análises conduzidas com estes dados simplesmente desconsiderariam a incerteza que está associada à não resposta. Isso é o que ocorre com os métodos de imputação descritos na Seção 2.3 neste capítulo, além de outros existentes na literatura. Eles ignoram o fato de que os valores perdidos são desconhecidos e as bases de dados imputadas apresentam apenas uma imputação para cada não resposta. Para tratar deste problema, foi proposta uma técnica por Rubin (1987) chamada de imputação múltipla, que inicialmente foi aplicada ao contexto de pesquisas amostrais. Dois motivadores para a proposição desta técnica estão em (Rubin, 1996):

1. os usuários finais e os responsáveis pela captação das informações, gestores e mantenedores de bases de dados são, em geral, entidades distintas que dispõem de diferentes capacidades de processamento e ferramentas para análise;

2. não existe uma justificativa plausível para se manter a não resposta, como é o caso das situações em que realmente não existe informação para uma dada questão[7].

Pode–se acrescentar a estes motivadores, um terceiro que estaria em mensurar a variabilidade, associada ao método de imputação, que afeta estimativas de quantidades de interesse calculadas a partir de dados imputados. Repetir um método de imputação algum número de vezes permite que seja quantificado o efeito do valor imputado sob a variância destas quantidades.

Segundo Herzog e Rubin (1983), a imputação múltipla foi planejada para gerar inferências adequadas sob um modelo para a não resposta. Para isso, utiliza–se somente de técnicas–padrão de análise estatística para dados completos. As imputações múltiplas combinam as várias análises para exibir a sensibilidade da inferência com relação aos dados faltantes.

A idéia da imputação múltipla está em, a partir de um determinado número de imputações em uma base de dados com itens faltantes, digamos m imputações, tratar estas m bases como se fossem completas (analisando–as com técnicas para dados completos) para se incorporar a variabilidade associada ao método de imputação às estimativas de

[7]Existem pesquisas em que a não resposta do entrevistado passa a ser uma característica de interesse ao pesquisador. Isso ocorre principalmente em estudos psicológicos (ver, por exemplo, (SILVA; SALOMÃO, 2002)).

interesse. A aplicação desta técnica requer três passos: a imputação propriamente dita, a análise e a combinação dos resultados. A Figura 2.3[8] a seguir ilustra estes passos.

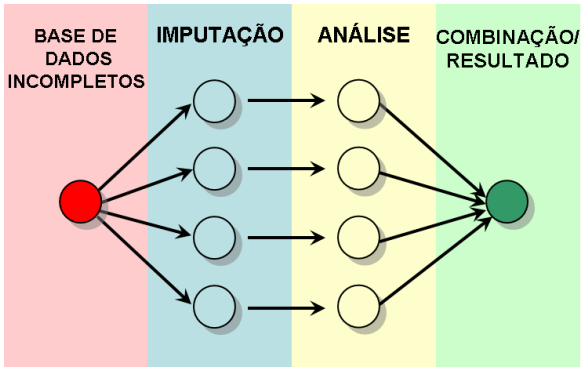


Figura 2.3: Esquema gráfico da realização de imputação múltipla para combinar resultados

Na fase de imputação, cada item faltante deve ser imputado m vezes segundo a mesma regra definida pelo método de imputação escolhido para ser aplicado. A fase de análise é conduzida por um procedimento padrão para dados completos onde se obtêm os parâmetros de interesse (de modelos de regressão linear, de regressão logística, análise fatorial, modelos de riscos proporcionais, estimação de componentes de variância, modelos de séries temporais, etc.) em cada uma das m bases de dados completadas. A terceira e última etapa da imputação múltipla é a combinação dos resultados que deve seguir algumas regras. A primeira delas está em combinar as m estimativas dos parâmetros. Seja θ uma quantidade a ser estimada da população. A estimativa de θ via imputação múltipla é dada por:

$$\widehat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\widetilde{\theta}_i, \tag{2.5.8}$$

onde $\widetilde{\theta}_i$ refere–se à estimativa do parâmetro θ na i–ésima base de dados imputada.

A segunda regra combina as estimativas da variância associadas a $\widehat{\theta}$. Considere que o interesse esteja em obter $Var(\widehat{\theta})$. Para isso, particiona–se a variância em dois componentes:

– a variabilidade dentro de cada base de dados imputada, chamada de $Var_d(\widehat{\theta})$;

– a variabilidade entre as bases de dados imputadas, chamada de $Var_e(\widehat{\theta})$.

[8]Figura reproduzida de www.multiple–imputation.com.

A variância dentro de cada base é dada pela média das variâncias estimadas,

$$Var_d(\widehat{\theta}) = \frac{1}{m}\sum_{i=1}^{m} Var(\widetilde{\theta}_i), \tag{2.5.9}$$

e a variância entre as bases de dados, ou variância entre as imputações, é a variância associada às m estimativas, ou seja, devido a imputação,

$$Var_e(\widehat{\theta}) = \frac{1}{m-1}\sum_{i=1}^{m} (\widetilde{\theta}_i - \widehat{\theta})^2. \tag{2.5.10}$$

A variância de $\widehat{\theta}$ será estimada pela soma dos dois componentes, com um adicional fator de correção para quantificar o erro para a simulação em $\widehat{\theta}$:

$$\widehat{Var}(\widehat{\theta}) = Var_t(\widehat{\theta}) = Var_d(\widehat{\theta}) + \Big(1 + \frac{1}{m}\Big) Var_e(\widehat{\theta}). \tag{2.5.11}$$

As inferências com relação a θ são baseadas na aproximação (SCHAFER, 1997a):

$$\frac{(\theta - \widehat{\theta})}{Var_t(\widehat{\theta})} \sim t_\zeta,$$

onde ζ é o número de graus de liberdade associado à distribuição de probabilidade t de Student, dado por:

$$\zeta = (m-1)\left[1 + \frac{Var_d(\widehat{\theta})}{(1+m^{-1})Var_e(\widehat{\theta})}\right]^2.$$

Segundo Schafer e Olsen (1998), a razão entre $Var_e(\widehat{\theta})$ e $Var_d(\widehat{\theta})$ é um reflexo de quanta informação está contida na parte faltante com relação à parte observada na análise de dados conduzida.

De acordo com Rubin (1987), quando λ é a fração de dados faltantes, o ganho em se obter estimativas baseado em m imputações múltiplas é de aproximadamente $(1+\frac{\lambda}{m})^{-\frac{1}{2}}$, então, por exemplo, para $m = 3$ e $\lambda = 0,10$, isso significa que três imputações terão um erro padrão somente 1,017 vezes o obtido com $m = \infty$. Quando $m \to \infty$ em (2.5.11) significa que a variância total tende a $Var_t(\widehat{\theta}) = Var_d(\widehat{\theta}) + Var_e(\widehat{\theta})$. Daí a eficiência com valores baixos de m. A Tabela 1 ilustra alguns cálculos desta relação.

Como qualquer método estatístico, a imputação múltipla está baseada em suposições que, se forem violadas, comprometem a análise (SCHAFER; OLSEN, 1998):

1. **o modelo para a população dos dados** – devem ser respeitados os modelos sobre sua origem e estes devem ser considerados na fase de imputação para garan-

Tabela 1: Razão entre o valor esperado do erro padrão quando calculado com $m < \infty$ e com $m \to \infty$

	λ					
m	**10%**	**20%**	**30%**	**50%**	**70%**	**90%**
3	1,017	1,033	1,049	1,080	1,111	1,140
5	1,010	1,020	1,030	1,049	1,068	1,086
10	1,005	1,010	1,015	1,025	1,034	1,044
20	1,002	1,005	1,007	1,012	1,017	1,022
30	1,002	1,003	1,005	1,008	1,012	1,015
50	1,001	1,002	1,003	1,005	1,007	1,009

tir a análise das imputações múltiplas posteriormente. Se a distribuição de X_j a ser imputada é normal, então essa variável deve ser imputada segundo sua característica, mesmo que para isso seja necessário usar–se de alguma transformação. Se a imputação é feita em X_j com o auxílio de uma variável aleatória X_k, onde uma posterior análise a ser conduzida considera X_k e X_l, então as conclusões obtidas para X_l podem ser viciadas.

2. **a distribuição a priori para os parâmetros do modelo** – esta característica envolve o teorema de Bayes (COX; HINKLEY, 1974) no que diz respeito às suposições que são feitas antes da análise. A priori deve ser não–informativa, ou seja, correspondente a um estado de desconhecimento sobre os parâmetros do modelo. De acordo com Rubin (1996) em algumas situações não usuais, como dados esparsos, amostras pequenas e grandes percentuais de não resposta, pode ser necessária a aplicação de uma distribuição informativa a priori.

3. **o mecanismo da não resposta** – supõe–se que o mecanismo de não resposta é do tipo ignorável, no sentido preciso definido por Rubin (1987) e registrado anteriormente neste capítulo. Sugere–se na literatura que o tipo MAR é o mecanismo onde se observam os melhores resultados, embora também seja aplicado ao mecanismo MCAR. No que diz respeito aos mecanismos do tipo não ignoráveis, não existem ainda métodos específicos para o seu tratamento e o uso de imputação múltipla pode trazer bons resultados se comparados com a exclusão das unidades que contêm itens faltantes ou imputados pela média (DURRANT, 2005). Essa suposição é considerada a mais forte de todas.

Durrant (2005) salienta que uma das vantagens da imputação múltipla está na possibilidade de se produzir arquivos de microdados que possam ser usados para uma variedade de análises. Este fato é especialmente útil quando o objetivo da imputação

múltipla é atender aos usuários finais de bases de dados de uso público, que desenvolvem diferentes tipos de análises. Além disso, aumenta–se a eficiência da estimação combinando os paradigmas Bayesiano (para criar as imputações) e frequentista (para avaliar os resultados) de maneiras complementares (RUBIN, 1996).

Como pontos limitantes podem ser citados alguns que passam da implementação à capacidade de processamento e armazenamento das m bases de dados imputadas. Rubin (1987) cita que em alguns casos as técnicas de imputação múltipla podem ser difíceis de implementar. Schafer e Olsen (1998) descrevem dispositivos para criar imputações múltiplas baseados em *data augmentation* (TANNER; WONG, 1987) e no algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977). Em contrapartida já existem pacotes computacionais que consideram análises a partir de múltiplas imputações em bases de dados, como é o caso do *mitools* (LUMLEY, 2004) e do *mix* (SCHAFER, 2003) para o software R[9], os procedimentos *PROC MI* e *PROC MIANALYZE* em SAS (YUAN, 2000) e outros como o *MICE* (BUUREN; OUDSHOORN, 2000), o *CAT*, *PAN* e *NORM* (SCHAFER, 1997a) (SCHAFER, 1997b). Com respeito às capacidades de armazenamento e processamento, devido a um número não muito grande de imputações necessárias, pode–se adequar o valor de m ao valor máximo do erro padrão que se aceita na análise.

Rubin (1987) também levanta o questionamento a respeito da aplicação da imputação múltipla em problemas que não envolvam pesquisas amostrais. O autor cita que, embora a principal motivação tenha surgido no contexto amostral, e é onde se encontram os maiores benefícios da técnica, nada impede que suas definições sejam aplicadas a contextos não amostrais e com bons resultados. Neste texto a imputação múltipla será utilizada para quantificar a incerteza associada ao método de imputação considerado nos capítulos seguintes.

[9] O *R* é um *software* livre para aplicações estatísticas e gráficas. Pode ser obtido em www.r–project.org.

3 *Avaliação do uso de redes Bayesianas discretas para imputação*

Redes Bayesianas são estruturas que combinam distribuições de probabilidade e grafos. Apesar das redes Bayesianas terem surgido na década de 80 e as primeiras tentativas em solucionar os problemas gerados a partir da não resposta em bases de dados datarem das décadas de 30 e 40, a utilização de estruturas deste tipo especificamente para imputação é bem recente: surgiram em 2002 em institutos oficiais de estatística e em 2003 no contexto de mineração de dados. Como existem questões teóricas e propriedades ainda não abordadas na literatura, além de poucas referências ao assunto, pretende–se avançar nas propriedades e tecer maiores considerações a respeito da aplicação de redes Bayesianas como um método de imputação. Apresenta–se neste capítulo um novo algoritmo para a imputação de dados baseado em redes Bayesianas discretas construídas a partir do conhecimento de especialistas[1]. Utilizam–se as avaliações sugeridas por Di Zio et al. (2004) e propõe–se um novo tipo de consistência: a consistência estrutural, que se relaciona à manutenção da estrutura da rede Bayesiana em sua classe de equivalência após a imputação. A importância desta medida de consistência está na avaliação da preservação dos relacionamentos de independência condicional entre as variáveis. Simulações e aplicações são feitas com o uso dos dados do Censo Demográfico Brasileiro e de dados de homicídios ocorridos no município de Campinas, em estruturas variadas de ajuste de redes. Para manter a coerência com as aplicações em Di Zio et al. (2004) e pela capacidade computacional nas simulações, as redes empregadas nos dados do Censo Demográfico são pequenas, limitadas a cinco variáveis.

[1]Parte deste capítulo corresponde a um artigo submetido à Revista Brasileira de Estatística. Dessa maneira, pode haver alguma parte do texto que se encontre repetida ao longo do capítulo.

3.1 Introdução

Na preocupação por se manter a qualidade nos dados e análises, institutos de pesquisa, gerenciadores dos processos de captação de dados, pesquisadores e responsáveis por bases de dados e informações estatísticas estudam metodologias e desenvolvem técnicas para identificar e minimizar problemas decorrentes das perdas de dados ou inserção de erros. Fazem parte deste cenário os processos de crítica e imputação, que permitem manter a qualidade dos produtos disponibilizados aos usuários e garantem a aplicação de procedimentos padrão de análise sem o comprometimento dos resultados devido a não resposta.

Imputação de dados já é uma forma bem conhecida de tratamento da não resposta que tem por objetivo "completar" os espaços vazios, de maneira que se possa utilizar de ferramentas para dados completos na busca por estatísticas ou características de uma dada população de interesse.

Em geral, as avaliações sobre os métodos de imputação contemplam apenas uma variável, referenciando–se a apenas um parâmetro de interesse. Textos recentes publicados pelo projeto EUREDIT[2] relatam a experiência de que o melhor método para imputação varia de acordo com a aplicação e, dependendo do método, existem bons resultados garantidos para o caso univariado.

Para realizar imputação no contexto multivariado, em 2002 surgem as primeiras aplicações de redes Bayesianas em estatísticas oficiais (THIBAUDEAU; WINKLER, 2002) (DI ZIO et al., 2004) e em 2003, em mineração de dados (HRUSCHKA–JR., 2003). Observa–se a necessidade de aprimoramento da teoria e discussão dos resultados até então obtidos.

O objetivo deste trabalho é, a partir de um novo algoritmo para imputação usando redes Bayesianas, avaliar este método utilizando–se das medidas de consistência propostas por Di Zio et al. (2004). Além disso, é proposta uma nova possibilidade de consistência, a consistência estrutural, que avalia a propriedade de manutenção da estrutura da rede após a imputação. São apresentadas simulações e aplicações com dados do Censo Demográfico Brasileiro do ano de 2000 e dados de homicídios em Campinas.

Este texto divide–se como segue. Na Seção 3.2 tem–se uma rápida revisão da teoria em redes Bayesianas discretas e na Seção 3.3 são apresentadas as redes Bayesianas

[2]Um projeto realizado para o desenvolvimento e avaliação de novos métodos para edição e imputação. Ver em www.cs.york.ac.uk\euredit\.

como ferramenta para imputação e se introduz um novo algoritmo. Na Seção 3.4 são apresentados quatro tipos de medidas de consistência para a avaliação das redes como método de imputação. Na Seção 3.5 têm–se alguns resultados obtidos com simulações nos dados do Censo Demográfico Brasileiro de 2000. A Seção 3.6 discorre sobre os dados de homicídios ocorridos no município de Campinas e a Seção 3.7 apresenta um relato das conclusões e futuros direcionamentos para pesquisas relacionadas.

3.2 Redes Bayesianas discretas

As redes Bayesianas surgiram na década de 80 e em pouco mais de duas décadas de existência, tornaram–se populares e difundiram–se em aplicações nas mais diferentes áreas, como por exemplo: aplicações na área médica (HECKERMAN, 1988) (RASMUSSEN, 1995) (SAHEKI, 2005), mineração de dados (HRUSCHKA–JR; HRUSCHKA; EBECKEN, 2005), reconhecimento de padrões (FREY, 1998) (SMYTH, 1997), em finanças (BINNER; KENDALL; CHEN, 2005).

Este seão trata de redes Bayesianas discretas e suas principais características de ajuste, como algoritmos, estruturas e sub–estruturas, etc. que venham a compor o cenário para a sua utilização em imputação. Apesar de haver uma enormidade de trabalhos disponíveis sobre o assunto, busca–se aqui relatar de forma resumida os conceitos e definições que serão aplicados ao longo deste texto. Referências mais aprofundadas serão citadas oportunamente.

Das definições para redes Bayesianas existentes na literatura, a que usamos neste texto considera dois elementos como seus componentes: o grafo e a distribuição de probabilidade associada a um conjunto de variáveis de interesse. A grande vantagem no uso deste tipo de estrutura está em conseguir representar incerteza de forma graficamente compacta (PEARL, 1988) (CHARNIAK, 1991) através de grafos.

Como componente fundamental da rede Bayesiana, o grafo será de destaque quando se tratar mais adiante da consistência estrutural, portanto o item a seguir aponta os grafos e seus elementos importantes.

3.2.1 Grafos e redes Bayesianas

Existem muitas aplicações de grafos na literatura e inicia–se esta seção com um exemplo dado na Figura 3.1. Esta figura representa a seguinte situação: uma dona de casa

resolve mapear os locais por onde ela deve passar para realizar as suas tarefas em um determinado período do dia (ver Quadro 1).

Tarefa	**Local**
1. limpar a casa	casa
2. levar o cachorro para tomar vacina	loja de animais
3. pagar as contas	banco
4. fazer compras	mercado
5. buscar as crianças na escola	escola
6. comprar o remédio do marido	farmácia

Quadro 1: Tarefas e seus respectivos locais de realização para a situação hipotética dos possíveis caminhos de uma dona de casa

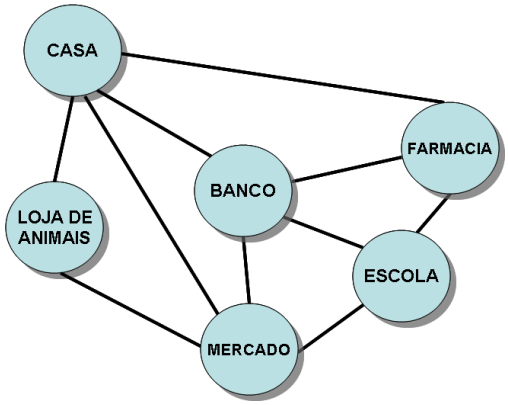


Figura 3.1: Exemplo de grafo para situação hipotética dos possíveis caminhos de uma dona de casa

Poder–se–ia pensar em minimizar o percurso em que a senhora vai desempenhar suas atividades ou então otimizar o tempo da realização das suas tarefas. Assim como estes objetivos, diversos outros podem ser modelados com o uso de um grafo em variadas áreas, como em informática, com a transmissão de informações; outro exemplo está na área de logística, com a modelagem da localização do armazenamento de mercadorias; em segurança pública, no fluxo de informações entre os sistemas de justiça; na arquitetura, com a definição de instalações elétricas domiciliares, etc. Principalmente para profissionais da matemática e ciência da computação, a teoria dos grafos apresenta vários tipos de aplicações.

Definição 1 – O grafo $\mathcal{G}$ é um composto de um conjunto de vértices ou nós (V) conectados por um conjunto de arcos (A), que representam as ligações entre os nós.

Para construir uma rede Bayesiana, necessitamos de um grafo orientado (ou di-

grafo) que para nós estabelecerá uma relação de dependência direta (em alguns casos poderemos chamar de causa e efeito) entre os vértices. No grafo orientado as arestas são chamadas de **arcos**, e a relação definida pelo conjunto A não é simétrica, existindo uma orientação na relação entre os nós. Imagine, por exemplo, que alguém que conheça a rotina da dona de casa do Exemplo 1 resolvesse prever a realização das suas tarefas. Sabe–se que, se a senhora estiver em casa, ela poderá ir ao banco ou à loja de animais com alguma probabilidade. Se ela estiver no banco, a tendência será de ela ir à farmácia comprar os remédios do marido ou ao mercado para realizar suas compras.

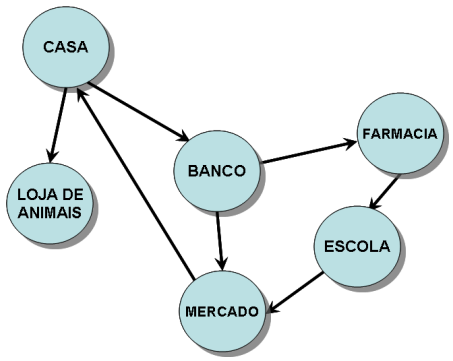


Figura 3.2: Grafo que identifica um possível caminho percorrido pela dona de casa no exemplo hipotético

A sequência dada na Figura 3.2 define um **caminho** (ou cadeia), que é aquele realizado do nó *casa* até o nó *mercado*, da forma [*casa, banco, farmácia, escola, mercado*]. Um **subcaminho** é qualquer conjunto de nós e arcos que esteja inserido no caminho, por exemplo o trajeto [*banco, escola*]. Um nó **pai** é aquele em que existe um arco partindo dele a qualquer outro nó no grafo e todo nó que recebe um arco a partir de um nó pai é chamado seu **descendente**. Nota–se da Figura 3.2 que o retorno da senhora à sua casa permite que haja um **ciclo** que se define como sendo o caminho de um vértice até si mesmo (NEAPOLITAN, 2004).

Definição 2 – Um grafo direcionado acíclico (GDA) é um grafo composto por nós e arcos no qual não existe a ocorrência de ciclos.

Para compor uma rede Bayesiana devemos ter um grafo direcionado acíclico. Tendo um conjunto de variáveis aleatórias $\mathbf{X} = X_1, \ldots, X_k$ associadas aos nós do grafo e uma distribuição de probabilidade conjunta definida em $\mathbf{X}$, podemos construir as seguintes definições:

Definição 3 – Uma rede Bayesiana é um par $\mathcal{B} = (\mathcal{G},\boldsymbol{\theta})$ definido sobre um conjunto de variáveis aleatórias $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$, onde cada X_i corresponde a um nó, $\mathcal{G}$ é um grafo direcionado acíclico que será chamado de estrutura e $\boldsymbol{\theta}$ é um conjunto de parâmetros que especificam distribuições de probabilidades condicionais que satisfaçam a condição de Markov:

$$P_\theta[X_i|X_j, pa(X_i)] = P_\theta[X_i|pa(X_i)],$$

onde $pa(X_j)$ é o conjunto de nós que são pais de X_j.

Em outras palavras, a condição de Markov para uma rede Bayesiana diz que qualquer nó na rede é condicionalmente independente de seus não descendentes condicionado a seus pais.

As redes Bayesianas discretas são aquelas construídas de forma que a cada nó esteja associada uma variável aleatória do tipo discreto. Nesse caso, a distribuição de probabilidade conjunta das variáveis aleatórias $\mathbf{X}$ é obtida pela fatoração

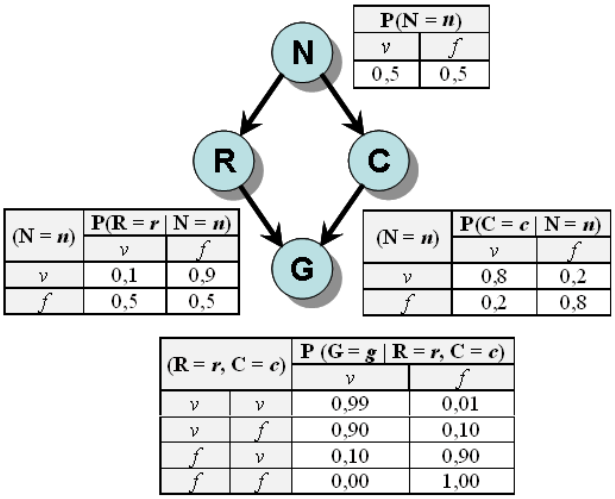
$$P(\mathbf{X}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \prod_{i=1}^{k} P(X_i = x_i|pa(X_i)),$$

e, de acordo com a Definição 3, a possível falta de arcos entre as variáveis X_i e X_j no grafo $\mathcal{G}$ significa independência entre as mesmas.

No Exemplo 2 ilustrado pela Figura 3.3 (MURPHY, 1998) observa–se uma rede Bayesiana discreta. A rede para este exemplo é formada pela estrutura gráfica ($\mathcal{G}$) e pelo conjunto de probabilidades associadas ($\boldsymbol{\theta}$) explicitadas nas tabelas. Esta rede especifica o relacionamento entre quatro variáveis aleatórias discretas, cada uma contendo dois valores possíveis (v = verdadeiro e f = falso). As variáveis são: a existência de nuvens no céu ($\boldsymbol{N}$), o regador ligado ($\boldsymbol{R}$), a ocorrência de chuva ($\boldsymbol{C}$) e a grama molhada ($\boldsymbol{G}$). Neste caso pode–se quantificar incertezas a partir da dependência entre a grama estar molhada e o regador ligado ou à ocorrência de chuva.

Observa–se na Figura 3.3 que, uma vez condicionado em $\boldsymbol{R}$ e $\boldsymbol{C}$, as variáveis $\boldsymbol{N}$ e $\boldsymbol{G}$ são independentes, ou seja, a grama estar molhada não depende de haver nuvens no céu dada a ocorrência de chuva e ao regador ligado. Os relacionamentos de causa e efeito entre as variáveis podem gerar as chamadas classes de equivalência que representam as mesmas relações de independência.

Neste mesmo exemplo, temos que $\{\boldsymbol{N}\}$ é o pai de $\boldsymbol{R}$ e $\boldsymbol{C}$ e $pa(\mathbf{N}) = \emptyset$. Na mesma rede, $pa(\mathbf{G}) = \{\mathbf{R}, \mathbf{C}\}$ e $P(\mathbf{G}, \mathbf{N}|pa(\mathbf{G})) = P(\mathbf{G}|pa(\mathbf{G}))$, pois $\mathbf{G}$ e $\mathbf{N}$ são independentes



P(N = *n*)	
v	*f*
0,5	0,5

(N = *n*)	P(R = *r* | N = *n*)	
	v	*f*
v	0,1	0,9
f	0,5	0,5

(N = *n*)	P(C = *c* | N = *n*)	
	v	*f*
v	0,8	0,2
f	0,2	0,8

(R = *r*, C = *c*)		P (G = *g* | R = *r*, C = *c*)	
		v	*f*
v	*v*	0,99	0,01
v	*f*	0,90	0,10
f	*v*	0,10	0,90
f	*f*	0,00	1,00

Figura 3.3: Rede para grama molhada citada no Exemplo 2 (MURPHY, 1998)

se condicionado aos pais de **G** (neste caso, **R** e **C**).

A Definição 3 leva a uma propriedade do GDA chamada d–separação. Se $\mathcal{B} = (\mathcal{G}, \theta)$ é uma rede Bayesiana que satisfaz a condição de Markov, então é possível identificar os nós condicionalmente independentes a partir da d–separação. Como exemplo, na rede da Figura 3.3, **N** e **G** são d–separados por $\{\mathbf{R}, \mathbf{C}\}$.

Para definir formalmente a d–separação, será necessária a utilização de alguns conceitos a mais sobre o relacionamento entre os nós em um grafo. Sejam X, Y e Z três nós que formam um subcaminho no grafo. Diz–se que (NEAPOLITAN, 2004):

1. Se $X \rightarrow Z \rightarrow Y$, temos um relacionamento do tipo *head–to–tail*;
2. Se $X \leftarrow Z \rightarrow Y$, temos um relacionamento do tipo *tail–to–tail*;
3. Se $X \rightarrow Z \leftarrow Y$, temos um relacionamento *head–to–head*.

Nos casos 1 e 2, ao se observar o nó Z, tem–se que ele bloqueia a cadeia. Em 3, se o nó Z não for observado e nenhum dos seus descendentes forem observados, ele bloqueará a cadeia. Maiores detalhes em Neapolitan (2004).Com estas informações pode–se definir formalmente d–separação:

Definição 4 – Seja $\mathcal{G} = (V, A)$ um grafo direcionado acíclico (GDA), e E um subconjunto de V, com X e Y nós distintos em $V - E$. Dizemos que X e Y são d–separados pelo conjunto E em $\mathcal{G}$ se todo caminho entre X e Y é bloqueado por E.

Pode–se construir para cada GDA um conjunto de equivalência, chamado de conjunto de equivalência de Markov, no sentido de que as redes preservariam as mesmas independências a partir das d–separações identificadas.

Segundo Neapolitan (2004), dois grafos são equivalentes se e somente se, baseado na condição de Markov, eles representam as mesmas independências condicionais. Em outras palavras, dois grafos são Markov–equivalentes se tiverem as mesmas ligações entre os nós sem considerar suas direções, além de preservar os mesmos relacionamentos *head–to–head* no grafo. Outras condições existem, mas para o desenvolvimento deste trabalho estas serão suficientes. O leitor que tiver necessidade de maior detalhamento teórico pode consultar Neapolitan (2004).

Do Exemplo 2, as Figuras 3.4(b) e 3.4(c) a seguir mostram a classe de equivalência do GDA da rede original (reproduzida na Figura 3.4(a)). Nenhum outro grafo é equivalente a eles.

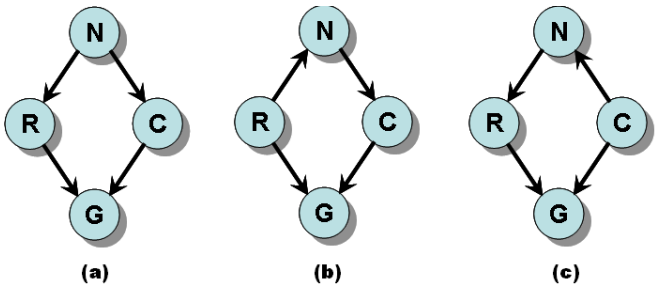


Figura 3.4: (a) Rede original e (b) e (c) Elementos da classe de equivalência da rede original para o exemplo da rede para a grama molhada

Definição 5 – Sejam $\mathcal{G}_1 = (V, A_1)$ e $\mathcal{G}_2 = (V, A_2)$ dois grafos direcionados acíclicos contendo o mesmo conjunto de nós V e A_1, A_2 seus respectivos conjuntos de arcos. $\mathcal{G}_1$ e $\mathcal{G}_2$ são ditos da mesma classe de equivalência se, para todos os subconjuntos mutuamente disjuntos $B, C, D \subseteq V$, B e C são d–separados por D em $\mathcal{G}_1$ e B e C são d–separados por D em $\mathcal{G}_2$. Isso implica que as mesmas independências condicionais entre B e C são observadas em $\mathcal{G}_1$ e $\mathcal{G}_2$.

O exemplo na Figura 3.5 a seguir é reproduzido de Neapolitan (2004) e exemplifica a identificação de elementos da classe de equivalência do grafo em (a).

Considere os grafos em (a), (b), (c) e (d) na Figura 3.5. Os grafos em (a) e (b) são equivalentes pois apresentam o mesmo conjunto de arestas, sem considerar suas direções

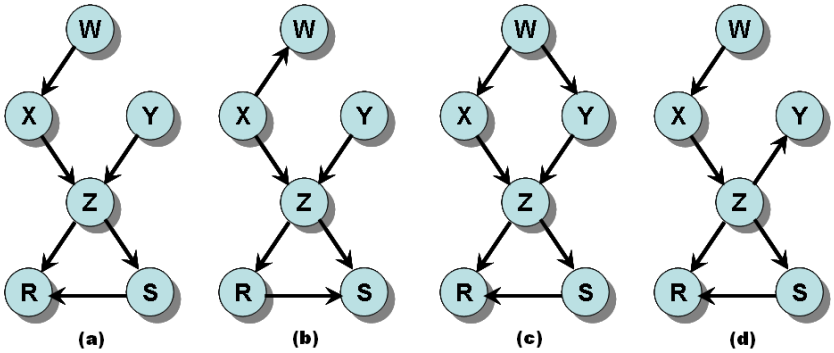


Figura 3.5: Exemplo de identificação de elementos de uma mesma classe equivalência (reproduzido de Neapolitan (2004))

e o mesmo relacionamento *head–to–head* $X \rightarrow Z \leftarrow Y$ no nó de bloqueio Z está presente nos dois GDAs. Considerando agora o conjunto {(a),(b)} e avaliando (c) e (d), temos que:

- (c) e {(a),(b)} não são equivalentes por causa da presença do arco em $W \rightarrow Y$ no grafo em (c), o que implica em conjuntos diferentes de arestas entre os dois grafos;
- (d) e {(a),(b)} não são equivalentes por não manter o relacionamento *head–to–head* no nó de bloqueio Z no grafo em (d).

Comparando agora (c) e (d), eles não são equivalentes entre si por causa da existência do arco $W \rightarrow Y$ em (c) e pelo não relacionamento *head–to–head* em (d).

Uma vez definida a estrutura da rede e seu conjunto de parâmetros $\boldsymbol{\theta}$, pode–se calcular uma função que atribui um valor para cada GDA (ou a um determinado padrão de GDAs) baseado nos dados. Esta quantidade atribuída ao grafo é denominada de função escore e seu cálculo depende da distribuição de probabilidade associada às variáveis aleatórias do problema em questão. A próxima seção trata especificamente desta função.

3.2.2 O ajuste da rede e o escore Bayesiano

Existem na literatura, diversas formas de ajuste para uma rede Bayesiana. Grosso modo, pode–se dispor de uma estrutura prévia e o objetivo passa a ser obter os parâmetros associados aos nós, ou então sabe–se de antemão, pelas variáveis aleatórias do problema, da existência de um conjunto $\boldsymbol{\mathcal{G}}=\{\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(p)}\}$ de possíveis estruturas que podem ser aplicadas aos dados e tem–se interesse em obter a estrutura que melhor se adeque a um dado padrão de GDA. Nas duas situações é possível obter um valor chamado escore Bayesiano, que é uma função calculada com base nos dados e que representa a distribuição de probabilidade conjunta das variáveis aleatórias sob uma dada estrutura (NEAPOLITAN, 2004).

O mais comum quando tratamos de redes discretas é considerar um espaço de estruturas de redes Bayesianas multinomiais, em que a função escore se obtém a partir de uma família de distribuições de Dirichlet, que são as conjugadas para observações multinomiais:

$$S(\mathcal{G}) = P(\mathcal{G}, d) = P(d|\mathcal{G})P(\mathcal{G}) = \prod_{i=1}^{l} \prod_{j=1}^{q_i^{(\mathcal{G})}} \frac{\Gamma(m_{ij}^{(\mathcal{G})})}{\Gamma(m_{ij}^{\mathcal{G}} + n_{ij}^{\mathcal{G}})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}^{(\mathcal{G})} + s_{ijk}^{(\mathcal{G})})}{\Gamma(\alpha_{ijk}^{(\mathcal{G})})}, \tag{3.2.1}$$

onde d representa a base de dados composta por n observações e m parâmetros (ou probabilidades) a calcular, r_i é o número de classes de X_i em $\mathcal{G}$, q_i é o número de diferentes combinações das classes dos pais de X_i em $\mathcal{G}$, $\alpha_{ijk} = \frac{m_{ij}}{r_i q_i}$, s_{ijk} é o número de vezes na qual x_i é igual a k e $\Gamma(.)$ é a função gama.

Os parâmetros da rede $\theta_{X_i|pa(X_i)}$ podem ser definidos por especialistas (e neste caso a estrutura da rede e a função escore estão dependentes destes valores) ou então estimados a partir dos dados por meio de contagens em cada uma das combinações entre as classes das variáveis.

Assim como na seção anterior foi descrita a equivalência entre duas redes a partir de suas estruturas, uma outra forma consistente de se avaliar equivalência é feita com base na função escore. Segundo Bøttcher e Dethlefsen (2003), pode–se mostrar que os escores para duas estruturas de redes equivalentes são iguais. Embora exista a possibilidade de se tratar da equivalência entre redes a partir da função escore, conforme em Neapolitan (2004) e Bøttcher (2004), esta verificação se torna complexa sob o contexto de imputação. Deve–se observar que existe um problema combinatório que depende do número de nós e arcos da estrutura original. Isso porque a obtenção das propriedades da função escore não é trivial e deve–se incorporar a variabilidade associada ao dado imputado.

A equivalência entre as redes pode ser avaliada a partir de $S(\mathcal{G})$ e, apesar de terem sido verificadas empiricamente algumas propriedades, estas não serão apresentadas neste texto por não originarem conclusões específicas.

3.2.3 Algoritmos para ajuste de estruturas

O objetivo desta seção está em apresentar os algoritmos para o ajuste de estruturas e estimação dos parâmetros que são utilizados para imputação a partir de redes Bayesianas. Segundo Hruschka–Jr. (2003), os algoritmos podem ser divididos de acordo com a forma de ajuste da rede: (1) segundo uma busca heurística da estrutura, ou (2) pelo conceito de independência condicional dos atributos da rede. Estas duas formas de

ajuste englobam a estrutura e os parâmetros da rede Bayesiana.

O fundamento da busca heurística é utilizado pelos algoritmos que propõem a rede Bayesiana para imputação, conforme será visto na próxima seção. A busca heurística tem como principal característica a dependência do processo de aprendizado da estrutura com relação à ordenação das variáveis. Isso significa que a ordem das variáveis também passa a ser um parâmetro de entrada para este tipo de algoritmo. Esse aspecto facilita em muito o ajuste de uma estrutura, principalmente se a base de dados tem um grande número de atributos. A idéia é de que se existirem $X_{(k)}$ variáveis ordenadas, a primeira delas, ou $X_{(1)}$, não terá pais mas poderá ser pai de todas as demais. Já $X_{(2)}$ não poderá ser pai de $X_{(1)}$ mas o poderá ser de todas as subseqüentes e assim sucessivamente. A última delas não terá descendentes. Dessa maneira, em um vetor linha que contenha a ordenação de todos os atributos de uma base de dados, uma variável não poderá ser pai de uma variável à sua esquerda, mas pode ser de qualquer uma que esteja à sua direita.

Os algoritmos que usam do conceito de independência das variáveis seguem a Definição 3, que diminui o número de parâmetros a serem calculados e necessários para se definir uma rede.

Uma parte dos algoritmos tem como objetivo o aprendizado apenas da estrutura da rede, e dentre estes existem basicamente três tipos:

1. os que fazem aprendizado baseado em restrições, por exemplo o algoritmo PC (SPIRTES; GLYMOUR; SCHEINES, 1993);

2. os algoritmos que buscam a estrutura para otimizar uma função alvo, por exemplo o K2 (COOPER; HERSKOVITS, 1992);

3. os algoritmos que combinam os dois procedimentos;

Quando o interesse está em se obter a estrutura da rede Bayesiana pode–se agir de duas formas: ou se encontra um grafo que satisfaz todas as restrições em alguma medida de independência condicional empírica nos dados, ou então usa–se alguma métrica para ser avaliada com base em um espaço de grafos retornando o modelo de mais alto escore encontrado. Os híbridos utilizam–se de uma composição dos dois tipos de algoritmos.

A título de informação, Ide (2005) em seu texto apresenta uma seção onde resume os algoritmos para inferência em redes Bayesianas e cita os principais para realizar inferência exata (propagação de crença, reversão de arcos, condicionamento de variáveis, etc.) e inferência aproximada em redes multiconectadas (baseados em métodos Monte

Carlo, avaliações parciais, propagação de mensagens em ciclos, etc.). Estes serão particularmente úteis se o interesse estiver em executar imputações múltiplas quando não se dispõe de uma rede Bayesiana de entrada.

Neste trabalho o ajuste das redes Bayesianas foi realizado usando o *deal* (BØTTCHER; DETHLEFSEN, 2003), que é um pacote desenvolvido para ajuste de redes discretas e/ou mistas no *software* R[3] e que, tanto ajusta redes a partir de restrições como busca a melhor estrutura para os dados.

Os algoritmos PC e K2 também serão citados neste texto por serem utilizados em dois contextos diferentes para imputação de dados a partir de redes Bayesianas. Ambos ajustam a estrutura da rede e são baseados em busca heurística, mas apresentam características distintas: o PC, implementado por exemplo no *Hugin Tool* (MADSEN et al., 2003), faz o aprendizado com base em restrições e o K2, implementado por exemplo no *Weka* (BOUCKÆRT, 2004), busca uma estrutura para otimizar uma determinada métrica.

3.3 Redes Bayesianas como ferramenta para imputação

Apesar de as redes Bayesianas já serem amplamente utilizadas em diversas áreas, em estatísticas oficiais as aplicações são mais recentes; ver por exemplo Getoor, Taskar e Koller (2001). Em imputação, a utilização de redes Bayesianas data de poucos anos atrás. Thibaudeau e Winkler (2002) são os primeiros que mencionam a aplicação de redes Bayesianas para imputação em tabelas de contingência.

Nesta seção são apresentados os algoritmos existentes para a aplicação de redes Bayesianas para imputação, onde se discutem os pontos positivos e limitantes em cada um deles e seu recente histórico. Também se descreve o algoritmo proposto para imputação que será utilizado ao longo do trabalho.

3.3.1 Histórico recente

No contexto de estatísticas oficiais, o projeto EUREDIT apresenta um apêndice (DI ZIO; SCANU, 2003) com uma primeira avaliação do uso de redes Bayesianas para imputação, ressaltando que não haveria comparação com os demais métodos estudados

[3] O R bem como o pacote *deal* podem ser otidos em www.r–project.org.

pelo projeto por serem estes apenas avaliações preliminares e não terem sido conduzidos sob os protocolos de experimentos do conjunto de estudos. Scanu, Di Zio e Vicard (2003) apresentam, ainda com o trabalho em progresso, alguns aspectos computacionais de execução da imputação utilizando redes Bayesianas e Di Zio, Scanu e Vicard (2003) relacionam os problemas em aberto e novas perspectivas do uso de redes Bayesianas para imputação.

Posteriormente, Di Zio et al. (2004) publicam o primeiro artigo com alguns resultados do método, aplicado em parte de uma base de dados não identificados do Censo Demográfico de 1991 da Inglaterra. Estes dados foram os mesmos utilizados no projeto EUREDIT sendo simuladas não respostas do tipo MAR e MCAR em 5% e 10% do total e os resultados foram comparados com procedimentos hot–deck.

Já na área de mineração de dados, Hruschka–Jr. (2003) propõe dois algoritmos para imputação Bayesiana, dos quais um deles executa um teste baseado na estatística de qui–quadrado para a ordenação de variáveis na construção da estrutura. Os resultados são apresentados para problemas de classificação, agrupamento e seleção de atributos e todas as simulações foram realizadas em sete bancos extraídos de repositórios de dados e dois bancos de dados gerados aleatoriamente com caráter didático.

Apesar dos contextos diferentes para o surgimento dos métodos de imputação usando redes Bayesianas, nenhum dos textos apresenta detalhamento teórico da metodologia em questão. Em todos eles sugere–se que estudos mais aprofundados são necessários. As duas maneiras de se aplicar redes Bayesianas para imputação diferem tanto na composição do algoritmo quanto na aplicação ao tipo de mecanismo de não resposta dos dados.

3.3.2 Algoritmos existentes

Atualmente tem–se registrado na literatura dois contextos distintos da utilização de redes Bayesianas para imputação: estatísticas oficiais, com os algoritmos de Thibaudeau e Winkler (2002), Di Zio, Scanu e Vicard (2003) e Di Zio et al. (2004), e em mineração de dados, com dois algoritmos propostos por Hruschka–Jr. (2003). Esta seção apresenta alguns aspectos importantes sobre estes algoritmos.

Considere o objetivo de imputar valores faltantes em $X_1, X_2, \ldots, X_k$ variáveis aleatórias em um conjunto de $a = 1, 2, \ldots, n$ unidades. Em seu algoritmo, Thibaudeau e Winkler (2002) usam a rede Bayesiana como uma representação alternativa para o método hot–deck. Já Di Zio, Scanu e Vicard (2003) e Di Zio et al. (2004) utilizam–se

da ordenação das variáveis em ordem descrescente de confiabilidade para a construção da rede Bayesiana que definirá as regras de imputação. Ambos os textos sugerem como medida de confiabilidade o percentual de valores presentes em cada variável, onde mais confiável implica em menor número de itens faltantes. Embora citem que maiores estudos são necessários neste sentido.

No artigo de Di Zio, Scanu e Vicard (2003), a imputação de itens faltantes em uma variável X_k qualquer se dá a partir do seu cobertor de Markov, que é o conjunto composto pelos pais e filhos de X_k, além dos pais dos filhos de X_k. Neste tipo de imputação, nenhuma variável aleatória estaria em um conjunto sem pais, ou seja, a cada variável a ser imputada seria ajustada uma nova rede em que apenas as informações relevantes influenciariam para gerar os valores a serem imputados em X_k. Segundo os autores, este procedimento traria menor variância de imputação, mas nenhum resultado da aplicação deste método foi apresentado formalmente. Um ponto limitante está no ajuste de uma nova rede a cada variável aleatória a ser imputada, produzindo um trabalho computacional intensivo.

No algoritmo de Di Zio et al. (2004), além de ordenar as variáveis de acordo com a sua confiabilidade, separam–se as mesmas em subconjuntos de tal maneira que o primeiro subconjunto P_0 contenha somente as variáveis órfãs, o subconjunto P_1 conterá aquelas cujos pais encontram–se em P_0, o conjunto P_2 conterá as variáveis com pais em $P_0 \bigcup P_1$, e assim sucessivamente até o último subconjunto $j = \nu - 1$, onde ν é o número de subconjuntos formados.

Após ordenadas as variáveis, ajusta–se uma rede Bayesiana de tal forma que aquelas mais confiáveis não podem ser filhas de variáveis menos confiáveis. Para a primeira observação ($a = 1$), observa–se a variável a imputar. Se esta estiver em P_0, então um valor é gerado de acordo com a sua distribuição marginal. Se esta pertencer a qualquer outro conjunto, então um valor é gerado de acordo com a sua posição na rede, respeitando a estrutura hierárquica da rede construída. Segue–se este procedimento até atingir $a = n$. Segundo Di Zio et al. (2004), neste algoritmo o percentual de itens corretamente imputados aumenta com relação às variáveis menos confiáveis da base de dados. Os piores desempenhos deste método encontram–se nos atributos pertencentes aos subconjuntos P_0 e P_1. O Quadro 2 explicita este algoritmo.

Em mineração de dados, Hruschka–Jr. (2003) desenvolve um método Bayesiano genérico de predição de valores ausentes em bancos de dados, e como contribuição propõe dois algoritmos para imputação. A base para estes dois algoritmos está no algoritmo K2

Algoritmo de Di Zio et al. (2004)
Entrada: Base de dados com valores faltantes.
Saída: Base de dados imputada.
1. Ordene as variáveis $X_1, X_2, \ldots, X_k$ em ordem decrescente de confiabilidade formando uma partição de ν subconjuntos mutuamente exclusivos;
2. Ajuste uma rede Bayesiana que respeite a ordenação definida em 1.;
3. Para cada item faltante de $a = 1$, observe a partição à qual a variável pertence;
   3.1. Se a variável pertencer ao primeiro conjunto (ou conjunto das variáveis sem pais), gere aleatoriamente um dado a ser imputado de acordo com a distribuição marginal da mesma;
   3.2. Se não, gere um dado a ser imputado de acordo com a estrutura da rede ajustada em 2.;
4. Repetir 3. para $a = 2, \ldots, n$.

Quadro 2: Algoritmo proposto por Di Zio et al. (2004)

(COOPER; HERSKOVITS, 1992), que é um método baseado em busca heurística para aprendizado de estruturas de redes probabilísticas a partir de uma base de dados que contenha apenas variáveis discretas e no algoritmo *Global Bayesian Conditioning* (GBC) (PEARL, 1988) usado para inferir os valores mais adequados para substituir os valores ausentes.

Em seu primeiro algoritmo, chamado de K2I (ou K2 para Imputação), Hruschka–Jr. (2003) sugere que se deve identificar o subconjunto $X_{(1)}, \ldots, X_{(p)}$, $p \leq k$, de variáveis com valores ausentes, definindo–as como alvo e excluindo todas as unidades que contenham pelo menos um item faltante. Ordena–se a base de dados resultante de tal forma que a primeira variável $X_{(1)}$ torna–se o primeiro atributo alvo. Aplica–se o algoritmo K2 para gerar uma rede Bayesiana para $X_{(1)}$ e com o algoritmo GBC, usando a rede obtida, estimam–se os valores mais adequados para os ausentes nesta variável. Repete–se este procedimento para todas as variáveis alvo restantes, ou seja, para $X_{(i)}$, $i = 2, \ldots, p$. Estes procedimentos estão apresentados no Quadro 3.

Neste primeiro algoritmo, a variável a ser imputada é sempre a base inicial e todo o processo está direcionado a este atributo, repetindo–se o procedimento tantas vezes quantas forem as variáveis com valores faltantes. Por ser baseado em uma busca heurística, isso faz com que o desempenho do método para imputação seja afetado pela ordenação das variáveis que geram a rede Bayesiana a partir do K2. Por este motivo, Hruschka–Jr. (2003) propõe no mesmo trabalho, um outro algoritmo para imputação, chamado de K2Iχ^2 , que obtém a melhor ordenação das variáveis com base em um teste

Algoritmo K2I

Entrada: Base de dados com valores faltantes.
Saída: Base de dados imputada.

1. Identificar o subconjunto $X_1^{'}, \ldots, X_p^{'}$ de variáveis com valores faltantes na base de dados a ser trabalhada;
2. Criar uma nova base de dados, excluindo as unidades que contenham pelo menos um item faltante;
3. Para cada atributo alvo, faça:
   - 3.1. Defina uma base de dados de aprendizado, tendo como objetivo gerar uma estrutura de classificação para a variável alvo;
   - 3.2. A partir do algoritmo K2 gerar uma rede Bayesiana para a base de dados em 3.1.;
   - 3.3. A partir do algoritmo GBC da rede em 3.2., estimar os valores adequados para imputar na variável alvo;
   - 3.4. Substituir os valores ausentes pelos encontrados em 3.3.

Quadro 3: Algoritmo K2I proposto por Hruschka–Jr. (2003)

de qui–quadrado.

Os procedimentos para gerar a rede e para estimar os valores a serem imputados são os mesmos nos dois algoritmos, mas a diferença principal entre o K2I e o K2Iχ^2 está na construção da rede a partir de ordenações distintas das variáveis usando o teste de qui–quadrado. No primeiro algoritmo a ordenação das variáveis é determinada pela seqüência em que se apresentam as mesmas na base de dados. No K2Iχ^2, o teste de qui–quadrado é usado para pares de variáveis de maneira que são obtidos os maiores graus de associação entre as variáveis e definida uma ordenação de acordo com estes cálculos. É realizada uma seqüência de testes de qui–quadrado, elencada a maior associação entre duas variáveis, por exemplo X_i e X_j que passam para as primeiras posições, $X_{(1)}$ e $X_{(2)}$ das variáveis ordenadas. A seguir, é realizado o teste entre $X_{(2)}$ e uma X_k qualquer para a escolha de $X_{(3)}$. Este procedimento é realizado até que todas as variáveis aleatórias estejam na base de dados ordenada.

Nos algoritmos propostos para o contexto de estatísticas oficiais, a aprendizagem das redes está baseada no conceito de independência condicional e faz–se necessário algum artifício extra (por exemplo, o auxílio de um especialista) para completar a rede a ser usada na imputação. Isso porque a ordenação das variáveis segundo o conceito da confiabilidade limita o relacionamento entre elas[4]. Ainda se poderia deparar com um possível "empate" neste tipo de ordenação, o que traria uma escolha arbitrária da dependência entre os atributos.

[4]Neste caso, uma variável mais confiável só pode ser pai de uma menos confiável que ela.

Com relação aos algoritmos de Hruschka–Jr. (2003), estes também são influenciados pela ordenação das variáveis, ou seja, uma ordem que não mapear o relacionamento entre os atributos de forma correta poderá inserir erros no processo de aprendizagem e conseqüentemente conduzir a um mau desempenho para imputação.

3.3.3 Algoritmo proposto para imputação

Considere que já exista uma rede Bayesiana $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$ ajustada de pesquisas anteriores ou construída a partir de uma base de dados contendo não resposta em pelo menos uma das variáveis $X_1, X_2, \ldots, X_k$ associadas aos nós da rede e onde se conhecem os parâmetros $\boldsymbol{\theta}$ da sua distribuição. É importante observar que, para justificar o uso da rede Bayesiana para imputação, esta deve manter as relações de causa e efeito das variáveis consideradas, ou seja, os critérios para construir as relações de independência condicional devem estar identificados na rede. Registra–se que o objetivo de usar uma rede Bayesiana para imputação está em preservar o relacionamento existente entre as variáveis, então o ajuste inicial (ou rede original) será fundamental para esta verificação. Se esta rede contiver algum problema de ajuste, a imputação ou qualquer procedimento inferencial a partir dela conterá problemas também.

Dispõe–se de n observações das quais n^*, $n^* \leq n$, contêm pelo menos um item faltante. Dividem–se as X_i, $i = 1, 2, \ldots, k$ variáveis em grupos disjuntos conforme identificados na estrutura $\mathcal{G}$ ajustada (a exemplo do que é conduzido por Di Zio et al. (2004), só que em sua proposta a identificação é feita antes do ajuste da rede): o subconjunto P_0 contém as variáveis órfãs (sem pais) em $\mathcal{G}$; o subconjunto P_1 contém aquelas com pais somente em P_0; já o subconjunto P_2 contém pais em $P_0 \bigcup P_1$, e assim sucessivamente até o último subconjunto $j = \nu - 1$, que contém os pais das variáveis em ν.

Se a variável estiver em P_0, imputa–se um valor (ou classe) de acordo com a distribuição marginal da variável em questão. Se a variável estiver em P_1, a imputação dar–se–á em função da distribuição condicionada em P_0, ou seja, é gerado um valor da distribuição $P(X_i|pa(X_i))$, sabendo que $X_i \subseteq P_1$ e $pa(X_i) \subseteq P_0$. Este procedimento prossegue até que a imputação nas variáveis do subconjunto ν, que têm pais no conjunto $\bigcup_{i=1}^{j+1} P_i$, sejam imputadas. O Quadro 4 resume este algoritmo.

A idéia de se particionar as variáveis em subconjuntos de acordo com a estrutura $\mathcal{G}$ da rede Bayesiana para imputação tem como principal objetivo considerar que, o que acontecer para um determinado subconjunto possa influenciar na imputação das variáveis do subconjunto seguinte. As imputações que forem conduzidas por este método

Algoritmo proposto
Entrada: Rede Bayesiana ajustada e
Base de dados com valores faltantes.
Saída: Base de dados imputada.
1. Identifique os subconjuntos $P_0, P_1, \ldots, P_\nu$ na rede Bayesiana de entrada;
2. Defina uma ordem de imputação em cada para as variáveis em cada subconjunto de acordo com algum critério;
3. Para cada subconjunto $j = 1, 2, \ldots, \nu$, faça:
   3.1. Se a variável pertencer ao primeiro conjunto (ou conjunto das variáveis sem pais), gere aleatoriamente um dado a ser imputado de acordo com a distribuição marginal da mesma;
   3.2. Se não, gere um dado a ser imputado de acordo com a estrutura da rede ajustada em 2.;
4. Retorne a base imputada.

Quadro 4: Algoritmo proposto para uso de redes Bayesianas discretas em imputação

nas variáveis em P_0 serão equivalentes àquelas produzidas pela imputação geral aleatória pois não estarão condicionadas a nenhuma variável aleatória na rede. Já as imputações realizadas em $P_1, P_2, \ldots$, serão equivalentes às imputações aleatórias dentro de classes, pois são obtidas a partir das distribuições de probabilidade condicionadas em seus pais. As classes de imputação, porém, são delimitadas pela rede $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$.

A avaliação do método também se torna facilitada, pois uma imputação em P_1 depende de menor número de parâmetros que aquelas em P_2, e esse pode ser um fator importante para se considerar as consistências em subconjuntos distintos de variáveis.

Um ponto limitante do algoritmo proposto no Quadro 4 está em não permitir alterar os valores dos parâmetros da rede, o que garante que todos os itens imputados em uma mesma variável apresentam a mesma distribuição descrita inicialmente. Isso significa que não se considera uma atualização de $\boldsymbol{\theta}$. No caso de a rede de entrada no algoritmo ser ajustada com a presença de itens faltantes em alguma variável, os parâmetros em cada nó só poderiam ser atualizados antes da fase de imputação.

Este método de imputação permite o tratamento de forma simples dos chamados zeros estruturais. Os zeros estruturais são aqueles casos em que existem blocos de perguntas que não se aplicam a determinados respondentes no questionário (como por exemplo, questões relacionadas ao trabalho para pessoas não inseridas no mercado). A rede Bayesiana funciona aqui como um dispositivo que, com probabilidades conhecidas em $\boldsymbol{\theta}$, permite gerar ocorrências para variáveis a partir de $\boldsymbol{\theta}$. Se um zero estrutural for considerado como uma classe da variável aleatória que a contém, então associa–se a este uma probabilidade nula (ou quase nula) se condicionada a uma classe que a determine (no

caso do exemplo, se a pessoa não estiver no mercado, então a probabilidade de pertencer a uma classe zero estrutural às questões do trabalho é unitária).

Di Zio et al. (2004) defendem a aplicação das redes Bayesianas para o processo de identificação dos zeros estruturais antes de considerar a imputação. Juntamente com este passo, os autores sugerem que seja incorporado o processo de crítica dos dados, que identificaria algum valor incoerente que passaria a ser um dado faltante. Uma vez realizado o processo de crítica, os itens faltantes seriam previamente classificados em válidos, de acordo com suas características, ou em não respostas a serem imputadas, evitando que imputações fossem realizadas em locais indevidos.

O algoritmo proposto no Quadro 4 permitiria manter a distribuição multivariada dos dados representada pela rede Bayesiana de entrada. Só que no algoritmo proposto nesta seção, as relações existentes entre as variáveis são aquelas que determinam sua probabilidade de imputação e tornam estas informações mais fidedignas à realidade dos dados. O ponto limitante está na necessidade de algum conhecimento prévio do comportamento dos dados. Isso implica que a utilização da experiência de especialistas na formação das redes e obtenção dos parâmetros é imprescindível para a funcionalidade do método. Na próxima seção estão sugeridas quatro maneiras para avaliação dos resultados da imputação.

3.4 Avaliação do uso da rede discreta para imputação

Os estudos atuais sobre a imputação a partir de redes Bayesianas levam à conclusão de que estas preservam a distribuição multivariada dos dados por causa da utilização da estrutura de independência na obtenção de um dado a ser imputado, embora esta afirmação ainda esteja sem confirmação teórica e existam poucos experimentos neste sentido. Devido à especificidade de cada base de dados e de cada problema, os resultados de uma avaliação comparativa a outros métodos de imputação dependeriam das variáveis aleatórias e percentuais de não resposta em cada uma delas, número de observações, grau de dependência entre as variáveis e, especificamente no caso das redes Bayesianas, da ordenação em que as variáveis se encontram na base de dados.

Di Zio et al. (2004) sugerem três parâmetros para a avaliação do uso de redes Bayesianas discretas para imputação em aspectos distintos:

– consistência da base de dados, que avalia a preservação dos microdados;

– consistência lógica, para avaliar a preservação das restrições lógicas;

– consistência estatística, com a preservação dos parâmetros e quantidades associadas à base de dados.

A primeira forma de avaliação diz respeito à preservação dos microdados que seria a propriedade de recuperar exatamente a informação perdida. A preservação das restrições lógicas está na idéia de que o método resgata a plausibilidade dos valores imputados com respeito às restrições lógicas. Isso pode ser feito de forma direta para variáveis discretas e um bom exemplo está na variável faixa etária. Não seria muito razoável que uma faixa etária imputada de "menos de dez anos de idade" fosse feita em um estado civil observado "casado". E a consistência estatística analisa a preservação dos parâmetros da distribuição conjunta a partir de índices descritivos simples. Expande–se neste texto a consistência estatística para parâmetros que possam mensurar o relacionamento entre as variáveis aleatórias, como é o caso do coeficiente de correlação entre estas.

Neste texto são avaliadas três medidas de consistência sugeridas em Di Zio et al. (2004): da base de dados, lógica e estatística. É proposta mais um tipo de consistência, esta com relação à preservação da estrutura da rede Bayesiana após a imputação. A idéia destas quatro maneiras de verificação está em identificar se o conhecimento à priori com respeito à base de dados se mantém e se não existem alterações bruscas, principalmente nas consistências estrutural e estatística, que resumiriam os principais interesses dos mantenedores da base de dados e dos usuários respectivamente.

3.4.1 Consistência da base de dados

O primeiro ponto a ser avaliado relaciona–se com a chamada consistência da base de dados. Nesta consistência se observa a propriedade de manutenção dos registros individuais dos dados após a imputação. Neste caso, só pode ser conduzida quando do conhecimento da base original para testar os valores imputados e o método, ou para verificação de resultados ou ainda para confirmação posterior dos itens imputados nas unidades.

Para a verificação da consistência dos microdados, Di Zio et al. (2004) propõem, para o caso univariado:

$$I_{x_{ij}}(\widetilde{x}_{ij}) = \begin{cases} 1, & se\ \widetilde{x}_{ij} = x_{ij} \\ 0, & se\ \widetilde{x}_{ij} \neq x_{ij} \end{cases}, i = 1, \ldots, n^*, \tag{3.4.2}$$

onde x_{ij} é o valor de X_{ij} na base original, $\widetilde{x}_{ij}$ é o correspondente valor imputado e n^* é o

número de unidades não respondentes na variável X. Neste texto, a extensão para o caso multivariado dar–se–á da forma:

$$I_{u_i}(\widetilde{u}_i) = \begin{cases} 1, & se\ \widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij} \\ 0, & c.c. \end{cases} , i = 1, \ldots, n^*, j = 1, \ldots, k, \tag{3.4.3}$$

onde u_i é o conjunto de valores correspondentes aos imputados da unidade i na base original e $\widetilde{u}_i$ é o conjunto de valores imputados da unidade i na base imputada. Aqui, se pelo menos um item imputado for diferente do real na unidade, então a imputação na unidade será considerada incorreta. Poder–se–ia pensar em atribuir pesos diferentes para números diferentes de acertos quando a imputação fosse conduzida em mais de uma variável por unidade. Este procedimento não foi realizado neste texto e fica como sugestão para trabalhos futuros.

Apesar de as variáveis aleatórias serem imputadas conforme a rede Bayesiana de entrada, seguindo a ordenação dos subconjuntos, considera–se a unidade de observação (composta por seus $\mathbf{x}_{ij}$ itens) o elemento que garante de fato as características de fidedignidade e as demais consistências a que se referem este texto. Embora seja uma restrição forte não admitir a imputação de um item incorreto em uma unidade, do ponto de vista da qualidade dos dados seria importante conhecer, de todas as unidades observadas, qual o percentual de unidades incorretamente imputadas para um dado método.

Um indicador da consistência é então

$$\xi_{Xj} = \sum_i \frac{I_{x_{ij}}(\widetilde{x}_{ij})}{n^*}, \tag{3.4.4}$$

que representa o valor esperado do percentual de acertos na variável X_j para o caso univariado e

$$\xi_U = \sum_i \frac{I_{u_i}(\widetilde{u}_i)}{n^*}, \tag{3.4.5}$$

que representa o valor esperado da proporção de unidades imputadas corretamente para o caso multivariado.

O cálculo de ξ_{X_j} no caso univariado é correspondente à verificação da imputação correta na unidade quando apenas se tem o problema da não resposta em uma variável aleatória. O valor de ξ_U depende do número de variáveis a serem imputadas por unidade, do número de classes das variáveis e da relação de dependência entre as variáveis na rede.

Uma forma de se conduzir um teste de verificação para a preservação dos microdados é aquela proposta por Chambers (2000) e que está descrita neste texto na Seção

2.4.1.

3.4.2 Consistência estrutural

É proposta neste texto uma medida de consistência estrutural, que é a propriedade que o método tem de manter a estrutura construída a partir dos dados imputados na mesma classe de equivalência da rede original.

A estrutura da rede é necessária para a representação do relacionamento de independência condicional entre as variáveis, portanto, a verificação de sua manutenção após a imputação é de fundamental importância para a confiabilidade do método. A consistência estrutural pode ser avaliada de duas formas: a partir do grafo (ou da estrutura propriamente dita) ou a partir de uma medida resumo do ajuste da estrutura, por exemplo a função escore calculada para a rede. No caso da consistência estrutural pela avaliação do grafo, partimos da observação após a imputação, dos mesmos nós e mesma sequência de arestas (independentes da direção) e relações *head–to–head* entre a rede original $\mathcal{B} = (\mathcal{G},\boldsymbol{\theta})$ e a rede ajustada após a imputação $\widetilde{\mathcal{B}} = (\widetilde{\mathcal{G}},\boldsymbol{\theta})$, conforme a equivalência dos grafos descrita na Seção 3.2.1.

Nesta forma de verificação tem–se como hipótese que uma imputação que mantém o relacionamento de independência nos dados é aquela que consegue reproduzir a estrutura $\mathcal{G}$ original dos dados após a construção da estrutura a partir dos dados imputados. Como parâmetros, perseguimos alguns aspectos que permitem identificar se o grafo gerado a partir da base imputada pertence à mesma classe de equivalência da rede original.

Nesse tipo de consistência existem algumas considerações a serem feitas sobre a forma de ajuste e as características dos dados que podem influenciar nos resultados. A primeira a ser considerada refere–se ao número de nós, ou variáveis aleatórias. Quanto maior o número de nós, maior a possibilidade de existência de arestas entre eles, o que pode levar a um maior número de redes na classe de equivalência da rede original. Certamente, uma classe de equivalência de uma rede para uma base que possui três variáveis aleatórias, e que pode apresentar no máximo três arestas em uma rede construída a partir delas, é muito menos sensível do que uma classe de equivalência para uma estrutura que contenha quatro variáveis e que pode conter no máximo seis arestas.

O mesmo ocorre com a quantidade de observações disponíveis e faltantes na base de dados. Parece intuitivo que, quanto maior o número de observações da base de dados, mais estável é a rede Bayesiana $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$ com relação à estrutura e aos parâmetros asso-

ciados. Já quanto maior o percentual de dados faltantes, maior a incerteza em se manter a estrutura dos dados originais. Isso porque, mesmo em se obtendo os mesmos parâmetros para as distribuições univariadas, as imputações em registros individuais podem ocasionar alterações nas relações de independência condicional entre as variáveis.

Para se construir uma medida de avaliação da consistência estrutural, considera–se o caso em que o número de nós da estrutura $\mathcal{G}$ de entrada é o mesmo a ser trabalhado na construção da rede após a imputação. Mesmo que não existam itens faltantes em alguma variável, é imprescindível que a estrutura $\widetilde{\mathcal{G}}$ tenha as mesmas variáveis da estrutura original para que esteja em sua classe de equivalência. Além disso:

- para que a estrutura $\widetilde{\mathcal{G}}$ esteja na mesma classe de equivalência de $\mathcal{G}$ é necessário que exista o mesmo número de arestas entre a rede original e a rede ajustada a partir dos dados imputados e;
- deve ser considerado que a estrutura construída a partir dos dados imputados mantenha os mesmos relacionamentos *head–to–head* em nós de bloqueio.

Um número de arestas menor que o da estrutura original implica que o método modifica o relacionamento inicial na direção da independência entre as variáveis e um número de arestas maior que o da estrutura original implica em modificar o relacionamento inicial na direção de maior dependência entre as variáveis.

Para se obter uma indicação do que ocorre neste sentido, define–se um conjunto de informações que expresse estas quantidades. Seja B uma variável aleatória que especifica os seguintes eventos:

$$B = \begin{cases} b_1, & se\ n(\widetilde{\mathcal{G}}) < n(\mathcal{G}) \\ b_2, & se\ n(\widetilde{\mathcal{G}}) = n(\mathcal{G}) \\ b_3, & se\ n(\widetilde{\mathcal{G}}) > n(\mathcal{G}) \end{cases}, \tag{3.4.6}$$

onde $n(\mathcal{G})$ representa o número de arestas em $\mathcal{G}$ originalmente ajustada e $n(\widetilde{\mathcal{G}})$ representa o número de arestas em $\widetilde{\mathcal{G}}$.

As redes pertencentes à mesma classe de equivalência de $\mathcal{G}$ estão contidas no conjunto $\{B = b_2\}$. O procedimento seguinte consiste em verificar as posições dos arcos independente de suas direções. Para cada estrutura pertencente ao conjunto de possíveis redes equivalentes, defina a seguinte variável aleatória:

$$C = \begin{cases} c_1, & se\ l_i(\widetilde{\mathcal{G}}) = l_i(\mathcal{G}) \\ c_2, & se\ l_i(\widetilde{\mathcal{G}}) \neq l_i(\mathcal{G}) \end{cases}, i = 1, \ldots, n(\widetilde{\mathcal{G}}), \tag{3.4.7}$$

onde $l_i(\mathcal{G})$ é a posição da aresta i na estrutura construída a partir da base original e $l_i(\widetilde{\mathcal{G}})$ é a posição da aresta i na estrutura construída a partir da base imputada.

Se é válida a suposição de que a imputação a partir de redes Bayesianas mantém a distribuição multivariada dos dados, então a consistência estrutural a partir do grafo deve manter–se após a construção da rede com a base imputada. Esta suposição será verificada neste texto através de simulação. Sugere–se que seja conduzido um estudo futuro sobre as propriedades teóricas destas variáveis.

A segunda forma de verificação da consistência estrutural é conduzida com base em uma quantidade calculada sob a rede original e o equivalente com a rede construída após a imputação. Isso pode ser feito a partir do valor de qualquer função, por exemplo o escore da rede, que é usada como métrica para o aprendizado da estrutura $\mathcal{G}$ e compõe a base de uma estratégia de busca heurística no seu ajuste. Segundo Bøttcher e Dethlefsen (2003), esta quantidade representa as independências condicionais entre as variáveis aleatórias a partir da probabilidade relativa:

$$S(\mathcal{G}) = P(\mathcal{G}, d) = P(d|\mathcal{G})P(\mathcal{G}),$$

que é calculada para a rede discreta pela Expressão (3.2.1). Neapolitan (2004) mostra que a função escore é consistente para a classe de modelos de uma distribuição representada pela rede Bayesiana que identifica o mapa de independências ótimo dos parâmetros de uma base de dados.

A hipótese levantada neste trabalho é de que, se a rede Bayesiana para a imputação mantém a relação observada na rede original, então os valores da função escore devem se manter os mesmos com a rede ajustada após a imputação. Para avaliar esta afirmação constrói–se uma medida simples de forma que:

$$\tau = S(\mathcal{G}) - S(\widetilde{\mathcal{G}}),$$

onde $S(\widetilde{\mathcal{G}})$ é o escore da rede calculado após a imputação. Espera–se que, para duas redes da mesma classe de equivalência, esta diferença não seja muito afastada de um determinado ϵ que varia conforme características próprias da rede como número de nós, de arestas, de classes em cada variável, de observações, etc.

Torna–se interessante aqui a obtenção das propriedades de τ para verificar a sua

composição no contexto de imputação:

$$E[\tau|d] = E\{[S(\mathcal{G}) - S(\widetilde{\mathcal{G}})] \mid d\}$$

e

$$Var[\tau|d] = Var\{[S(\mathcal{G}) - S(\widetilde{\mathcal{G}})] \mid d\}.$$

Na próxima seção, descreve–se a consistência lógica proposta por Di Zio et al. (2004).

3.4.3 Consistência lógica

A consistência lógica refere–se à capacidade que o método tem de identificar, em uma variável, situações não prováveis de imputação. São características destes casos os zeros estruturais, as restrições de crítica e limitações definidas pelas próprias variáveis (como é o caso da combinação entre trabalho remunerado e faixa etária menor que dez anos de idade em pesquisas sobre emprego e rendimento, por exemplo). Para que seja possível quantificar essa medida de consistência é necessário que os parâmetros da rede reflitam estas circunstâncias.

Por exemplo, se uma dada resposta no questionário de uma pesquisa implicar o não preenchimento de uma ou mais perguntas, então estas não seriam tratadas pelo método de imputação como um dado faltante e sim, como sendo um zero estrutural. Quando o ajuste da rede identifica estas situações, a imputação nestas classes têm probabilidade nula, então é esperado que nenhuma imputação seja feita. Uma outra possibilidade está em que, dependendo do instrumento computacional de ajuste da rede, possam ser criadas restrições específicas que impeçam a imputação nestas classes.

Di Zio et al. (2004) avaliam a consistência lógica em seu artigo a partir das probabilidades estimadas de se pertencer a uma classe característica de zero estrutural condicionado às suas restrições. A forma de obter informações sobre a consistência lógica é análoga à maneira de tratar a consistência da base de dados, mas só que agora aplicada apenas às variáveis (ou classes) que representam zeros estruturais. Isso significa obter a quantidade:

$$\xi_z = \frac{\sum_{i=1}^{n^*} I_{x_{z,i}}(\widetilde{x}_{z,i})}{n^*}, \tag{3.4.8}$$

onde $x_{z,i}$ é o valor real para a classe que contém um zero estrutural, $\widetilde{x}_{z,i}$ é o valor imputado na mesma classe e $I_{x_{z,i}}(\widetilde{x}_{z,i}) = 1$ se o valor imputado coincide com o valor real.

3.4.4 Consistência estatística

Com relação à consistência estatística, existem diversas formas de fazê-la. Di Zio et al. (2004) avaliam este tipo de consistência do ponto de vista dos parâmetros da rede em cada nó, mas aqui esta consistência é tratada de forma mais abrangente. Além dos parâmetros da rede, também são consideradas medidas importantes que decorrem diretamente do relacionamento entre as variáveis, como medidas de associação e parâmetros de modelos ajustados após os dados imputados.

De uma maneira geral, a consistência estatística após a imputação é a propriedade que o método tem de manter os parâmetros da rede, características das distribuições univariadas e multivariadas e quantidades de interesse.

Para avaliar os parâmetros da rede, uma medida simples proposta no texto de Di Zio et al. (2004) está em considerar uma distância entre as freqüências relativas da classe z antes e depois da imputação. Seja

$$\Delta = \frac{1}{2} \sum_{z} |f_z - \widetilde{f}_z|, \tag{3.4.9}$$

onde f_z denota a freqüência relativa da categoria z de X no conjunto de dados reais dos n^* itens faltantes. Da mesma forma, $\widetilde{f}_z$ representa a freqüência relativa depois da imputação. O valor de Δ será um resultado entre 0 (que significa a igualdade nas duas distribuições) e 1, e pode ser extendido facilmente para o caso multivariado onde, neste caso, os autores observam que se podem avaliar diferentes estratégias de imputação para o mesmo número de variáveis e categorizações.

Especificamente neste texto trata–se, além dos valores de Δ, de uma medida de associação entre duas variáveis, aqui denotada por $V(X_i, X_j)$. Existe uma ampla variedade de medidas apropriadas para avaliar a associação entre variáveis nominais, tais como o teste de qui–quadrado, o qui–quadrado da razão de verossimilhança, o qui–quadrado de Mantel–Haenszel, o valor *Phi*, a medida V de Cramer, o valor *Tau* de Goodman e Kruskal e o coeficiente Kappa (ver, por exemplo (GOODMAN; KRUSKAL, 1954) (HINKLE; WIERSMA; JURS, 1994) (AGRESTI, 2002)). Será considerada neste texto a medida V de Cramer, que é uma medida conceitualmente similar ao coeficiente de correlação (ver a Seção 2.4.1 no Capítulo 2), que se refere ao grau de associação linear entre duas variáveis do tipo contínuo. Os valores possíveis para o coeficiente de correlação estão no intervalo entre -1 e 1. Já para as medidas de associação, estas variam, mas a maior parte delas resulta um valor entre 0 e 1. Cabe observar que estas medidas de associação não

assumem valores negativos devido a não ordenação das classes de uma variável nominal. Para variáveis ordinais, por exemplo, as medidas de associação podem ser negativas.

A idéia de aplicação da consistência estatística neste caso é a de avaliar se a existência de associação entre duas variáveis após a imputação permanece a mesma ou é influenciada por algum fator inerente ao método que usa a rede Bayesiana.

Apesar de os testes de associação entre duas variáveis apresentarem significância com base na distribuição de qui–quadrado, o valor da estatística é difícil de ser interpretado por ser uma função do tamanho da amostra, da independência entre as variáveis e dos graus de liberdade. O valor V de Cramer, assim como o de outras medidas de associação, mostra–se como solução para esta dificuldade de interpretação, embora também seja construído com base no qui–quadrado (AGRESTI, 2002). A fórmula do qui–quadrado é:

$$\chi^2(X_i, X_j) = \sum_{k=1}^{c} \frac{(o_k - e_k)^2}{e_k},$$

onde o_k é a freqüência observada e e_k é a freqüência esperada nas c combinações de classes das variáveis X_i e X_j consideradas em uma tabela de contingência. O valor V de Cramer é calculado por (CRAMER, 1999):

$$V(X_i, X_j) = \sqrt{\frac{\chi^2(X_i, X_j)}{n(l-1)}}, \tag{3.4.10}$$

onde n é o número total de casos na tabela de contingência e l é o menor entre os números de linhas e colunas da tabela. Esta medida é a mais popular das medidas de associação para variáveis nominais e é apropriada para tabelas que são maiores que as do tipo 2×2 (ou seja, tabelas para variáveis do tipo dicotômicas).

Tão importante quanto a obtenção do valor da associação após a imputação por redes Bayesianas, está a verificação de alguma possível alteração na conclusão a partir do seu valor calculado. Para avaliar os valores de $V(X_i, X_j)$ tomaremos por base a Tabela 2 que identifica a interpretação referente a cada intervalo do valor de $\rho(X_i, X_j)$ (SHIMAKURA, 2006). Uma associação que estiver, por exemplo, entre 0 e 0,19 é interpretada como bem fraca e uma mudança neste patamar poderia implicar numa alteração brusca na interpretação do valor de $V(X_i, X_j)$.O valor nulo implica em não associação entre as variáveis.

Para avaliar o comportamento do coeficiente de correlação após a imputação a partir de redes Bayesianas é feita a suposição de que os relacionamentos entre as variáveis mapeadas pela estrutura $\mathcal{G}$ permanecem inalterados quando obtido $\widetilde{\mathcal{G}}$, e que a variação que

Tabela 2: Interpretação dos valores do coeficiente de correlação ($\rho(X_i, X_j)$) entre duas variáveis que servem de base para a interpretação do valor $V(X_i, X_j)$ de Cramer

Valor de $\|\rho(X_i, X_j)\|$	**Interpretação da correlação**
0,00 a 0,19	Bem fraca
0,20 a 0,39	Fraca
0,40 a 0,69	Moderada
0,70 a 0,89	Forte
0,90 a 1,00	Muito forte

ocorrer de $V(X_i, X_j)$ para o valor de $\widetilde{V}(X_i, X_j)$ não altera a interpretação da associação entre X_i e X_j.

As próximas seções trazem resultados destas avaliações em algumas redes obtidas a partir dos dados do Censo Demográfico Brasileiro e de uma base de casos de homicídios ocorridos no município de Campinas.

3.5 Aplicação aos dados do Censo Demográfico

Como aplicação, foram utilizadas redes Bayesianas para imputação com configurações de redes de três, quatro e cinco nós, nas características do domicílio do questionário básico CD–01 do Censo Demográfico Brasileiro.

O Censo Demográfico no Brasil compreende um grande conjunto de operações de coleta, processamento, análise e disseminação de dados populacionais que ocorrem a cada e durante dez anos. Os Censos Demográficos produzem informações imprescindíveis para a definição e acompanhamento de políticas públicas e tomada de decisões de investimento, sejam elas de caráter público ou privado. Todas as etapas de planejamento, treinamento, coleta e outras relacionadas ao Censo Demográfico Brasileiro podem ser consultadas no documento *Metodologia do Censo Demográfico 2000* (IBGE, 2003), sendo importante destacar aqui apenas alguns pontos.

O Censo Demográfico é a pesquisa responsável pela atualização das estatísticas demográficas oficiais do país. É realizado pelo IBGE em todo o território nacional e é coletado em dois questionários: o questionário básico (CD–01), que possui itens pesquisados para todos os habitantes e o questionário da amostra (CD–02), que é aplicado a amostras de 10% ou 20% dos domicílios de cada um dos municípios brasileiros. O questionário básico contém duas partes: uma com dez perguntas sobre características do domicílio e uma segunda contendo nove perguntas sobre as características dos moradores a ser preenchido para cada morador. Já o questionário da amostra é aplicado a parte dos domicílios e é

Tabela 3: Características de domicílios consideradas no ajuste das redes discretas para imputação

Nome da variável	**Descrição da variável**	**Classes**	**(%) do total**
TIPODOM	Tipo do domicílio	1 – casa	0,964
		2 – apartamento	0,024
		3 – cômodo	0,012
CONDDOM	Condição do domicílio	1 – próprio (já pago)	0,634
		2 – próprio (ainda pagando)	0,106
		3 – alugado	0,160
		4 – cedido por empregador	0,004
		5 – cedido de outra forma	0,066
		6 – outra condição	0,030
CONDTER	Condição do terreno	1 – próprio	0,703
		2 – cedido	0,020
		3 – outra condição	0,017
		Z – zero estrutural	0,259
ABASTEC	Forma de abastecimento de água	1 – rede geral	0,940
		2 – poço ou nascente	0,020
		3 – outra	0,040
TIPOCAN	Como a água chega no domicílio	1 – canalizada	0,853
		2 – canalizada no terreno	0,097
		3 – não canalizada	0,050

mais extenso que o CD–01. Apresenta maior variedade de perguntas, e em seu conteúdo também contém as questões encontradas no questionário básico. No CD–02, a cada observação é associado um peso de expansão da amostra.

Nessa seção avalia–se o uso de redes Bayesianas discretas para imputação conforme descrito na seção anterior. São construídos os indicadores de consistência da base de dados, consistência estrutural, consistência estatística e consistência lógica para redes de domicílios do município de Natal com renda domiciliar (informada ou imputada) nula na base de dados original. Isso totaliza 15.225 de um total de 179.822 domicílios.

As variáveis consideradas para a construção das redes discretas estão listadas na Tabela 3. Essas variáveis foram escolhidas dentre as dez pelo relacionamento existente entre elas e pela existência de zeros estruturais, ou seja, em algum momento uma dada resposta leva o respondente a não responder uma ou mais variáveis do questionário, sem que isso seja caracterizado como não resposta. Todas as variáveis consideradas referem–se a características do domicílio, por exemplo, tipo do domícilio, identificada por TIPODOM, que registra se o mesmo é uma casa, apartamento ou cômodo. Ainda na Tabela 3 pode–se conhecer as classes de cada variável bem como o percentual de cada uma delas do total de 15.225 domicílios.

Tabela 4: Número e percentual de não resposta das variáveis na base de dados original

Nome da variável	**Número de imputados**	**(%) de imputados**
TIPODOM	165	0,011
CONDDOM	222	0,015
CONDTER	209	0,014
ABASTEC	171	0,011
TIPOCAN	392	0,026

A título de informação, na Tabela 4 são descritos os percentuais de não resposta em cada uma destas variáveis na base original. Percebe–se que o percentual de não resposta nestas variáveis é baixo, por exemplo 0,011% de não resposta para tipo de domicílio. Por este motivo, para todas as aplicações construídas foram geradas várias situações de não resposta do tipo MCAR em percentuais variados (1%, 3%, 5%, 7%, 10%, 20%, 30%, 40% e 50%) para cada variável. Foram avaliadas as consistências da base de dados, lógica, estrutural e estatística em cada caso após quinhentos processos de imputação em cada uma delas.

Com os dados do Censo Demográfico avaliou–se a imputação em redes discretas de três, quatro e cinco nós, que foram construídas agregando o conhecimento de especialistas, impedindo que relações de causa e efeito pouco coerentes com a realidade fossem obtidas no ajuste da rede, por exemplo, a condição do domicílio (CONDDOM) ser influenciada pela condição do terreno (CONDTER). Neste caso, a determinante para se responder a condição do terreno no questionário é dada a partir da variável condição do domicílio. Se fossem registradas como resposta algumas das classes (*alugado, cedido por empregador, cedido de outra forma* ou *outra condição*) em CONDDOM, então CONDTER teria obrigatoriamente uma não resposta do tipo zero estrutural por não se aplicar esta pergunta a estes casos. Dessa forma, não seria coerente considerar um arco de CONDTER para CONDDOM.

Funcionaram como especialistas na definição destas redes os técnicos que trabalharam diretamente com a metodologia e análise dos dados do Censo Demográfico 2000, os técnicos que trabalharam na definição do método para imputação da variável de renda e a bibliografia disponível na área demográfica.

3.5.1 Rede com três nós

Inicialmente procedeu–se com a simulação em uma rede composta por três variáveis: tipo do domicílio (TIPODOM), condição do domicílio (CONDDOM) e condição

do terreno (CONDTER). A escolha destas deu–se pelo relacionamento que elas possuem entre si, pela ordenação no questionário e pela ocorrência de zero estrutural na variável CONDTER, onde o zero estrutural a essa questão decorre da resposta das classes (*alugado, cedido por empregador, cedido de outra forma* ou *outra condição*) em CONDDOM.

Foram mantidas as seis categorias da variável CONDDOM para a avaliação da imputação naquelas que contivessem muitas classes. Adiante vê–se que a quantidade de classes de uma variável pode ser um fator que determine a consistência da base (dos microdados) e a consistência estatística após a imputação.

O grafo da rede gerada (ajustada pelo *deal*, (BØTTCHER; DETHLEFSEN, 2003)) com estas três variáveis pode ser visto na Figura 3.6 a seguir.

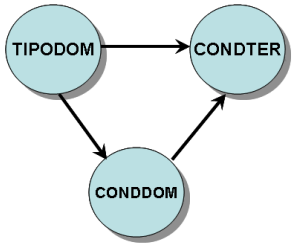


Figura 3.6: Grafo da rede ajustada para três variáveis discretas

Ao identificar os conjuntos de variáveis no grafo temos que TIPODOM pertence ao conjunto P_0 das variáveis sem pais, CONDDOM ao conjunto P_1 tendo como pai a variável em P_0, e em P_2 está CONDTER cujos pais estão no conjunto $P_0 \bigcup P_1$. A sequência de imputação aqui é dada pela distribuição condicional das variáveis de acordo com a relação descrita no grafo. É imputado TIPODOM de acordo com a sua distribuição marginal, depois é imputado CONDDOM de acordo com P(CONDDOM|TIPODOM) e finalmente, CONDTER segundo P(CONDTER|CONDDOM,TIPODOM).

Os resultados das 500 simulações nesta rede em várias perturbações entre as variáveis, nas medidas de consistência da base de dados, lógica e estatística, encontram–se nas Tabelas A.1 a A.4 no Apêndice A. Não foram registrados no Apêndice A os valores da consistÊncia estrutural para esta aplicação pois todas as estruturas $\widetilde{\mathcal{G}}$ construídas após a imputação pertenceram à classe de equivalência da estrutura original.

Na consistência da base de dados é observado o percentual de unidades corretas daquelas que foram imputadas conforme (3.4.5). Quando a imputação é feita em uma variável, tem–se o equivalente ao percentual de observações corretas imputadas naquela variável de acordo com (3.4.4). A Figura 3.7 mostra os resultados das simulações para a consistência da base de dados. O que se observa nos gráficos dessa figura é que, na

imputação em TIPODOM e em CONDTER, que são variáveis que apresentam poucas classes de imputação, o percentual de manutenção dos dados originais é alto, independente de sua posição na rede. Curioso observar que este resultado ocorreu também em CONDTER onde a imputação respeitou a probabilidade condicional observada no grafo da Figura 3.6. Já para a variável CONDDOM, que possui muitas categorias, o percentual de unidades corretamente imputadas não chegou à metade do total de informações perdidas, e este fato pode influenciar negativamente a imputação multivariada como pode ser observado nos gráficos CONDDOM–CONDTER e TIPODOM–CONDDOM–CONDTER na Figura 3.7.

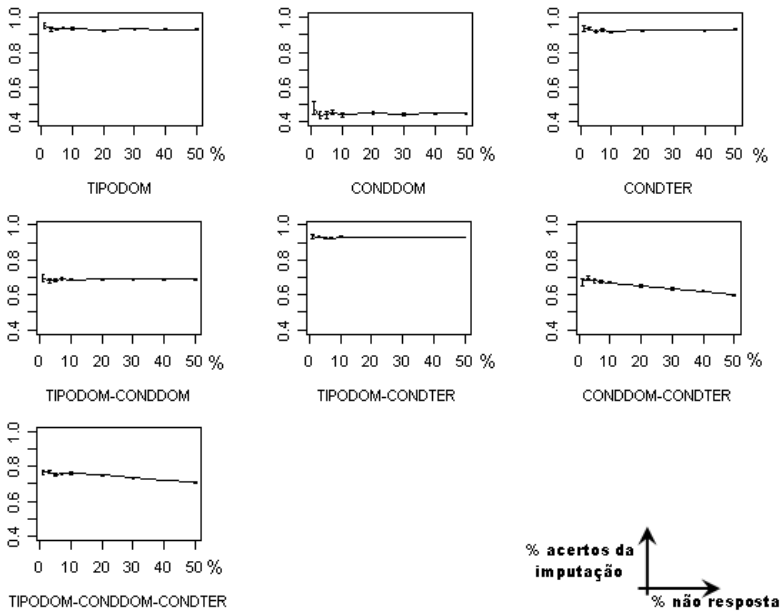


Figura 3.7: Resultado da avaliação da consistência da base de dados (microdados) após 500 imputações a partir de rede com três nós e em diferentes percentuais de não resposta

Uma observação importante a ser feita é de que não é observada uma dependência marcante entre o percentual de não resposta na variável (ou combinação de variáveis) a ser imputada e o seu desempenho na consistência da base de dados.

Todas as 500 imputações conduzidas sob os percentuais de não resposta considerados com respeito à consistência estrutural para a rede de três nós, mantiveram a estrutura da rede imputada na mesma classe de equivalência da rede original. Associa–se este resultado ao fato de ser esta uma estrutura simples no qual sua equivalência é comparada a poucos elementos. Tendo em vista esse resultado, não foi possível registrar os valores de (3.4.6) e (3.4.7).

No que diz respeito à consistência lógica, foram avaliadas as imputações em CONDTER e em CONDTER condicionado à variável CONDDOM. Isso porque as avaliações foram realizadas apenas para a verificação dos zeros estruturais. Quando estas limitações estão identificadas nos parâmetros de entrada da rede, não se observaram itens imputados erroneamente. Já quando não se estabelecem regras de imputação sobre os zeros estruturais, a baixa probabilidade para estas ocorrências na rede Bayesiana de entrada permitiu uma alta proporção de acertos nas unidades imputadas. Quando é imputada apenas CONDTER, o percentual de acertos nos itens imputados como zeros estruturais é de mais de 93,5%. Quando são imputados CONDTER e CONDDOM, o percentual de imputações corretas nestes casos é de mais de 65,0%. Este valor é influenciado pela imputação em CONDDOM, que não é eficiente conforme visto na medida de consistência da base de dados. Estes resultados confirmam a hipótese de Di Zio et al. (2004) de que a rede Bayesiana pode ser utilizada como instrumento de imputação combinado aos processos de edição. Os resultados em todos os percentuais de não resposta simulados podem ser vistos na Tabela A.2 do Apêndice A.

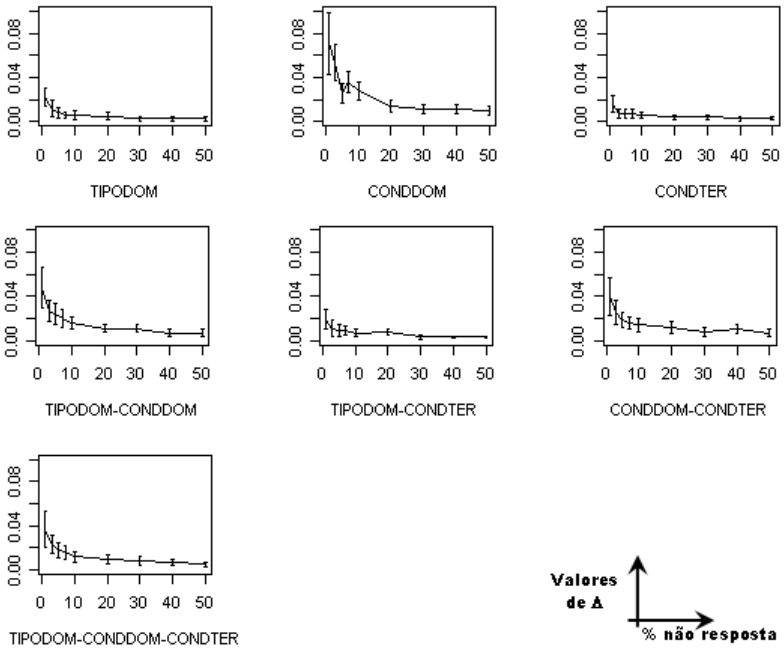


Figura 3.8: Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir de rede com três nós e em diferentes percentuais de não resposta

Partindo para a avaliação da consistência estatística, a Figura 3.8 mostra o comportamento dos valores de Δ para as 500 imputações em diversos percentuais de não resposta. Percebe–se que esta não está influenciada pelos resultados obtidos pela consistência da base de dados e seu desempenho depende, em alguns casos, do percentual de

não resposta na variável ou conjunto de variáveis. Conforme apresentado na seção anterior, no caso da manutenção dos parâmetros da rede, quanto mais o valor de Δ aproximar–se de zero, mais próxima está a igualdade entre as distribuições original e a obtida a partir da base de dados imputada. Todos os valores observados nos gráficos da Figura 3.8 estão próximos de zero, e isso permite concluir que em todas as situações os parâmetros da rede Bayesiana para imputação mantêm–se muito próximos dos parâmetros da rede original. Os resultados para todos os percentuais de não resposta podem ser vistos na Tabela A.3 no Apêndice A.

Na direção da avaliação da medida de associação, ainda na consistência estatística, a Tabela 5 apresenta o padrão de comparação do valor V de Cramer das variáveis duas a duas. A associação entre as variáveis CONDDOM e CONDTER, ou seja, entre a condição do domicílio e a condição do terreno, é considerada forte segundo a Tabela 2 dada na Seção 3.4.4. Já aquelas entre TIPODOM e as demais variáveis são bem fracas. A condução desta avaliação foi realizada apenas nas imputações das variáveis duas a duas a partir da rede Bayesiana original. Imputações em apenas uma variável ou nas três simultaneamente não alteram o comportamento das medidas de associação na Tabela 5.

Tabela 5: Medidas de associação (V de Cramer) entre duas variáveis na rede original com três nós

	TIPODOM	**CONDDOM**	**CONDTER**
TIPODOM	1		
CONDDOM	0,099	1	
CONDTER	0,077	0,708	1

O que se nota após a imputação é que esta é mais afetada negativamente para valores mais altos do valor V de Cramer à medida que se aumenta o percentual de não resposta. Isso leva a crer que o relacionamento não identificado na rede (como é o caso do grau da associação entre as variáveis) pode ser modificado após a imputação a partir de redes Bayesianas em altos percentuais de não resposta.

Na Figura 3.9 observa–se o comportamento do valor V de Cramer calculado após a imputação a partir de diferentes percentuais de não resposta. O que se observa é que a medida de associação mais alta parece apresentar maior fragilidade após a imputação em todos os percentuais de não resposta, mas de forma mais expressiva nas bases com mais de 10% de itens faltantes. Observando os valores de $V(X_{ij}, \widetilde{X}_{ij})$ na Figura 3.9 nota–se que, após a imputação em determinados percentuais de não resposta o método altera a associação entre CONDDOM e CONDTER de forte para moderada. Os resultados destas simulações estão na Tabela A.4 do Apêndice A.

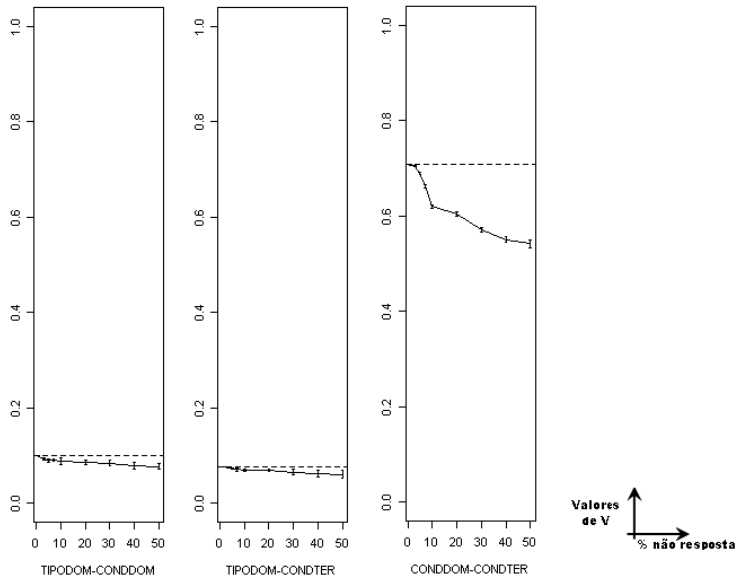


Figura 3.9: Gráficos do valor V de Cramer calculado após 500 imputações a partir de rede com três nós e em diferentes percentuais de não resposta (a linha tracejada refere–se à medida de associação calculada a partir da base original)

A seguir desenvolvem–se os mesmos critérios de avaliação mas para uma rede de quatro nós ajustada a partir dos mesmos dados.

3.5.2 Rede com quatro nós

Para avaliar a rede discreta construída com quatro nós, aproveitamos a rede no item anterior acrescida da variável forma de abastecimento de água, representada por ABASTEC. Os conjuntos de variáveis P_0, P_1 e P_2 permanecem os mesmos da rede ajustada com três nós, mantendo a mesma sequência de imputação da seção anterior. Com a inclusão da variável ABASTEC, define–se o conjunto P_3, cujos pais encontram–se no conjunto $\bigcup_{i=0}^{2} P_i$, e sua imputação segue a regra descrita no grafo da Figura 3.10, que identifica a probabilidade condicional P(ABASTEC|TIPODOM,CONDTER), ou seja, para imputar um item faltante na variável forma de abastecimento de água necessita–se de informação sobre o tipo de domicílio e a condição do terreno.

São obtidos resultados aqui para as consistências da base de dados, estrutural e estatística, uma vez que a consistência lógica foi avaliada na rede anterior e o nó acrescentado nesta rede não apresenta zeros estruturais.

A avaliação do desempenho da imputação a partir da rede Bayesiana após a inclusão da variável ABASTEC assemelha–se ao observado na rede anterior. Os valores das medidas de consistência observados após as simulações a partir da rede com quatro nós podem ser encontrados nas Tabelas A.5 a A.8 do Apêndice A. As medidas de consistência

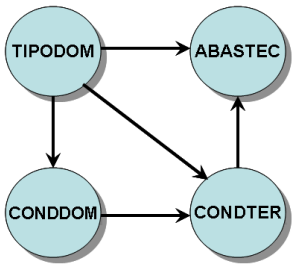


Figura 3.10: Grafo da rede ajustada para quatro variáveis discretas

da base de dados ao se imputar pela rede definida no grafo da Figura 3.10 podem ser observadas graficamente na Figura 3.11. A pior situação na rede de quatro nós é aquela observada na perda de informação entre CONDDOM e ABASTEC, resultando em um percentual de manutenção de cerca de 67% das unidades imputadas, independente do percentual de não resposta apresentado.

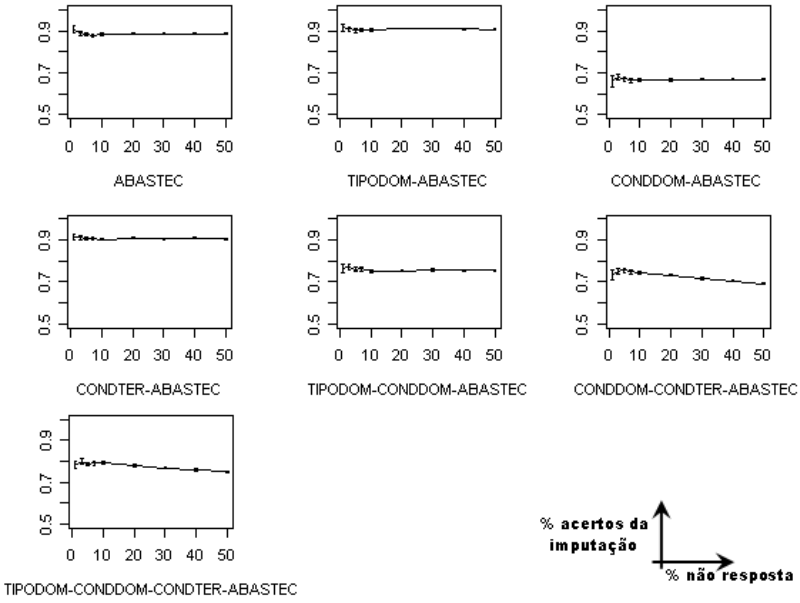


Figura 3.11: Resultado da avaliação da consistência da base de dados (microdados) após 500 imputações a partir de rede com quatro nós e em diferentes percentuais de não resposta

Com relação à consistência estrutural, sendo esta uma estrutura um pouco mais complexa que aquela observada da rede com três nós, a manutenção da estrutura da rede é alta, independente do percentual de não resposta e do número de variáveis imputadas. Um exemplo está na pertinência à classe de equivalência da rede original de todas as estruturas construídas após a imputação, nos percentuais de não resposta de 1% e 3%. Nessa medida de consistência não se observa a dependência entre a consistência estrutural

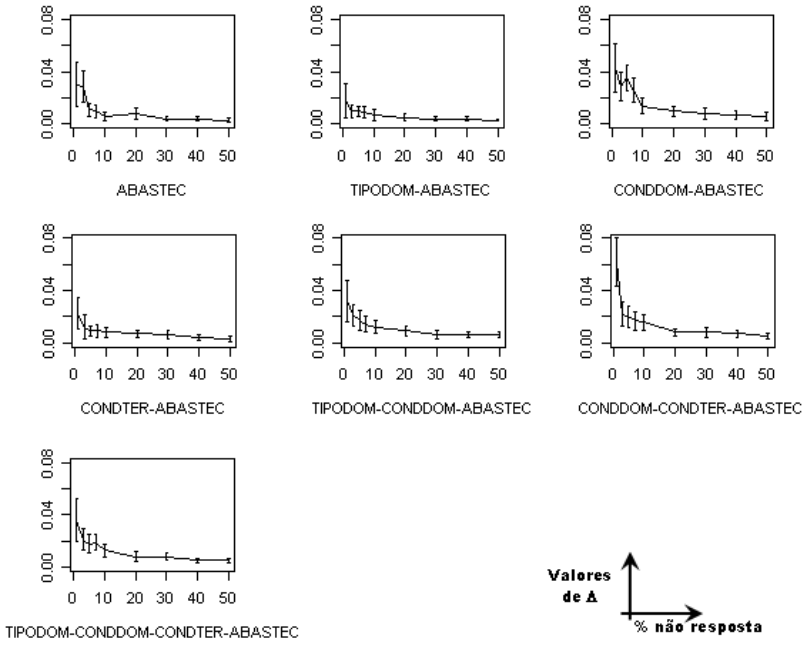


Figura 3.12: Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir de rede com quatro nós e em diferentes percentuais de não resposta

e o percentual de não resposta. Os resultados desta simulação encontram–se na Tabela A.6 do Apêndice A.

Na consistência estatística, o mesmo observado na rede de três nós ocorre para a avaliação de Δ nos valores dos parâmetros da rede Bayesiana. Independente da combinação de variáveis imputadas, o melhor desempenho na manutenção desta consistência é observado para os maiores percentuais de não resposta na variável, embora seja importante ressaltar que nesta situação, todos os resultados mantiveram–se próximos da igualdade entre as duas distribuições (valores de Δ próximos de zero). A Figura 3.12 mostra o comportamento dos valores de Δ após as 500 imputações a partir da rede Bayesiana com quatro nós.

Tabela 6: Medidas de associação (V de Cramer) entre duas variáveis na rede original com quatro nós

	TIPODOM	**CONDDOM**	**CONDTER**	**ABASTEC**
TIPODOM	1			
CONDDOM	0,099	1		
CONDTER	0,077	0,708	1	
ABASTEC	0,052	0,325	0,118	1

Na avaliação da medida de associação na consistência estatística após a imputação, a Tabela 6 apresenta o padrão de comparação para os resultados observados na Figura 3.13. De acordo com a Tabela 6, todas as associações calculadas entre TIPODOM

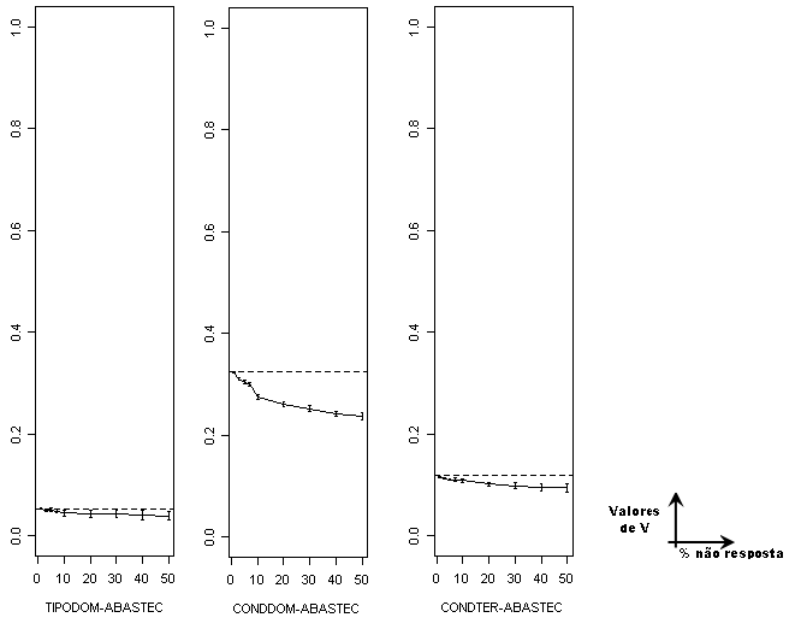


Figura 3.13: Gráficos do valor V de Cramer calculado após 500 imputações a partir de rede com quatro nós e em diferentes percentuais de não resposta (a linha tracejada refere–se à medida de associação calculada a partir da base original)

e CONDTER com a variável ABASTEC, estão caracterizadas como bem fracas, enquanto que o valor V de Cramer entre CONDDOM e ABASTEC é interpretado como associação fraca entre estas duas variáveis. Estas classificações não mudam após a imputação, embora ainda se observe uma leve diminuição nas associações com TIPODOM e CONDTER à medida que se aumenta o percentual de não resposta na variável. Mais uma vez, observa–se na Figura 3.13 que a associação mais afetada pela imputação foi aquela de maior grandeza (ou seja, entre CONDDOM e ABASTEC).

A seção a seguir trata da avaliação das consistências ao utilizar–se uma rede Bayesiana de cinco nós para imputação nos dados do Censo Demográfico.

3.5.3 Rede com cinco nós

Para os estudos conduzidos com a rede de cinco nós considerou–se a rede de quatro nós tratada anteriormente, mas com a inclusão da variável tipo de canalização da água até o domicílio, representada por TIPOCAN. A Figura 3.14 traz o grafo da rede para estas cinco variáveis ajustada pelo *deal* (BØTTCHER; DETHLEFSEN, 2003). A Tabela 3 mostra que a variável TIPOCAN possui características semelhantes às variáveis TIPODOM e ABASTEC, com poucas classes e uma delas com grande percentual de ocorrência nos dados.

Os conjuntos de variáveis P_0, P_1, P_2 e P_3 permanecem os mesmos da rede ajustada

com quatro nós, mantendo portanto a mesma sequência de imputação da seção anterior. Ao se incluir a variável TIPOCAN, define–se o conjunto P_4, cujos pais estão em $\bigcup_{i=0}^{3} P_i$. A imputação em TIPOCAN será dada pela relação P(TIPOCAN|TIPODOM, CONDTER, ABASTEC) identificada pelo grafo na Figura 3.14.

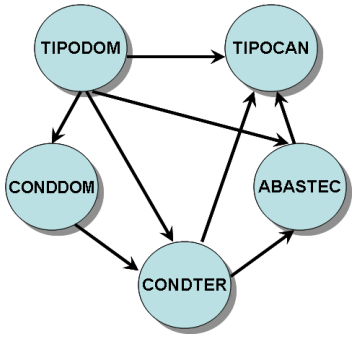


Figura 3.14: Grafo da rede ajustada para cinco variáveis discretas

A avaliação do desempenho da rede Bayesiana após a inclusão da variável TIPOCAN assemelha–se ao observado para as redes de três e quatro nós, em todas as medidas de consistência. A exemplo da rede de quatro nós, a avaliação da consistência lógica foi suprimida por ter sido observada no zero estrutural da variável CONDTER na rede de três nós.

No Apêndice A, as Tabelas A.9 a A.12 descrevem os resultados das simulações conduzidas para as medidas de consistências na rede de cinco nós.

A consistência da base de dados tem seu comportamento apresentado na Figura 3.15. Mais uma vez, tem–se que o percentual de observações corretamente imputadas após 500 imputações simuladas na base de dados original não depende dos percentuais de não resposta estudados (1%, 3%, 5%, 7%, 10%, 20%, 30%, 40% e 50%). A pior situação na rede com cinco nós é aquela na qual TIPOCAN depende de CONDDOM, resultando num percentual de manutenção em cerca de 66,0% das unidades imputadas.

Com uma estrutura um pouco mais complexa do que as outras duas, a avaliação da consistência estrutural apresentou resultados interessantes. Para baixos percentuais de não resposta a estrutura mantém–se preservada independente da combinação de variáveis que será imputada. O mesmo não ocorreu para percentuais mais altos de dados faltantes. Este foi o caso, por exemplo, das imputações conduzidas em percentuais acima dos 10% de não resposta, onde se chegou a até 40% de alteração da estrutura da rede após o seu ajuste com os dados imputados.

Partindo para a avaliação da consistência estatística percebe–se, a partir dos resultados na Figura 3.16, que esta se assemelha ao observado anteriormente, ou seja, os

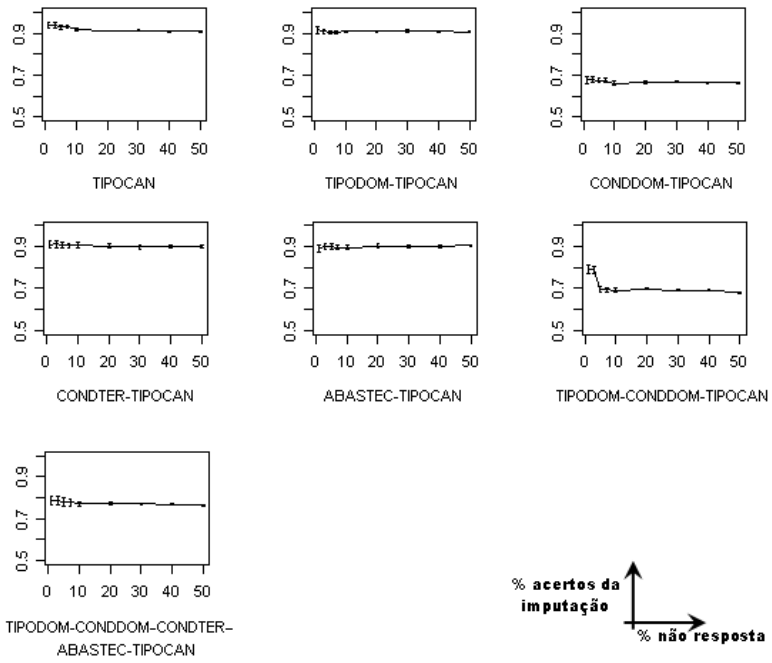


Figura 3.15: Resultado da avaliação da consistência da base de dados (microdados) após 500 imputações a partir de rede com cinco nós e em diferentes percentuais de não resposta

resultados na consistência dos dados e estrutural não parecem influenciar a consistência estatística. Com relação aos valores de Δ, observa–se que estes melhoram à medida que se aumenta o percentual de não resposta simulado. A Figura 3.16 explicita os valores de Δ para os diferentes percentuais de não resposta estudados e em diferentes combinações de variáveis. Além da melhoria na consistência estatística associada aos valores dos parâmetros da rede, observa–se uma diminuição dos correspondentes desvios à medida que se aumenta o percentual de não resposta simulado.

Tabela 7: Medidas de associação (V de Cramer) entre duas variáveis na rede original com cinco nós

	TIPODOM	CONDDOM	CONDTER	ABASTEC	TIPOCAN
TIPODOM	1				
CONDDOM	0,099	1			
CONDTER	0,077	0,708	1		
ABASTEC	0,052	0,325	0,118	1	
TIPOCAN	0,094	0,298	0,124	0,638	1

Ainda na consistência estatística, mas observando agora o valor V de Cramer entre duas variáveis após a imputação, observa–se que, à medida que se aumenta o percentual de não resposta, tende–se a diminuir a associação entre as variáveis consideradas. A Figura 3.17 traz o resultado dos cálculos das medidas de associação após as 500 simulações nos diferentes percentuais de não resposta estudados. Para a comparação, utilizou–se dos

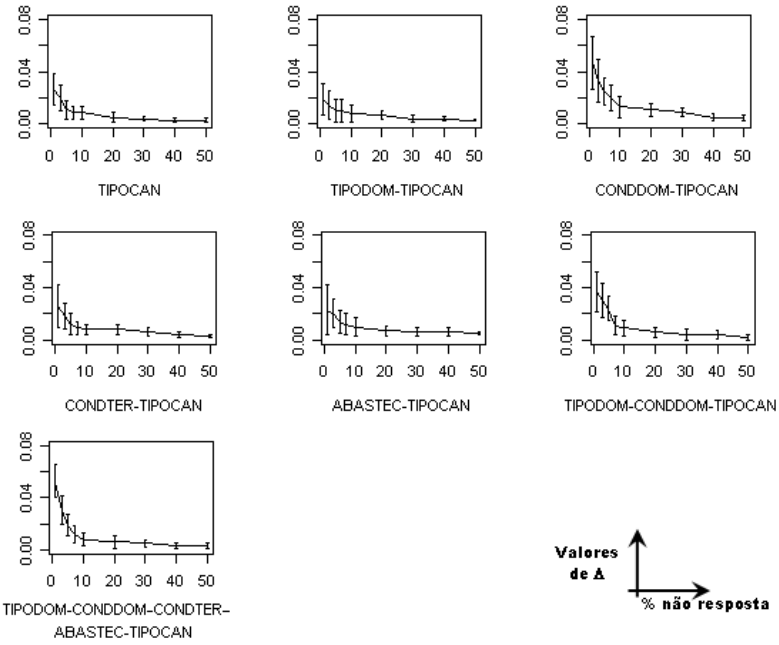


Figura 3.16: Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir de rede com cinco nós e em diferentes percentuais de não resposta

valores calculados a partir da base original que podem ser observados na Tabela 7.

A seção a seguir apresenta a avaliação das mesmas medidas de consistência aplicadas aos dados de homicídios ocorridos no município de Campinas.

3.6 Aplicação aos dados de homicídios em Campinas

Estes dados foram produtos de uma grande pesquisa realizada para investigar metodologias de tratamento do tempo da justiça criminal utilizando o município de Campinas no Estado de São Paulo como estudo de caso (VARGAS, 2000) (VARGAS, 2004) (VARGAS, 2006). A base foi formada a partir da compilação de questionários preenchidos em fóruns do município por coletores treinados. São noventa e três ocorrências registradas e cento e quarenta e oito variáveis.

O objetivo do estudo consistia em determinar atributos que afetassem os tempos da justiça criminal em suas variadas fases (policial e de investigação, de denúncia e interrogatório e de recurso e sentenciamento). Para isso, as técnicas utilizadas na ocasião foram a modelagem das variáveis de tempo decorrido em cada fase, especialmente partindo–se de modelos de regressão e modelos de sobrevivência. Estas análises foram facilitadas pela divisão dos atributos em blocos correspondentes às fases de interesse, podendo cada etapa

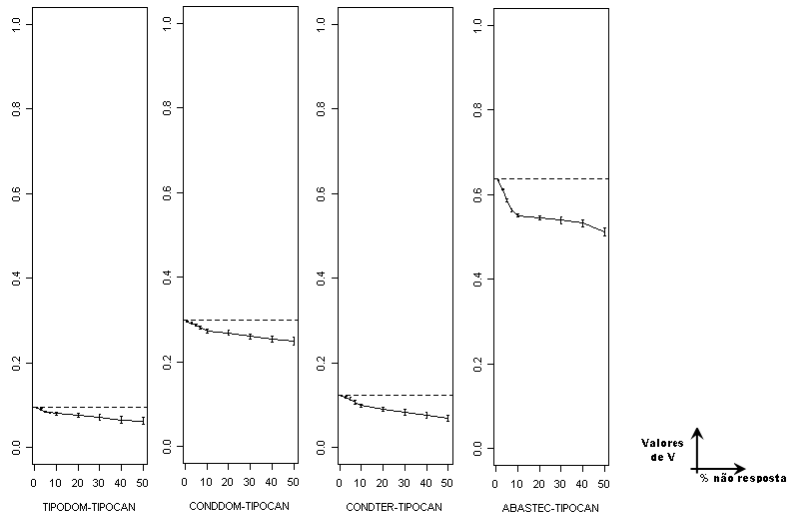


Figura 3.17: Gráficos do valor V de Cramer calculado após 500 imputações a partir de rede com cinco nós e em diferentes percentuais de não resposta (a linha tracejada refere–se à medida de associação calculada a partir da base original)

exercer alguma influência nas posteriores (por exemplo a fase policial, que é anterior à fase de denúncia e interrogatório, pode afetá–la, embora a recíproca não seja verdadeira devido ao curso temporal).

Para esta aplicação foram excluídos os atributos específicos como o nome do réu, as datas, os artigos judiciais e as observações de cada processo. Como o tratamento aqui é dado a redes discretas, as variáveis contínuas foram desconsideradas.

Optou–se por delimitar as variáveis em blocos de acordo com a sua referência (vítima, réu ou processo) e partiu–se da suposição de que os blocos vítima e réu podem definir atributos dos processos, mas a estes atributos não são característicos das variáveis de vítima ou réu. Este pressuposto resguarda a classificação de algum (ou mais de um) dos autores do crime a partir das características dos processos, embora permita que as mesmas sejam obtidas com base nas informações de réu e vítima. Objetiva–se com isso evitar imputação que possa significar algum pré–julgamento (VARGAS, 2004). A Tabela 8 apresenta as variáveis selecionadas para o desenvolvimento desta aplicação. Nesta tabela estão, além das classes possíveis para cada variável, os percentuais de ocorrência em cada uma delas.

Utilizando a informação de influência dos blocos de atributos, tem–se que a imputação para variáveis de processo é realizada a partir de informações da vítima, do réu e do processo. Variáveis da vítima e do réu só podem ser imputadas a partir de atributos de

seus próprios blocos. Na prática, este pressuposto e a estrutura do sistema de justiça atuam diretamente no método de ajuste da rede pois, se for utilizada busca heurística para sua modelagem, a ordenação das variáveis exercerá influência no processo de imputação que contrarie o pressuposto do pré–julgamento.

O relacionamento de independência condicional entre as variáveis nas redes foi definido pelas limitações impostas pelos blocos, pelo conhecimento de especialistas da área de Processo Penal e pelo estudo das correlações entre os atributos[5].

São avaliadas aqui as consistências:

- da base de dados (tendo em vista o interesse em analisar os dados a partir de fluxo de papéis ou pessoas ou então para se utilizar algum método padrão de análise estatística);
- estrutural (para se observar o que ocorre com poucas observações na base de dados);
- estatística (com relação aos parâmetros).

A consistência lógica não será abordada nesta aplicação em razão da não existência de zeros estruturais nas variáveis consideradas.

Serão observadas aqui duas redes formuladas com as variáveis que estão listadas na Tabela 8: a rede **réu–prisão**, com variáveis características do réu, ocorrência e prisão; e a rede **vítima–prisão**, com características da vítima e a variável prisão. Estas duas redes apresentam modelagens distintas, mas possuem em comum a importância na determinação das probabilidades associadas ao nó prisão.

Dentre as variáveis características da ocorrência estão o tipo de crime (identificada por CRIME), o tipo de arma utilizada pelo réu (ARMA) e a relação de conhecimento entre réu e vítima (RELACAO). No bloco de variáveis relacionadas ao réu estão o sexo (SEXOREU), a cor (CORREU), o estado civil (CIVIREU) e o grau de instrução (INSTREU). Para a vítima, registraram–se o sexo (SEXOVIT), estado civil (CIVIVIT) e cor (CORVIT). A variável prisão durante o processo (PRISAO) é uma informação que está relacionada com o réu e a ocorrência, por isso os especialistas a definem como complementar a todos os blocos.

[5]Foram utilizados os conhecimentos compartilhados com pesquisas em estudos de consultoria (VARGAS, 2004) (VARGAS, 2006), experiências literárias relacionadas (VARGAS, 2000) e informações coletadas de um Fórum de Segurança Pública, realizado quando da apresentação dos trabalhos premiados no *Concurso Nacional de Pesquisas Aplicadas em Justiça Criminal e Segurança Pública*, em Brasília entre os dias 02 a 06 de maio de 2005.

Tabela 8: Variáveis de características de réu, vítima e crime para dados de homicídios em Campinas

Nome da variável	**Descrição da variável**	**Classes**	**(%) do total**
CRIME	Tipo de crime	0 – tentativa de homicídio	0,431
		1 – homicídio	0,569
ARMA	Tipo de arma	1 – arma branca	0,301
		2 – arma de fogo	0,688
		3 – outra	0,011
PRISAO	Prisão durante o processo	0 – não	0,398
		1 – sim	0,602
SEXOREU	Sexo do réu	0 – masculino	0,935
		1 – feminino	0,065
CORREU	Cor do réu	1 – branco	0,613
		2 – pardo	0,237
		3 – preto	0,150
CIVIREU	Estado civil do réu	1 – casado	0,301
		2 – solteiro	0,452
		3 – amasiado	0,172
		4 – outro	0,075
INSTREU	Grau de instrução do réu	1 – sem instrução	0,151
		2 – ensino básico	0,785
		3 – ensino médio	0,064
		4 – ensino superior	0,000
SEXOVIT	Sexo da vítima	0 – masculino	0,882
		1 – feminino	0,118
CIVIVIT	Estado civil da vítima	1 – casado	0,237
		2 – solteiro	0,441
		3 – amasiado	0,183
		4 – outro	0,139
CORVIT	Cor da vítima	1 – branca	0,559
		2 – parda	0,323
		3 – preta	0,118
RELACAO	Relação de conhecimento entre réu e vítima	1 – desconhecido	0,118
		2 – conhecido	0,591
		3 – ex/rival	0,129
		4 – negócios/outro	0,162

As simulações conduzidas nesta seção consideram os percentuais de não resposta em 5%, 7%, 10%, 20%, 30%, 40% e 50% e as avaliações dar–se–ão a partir da imputação em subconjuntos de variáveis observados na estrutura da rede conforme proposto na Seção 3.3.3. As redes também são ajustadas nas seções seguintes com a utilização do *deal* (BØTTCHER; DETHLEFSEN, 2003). Todos os resultados das medidas de consistência avaliadas encontram–se nas Tabelas A.13 a A.18 do Apêndice A.

3.6.1 Rede réu–prisão

Na rede réu–prisão, observada pelo grafo da Figura 3.18, modelam–se as relações de causa e efeito entre as variáveis associadas ao réu e as características do crime. Para o ajuste da estrutura e obtenção dos parâmetros seguiu–se a limitação imposta pelos blocos de variáveis. Esse procedimento também evita situações de pré–julgamento. Ainda a compor a informação a priori sobre esta rede, contou–se com a experiência de pesquisadores da área do Código do Processo Penal (CPP). Foram identificados os seguintes conjuntos de variáveis: em P_0, que são as variáveis sem pais na rede (sexo do réu, cor do réu, grau de instrução do réu e tipo de crime), as variáveis em P_1, que são aquelas que possuem pais em P_0 (estado civil do réu e tipo de arma), e em P_2, a variável prisão, que tem pais no conjunto $P_0 \bigcup P_1$. As avaliações seguirão estes blocos de variáveis, diferentemente do que foi desenvolvido na seção anterior com os dados do Censo Demográfico, onde foi avaliada a imputação em variáveis isoladas e em combinações de variáveis.

Pelo número menor de observações e maior número de nós, os resultados serão separados por blocos. Todas as variáveis foram perturbadas no bloco considerado e os resultados foram obtidos com base em 500 imputações feitas em cada percentual de não resposta e para cada situação.

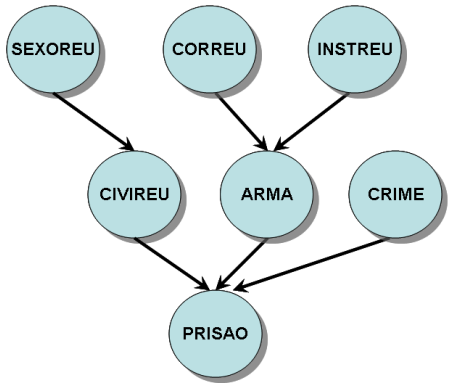


Figura 3.18: Grafo da rede ajustada para variáveis do réu e características do crime nos dados de homicídios em Campinas

No que diz respeito à consistência da base, a Figura 3.19 mostra o comportamento

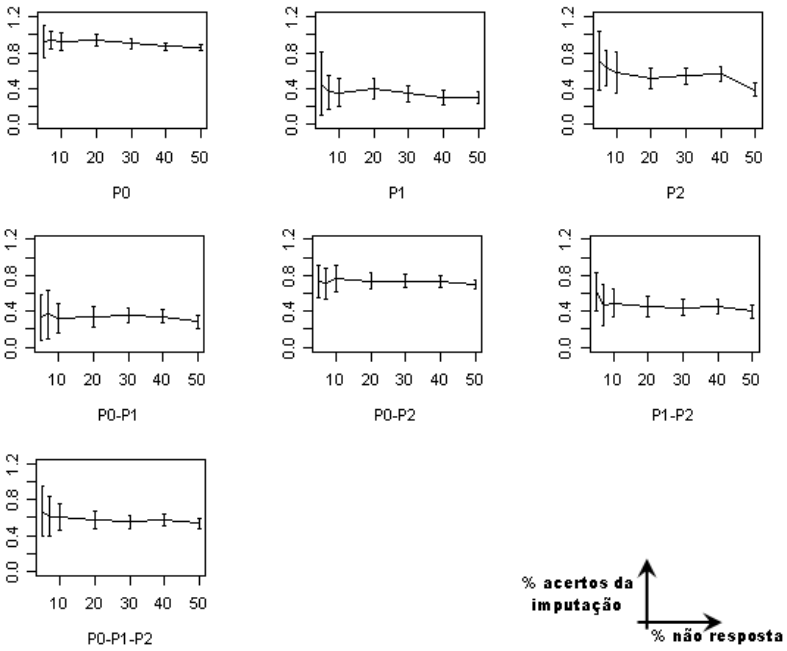


Figura 3.19: Resultado da avaliação da consistência da base de dados (microdados) após 500 imputações a partir da rede réu–prisão e em diferentes percentuais de não resposta

da consistência dos dados das 500 imputações nos blocos de variáveis e para os diversos percentuais de não resposta considerados. A imputação no bloco P_0 de variáveis, feita pela distribuição marginal de cada uma delas, garante a imputação do registro individual na base de dados em mais de 85% dos dados faltantes. Já as imputações feitas nos blocos subsequentes P_1 e P_2 e naquelas realizadas em combinações dos blocos de variáveis, apresentam baixo desempenho no que diz respeito à manutenção do dado individual (com exceção para a combinação P_0–P_2, que manteve cerca de 72,0% dos dados originais). Observando o comportamento dos gráficos na Figura 3.19, não se nota influência do percentual da não resposta nos resultados da consistência na base de dados. A proporção de acertos entre as classes imputadas e suas equivalentes na base original é alta para perturbações em P_0 e em P_0–P_2. As imputações que contêm alterações em P_1, que inclui a variável CIVIREU com muitas classes, apresentam percentual de coincidência entre os registros original e imputado da ordem de 40,0%.

Com relação à consistência estrutural, percebe–se que, quanto menor o percentual de não resposta na variável, maior o percentual de estruturas construídas após a imputação que se mantêm na mesma classe de equivalência da rede original. Uma observação a se destacar aqui refere–se à estrutura dessa rede ser mais complexa que aquelas tratadas nas seções anteriores e ao pequeno número de observações da base de dados em que foi trabalhada essa estrutura. Na não resposta simulada nos blocos P_1 e P_2, por exemplo,

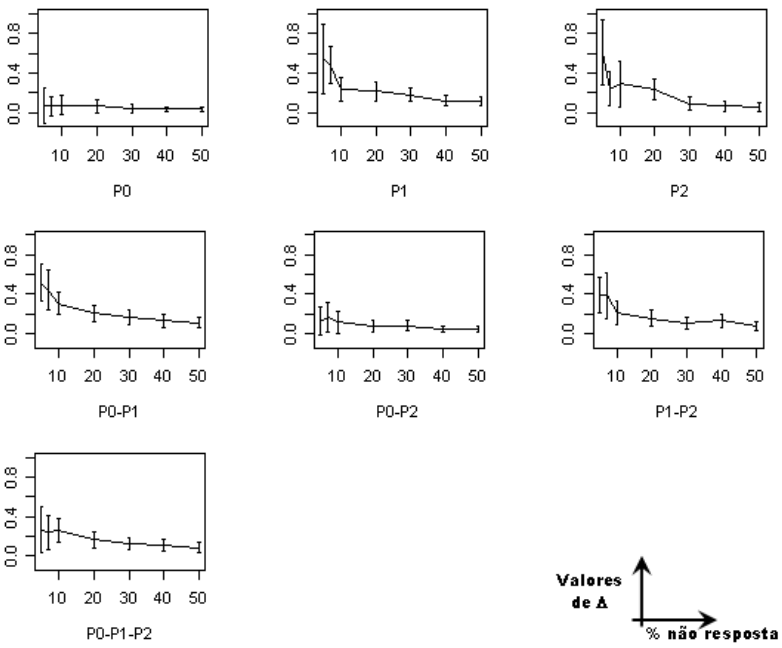


Figura 3.20: Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir da rede réu–prisão e em diferentes percentuais de não resposta

88,8% das estruturas construídas após a imputação fazem parte da classe de equivalência da estrutura original. Este percentual decai para 37,4% quando a não resposta é de 50%.

A consistência estatística sob os parâmetros da rede apresentou conclusões semelhantes às redes ajustadas para o Censo Demográfico Brasileiro. As estimações de Δ, para imputações somente em P_0, estão próximas de zero, independente do percentual de não resposta. Nas demais perturbações, a tendência é de melhoria na consistência estatística, à medida que se aumenta o percentual de não resposta na variável. Isso implica que mais próximos estão os parâmetros da rede construída a partir dos dados imputados daqueles calculados com a rede original. Um exemplo está na avaliação da consistência estatística após perturbação na variável P_2, no qual o valor de Δ para 5% de não resposta é de 0,610 e decresce até atingir 0,054 em 50% de não resposta. A Figura 3.20 ilustra os resultados na avaliação dos valores de Δ nestas simulações, que podem ser visualizados na Tabela A.15 no Apêndice A.

3.6.2 Rede vítima–prisão

O grafo da rede que associa características da vítima e a variável prisão pode ser visualizado na Figura 3.21. Esta é uma rede mais simples que a anterior, e onde se observa que a variável cor da vítima não faz parte das relações de dependência estabelecidas entre

as demais variáveis. Observando a Figura 3.21 identificam–se os subconjuntos P_0, formado pelas variáveis sexo e cor da vítima, P_1, composto por estado civil da vítima e relação de conhecimento entre réu e vítima, e P_2, que traz a variável prisão, cujos pais estão em $P_0 \bigcup P_1$.

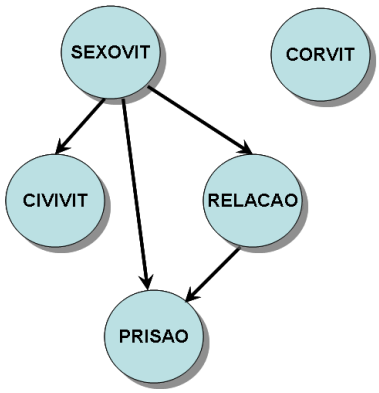


Figura 3.21: Grafo da rede ajustada para variáveis da vítima e prisão nos dados de homicídios em Campinas

Nota–se que o pressuposto para se evitar o pré–julgamento é atendido, não sendo observados arcos direcionados no grafo que partam das variáveis características do crime ou prisão para as características da vítima.

Mais uma vez, as avaliações seguirão as delimitações formadas por P_0, P_1 e P_2 para as consistências da base de dados, estrutural e estatística nos valores de Δ.

Partindo–se da consistência na base de dados, o que se observa é que, a exemplo do que ocorreu nas demais redes, o percentual de registros corretamente imputados não depende em geral, do percentual de não resposta, mas sim do número de classes na variável (ou variáveis) considerada, exceto em P_2. As simulações para a consistência da base de dados resultantes das 500 imputações na rede vítima–prisão podem ser conferidas na Figura 3.22 a seguir. As imputações conduzidas sob P_1, que contém variáveis que apresentam várias classes, apresentam o pior desempenho na manutenção dos microdados quando comparadas com as demais. Em um percentual simulado de 50% de não resposta atingiu–se uma proporção de 0,285 de manutenção dos dados individuais.

Avaliando a grandeza que define a consistência estrutural, percebe–se que, à medida que se aumenta o percentual de não resposta, decresce a proporção de estruturas $\widetilde{\mathcal{G}}$ pertencentes à classe de equivalência da estrutura $\mathcal{G}$ original.

Na consistência estatística, observa–se o análogo obtido nas demais redes: os melhores valores de Δ (os mais próximos de zero) são obtidos nos maiores percentuais de não resposta simulada. As variáveis em P_1 e P_2 apresentam piores desempenhos em Δ quando

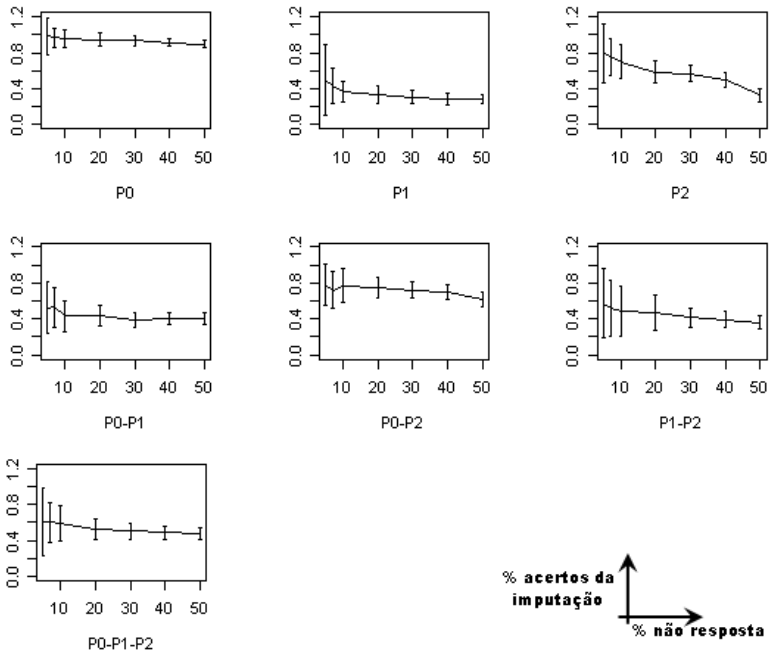


Figura 3.22: Resultado da avaliação da consistência da base de dados (microdados) após 500 imputações a partir da rede vítima–prisão e em diferentes percentuais de não resposta

comparados aos demais. A Figura 3.23 mostra o resultado da avaliação da consistência estatística para a rede vítima–prisão.

3.7 Conclusões e futuros direcionamentos

Neste capítulo foi proposto um novo algoritmo para o uso de redes Bayesianas para imputação em variáveis aleatórias discretas que tem como finalidade agregar o conhecimento de especialistas sobre as variáveis e seus relacionamentos de independência condicional na construção da rede. Essa informação evita uma ordenação prévia dos atributos que implique em uma alteração na distribuição multivariada das variáveis.

Na avaliação, foram conduzidas três medidas de consistência propostas por Di Zio et al. (2004): consistência dos dados, lógica e estatística. Além disso, propôs–se neste texto a consistência estrutural, com o objetivo de auxiliar na verificação da hipótese de que a imputação conduzida sob redes Bayesianas preservaria o relacionamento multivariado entre as variáveis da rede. Considera–se fundamental que outros aspectos sejam observados ao se examinar a consistência estrutural, como os parâmetros da rede e a manutenção dos registros individuais. Por isso, a consistência estatística foi estendida para além dos parâmetros da rede com o cálculo de uma medida de associação entre as variáveis. As

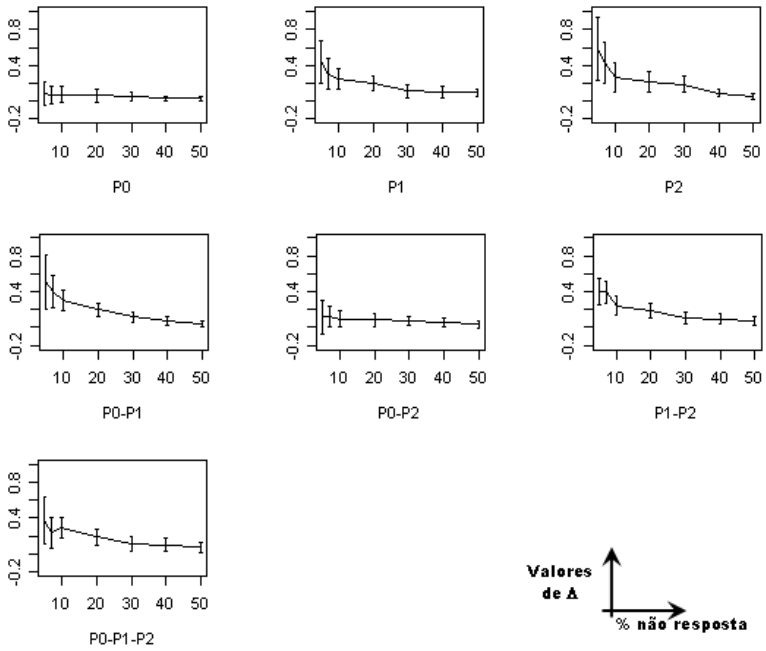


Figura 3.23: Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir da rede vítima–prisão e em diferentes percentuais de não resposta

redes estudadas com os dados do Censo Demográfico foram pequenas (em no máximo cinco nós) para se equipararem ao estudo realizado por Di Zio et al. (2004), embora o objetivo neste trabalho não fosse a comparação em si. Com os dados de homicídios em Campinas, não era possível a construção de redes maiores devido a limitação no número de observações disponíveis.

Em suma, os resultados obtidos com as redes estudadas neste texto serviram para a percepção de aspectos conjuntos que se buscam ao se definir um método de imputação e para balizar futuros direcionamentos no tema. Por ser esta uma aplicação relativamente recente das redes Bayesianas, ainda muito existe para ser estudado e as avaliações que se seguem abaixo indicam que o estudo e aplicação deste método para imputação mostram–se promissores:

- a preservação da estrutura, em geral, não depende do percentual de não resposta quando o número de observações da base de dados é relativamente alto;

  Conforme em Neapolitan (2004), a classe de equivalência de uma dada estrutura pode ser obtida para $n \to \infty$ em determinadas classes de modelos. As simulações conduzidas com redes discretas consideraram estruturas simples, com poucas variáveis e nas situações de grande número de unidades na base (o caso do Censo Demográfico) e um número mais restrito (o caso de homicídios em Campinas). Para

a base de dados maior, mais de 90% das redes Bayesianas aprendidas sob os dados imputados em uma rede com quatro nós e sob 50% de não resposta, por exemplo, pertenciam à classe de equivalência da rede original. Um fator que exerce influência na medida de consistência estrutural é o número de nós, o que implica também na complexidade da rede. No caso dos dados do Censo Demográfico na rede construída com cinco nós, em altos percentuais de não resposta a estrutura construída após a imputação encontrou–se na mesma classe de equivalência da estrutura original em uma ordem de 60% das vezes. Já a estrutura da rede Bayesiana mostrou–se sensível quando a imputação é conduzida sob um percentual alto de não resposta em base de dados pequenas.

- com relação aos parâmetros da rede, estes se mostram mais próximos dos parâmetros originais à medida que se aumenta o percentual de não resposta na base, independentemente da estrutura;

  Este resultado entra em contradição com o anterior para bases de dados pequenas. Nos dados do Censo Demográfico, em todos os percentuais de não resposta simulados, os valores de Δ mostraram–se próximos de zero, ou seja, os parâmetros da rede ajustada após a imputação não sofreram alterações bruscas, salvo para aquelas variáveis com muitas classes, que apresentaram desempenho menor para esta consistência. Já nos dados de homicídios de Campinas, aos maiores percentuais de não resposta associam–se os melhores desempenhos dos valores de Δ mas, em contrapartida, é onde se observa o menor percentual de redes na classe de equivalência da rede original. Este fato leva a crer que, para bases pequenas, a imputação por redes Bayesianas induz a manter o relacionamento entre as variáveis representado pelo grafo a partir de uma compensação nos parâmetros.

- altos valores de medidas de associação entre as variáveis são afetadas pelos maiores percentuais de não resposta após a imputação via rede Bayesiana;

  Nas situações avaliadas, os baixos valores calculados da medida V de Cramer não sofreram alterações após a imputação. Este foi o caso, por exemplo, da associação entre TIPODOM e TIPOCAN na rede com cinco nós. Já as medidas de associação mais altas (em comparação com as demais calculadas na mesma rede), são mais sensíveis ao aumento do percentual de não resposta.

  Possivelmente, esta mudança no nível da medida de associação pode estar relacionada com a mudança na relação observada entre as variáveis que é explicitada na estrutura da rede. As relações de dependência entre os nós não são necessariamente

as mesmas identificadas pela medida de associação V de Cramer calculada entre as variáveis. Neste caso, a imputação a partir de redes Bayesianas como método tende a diminuir a associação existente entre as variáveis à medida que se aumenta o percentual de não resposta.

- a consistência da base de dados é afetada pelo número de categorias das variáveis e independe do percentual de não resposta e da posição da variável na rede;

  Nas redes consideradas neste texto, o desempenho da preservação dos dados não foi afetada pelo percentual de não resposta, o que ocorreu em função do número de classes das variáveis. Segundo Chambers (2000), esta é a propriedade mais difícil de se obter, mas que é de fundamental importância para analistas que trabalham com mineração de dados. Se o interesse do pesquisador está em maximizar a coincidência entre o valor real e o imputado, então deve–se investir em um método de imputação que mantenha esta característica. Como a rede Bayesiana apresentou o mesmo desempenho para resgatar o registro em todos os percentuais de não resposta e posição da variável na rede, deve–se estabelecer algum critério (possivelmente algum tipo de algoritmo específico) que proporcione desempenho máximo para classificação. Outros métodos de imputação não são usados para preservar os microdados e sim, para manter os parâmetros para alguma análise de interesse. Possivelmente, a distribuição das proporções entre as classes afeta o desempenho da medida de consistência da base de dados, embora não foram conduzidos ensaios que verifiquem esta afirmação.

Estes resultados foram algumas observações que se tornaram recorrentes para as bases de dados, estruturas e parâmetros diferentes, portanto percebeu–se a necessidade de registrá–los. Outros aspectos, como o decaimento não linear do coeficiente de correlação em função do percentual de não resposta ou a diminuição do desvio das estimativas também em função do percentual de não resposta para os valores de Δ e do percentual de acertos nos registros imputados, são itens específicos e dependentes do modelo real e suposto aos dados. Estes aspectos carecem de maiores estudos e apresentam–se como futuros direcionamentos deste trabalho.

4 *Imputação de dados quantitativos a partir de redes Bayesianas mistas*

A realidade de muitas bases de dados contém variáveis discretas e contínuas em seus domínios e os atuais algoritmos que realizam a imputação a partir de redes Bayesianas consideram apenas dados discretos. Apesar de as variáveis contínuas poderem ser discretizadas, este procedimento limita a sua aplicação pela perda da informação original, que pode ser de interesse para o usuário.

Este capítulo trata de redes Bayesianas mistas para a imputação em variáveis discretas e contínuas, onde é apresentado um algoritmo para a sua aplicação. Como uma extensão do caso discreto, figuram as quatro medidas de consistência traçadas no capítulo anterior[1].

4.1 Introdução

As aplicações de redes Bayesianas para imputação são bem recentes, datam de poucos anos apenas e destinam–se somente a dados discretos. Apesar de ser justificável em alguns casos que as variáveis contínuas possam ser discretizadas no emprego de redes Bayesianas discretas para imputação, é bem sabido que existem limitações neste procedimento pela perda das informações originais. O objetivo deste trabalho é apresentar uma nova forma de tratamento da não resposta em dados quantitativos, esta utilizando as redes Bayesianas, que aparecem na literatura como bastante promissoras por manterem características da distribuição e do relacionamento entre as variáveis após a imputação

[1]Parte deste texto será submetido à publicação ao *The Imputation Bulletin* do *Statistics Canada*, o instituto oficial de estatísticas do Canadá. Dessa maneira, algumas partes podem se encontrar repetidas.

dos dados discretos.

Neste texto, expandem–se as idéias descritas no capítulo anterior para o caso em que a rede Bayesiana é aprendida a partir de uma base de dados contendo variáveis discretas e contínuas. As principais motivações para propor redes Bayesianas mistas para imputação estão em: (a) aplicabilidade a um maior número de problemas reais, tendo em vista que em muitas situações, além de variáveis categorizadas, registram–se variáveis como idade, rendimentos, peso, altura, níveis de colesterol, etc.; (b) manutenção da informação em sua unidade original para o caso da variável contínua (no caso da discretização da variável aleatória contínua, este procedimento torna–se limitado, por tornar a análise mais restritiva); (c) analogia que este método apresenta com a formação de classes de imputação nos métodos hot–deck e árvores de regressão (BREIMAN et al., 1984) para imputação em variáveis contínuas e (d) manutenção da distribuição multivariada associada aos dados.

Para isso, são utilizadas redes mistas conforme descrito em Bøttcher e Dethlefsen (2003) e propõe–se um algoritmo para executar esta tarefa. São propostas também, a exemplo do desenvolvido para o caso discreto, quatro medidas para avaliar a consistência após a imputação, a saber: consistência da base de dados, consistência estrutural, consistência lógica e consistência estatística.

4.2 Redes Bayesianas mistas

Geiger e Heckerman (1994) desenvolvem o aprendizado em redes Bayesianas Gaussianas considerando apenas variáveis contínuas e Murphy (1998) descreve inferência e aprendizado em redes Bayesianas híbridas. Bøttcher e Dethlefsen (2003) implementam o *deal*, um pacote em R para ajuste de redes Bayesianas mistas no qual suas partes discretas e contínuas equivalem respectivamente aos trabalhos de Heckerman et al. (1995) e Geiger e Heckerman (1994).

Esta seção descreve a teoria relacionada a redes Bayesianas mistas, que são propostas posteriormente para imputação em variáveis discretas e contínuas de uma mesma base de dados.

Nas redes Bayesianas discretas, faz–se a suposição de distribuição de Dirichlet como priori conjugada ao se ajustar uma estrutura no *deal* (BØTTCHER; DETHLEFSEN, 2003). No caso de existirem nós discretos e contínuos na base de dados, o ajuste deve ser conduzido conforme a suposição de que a distribuição conjunta sob todos os nós é normal

condicional (ou condicional Gaussiana) em cada combinação i das classes de nós discretos que são seus pais.

Considere um conjunto de variáveis aleatórias $\mathbf{X}= \{X_1, \ldots, X_k\}$, que contenha d variáveis discretas e c variáveis contínuas. Dessa forma, $\mathbf{X}$ está particionada em $(\mathbf{X}_D, \mathbf{X}_C)$, onde D é o conjunto das discretas e C o das contínuas. Associados às variáveis aleatórias estão os vértices de um grafo direcionado acíclico (GDA) $\mathcal{G} = (V, A)$, onde $V = (V_d, V_c)$, V_d e V_c representando respectivamente os nós discretos e contínuos no grafo e A é o seu conjunto de arestas. A distribuição de probabilidade conjunta $P(\mathbf{X})$, que compõe a rede Bayesiana delimitada por $\mathcal{G}$, fatora da forma (BØTTCHER, 2004):

$$P(\mathbf{X}) = P(\mathbf{X}_D, \mathbf{X}_C) = \prod_{i \in D} P(X_i | pa_D(X_i)) \prod_{j \in C} P(X_j | pa_D(X_j), pa_C(X_j)), \tag{4.2.1}$$

onde $pa_D(X_j)$ e $pa_C(X_j)$ são respectivamente os conjuntos de variáveis discretas e contínuas que são pais de X_j. Não será considerado o caso em que variáveis discretas apresentem pais do tipo contínuo, e esta limitação assegura a existência de métodos para cálculos exatos na obtenção das soluções. Para maiores detalhes a esse respeito consultar Lauritzen (1992) e Lauritzen e Jensen (2001).

Uma vez fatorada a distribuição conjunta em (4.2.1), a parte discreta obtém–se como no capítulo anterior, em geral supondo priores de Dirichlet como conjugadas à distribuição multinomial nesta parcela. Já à parte mista, supõe–se que as distribuições de probabilidade locais são regressões lineares normais (ou equivalentemente, lineares Gaussianas), com parâmetros dependentes das configurações dos pais discretos. Isso significa que, para uma variável X_j, $j \in C$, tem–se:

$$(X_j | pa_D(X_j), pa_C(X_j)) \sim N(\mu_j, \sigma^2_{X_j | pa_D(X_j)}), \tag{4.2.2}$$

onde $\mu_j = \beta_{0,X_j|pa_D(X_j)} + \beta_{i,Xj|pa_D(X_j)} x_{i,pa_C(X_j)}$; $\beta_{0,X_j|pa_D(X_j)}$ é a média da variável X_j na configuração de seus pais discretos, ou em outras palavras, o intercepto da regressão linear; $\beta_{i,X_j|pa_D(X_j)}$ são os coeficientes da regressão, que definem a contribuição de cada pai contínuo em cada nível da variável discreta e $\sigma^2_{X_j|pa_D(X_j)}$ é a variância condicional de X_j em seus pais discretos.

Murphy (1998) caracteriza os coeficientes $\beta_{i,X_j|pa_D(X_j)}$ como pesos para as arestas que chegam ao nó X_j a partir de seus pais. Em (4.2.2) existem três situações possíveis do relacionamento da variável aleatória contínua com seus pais, que podem diferenciar a estimação dos parâmetros na parte mista da rede:

1. o nó X_j não tem nenhuma variável do tipo discreto como pai, ou seja, não existe alteração na média de X_j decorrente da mudança de categorias:

$$\begin{aligned}(X_j|pa_D(X_j), pa_C(X_j)) &= (X_j|pa_C(X_j)) \\ &\sim N(\mu_j = \beta_{0,X_j} + \beta_{i,X_j} x_{i,pa_C(X_j)}, \sigma^2_{X_j}); \end{aligned} \quad (4.2.3)$$

2. o nó X_j não tem nenhuma variável do tipo contínuo como pai, ou seja, não existe um acréscimo quantitativo na média de X_j calculada em cada configuração de seus pais discretos:

$$\begin{aligned}(X_j|pa_D(X_j), pa_C(X_j)) &= (X_j|pa_D(X_j)) \\ &\sim N(\mu_j = \beta_{0,X_j|pa_D(X_j)}, \sigma^2_{X_j|pa_D(X_j)}); \end{aligned} \quad (4.2.4)$$

3. o nó X_j não tem pais:

$$(X_j|pa_D(X_j), pa_C(X_j)) = X_j \sim N(\mu_j = \beta_{0,X_j}, \sigma^2_{X_j}). \quad (4.2.5)$$

Neste trabalho não serão considerados os nós contínuos tendo como pais apenas variáveis discretas. Isso porque, para o enfoque da imputação, existem outros métodos já estabelecidos na literatura que se tornam equivalentes (imputação pela média geral, imputação por regressão preditiva ou por regressão aleatória, todos estes citados no Capítulo 2 deste trabalho; justifica–se esta equivalência na seção seguinte).

Desse modo, a distribuição de probabilidade conjunta em $\mathbf{X}$ sob a estrutura $\mathcal{G}$ é definida como:

$$\begin{aligned}P(\mathbf{X}|\mathcal{G}) &= P(\mathbf{X}_D, \mathbf{X}_C|\mathcal{G}) = \\ &= P(\mathbf{X}_D)|2\pi\Sigma_D|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{X}_C - \mu_d)'\Sigma_D^{-1}(\mathbf{X}_C - \mu_d)\right\}, \end{aligned} \quad (4.2.6)$$

onde, para cada $d \in D$, μ_d representa a média incondicional nas variáveis contínuas e Σ_D é a matriz de covariâncias para todas as variáveis contínuas contidas na estrutura em $\mathcal{G}$.

A estimação dos parâmetros em (4.2.6) supondo $\mathcal{G}$ conhecida pode ser feita por um especialista ou através da especificação de distribuições a priori que posteriormente são atualizadas a partir dos dados de acordo com o teorema de Bayes (COX; HINKLEY, 1974). Para essa estimação, são consideradas algumas propriedades que foram introduzidas por Spiegelhalter e Lauritzen (1990), Geiger e Heckerman (1994), Bøttcher (2004) e Bøttcher (2005):

– a *independência global dos parâmetros* é a suposição de que os parâmetros associados a uma variável são independentes dos parâmetros referentes às outras variáveis;

– a *independência local* especifica que os parâmetros são independentes para cada configuração dos pais discretos;

– propriedades de *conjugação local e global*, decorrentes das duas propriedades anteriores, são as propriedades de que a distribuição dos parâmetros pertençam à mesma família de distribuições conjugadas, tanto local quanto globalmente;

– a *modularidade dos parâmetros* relaciona–se ao relacionamento entre as distribuições a priori dos parâmetros para diferentes estruturas de redes Bayesianas;

– a propriedade de *equivalência de eventos* diz que duas estruturas de redes Bayesianas que representam o mesmo conjunto de independências e dependências devem corresponder aos mesmos eventos, e portanto receber o mesmo escore;

– a *independência dos parâmetros a posteriori* diz que os parâmetros permanecem independentes condicionados aos dados.

Como o *deal* (BØTTCHER; DETHLEFSEN, 2003) tem implementado priores de distribuições conjugadas e considera estas propriedades, então as posteriores são calculadas de forma direta. A parte mista da rede, definida sob seus nós contínuos, tem como família de conjugadas padrão as distribuições Gama–inversas Gaussianas e encontram–se em Bøttcher (2005) e Bøttcher e Dethlefsen (2003) a descrição dos estimadores de todos os parâmetros necessários para calcular as priores e as posteriores no ajuste da rede.

Uma medida resultante do ajuste da rede mista está na função escore, que se refere à probabilidade de um GDA $\mathcal{G}$ representar as independências condicionais entre as variáveis aleatórias, medindo o quão provável é o grafo condicionado a uma base de dados observados. Como $P(\mathbf{X})$ fatora da forma dada em (4.2.1), a função escore calculada no *deal* (BØTTCHER; DETHLEFSEN, 2003), supondo que se conhece a estrutura $\mathcal{G}$ da rede mista de interesse, também é calculada a partir da fatoração em uma parte discreta e uma contribuição da parte contínua condicionada na sua parte discreta. Bøttcher (2004) obtém essa função e é esta que está implementada no *deal*. Esta será utilizada neste trabalho para o ajuste da rede original e na obtenção das redes após a imputação.

4.3 Imputação a partir de redes Bayesianas mistas

Nesta seção é proposto o uso da rede Bayesiana mista para imputação em variáveis discretas e contínuas, de acordo com a distribuição conjunta especificada pela rede $\mathcal{B}$ que é delimitada pelo grafo $\mathcal{G}$. O que se pretente é estender o algoritmo definido no Capítulo 3 para redes mistas. A forma com que a rede é aprendida no algoritmo proposto nesta seção evita que sejam construídos relacionamentos de independência condicional que possam não estar coerentes com a realidade das variáveis. Além disso, agrega–se o conhecimento de especialistas na formação da estrutura da rede a ser modelada para imputação. Esta informação tem a possibilidade de reunir elementos que poderiam ser desconsiderados caso fosse conduzida uma ordenação dos atributos a partir de um critério menos flexível (como aquele em Di Zio et al. (2004)).

Algumas observações são necessárias antes do prosseguimento do texto. Em primeiro lugar, sabe–se que, em geral, os métodos de imputação fazem uso de diversas informações disponíveis, mas se não há um relacionamento entre tais informações e os itens a serem imputados, ou se o método de imputação não for o mais adequado à situação em questão, então este pode significar uma fonte a mais de vício aos dados. Esta observação é importante pelo fato de ser a rede Bayesiana uma representação gráfica de uma distribuição conjunta de probabilidade, que muitas vezes mapeia as relações de causa–e–efeito entre variáveis a partir da experiência de pessoas da área em questão. Nem sempre se tem certeza de que as suposições feitas por estes sejam válidas. Se a rede não for definida de forma a preservar o relacionamento entre as variáveis, a imputação a partir dela pode surtir um efeito não desejado.

Uma segunda observação está no mecanismo de não resposta associado a cada variável. Segundo Coppola et al. (2002), as imputações são em geral conduzidas sob a suposição de que os mecanismos de não resposta são do tipo MAR (ou *Missing at Random*, conforme descrito no Capítulo 2). Thibaudeau e Winkler (2002) registram que, ao utilizar as redes Bayesianas como método para imputação, as restrições dadas pela estrutura fazem com que a imputação seja sempre aplicada para o mecanismo do tipo não ignorável. Esta afirmação pode ser confirmada ao se observar, por exemplo, X_1 e X_2, onde a imputação é conduzida segundo a probabilidade condicional $P(X_2|X_1)$, sendo X_1 a variável idade e X_2 o estado civil. Se ($X_1 < 10$ anos), e não existe uma probabilidade positiva de que ocorra o evento ($X_2 =$ *casado* $|X_1 < 10$ anos), então imputações do tipo *casado* não serão efetuadas no grupo *menor de 10 anos de idade*. Di Zio et al. (2004) apontam este fato como sendo uma vantagem do método por facilmente ser possível de se combinar regras

de crítica e edição à estrutura da rede Bayesiana ajustada aos dados.

Além das observações feitas na seção anterior, um ponto a mais refere–se à ordenação das variáveis. Mais adiante explica–se o porquê de a ordenação sugerida ter importância para o método proposto.

Em se tratando de ajuste de redes Bayesianas, existem duas situações possíveis: (1) a rede já se encontra disponível de situações anteriores ou do conhecimento de especialistas; (2) é necessário construir a rede com base em um conjunto de dados que contém a não resposta.

Para se construir uma rede Bayesiana, conta–se com uma diversa gama de algoritmos que aprendem estrutura e/ou parâmetros sob diversos aspectos. Se o objetivo está em construir uma rede a partir de dados que contenham itens faltantes, então um conjunto de métodos que se utilizam de *Markov Chain Monte Carlo* (MCMC) (TANNER, 1993) ou aproximação para grandes amostras podem ser aplicados (ver, por exemplo, (NEAPOLITAN, 2004) para uma detalhada descrição dos mesmos). Além destes, pode–se aplicar o algoritmo *Structural EM* (SEM) (FRIEDMAN, 1998) para se aprender a estrutura da rede nestes casos. Aqui a estimação será dada conforme em Bøttcher e Dethlefsen (2003) a partir do *deal*.

Uma vez obtida a estrutura e seus parâmetros associados, a rede passa a funcionar como instrumento para preencher os espaços vazios da base de dados. Partindo deste raciocínio, tem–se então um conjunto de parâmetros de entrada que descrevem o relacionamento existente entre as variáveis. Neste trabalho, parte–se do pressuposto de que a rede Bayesiana é conhecida antes de iniciar o processo de imputação, ou seja, considera–se que a rede deve estar definida para a sua aplicação como método de imputação.

Di Zio et al. (2004) e Di Zio, Scanu e Vicard (2003), quando propõem seus algoritmos para imputação em variáveis categorizadas, utilizam–se dos dados incompletos no ajuste da estrutura e obtenção dos parâmetros. Para isso, têm como instrumento o pacote comercial *Hugin Tools*[2] (MADSEN et al., 2003), que traz implementado o algoritmo PC (SPIRTES; GLYMOUR; SCHEINES, 1993) permitindo empregar a não resposta a partir do algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977). Em uma etapa preliminar deve–se ordenar as variáveis de acordo com a sua confiabilidade[3]. Esta ordenação deve ser respeitada na definição dos arcos orientados no grafo $\mathcal{G}$ para que as variáveis de menor

[2]Existe uma versão com capacidade e desempenho limitadas, disponível gratuitamente para estudantes para uso em teste por um período de trinta dias. Informações podem ser obtidas em www.hugin.com.

[3]O autor sugere como medida de confiabilidade o percentual de valores faltantes, mas cita que maiores estudos são necessários neste sentido.

confiabilidade sejam imputadas condicionadas àquelas de maior confiabilidade na base de dados. Apesar de garantir que a imputação em variáveis dependa em sua maior parte de outras com maior percentual de resposta, essa maneira de construir a rede Bayesiana pode introduzir vícios para o cálculo de quantidades básicas a partir dos dados e alterar o relacionamento conjunto entre as variáveis após a imputação.

O exemplo simples a seguir demonstra que a ordem das variáveis na construção da rede Bayesiana pode influenciar em quantidades calculadas com base nos dados após a imputação. A estimação de coeficientes de correlação ou contagens em tabelas de contingência estariam afetados pela ordem do atributo na rede. Considere três variáveis aleatórias A, B e C e o grafo em (1) na Figura 4.1, descrevendo o melhor relacionamento entre elas. O grafo em (2) na mesma figura representa a ordenação sob o aspecto da confiabilidade definido por Di Zio et al. (2004). Dessa maneira, a rede Bayesiana utilizada como método para imputação nesta base, seria a que vemos em (2) na Figura 4.1.

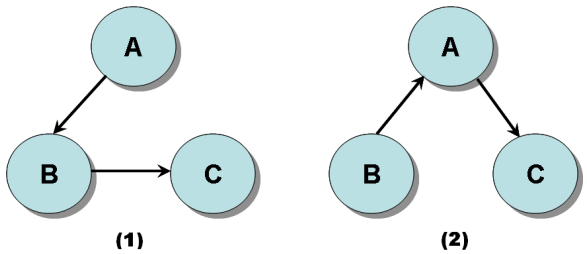


Figura 4.1: Grafos do exemplo da importância da ordenação das variáveis para imputação: (1) grafo da rede que descreve o melhor relacionamento entre as variáveis (2) grafo da rede que ordena conforme a confiabilidade das variáveis

Ao se desejar imputar a variável C nos dados do Apêndice B, uma tabela de contingência entre B e C após a imputação usando (2) teria um acréscimo de quase 20% em um cruzamento de categorias, por exemplo, $(B = b_2, C = c_2)$. Por este motivo, defendemos aqui que a ordenação a ser utilizada para a construção da estrutura da rede para o método de imputação deva ser aquela em que melhor se obtenha do relacionamento entre as variáveis, representado pelos arcos orientados no grafo. A experiência de um perito no assunto específico, inserida como informação a priori na construção da estrutura da rede Bayesiana $\mathcal{B}$ permite maior aproveitamento da ligação entre os atributos da base de dados. Nesse caso, as redes de entrada do algoritmo proposto para imputação refletirão o conhecimento de especialistas. Em redes mais complexas, esta característica pode significar a inserção de vícios em cadeia (a variável imputada posteriormente pode carregar os erros de uma imputação anterior a ela).

Os algoritmos propostos por Hruschka–Jr. (2003) obtêm parâmetros e estrutura baseados em dados completos, sendo necessário que se exclua as unidades que contenham pelo menos um item faltante. Este procedimento limita o método de imputação para aplicações apenas às não respostas do tipo MCAR pois, ao excluir unidades da base de dados antes do aprendizado da estrutura, supõe–se que a distribuição associada às variáveis não se altera. Em seus algoritmos, a ordenação das variáveis mostra–se importante de tal forma para os resultados, que o autor desenvolve um teste de ordenação baseado no teste de qui–quadrado para obter melhor desempenho de seu método.

A ordenação das variáveis e a construção da estrutura da rede são elementos que representam um modelo desconhecido do qual são originados os dados. Como não se sabe da veracidade deste modelo, a rede Bayesiana para o método de imputação específico será considerada conhecida, e reforça–se a necessidade de estudo mais aprofundado sobre a adequação de uma estrutura aos dados no contexto de imputação.

4.3.1 Algoritmo proposto para imputação

Seja o interesse em imputar em $\mathbf{X} = \{\mathbf{X}_D, \mathbf{X}_C\}$, uma ou mais variáveis para a qual se dispõe de uma rede Bayesiana mista $\mathcal{B} = (\mathcal{G}, \theta)$ representando o comportamento conjunto de $\mathbf{X}$. Definem–se então os seguintes subconjuntos disjuntos de variáveis a partir de $\mathcal{G}$: P_0, que contém as variáveis sem pais na rede; P_1, cujas variáveis têm como pais somente variáveis no conjunto P_0; P_2, onde os pais se encontram em $P_0 \bigcup P_1$, e assim sucessivamente até o subconjunto P_j, $j = \nu - 1$, que contém os pais das variáveis em ν.

Procede–se com o método de imputação da seguinte maneira: se a variável estiver em P_0, gera–se um valor a ser imputado de sua distribuição marginal, se a variável for discreta, ou de acordo com (4.2.5), se for contínua. Se X_j a imputar estiver em P_1, então um valor será gerado de acordo com $P[X_j|pa_D(X_j) \subset P_0]$ se a variável for discreta e de acordo com $P[X_j|(pa_D(X_j), pa_C(X_j)) \subset P_0]$ se for do tipo contínua. Segue–se com o mesmo procedimento até o último subconjunto respeitando–se a ordem estabelecida de acordo com a rede. O Quadro 5 resume este método.

A idéia de separar subconjuntos de variáveis está em controlar as imputações bem como seus resultados. Este procedimento evita que valores gerados para variáveis em $P_1, P_2, \ldots, P_\nu$ sejam condicionados à não resposta em seus pais e permite avaliar o desempenho do método sob várias situações. A Figura 4.2 mostra um grafo $\mathcal{G}_h$ de uma rede Bayesiana hipotética com seus subconjuntos de variáveis identificados. Mais adiante vê–se que os resultados de avaliação das consistências quando apenas variáveis em P_2 são

Algoritmo proposto

Entrada: Rede Bayesiana ajustada e
Base de dados com valores faltantes.
Saída: Base de dados imputada.

1. Identifique os subconjuntos $P_0, P_1, \ldots, P_\nu$ na rede Bayesiana de entrada
2. Defina uma ordem de imputação em cada subconjunto de forma que as variáveis discretas sejam imputadas antes das contínuas. Entre as discretas ou contínuas de um mesmo subconjunto, pode ser utilizado qualquer critério de ordenação;
3. Para cada subconjunto $j = 1, 2, \ldots, \nu$, faça:
   - 3.1. Se a variável pertencer ao primeiro conjunto (ou conjunto das variáveis sem pais), gere aleatoriamente um dado a ser imputado de acordo com a distribuição marginal da mesma, se discreta ou de acordo com (4.2.5) se contínua;
   - 3.2. Se não, gere um dado a ser imputado de acordo com a estrutura da rede ajustada em 2., se discreta, ou de acordo com (4.2.3) ou (4.2.4) dependendo do conjunto de pais da variável contínua;
4. Retorne a base imputada.

Quadro 5: Algoritmo proposto para uso de redes Bayesianas mistas em imputação

imputadas diferem daquelas obtidas quando, por exemplo, P_1 e P_2 são imputadas.

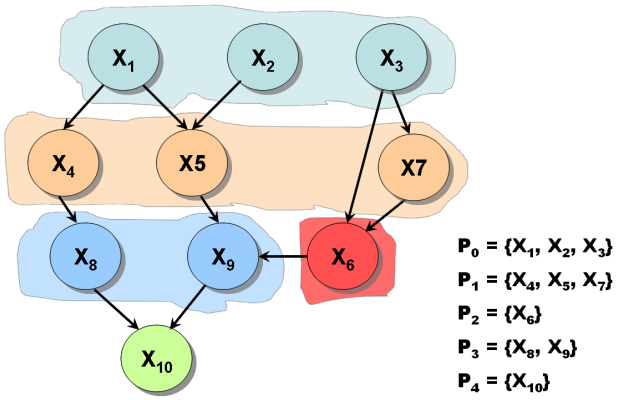


Figura 4.2: Exemplo da formação de subconjuntos de variáveis após a estrutura ajustada

No método de imputação conduzido neste texto para variáveis discretas e contínuas considera–se, quando necessário, uma ordenação das variáveis dentro dos subconjuntos que segue aquela observada na própria base de dados. Qualquer outra pode ser utilizada, desde que seja seguida a ordenação definida pelos subconjuntos identificados no grafo.

Realizou–se inicialmente a imputação nas variáveis discretas e logo após nas variáveis contínuas de cada subconjunto, mas a escolha da ordem de imputação dentro de cada um deles deu–se arbitrariamente[4]. A rede Bayesiana de estrutura $\mathcal{G}$ como

[4]Não foi determinada a melhor ordem enre as variáveis e sabe–se da importância deste procedimento

método de imputação para variáveis qualitativas permite que, dependendo da posição do nó, a imputação seja conduzida de forma análoga a algum dos métodos descritos na Seção 2.3 do Capítulo 2. Essas equivalências são registradas na seção a seguir.

4.3.2 Observações em variáveis quantitativas

Após se definir o método de imputação a partir de redes mistas, cabem alguns comentários com relação às variáveis quantitativas. O primeiro refere–se àquelas que estão em P_0, ou seja, que não têm pais na rede. Isso significa que a imputação dar–se–á por (4.2.5), ou seja, a partir da sua média geral. Este caso é análogo à imputação aleatória geral (vide Capítulo 2) onde o valor de X_j a ser imputado seguirá:

$$\widetilde{x}_{ij} = \mu_j + e_{ij}, \qquad i = 1, 2, \ldots, n^*,$$

onde n^* é o número de observações ausentes em X_j e $e_{ij} \sim N(0, \sigma^2_{X_j})$, pois $X_j \sim N(\mu_j = \beta_{0,X_j}, \sigma^2_{X_j})$. Uma potencial desvantagem dessa forma de imputar X_j, quando esta não tiver pais na rede, está na distorção da forma de sua distribuição e da correlação entre esta e as demais variáveis (DURRANT, 2005).

Se X_j tem somente pais discretos na rede, então a imputação nesta variável seguirá análoga à imputação aleatória dentro de classes, ou seja, o valor imputado seguirá

$$\widetilde{x}_{ij} = \beta_{0,X_j|pa_D(X_j)} + e_{ij}, \qquad i = 1, 2, \ldots, n^*,$$

onde $e_{ij} \sim N(0, \sigma^2_{X_j|pa_D(X_j)})$, pois $X_j \sim N(\mu_j = \beta_{0,X_j|pa_D(X_j)}, \sigma^2_{X_j|pa_D(X)})$. A imputação nestas variáveis, segundo Kalton (1983) e Albieri (1989) causa distorções nas caudas da distribuição. Este efeito pode ser indesejado caso a variável em questão seja a renda, pois os valores muito baixos (associados à pobreza) estariam subestimados.

Se X_j tem pais discretos e contínuos na rede, então a imputação será análoga à imputação por regressão mais efeito de aleatoriedade, onde um valor para X_j será gerado de acordo com uma regressão ajustada nos pais de X_j, ou seja,

$$\begin{aligned} \widetilde{x}_{ij} &= \mu_j + e_{ij}, \qquad i = 1, 2, \ldots, n^*, \\ &= \beta_{0,X_j} + \beta_{i,X_j} x_{i,pa_C(X_j)} + e_{ij}, \end{aligned}$$

onde $e_{ij} \sim N(0, \sigma^2_{X_j|pa_C(X_j),pa_D(X_j)})$. De acordo com Durrant (2005), a imputação por regressão aleatória também é referida na literatura como imputação por distribuição condi-

no processo de imputação. O estudo da ordenação é sugerido como trabalho futuro.

cional e uma grande vantagem deste método está em manter a distribuição das variáveis e permitir a estimação de quantidades distribucionais.

De acordo com essas observações, chega–se à conclusão de que a rede Bayesiana mista para imputação em variáveis contínuas funciona como uma coleção de métodos de imputação representados pelos parâmetros da rede. Este fato permite combinar as vantagens e limitações de cada um deles no intuito de manter ao máximo possível as características do modelo de imputação que representa o relacionamento associado às variáveis. No sentido de verificar o quanto a imputação por este método torna–se promissora a diversos objetivos, descrevem–se na próxima seção algumas formas para a sua avaliação.

4.4 Avaliação do uso da rede mista para imputação em dados quantitativos

Assim como foram tratadas algumas medidas de consistência para o caso das redes Bayesianas discretas para imputação, esta seção também cita uma extensão para a parte da rede que reflete as variáveis contínuas. Os resultados partem do pressuposto considerado por Bøttcher (2004) de fatoração da distribuição de probabilidade conjunta sob as variáveis aleatórias, que é dada em (4.2.1). Devido a limitação na construção das redes de variáveis discretas não apresentarem pais contínuos e pela fatoração da distribuição conjunta, a avaliação da parte discreta da rede mista ocorre de forma independente da parte contínua. Como os nós contínuos podem estar condicionados aos nós discretos, o desempenho na imputação de variáveis quantitativas pode ser influenciado pelo que ocorre às variáveis qualitativas.

Chambers (2000) cita alguns critérios para comparação entre métodos de imputação e algumas propriedades desejadas para o seu desempenho e Di Zio et al. (2004) estabelecem três tipos de consistências que permitem uma comparação com outros métodos para imputação em variáveis discretas. Para a aplicação de redes Bayesianas mistas para imputar em variáveis aleatórias contínuas, estabelecem–se aqui alguns tipos de consistências, que se definem como extensão daquelas consideradas para o caso discreto, além da proposição da consistência estrutural, que permite avaliar a manutenção dos relacionamentos de independências condicionais descritas pelas estruturas das redes ajustadas após a imputação.

Apesar de serem quantidades e indicadores simples, a inexistência de resultados

na literatura que dessem um direcionamento acerca da aplicação de redes Bayesianas para imputação em variáveis contínuas torna este trabalho uma contribuição que pode servir como base para outras proposições e aplicações semelhantes. A seguir, descrevem–se as consistências da base de dados, estrutural, lógica e estatística.

4.4.1 Consistência da base de dados

Principalmente para gestores de grandes bases de dados, existe interesse em se obter uma informação estimada para a não resposta, como se esta não tivesse sido imputada. É certo que, se a informação for perdida por algum problema mecânico ou de captação do dado, existem situações em que é possível recuperar a informação original e compará–la com o valor imputado. Mas na maioria dos casos não é isso o que ocorre. Por exemplo, se a não resposta se dá pela recusa no fornecimento da informação, então uma imputação a este campo carrega uma incerteza que está relacionada com a probabilidade da não resposta associada àquela variável.

No caso da variável discreta, existem facilidades em se comparar o valor real e o dado imputado, e um índice deste tipo de consistência está no valor esperado do percentual de acertos na base. Já no caso da variável contínua, o ideal seria do valor imputado coincidir com o valor real, mas pequenas variações (estas a serem definidas pelos mantenedores das bases ou pesquisadores) são aceitáveis, desde que não modifiquem a distribuição original da variável.

Definição 6 – A consistência da base de dados é a propriedade que mantém os registros individuais da mesma após a imputação, ou os torna muito próximos dos observados, no caso da variável aleatória ser do tipo contínua.

Para avaliar a consistência da base de dados para o caso contínuo em imputação usando redes mistas, utilizamos um recurso simples:

$$\xi_{X_j} = \frac{\sum_{i=1}^{n^*}(x_{ij} - \widetilde{x}_{ij})^2}{n^*}, \tag{4.4.7}$$

onde n^* é o número de itens faltantes na variável X_j, x_{ij} é o valor observado e $\widetilde{x}_{ij}$ é o seu equivalente imputado.

O que se espera é que, se o modelo de imputação usando rede Bayesiana mista refletir a propriedade de consistência da base, então o valor de ξ_{X_j} para variáveis contínuas esteja próximo de zero.

Chambers (2000) utiliza como ferramenta para avaliação da consistência da base o coeficiente de correlação entre os dados imputados em uma determinada variável aleatória X_j e o seu correspondente valor observado, ou seja, $\rho(X_j, \widetilde{X}_j)$. As conclusões sobre essa consistência são balizadas pela interpretação do coeficiente calculado a partir da Tabela 2 no Capítulo 3. Quanto mais próximos estiverem os valores imputados de seus equivalentes observados, mais forte será essa correlação. Uma outra sugestão do mesmo autor está em ajustar um modelo de regressão linear normal tendo como variável dependente o valor imputado $\widetilde{X}_j$, a observada como preditora e o intercepto nulo. Outras quantidades poderiam ser pensadas, como a função módulo ou uma indicadora de pertinência do valor imputado a uma região previamente definida como plausível para a imputação no registro. Esta última seria útil principalmente para o caso de serem criadas classes de imputação, o que ocorre quando o nó contínuo apresenta um ou mais nós discretos como pais. Como observação, só será avaliada no caso contínuo a consistência univariada para a base de dados.

4.4.2 Consistência estrutural

A estrutura $\mathcal{G}$ de uma rede Bayesiana retrata as relações de independência condicional entre variáveis de interesse (NEAPOLITAN, 2004). O comportamento de $\mathcal{G}$, construído a partir dos dados imputados, permite verificar se a rede Bayesiana como método para imputação apresenta como característica a manutenção dos relacionamentos de independência condicional existentes entre as variáveis. Para isso é proposta a consistência estrutural definida a seguir.

Definição 7 – A consistência estrutural é a propriedade que a rede tem de manter–se na mesma classe de equivalência da rede original após o seu ajuste a partir dos dados imputados.

Neste caso, o padrão de comparação é a rede original, ou a rede que representa as probabilidades de imputação em cada variável. A exemplo da rede discreta, esta avaliação pode ser realizada a partir da estrutura $\mathcal{G}$ propriamente dita ou a partir de alguma medida resumo, como é o caso da função escore. Neste texto será apresentada a consistência com base em $\mathcal{G}$.

Fundamentalmente, para se obter a consistência estrutural avaliando a estrutura da rede, seguem–se os passos apresentados na Seção 3.4.2. O fato de a rede conter variáveis discretas e contínuas numa mesma estrutura não modifica a condução desta avaliação. Em

se tratando de avaliar a consistência estrutural a partir do grafo, observa–se se a estrutura $\widetilde{\mathcal{G}}$ ajustada após a imputação pertence à mesma classe de equivalência da estrutura $\mathcal{G}$ original dos dados conforme descrito na Seção 3.2.1 no Capítulo 3.

Para se construir um índice da presença de $\widetilde{\mathcal{G}}$ na classe de equivalência de $\mathcal{G}$, partimos da suposição de que existe o mesmo número de variáveis (e conseqüentemente o mesmo número de nós) entre a base original e a imputada. Além disso, é necessário que a estrutura mantenha os mesmos relacionamentos *head–to–head* em nós de bloqueio e que apresente o mesmo número de arestas entre $\widetilde{\mathcal{G}}$ e $\mathcal{G}$. Considere as variáveis B e C, respectivamente obtidas por (3.4.6) e (3.4.7) no Capítulo 3. A combinação entre estas duas variáveis representa o que ocorre com as estruturas construídas após a imputação. De interesse para formar a classe de equivalência de $\mathcal{G}$ estão as redes $\widetilde{\mathcal{B}} = (\widetilde{\mathcal{G}}, \widetilde{\theta})$, que se encontram em $\{B = b_2\}$ e $\{C = c_1\}$.

Seja $\Psi_{\mathcal{G}}$ a classe de equivalência da rede original $\mathcal{G}$ e m o número de repetições de bases imputadas pela rede Bayesiana $\mathcal{B} = (\mathcal{G}, \theta)$. Um índice da presença de $\widetilde{\mathcal{G}}$ em $\Psi_{\mathcal{G}}$ é dado por:

$$\varpi = \frac{\sum_{i=1}^{m} I(\widetilde{\mathcal{G}}_i \in \Psi_{\mathcal{G}})}{m}, \tag{4.4.8}$$

onde $I(\widetilde{\mathcal{G}}_i \in \Psi_{\mathcal{G}})$ é a indicadora da presença da i–ésima estrutura construída após a imputação na classe de equivalência da rede original.

Altos valores de ϖ indicam que a imputação a partir da rede Bayesiana inicial preserva os mesmos relacionamentos que descreve. Observa–se a necessidade de se construirem redes coerentes com o relacionamento existente entre as variáveis para que um baixo valor de ϖ não seja confundido com o baixo desempenho na consistência estrutural, quando na verdade estaria refletindo o não ajuste ao modelo da rede original.

4.4.3 Consistência lógica

Para variáveis discretas, a medida de consistência lógica é construída principalmente a partir dos zeros estruturais. Tal qual a consistência lógica na variável aleatória discreta, nas variáveis contínuas também ocorrem os zeros estruturais. Esse é o caso, por exemplo, de linhas de produção em que se acompanha o tempo de funcionamento de máquinas de uma seção. Este atributo só é registrado se houver a informação de que a máquina não está quebrada ou em manutenção. Se estas condições forem observadas, então não existirá imputação da variável tempo nas unidades equivalentes.

Com relação aos zeros estruturais imputados em variáveis contínuas, a avaliação

é feita com base no valor de

$$\xi_z = \frac{\sum_{i=1}^{n^*} I_{x_{z,ij}}(\widetilde{x}_{z,ij})}{n^*},$$

onde n^* é o número de itens faltantes em X_j, z é a área de X_j característica por apresentar o zero estrutural e $I_{x_{z,ij}}(\widetilde{x}_{z,ij})$ é a indicadora de ocorrência de coerência na imputação em $\widetilde{x}_{z,ij}$, da mesma forma como em Di Zio et al. (2004) e o realizado no Capítulo 3.

O sentido da consistência lógica para variáveis contínuas pode ser ampliado para permitir a identificação de valores que seriam pouco prováveis à imputação de um dado item. A idéia deste tipo de grandeza está em avaliar, para X_j contínua, o efeito de distorções nas caudas da sua distribuição após a imputação. Segundo Kalton (1983) e Albieri (1989), o método de imputação aleatória dentro de classes altera a frequência dos valores muito baixos ou muito altos de X_j. O equivalente a este método dar–se–ia se o nó contínuo apresentasse somente pais discretos na rede (ver Seção 4.3.2).

Um exemplo desta situação refere–se à associação entre idade condicionado ao fato de a unidade se encontrar em atividade remunerada. Seria possível, devido aos parâmetros da rede, porém incoerente, imputar uma idade menor que dez anos neste caso.

Dessa maneira, a associação dos valores pouco prováveis a uma região R de aceitação direcionou a composição de uma medida de consistência lógica mais ampla. Para avaliar a consistência neste caso, é proposto um indicador de pertinência do item imputado à sua região de mais provável ocorrência. Seja então:

$$I_{x_{R,ij}}(\widetilde{x}_{R,ij}) = \begin{cases} 1, & se\ |x_{R,ij} - \widetilde{x}_{R,ij}| < \epsilon_j \\ 0, & \text{caso contrário.} \end{cases}, \tag{4.4.9}$$

onde $x_{R,ij}$ é o valor real da variável para a região R definida por uma área limitante, $\widetilde{x}_{R,ij}$ é o seu correspondente valor imputado e ϵ_j é o limite estabelecido para determinar o quanto se permite o valor imputado distanciar do valor real. O valor de ϵ_j pode ser estipulado conforme o interesse em ser mais ou menos rigoroso na imputação ou pré–definido por uma quantidade tal qual a (2.4.3) definida por Chambers (2000).

Um indicador da consistência lógica para a região R após aplicado o método de imputação é construído por:

$$\xi_R = \frac{\sum_{i=1}^{n^*} I_{x_{R,ij}}(\widetilde{x}_{R,ij})}{n^*},$$

onde n^* é o número de itens faltantes em X_j.

4.4.4 Consistência estatística

Na consistência estatística avalia–se a flutuação de quantidades básicas do ajuste da rede ou de parâmetros de interesse derivados como média, mediana, coeficiente de correlação, etc., para distribuições univariadas ou multivariadas decorrente dos dados.

Di Zio et al. (2004) registram a consistência estatística para os parâmetros da rede Bayesiana ajustada a partir dos dados imputados. Como a rede Bayesiana mista fatora em uma parte discreta e outra mista (BØTTCHER, 2004), então a avaliação dos parâmetros pode ser feita em cada uma das partes separadamente. Nesta seção apenas os nós contínuos serão tratados.

Definição 8 – A consistência estatística refere–se à propriedade que o método possui de manter os mesmos parâmetros da rede original e os parâmetros genéricos (ajuste de modelos, coeficiente de correlação e outros) a partir dos dados após a imputação usando redes Bayesianas.

Se o interesse estiver em avaliar os parâmetros da rede, uma primeira identificação deve estar na posição que o nó ocupa no grafo. Se a variável não apresentar pais na rede, então avaliar a consistência dos parâmetros de sua distribuição é o equivalente a observar o que acontece com suas média e variância após ter–se imputado a não resposta observada. Se a variável possuir apenas pais discretos, então avaliar os parâmetros da rede significa avaliar o que ocorre com os valores de β em $\mu_j = \beta_{0,X_j|pa_D(X_j)}$ sem considerar e_{ij}, ou seja, verificar se existem alterações nas contribuições individuais dos seus pais discretos. Isso é diferente de avaliar a média geral da variável, μ_{X_j}, embora esta seja uma composição dos valores de β de acordo com a rede Bayesiana.

O mesmo vale para o nó que possui pais discretos e contínuos em $\mathcal{G}$, onde também se avalia a contribuição dos pais contínuos a partir dos parâmetros $\beta^{'}$s, conforme $\mu_j = \beta_{0,X_j} + \beta_{i,X_j} x_{i,pa_C(X_j)}$.

Nas seções seguintes apresentam–se alguns estudos simulados onde se pretende avaliar quantidades referentes à distribuição das variáveis contínuas imputadas, especificamente a média e a mediana. Os parâmetros da rede, por se tratarem de regressões lineares nos pais dos nós contínuos, serão abordados no Capítulo 5 a seguir, que tratará da avaliação de quantidades associadas aos modelos ajustados com dados imputados, através da imputação múltipla.

4.5 Aplicação aos dados do Censo Demográfico

Para a aplicação da simulação em redes mistas utilizando os dados do Censo Demográfico foram considerados os registros do responsável pelo domicílio com rendimentos de mais de um e até dez salários mínimos no município de Natal, o que totalizou 20.686 domicílios de um total de 179.822. Este corte foi necessário devido à instabilidade da rede para rendimentos inferiores a um salário mínimo e aos valores muito díspares da distribuição da renda. Além disso, buscou–se atender a suposição de normalidade condicional nos nós. A título de informação, entre um salário mínimo (inclusive) e dez salários mínimos encontram–se 153.432 domicílios, o que representa 85,32% do total. Isso significa que entre a renda nula e um salário mínimo (inclusive) estão 132.746 domicílios. Essa alta frequência nas rendas inferiores a um salário mínimo desestabiliza a rede e provoca a não conformidade com a suposição de distribuição condicional Gaussiana nos nós contínuos exigida pelo *deal* (BØTTCHER; DETHLEFSEN, 2003).

A variável renda no questionário do universo (CD–01) do Censo Demográfico teve seu tratamento para a não resposta dado pela primeira vez nos dados do Censo Demográfico de 2000, a partir do método de árvores de regressão (BREIMAN et al., 1984) para a construção de classes de imputação, de onde se selecionavam doadores para as não respostas contidas em cada classe (PESSOA; SANTOS, 2004). O problema da não resposta na renda não é trivial por apresentar um grande ponto de massa no valor nulo, por se observar o maior percentual de perda de informações para maiores valores da renda, por apresentar distribuição assimétrica e irregular[5] e por se supor um mecanismo de não resposta do tipo ignorável (MAR ou MCAR, conforme citado no Capítulo 2) quando indícios se tem de caracterizar–se como perda do tipo não ignorável (NMAR).

A escolha da renda deve–se pela não existência de outro atributo contínuo viável para a aplicação do método. Das variáveis contínuas presentes no CD–01, a outra alternativa possível seria a de se imputar a variável idade contínua, mas não haveria relacionamento coerente ou compreensível para o aprendizado da rede. No questionário da amostra (CD–02) do Censo Demográfico 2000 disponibiliza–se de uma maior possibilidade em variáveis contínuas, mas o ajuste de uma rede Bayesiana a partir de planos amostrais, que levaria em conta o desenho, ainda é um aspecto em aberto na literatura.

Para o curso das simulações em redes Bayesianas mistas para imputação foram definidas três redes de diferentes características com o objetivo de se imputar a variável

[5]Por irregular entenda–se que existem pontos de alta frequência nos salários em múltiplos do salário mínimo.

Tabela 9: Características de domicílios e responsáveis pelos domicílios consideradas no ajuste das redes mistas para imputação

Nome da variável	**Descrição da variável**	**Classes**	**(%) do total**
CONDDOM	Condição do domicílio	1 – próprio (já pago)	0,571
		2 – próprio (ainda pagando)	0,163
		3 – alugado	0,205
		4 – cedido por empregador	0,010
		5 – cedido de outra forma	0,043
		6 – outra condição	0,008
QTDBAHN	Número de banheiros existentes no domicílio	0	0,074
		1	0,470
		2	0,316
		3 e mais	0,140
SEXO	Sexo do responsável pelo domicílio	1 – masculino	0,590
		2 – feminino	0,410
FXETARIA	Faixa etária do responsável pelo domicílio	1 – menor que 20 anos	0,015
		2 – entre 20 e 29 anos	0,242
		3 – entre 30 e 39 anos	0,245
		4 – entre 40 e 49 anos	0,169
		5 – entre 50 e 59 anos	0,134
		6 – maior que 60 anos	0,195
CURSOELV	Curso mais elevado que o responsável pelo domicílio frequentou	1 – nenhum	0,046
		2 – básico/fundamental	0,727
		3 – médio/ 2^o grau	0,176
		4 – superior/pós graduação	0,051
ANOSEST	Anos de estudo do responsável pelo domicílio	0 – sem estudo	0,076
		1 – de 1 a 5	0,328
		2 – de 6 a 11	0,505
		3 – 12 ou mais	0,091

renda. As variáveis utilizadas na construção destas redes podem ser vistas na Tabela 9, que apresenta, além das classes das variáveis categorizadas consideradas, os percentuais de ocorrência em cada uma delas para o corte de domicílios investigado.

Ao se avaliar a consistência estatística da variável renda, serão observadas quantidades pertencentes à sua distribuição, e a Tabela 10 a seguir mostra um resumo destas. Para se aproximar da suposição de normalidade condicional exigida pelo *deal* (BØTTCHER; DETHLEFSEN, 2003) para variáveis contínuas, aplicou–se a transformação logaritmo nos valores da variável renda, mas todos os resultados serão apresentados em sua unidade original.

As simulações conduzidas deram–se nas redes: domicílio–renda, pessoa–renda e domicílio–pessoa–renda, onde a denominação destas deve–se ao relacionamento de características de domicílios e responsáveis pelo domicílio à variável rendimentos. Na última

Tabela 10: Características da distribuição da renda nos dados do Censo Demográfico para domicílios do município de Natal no corte de renda de mais de um a dez salários mínimos (valores em R$)

Mínimo	1^o Quartil	Mediana	Média	3^o Quartil	Máximo
152,00	267,00	400,00	532,00	700,00	1.504,00

rede buscou–se uma equivalência com o modelo para imputação que foi utilizado para o tratamento da renda no IBGE (PESSOA; SANTOS, 2004)(PESSOA; MOREIRA; SANTOS, 2004). Não é conduzida uma comparação entre a rede Bayesiana e o modelo de Pessoa e Santos (2004) para a imputação na renda por serem modelos de composições distintas e por terem sido controladas as não respostas que foram geradas em diferentes percentuais (1%, 3%, 5%, 7%, 10%, 20%, 30%, 40% e 50%). Um estudo em que os dois métodos fossem aplicados sob as mesmas condições seria relevante para a área. Os resultados para as consistências avaliadas foram obtidos de 500 simulações em cada situação de não resposta especificada. Todos os resultados encontram–se listados nas Tabelas C.1 a C.17 do Apêndice C.

4.5.1 Rede domicílio–renda

Iniciam–se as avaliações do método de redes Bayesianas mistas para imputação com uma rede simples, que relaciona dois atributos do domicílio à variável renda: a condição do domicílio (CONDDOM) e a quantidade de banheiros existentes no domicílio (QTDBAHN). Apesar de parecer pouco lógico que a quantidade de banheiros no domicílio forneça indicações sobre a variável renda, os estudos conduzidos por Pessoa e Santos (2004) e Pessoa, Moreira e Santos (2004) identificaram uma grande influência da variável QTDBAHN na construção das árvores de regressão (BREIMAN et al., 1984) para imputação da variável renda. Por este motivo, esta foi incluída na rede. A Figura 4.3 mostra o grafo da rede Bayesiana ajustada pelo *deal* (BØTTCHER; DETHLEFSEN, 2003) para a imputação da renda a partir das variáveis do domicílio.

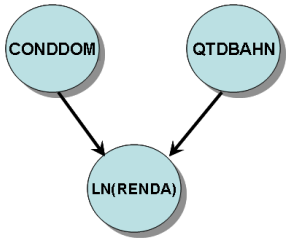


Figura 4.3: Grafo da rede ajustada para imputação a partir de variáveis de domicílio e renda nos dados do Censo Demográfico

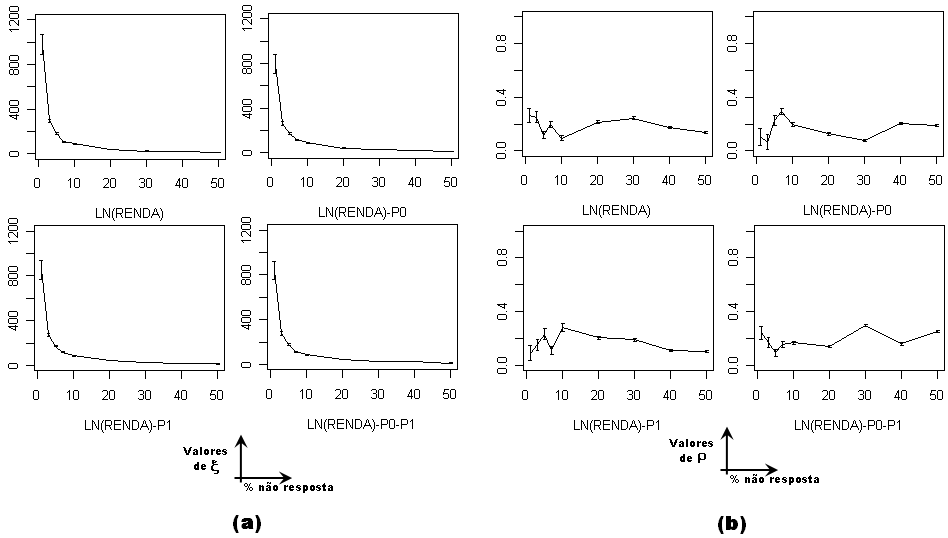


Figura 4.4: Resultados na consistência da base de dados a partir de 500 simulações da rede domicílio–renda para imputação na variável renda em diferentes percentuais de não resposta (a) valores de $\xi_{X_j} = \frac{\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ (b) valores de $\rho(X_{ij}, \widetilde{X}_{ij})$

Foram realizadas as imputações apenas na variável LN(RENDA) respeitando a estrutura em $\mathcal{G}$, ou seja, de acordo com P(LN(RENDA)|CONDDOM, QTDBAHN), sem considerar imputações realizadas em seus pais. Em sequência, seguiram–se as simulações com imputações em pares de variáveis, para se avaliar as consistências citadas na Seção 4.4. As perturbações geradas pela não resposta estavam em CONDDOM–LN(RENDA), QTDBAHN–LN(RENDA) ou em CONDDOM–QTDBAHN–LN(RENDA).

Partindo da consistência da base de dados, duas medidas foram coletadas das simulações: a quantidade em (4.4.7) e o coeficiente de correlação entre o valor observado e o imputado, conforme sugerido por Chambers (2000). As Figuras 4.4 (a) e 4.4 (b) mostram o comportamento das simulações nestas duas quantidades, para os percentuais de não resposta estudados. Dos valores de ξ_{RENDA} pôde–se constatar que existe uma diminuição brusca a partir de 10% de não resposta, o que sugeriria ou uma maior proximidade entre os valores imputados e os observados, ou uma compensação maior entre os registros imputados. No que diz respeito à correlação, percebeu–se que em todos os percentuais de não resposta considerados e em todas as imputações conduzidas, a associação linear entre os valores imputados e os observados é interpretada como fraca ou bem fraca de acordo com a Tabela 2. Esse fato indica que, para esta rede, o método de imputação não preserva os valores na variável (ou em outras palavras, não possui a consistência da base de dados).

Quando se avalia a consistência estrutural, em todas as situações observou–se 100% de ocorrência das estruturas $\widetilde{\mathcal{G}}_i$ ajustadas, na classe de equivalência da estrutura

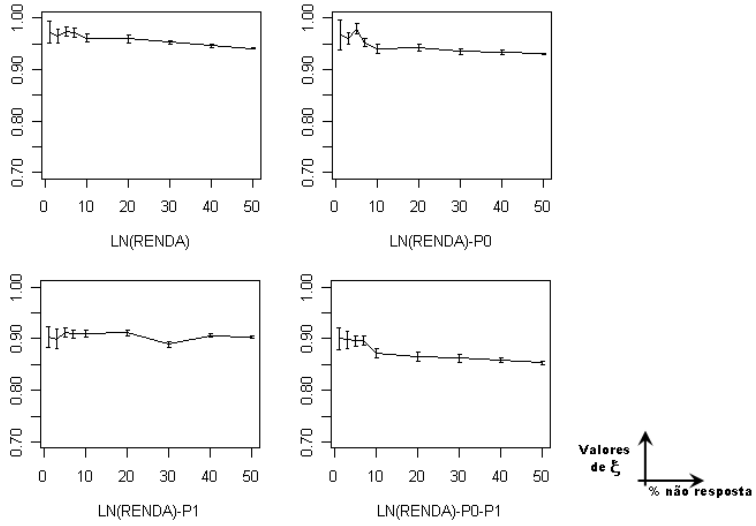


Figura 4.5: Resultados de ξ_R na consistência lógica a partir de 500 simulações da rede domicílio–renda para imputação na variável renda em diferentes percentuais de não resposta

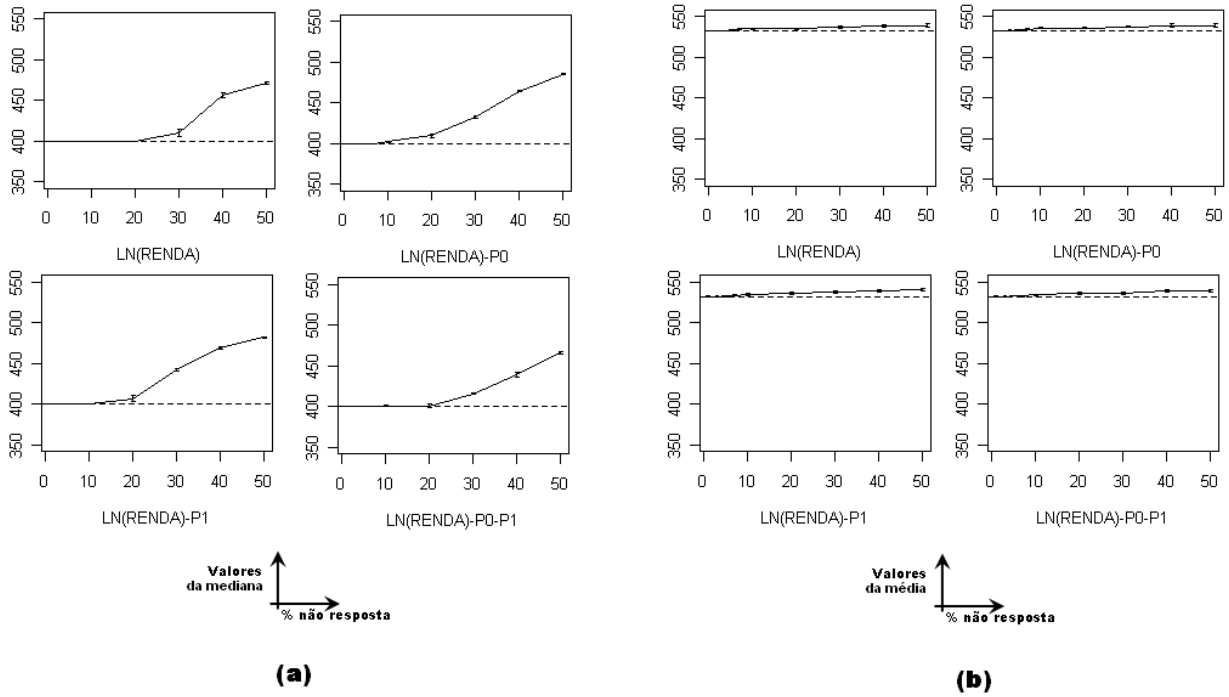


Figura 4.6: Resultados na consistência estatística a partir de 500 simulações da rede domicílio–renda para imputação na variável renda em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

$\mathcal{G}$ da rede original. Examinando–se os fatores possíveis para este resultado, atribuiu–se à simplicidade e ao pequeno número de nós e de arestas da rede.

Na consistência lógica, a expressão em (4.4.3) deu origem à proporção de registros pertencentes à região delimitada por ϵ_j, tendo sido definida para estas aplicações como o valor do desvio padrão da variável. O que se observou como resultados deste tipo de consistência está representado graficamente na Figura 4.5. Verificou–se aqui um leve decrescimento no percentual de valores pertencentes à região ϵ_j à medida que se aumenta o percentual de não resposta e quando todas as variáveis sofrem perturbação.

Já a medida de consistência estatística foi verificada para os valores da mediana e da média após a imputação. Estes resultados podem ser visualizados nas Figuras 4.6 (a) e 4.6 (b), respectivamente. O comportamento da mediana da renda não se altera para percentuais de não resposta inferiores a 10%. A partir dos 10%, para esta rede, a mediana tem uma tendência ascendente na direção do valor da média e em função do percentual de não resposta. Uma causa intuitiva para este comportamento está na suposição de normalidade condicional para o nó contínuo. Isso porque a distribuição conjunta no nó, sendo uma distribuição normal multivariada, conduziria a mediana a se aproximar da média à medida que mais valores fossem inseridos de acordo a uma distribuição simétrica. Apesar da assimetria ser minimizada pela transformação logaritmo, o retorno para a unidade original da variável reflete a distribuição que foi considerada no nó.

As variações que ocorreram nos valores da média foram em menor dimensão, o que leva a crer, neste caso, que a mediana é uma medida mais sensível ao método quando comparada à média.

4.5.2 Rede pessoa–renda

Na rede pessoa–renda, a aplicação do método utiliza variáveis dos responsáveis pelo domicílio para se imputar um valor para a não resposta na renda. Para isso, limitaram–se às variáveis sexo, anos de estudo (ANOSEST) e curso mais elevado que o responsável pelo domicílio frequentou (CURSOELV), tendo concluído pelo menos uma série. Registra–se que a variável ANOSEST não é perguntada diretamente ao respondente, sendo esta derivada de CURSOELV e de uma variável que registra a última série concluída com aprovação. Estas são as únicas variáveis disponíveis para moradores, com exceção da idade, que será acrescentada na próxima aplicação.

A Figura 4.7 mostra o grafo da rede Bayesiana ajustada para estas variáveis. São

identificados neste grafo os subconjuntos de variáveis $P_0 = \{\text{SEXO, CURSOELV}\}$, $P_1 = \{\text{ANOSEST}\}$ e $P_2 = \{\text{LN(RENDA)}\}$. O objetivo continua sendo avaliar a imputação na variável RENDA a partir dos relacionamentos identificados pelo grafo. Foram realizadas simulações em P_2 sem imputar as demais, e em P_2 com perturbações em P_0 e P_1.

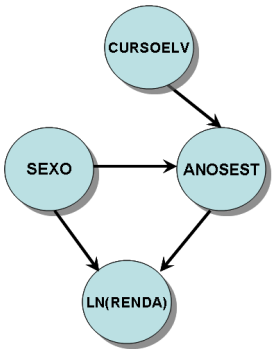


Figura 4.7: Grafo da rede ajustada para variáveis de pessoa e renda para imputação aos dados do Censo Demográfico

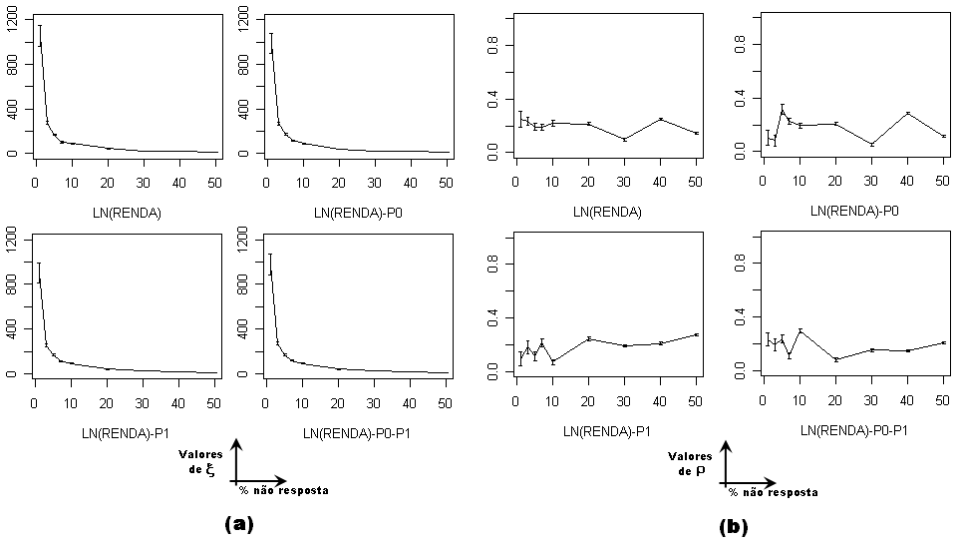


Figura 4.8: Resultados na consistência da base de dados a partir de 500 simulações da rede pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta (a) valores de $\xi_{X_j} = \frac{\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ (b) valores de $\rho(X_{ij}, \widetilde{X}_{ij})$

Primeiramente, observando os resultados para a consistência da base de dados, pôde–se concluir que existe baixa correlação entre o valor imputado e o seu correspondente observado. As Figuras 4.8 (a) e 4.8 (b) mostram respectivamente o comportamento de $\frac{\xi_{X_j}=\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ e $\rho(X_{ij}, \widetilde{X}_{ij})$ nas diferentes situações de perturbação e não resposta. Apesar de serem resultados obtidos a partir de redes Bayesianas diferentes, graficamente mostram–se semelhantes aos da aplicação na seção anterior, com o decaimento acentuado da diferença quadrática entre os valores observados e imputados até os 10% de não resposta

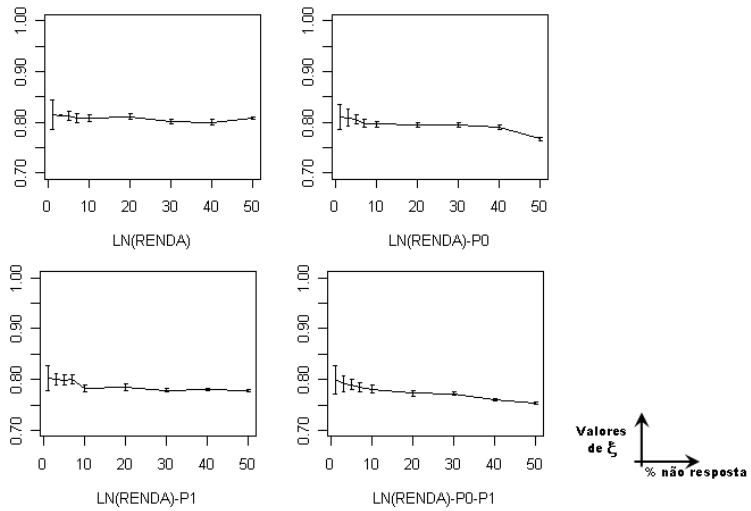


Figura 4.9: Resultados de ξ_R na consistência lógica a partir de 500 simulações da rede pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta

simulada. Podem ser visualizados os valores de ξ_{X_j} e de $\rho(X_{ij}, \widetilde{X}_{ij})$ nas Figuras 4.8 (a) e 4.8 (b), respectivamente.

Referente à consistência estrutural, temos o que segue. Com uma estrutura um pouco mais complexa que a da seção anterior, as proporções de estruturas $\widetilde{\mathcal{G}}_i$ na mesma classe de equivalência de $\mathcal{G}$ chegam ao mínimo de 78% com 50% de não resposta. A manutenção da estrutura é atingida para percentuais baixos de não resposta (1% e 3%) e não se percebe influência para os demais desempenhos de percentual ou do tipo de perturbação nesta medida de consistência.

Na consistência lógica, o desempenho da proporção de valores imputados contidos na região delimitada por ϵ_j em (4.4.3) foi ligeiramente afetado pelo percentual de não resposta nas perturbações que contiveram variáveis em P_0. Este aspecto confirma o aumento na frequência de observações situadas nas caudas da distribuição da variável renda, conforme descrito em Kalton (1983) e Albieri (1989). Estas medidas de consistência para diversos percentuais de não resposta podem ser visualizadas graficamente na Figura 4.9.

Avaliando a mediana e a média da renda na consistência estatística, percebeu–se o mesmo comportamento da mediana e da média na rede domicílio–renda. A mediana sofre um acréscimo na direção da média à medida que se aumenta o percentual de não resposta e independe do tipo de perturbação aplicada. As Figuras 4.10 (a) e 4.10 (b) ilustram graficamente o comportamento destas quantidades para as 500 imputações em vários percentuais de não resposta e em vários tipos de perturbação na rede pessoa–renda.

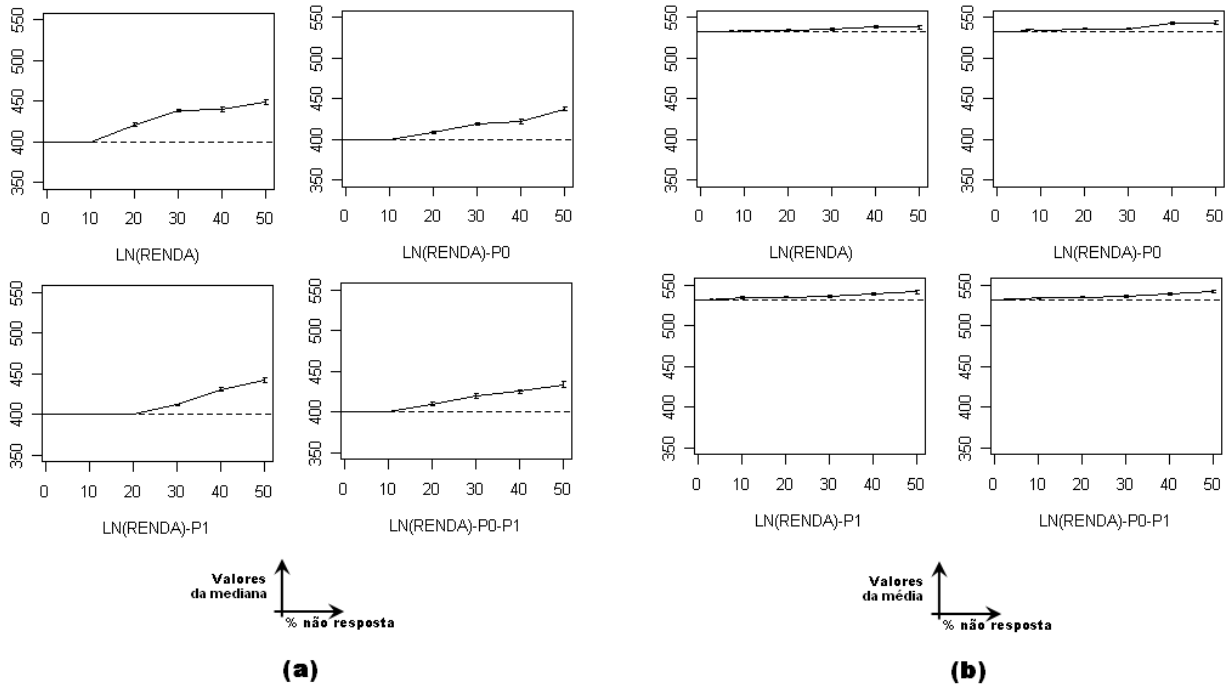


Figura 4.10: Resultados na consistência estatística a partir de 500 simulações da rede pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

4.5.3 Rede domicílio–pessoa–renda

A terceira rede mista ajustada para a imputação nos dados do Censo Demográfico combina variáveis de domicílios e responsáveis pelo domicílio e a renda. A estrutura desta rede pode ser vista na Figura 4.11, que identifica as independências condicionais consideradas para a imputação. A partir deste grafo observamos os subcontjuntos P_0, P_1 e P_2 de variáveis e seus relacionamentos.

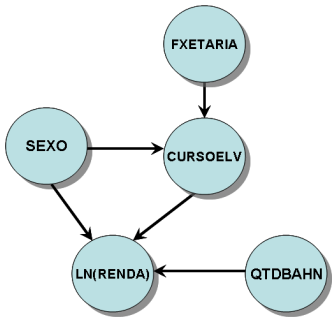


Figura 4.11: Grafo da rede ajustada para variáveis de domicílio, pessoa e renda para imputação aos dados do Censo Demográfico

Nas variáveis em P_0 imputam–se as variáveis sem pais FXETARIA, SEXO e QTDBAHN. No bloco de imputação em P_1, é imputada a variável CURSOELV de acordo

com sua distribuição condicional P(CURSOELV|SEXO, FXETARIA). Em P_2, a variável LN(RENDA) é imputada condicionada em seus pais. A combinação dos pais define as classes de imputação onde, em cada uma delas, se obtém a renda média e seu desvio e é gerado um valor a ser imputado de acordo com a distribuição correspondente.

Foram avaliadas as mesmas consistências das duas redes anteriores. Na consistência da base de dados, o valor de $\xi_{X_j} = \frac{\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ e de $\rho(X_{ij}, \widetilde{X}_{ij})$ são calculados. O que se percebe é que nesta rede não se obtêm valores imputados próximos dos valores observados. No caso dos valores de ξ_{X_j}, o que se nota da Figura 4.12 (a) é um acentuado decrescimento de seu valor até o percentual de 10% de não resposta. Observando a Figura 4.12 (b), os valores de $\rho(X_{ij}, \widetilde{X}_{ij})$ mostram que a correlação máxima obtida entre os valores imputados e observados é de aproximadamente 0,40, interpretada como correlação moderada de acordo com a Tabela 2 na Seção 3.4 deste texto. Este resultado corrobora com a afirmação de Chambers (2000) de que esta é a característica mais difícil de se atingir com um método de imputação.

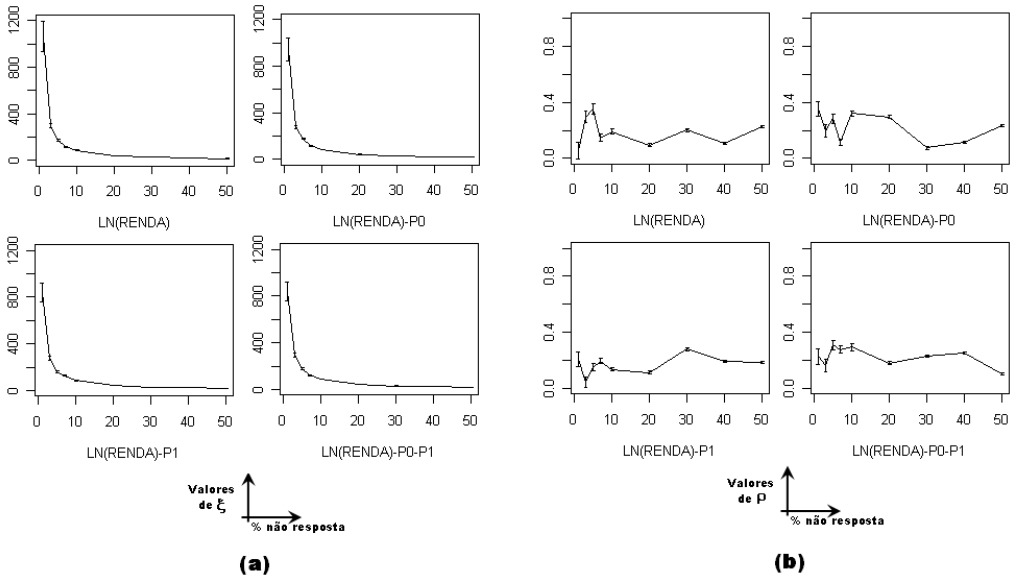


Figura 4.12: Resultados na consistência da base de dados a partir de 500 simulações da rede domicílio–pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta (a) valores de $\frac{\xi_{X_j}=\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ (b) valores de $\rho(X_{ij}, \widetilde{X}_{ij})$

Como medida de consistência estrutural, temos as proporções de redes obtidas após a imputação na mesma classe de equivalência da rede original. O que se observa, a exemplo do ocorrido com a rede pessoa–renda, é que estas proporções são altas, aparentemente não existindo dependência dos resultados com os tipos de perturbação ou o percentual de não resposta. Na consistência lógica, os valores de ξ_R têm apresentados o seu

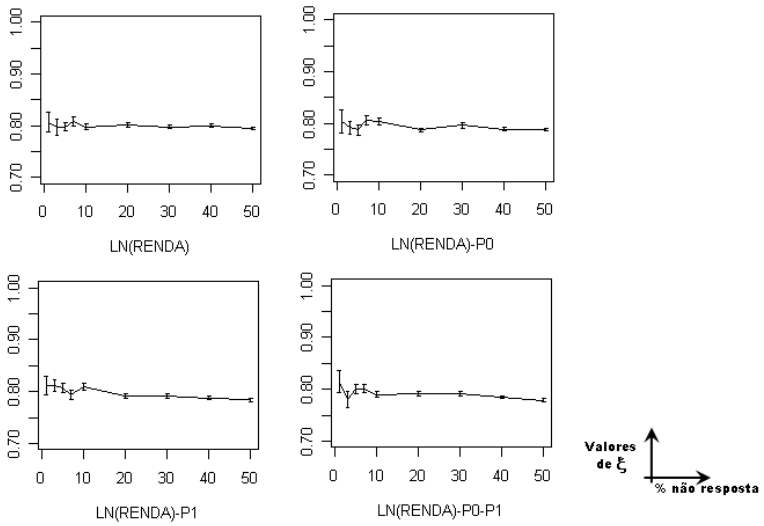


Figura 4.13: Resultados de ξ_R na consistência lógica a partir de 500 simulações da rede domicílio–pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta

comportamento na Figura 4.13. Estima–se que em todos os tipos de perturbação e para todos os percentuais de não resposta, 80% dos dados imputados se encontram na região delimitada por um desvio calculado sobre os dados observados na variável transformada.

Na medida de consistência estatística, analisando os valores da mediana e da média da variável renda após a imputação, também se obtêm conclusões semelhantes ao observado nas duas redes anteriores. A mediana sofre um acréscimo na direção da média à medida que se aumenta o percentual de não resposta e independe do tipo de perturbação aplicada. As Figuras 4.14 (a) e 4.14 (b) ilustram graficamente seus comportamentos dos resultados para a mediana e a média das 500 imputações.

O que se resume das medidas de consistência calculadas com base nestas três redes Bayesianas está resumidamente em três aspectos:

1. se o interesse no método de imputação está em preservar os dados individuais da variável contínua, a rede Bayesiana não é uma boa escolha;

2. é satisfatório imputar um atributo contínuo pela rede Bayesiana se o objetivo estiver em obter uma característica da sua distribuição (a média, por exemplo);

3. se a distribuição da variável contínua for assimétrica, a mediana pode ser afetada na direção da média geral à medida que se aumenta o percentual de não resposta.

A despeito dos resultados obtidos para redes mistas nos dados do Censo Demográfico, a próxima seção traz três aplicações de redes Bayesianas mistas para imputação

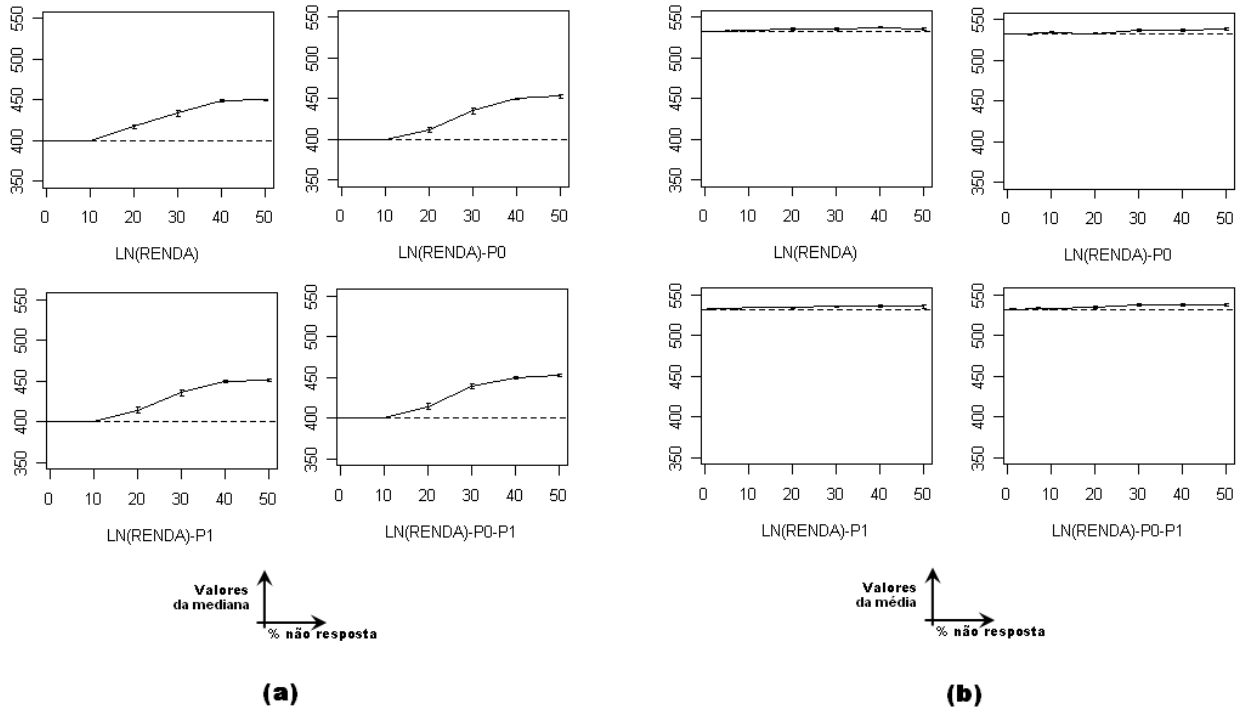


Figura 4.14: Resultados na consistência estatística a partir de 500 simulações da rede domicílio–pessoa–renda para imputação na variável renda em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

em variáveis de tempo nos dados de homicídios em Campinas.

4.6 Aplicação aos dados de homicídios em Campinas

Vargas (2006) conduz um estudo que se propõe a desenvolver e testar uma metodologia quantitativa de análise do tempo da justiça criminal. O principal objetivo do estudo é o de atentar pesquisadores e gestores de políticas públicas quanto às potencialidades da dimensão temporal, tanto para a explicação das atividades de controle exercidas pelos operadores da justiça criminal quanto à adoção das estratégias mais racionais no desempenho destas atividades.

É esse um estudo pioneiro pela utilização da técnica de análise de sobrevivência para tratamento das variáveis de tempo do sistema de justiça. Foi realizado com base em dados de crimes de homicídios ou tentativas de homicídios que foram arquivados no município de Campinas no ano de 2003. Estes dados compreendem noventa e três casos e uma extensa lista de cento e quarenta e nove variáveis particionadas em blocos de acordo com as suas características. Estes blocos podem ser vistos na Tabela 11. A própria autora reconhece que, pelo número restrito de ocorrências, não é possível identificar padrões de

Tabela 11: Variáveis de características de réu, vítima, papéis e crime para dados de homicídios e tentativas de homicídios, arquivados em Campinas no ano de 2003

Bloco	**Descrição**	**Algumas variáveis**
RÉU	Identificação do réu	sexo, estado civil, cor, profissão, grau de instrução, idade, etc.
VÍTIMA	Identificação da vítima	sexo, estado civil, cor, profissão, grau de instrução, idade, etc.
CRIME	Características do crime	tipo de crime, tipo de arma utilizada, local da ocorrência, motivo, data do delito, se houve mais de um réu ou vítima, etc.
PAPÉIS	Características do processo, boletim de ocorrência, decisões no Ministério Público, etc.	ano do processo, prisão durante o processo, existência de recurso, regime prisional, sentença do juri, cotas ao ministério público, pedidos de dilação de prazo, etc.
TEMPO	Variáveis de tempo observadas nos processos	tempo da fase policial, tempo do Ministério Público, tempo do juiz, tempo até a sentença intermediária, etc.

comportamento ou obter parâmetros precisos para relacionamentos mais complexos entre as variáveis (VARGAS, 2006). A utilização destes dados, para avaliar a imputação a partir de redes Bayesianas mistas, tem como objetivo computar os resultados advindos de bases pequenas.

A divisão entre as variáveis efetuada neste capítulo difere da realizada em Vargas (2006), onde estas são particionadas de acordo com as diversas fases que o sistema de justiça penal produz no fluxo das decisões tomadas nas diferentes organizações: ocorrências (cujo segmento organizacional relaciona–se à polícia militar), inquéritos (polícia civil), denúncias (Ministério Público) e processos (justiça). Os blocos definidos para estas aplicações foram compostos de tal maneira a facilitar o ajuste da rede Bayesiana original e evitar imputação que possa significar algum pré–julgamento[6].

Em virtude de uma grande quantidade de variáveis e um número restrito de observações, apenas algumas variáveis foram utilizadas para estas aplicações de redes Bayesianas mistas em imputação. Estas redes foram identificadas de acordo com cruzamentos entre blocos de variáveis e o objetivo estava em imputar a variável de tempo correspondente.

Para o emprego dos dados de homicídios nas redes desta seção serão consideradas as variáveis listadas na Tabela 12 em complemento às variáveis da Tabela 8 apresentada no Capítulo 3. Neste capítulo, a avaliação das medidas de consistência é desenvolvida para

[6]No texto em Vargas (2004) existe uma advertência sobre a possibilidade de se conferir um pré–julgamento a partir de dados de crimes de estupro, fato este que também será evitado neste texto.

Tabela 12: Adendo à Tabela 8, com variáveis de características de vítima e crime para os dados de homicídios em Campinas

Nome da variável	**Descrição da variável**	**Classes**	**(%) do total**
LOCAL	Local de ocorrência	1 – bar	0,172
		2 – rua	0,355
		3 – casa	0,215
		4 – outro	0,258
IDADEVIT	Faixa etária da vítima	1 – até 20 anos	0,118
		2 – de 20 a 29 anos	0,409
		3 – de 30 a 39 anos	0,269
		4 – de 40 a 49 anos	0,172
		5 – 50 anos ou mais	0,032
DILACAO	Se houve pedido de dilação de prazo	0 – não	0,237
		1 – sim	0,763

Tabela 13: Variáveis contínuas utilizadas na composição das redes mistas para dados de homicídios em Campinas

Nome da variável	**Descrição da variável**	**Mediana**	**Média do total**
TPOL1	tempo da fase policial	5 dias	20 dias
TMP	tempo do Ministério Público	40 dias	202 dias
TFASEFI	tempo até a sentença final	3.038 dias	3.690 dias

as variáveis contínuas citadas na Tabela 13.

Os tempos medianos e médios apresentados para as variáveis da Tabela 13 contêm uma informação além daquela associada à distribuição da mesma. De acordo com Vargas (2006), o Código do Processo Penal (CPP) preconiza prazos para o cumprimento de cada uma das fases constituintes do sistema de justiça penal. Se ao final destas não se houver concluído a etapa, é necessário compor em medida excepcional, elevando os tempos de referência de cada etapa. Para fins de avaliação na imputação das variáveis de tempo, isso significa que os parâmetros de entrada da rede podem estar adaptados a uma situação especial se forem estimados a partir de uma base de dados pequena. Como exemplo, cita–se o tempo do Ministério Público[7] (TMP) que deve ser de dez dias se o réu estiver preso e de trinta dias se estiver solto, enquanto que a média de tempo do total na amostra disponível foi de 202 dias (ver Tabela 13).

[7]Esse tempo é aquele decorrido entre a data da denúncia e a data do encerramento do inquérito policial.

As simulações conduzidas com os dados de homicídios em Campinas supõem que os parâmetros refletem a realidade das unidades coletadas, sem levar em consideração os prazos de referência estabelecidos pelo CPP.

As simulações dirigidas nesta seção consideram os percentuais de não resposta em 5%, 7%, 10%, 20%, 30%, 40% e 50% e as avaliações dar–se–ão a partir da imputação em subconjuntos de variáveis observadas na estrutura da rede conforme proposto na Seção 4.3.1. As redes também são ajustadas com a utilização do *deal* (BØTTCHER; DETHLEFSEN, 2003).

Avaliam–se para estas aplicações as consistências:

- da base de dados, mais precisamente a partir do coeficiente de correlação entre os valores observado e imputado;
- estrutural, para computar a proporção de redes mistas na mesma classe de equivalência da rede original;
- estatística, especificamente para os valores da mediana e da média nas variáveis de tempo após imputadas.

Foram examinadas as imputações em três redes distintas, a saber: rede vítima–crime–tempo, rede crime–papéis–tempo e rede réu–crime–papéis–tempo, combinando variáveis específicas de cada bloco. Os resultados das medidas de consistência resultantes das 500 simulações para estas três redes encontram–se no Apêndice C, entre as Tabelas C.18 e C.27 .

4.6.1 Rede vítima–crime–tempo

Na rede vítima–crime–tempo exploram–se as relações existentes entre as variáveis dos blocos vítima e crime com a variável tempo do Ministério Público (TMP). Este tempo está caracterizado por registrar o período decorrido em dias entre a data da denúncia e a data do encerramento do inquérito policial. O grafo na Figura 4.15 ilustra esta rede.

Na Figura 4.15 são identificados os subconjuntos P_0, P_1, P_2 e P_3 de variáveis. Em P_0, imputam–se o sexo da vítima (SEXOVIT) e a idade da vítima (IDAVIT) de acordo com suas respectivas distribuições marginais, por serem elementos do conjunto de nós sem pais na rede. Em P_1, é tratado o tipo de crime (CRIME) de acordo com a probabilidade condicional P(CRIME|SEXOVIT, IDAVIT). A variável local de ocorrência

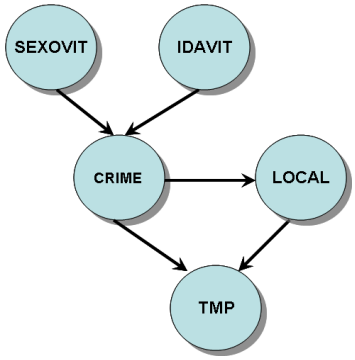


Figura 4.15: Grafo da rede ajustada para variáveis de vítima, crime e tempo para imputação aos dados de homicídio de Campinas

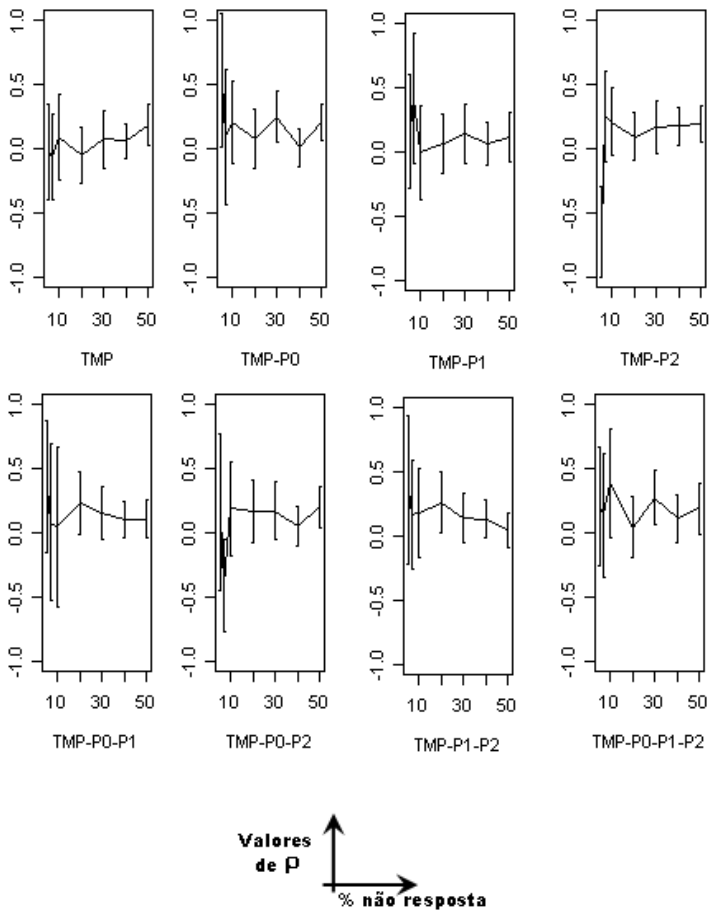


Figura 4.16: Resultados do coeficiente de correlação na consistência da base de dados a partir de 500 simulações da rede vítima–crime–tempo, para imputação na variável de tempo em diferentes percentuais de não resposta

do crime (LOCAL) em P_2 é condicionada a CRIME e finalmente em P_3, a variável TMP é imputada conforme a distribuição condicionada em CRIME e LOCAL.

Avaliando a consistência da base de dados para a variável contínua TMP imputada, foram estimados os valores do coeficiente de correlação entre o item observado na base original e o imputado conforme as relações entre as variáveis em diversos percentuais de não resposta simulados e em vários tipos de perturbações diferentes. Estes resultados têm seu comportamento registrado na Figura 4.16. O que se avalia neste tipo de consistência é que, para esta rede e na imputação dos valores individuais de TMP, a correlação é quase nula, em alguns pontos sendo interpretada como bem fraca.

Na consistência estrutural, chegou–se em alguns casos, a apenas 20% das estruturas $\widetilde{\mathcal{G}}$ obtidas na mesma classe de equivalência da estrutura $\mathcal{G}$ original. Não foram apresentados os valores do direcionamento na perda da estrutura (para os casos em que se aumentam ou diminuem o número de arestas), por não ter sido um resultado expressivo em determinada direção. Os resultados na consistência estrutural não parecem depender do tipo de perturbação ou do percentual de não resposta gerado para cada uma das simulações nesta rede.

A avaliação da medida de consistência estatística deu–se nos valores da mediana e da média e os resultados podem ser verificados respectivamente nas Figuras 4.17 (a) e 4.17 (b). O que se observa a partir do comportamento dos gráficos da mediana e da média é o fato de a mediana sofrer maior influência do método de imputação do que a média. Apesar de a imputação ter sido conduzida em diferentes tipos de perturbação, em todas as situações e até os 10% de não resposta simulada, os valores da mediana foram corretamente obtidos após a imputação. Acima deste percentual, afetou–se a mediana de forma ascendente à medida que se aumentava a não resposta. Com os valores da média, independente dos blocos de variáveis imputadas e do percentual de não resposta imputado, estes mantiveram–se preservados.

4.6.2 Rede crime–papéis–tempo

A rede Bayesiana crime–papéis–tempo combina variáveis de cada um destes blocos com o objetivo de se imputar a variável de tempo decorrido entre a data do registro da ocorrência e a data da sentença intermediária (TFASEFI). A Figura 4.18 a seguir representa o grafo da rede Bayesiana obtida para os dados trabalhados.

Observando a Figura 4.18 foram identificados os subconjuntos de variáveis tal

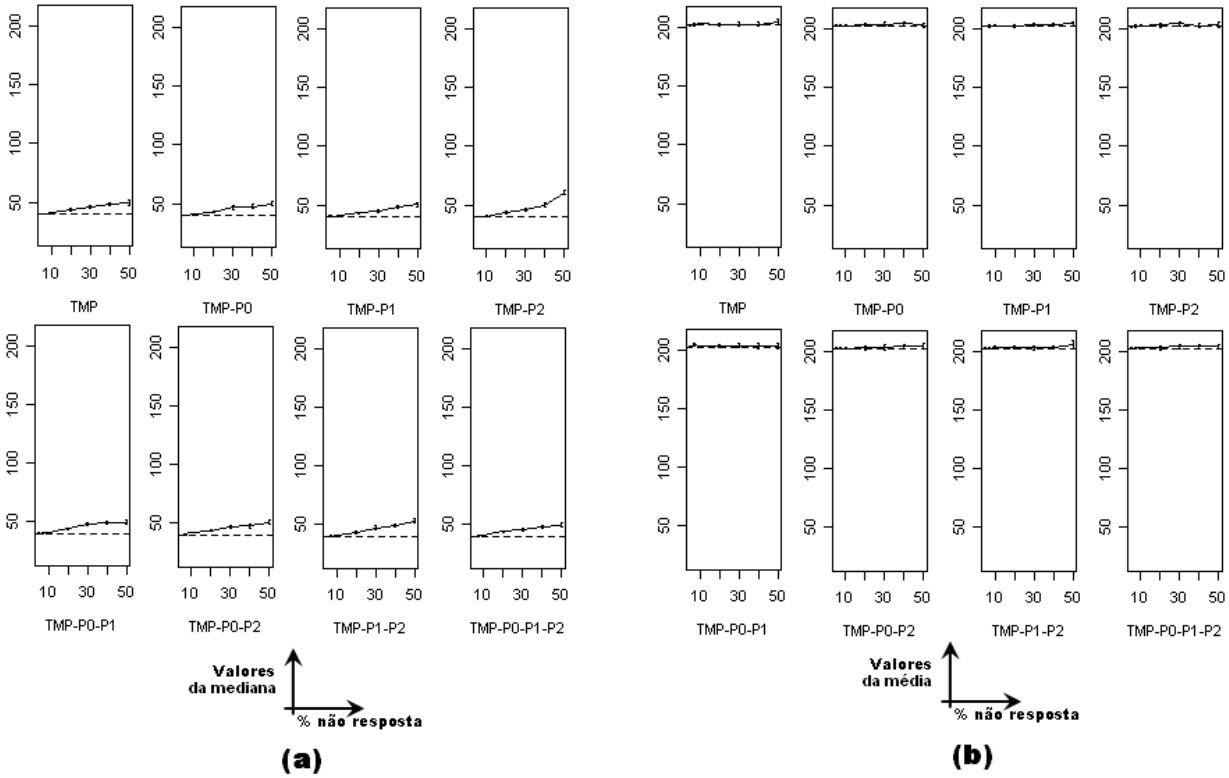


Figura 4.17: Resultados na consistência estatística a partir de 500 simulações da rede vítima–crime–tempo para imputação na variável de tempo em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

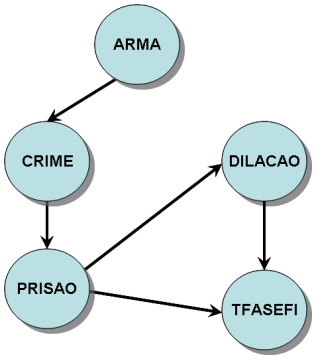


Figura 4.18: Grafo da rede ajustada para variáveis de crime, papéis e tempo para imputação aos dados de homicídio de Campinas

como sugerido na Seção 4.3. Imputa–se a variável tipo de arma utilizada no crime (ARMA) no conjunto P_0 a partir de sua distribuição marginal. A variável tipo de crime (CRIME) é imputada condicionada à variável ARMA em P_1 e a variável prisão durante o processo (PRISAO) é imputada em P_2 condicionada em CRIME. No subconjunto P_3 encontra–se a variável pedido de dilação durante o processo (DILACAO) e a variável TFASEFI em P_4, que será o objeto de estudo das avaliações da imputação usando a rede mista, é imputada condicionada em PRISAO e DILACAO.

A avaliação da consistência da base de dados é feita também a partir do cálculo do coeficiente de correlação entre o valor real observado e o valor imputado, e seu resultado para as 500 simulações em diversos percentuais de não resposta pode ser visto na Figura 4.19. A exemplo do observado na rede da seção anterior, a rede crime–papéis–tempo na imputação da variável TFASEFI não possui a propriedade de preservar os microdados. Novamente reforça–se a afirmação de Chambers (2000) de que esta é a propriedade mais difícil de se atingir com um método de imputação, e geralmente não se obtêm bons resultados para esta característica.

Na consistência estrutural, o pior desempenho deu–se em 18% de estruturas $\widetilde{\mathcal{G}}$ presentes na classe de equivalência da rede original. Apesar de este resultado ter ocorrido em 40% de não resposta, em maiores percentuais sobrevieram desempenhos melhores. Em algumas situações de perturbação, maiores percentuais de manutenção da estrutura deram–se em maiores percentuais de não resposta. Este foi o caso, por exemplo, da imputação em TFASEFI–P_0–P_2, que obteve 48% de consistência estrutural em 5% de não resposta, enquanto que para o mesmo tipo de perturbação, em 40% de não resposta, obtiveram–se 64% das estruturas $\widetilde{\mathcal{G}}$ na classe de equivalência de sua estrutura original.

Para a consistência estatística, avaliaram-se novamente a mediana e a média na distribuição de TFASEFI. Os comportamentos das 500 simulações para estas quantidades

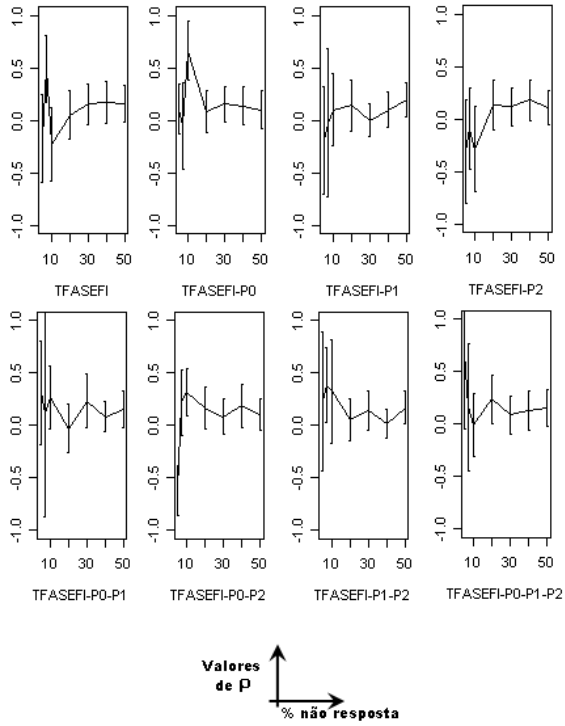


Figura 4.19: Resultados do coeficiente de correlação na consistência da base de dados a partir de 500 simulações da rede crime–papéis–tempo, para imputação na variável de tempo em diferentes percentuais de não resposta

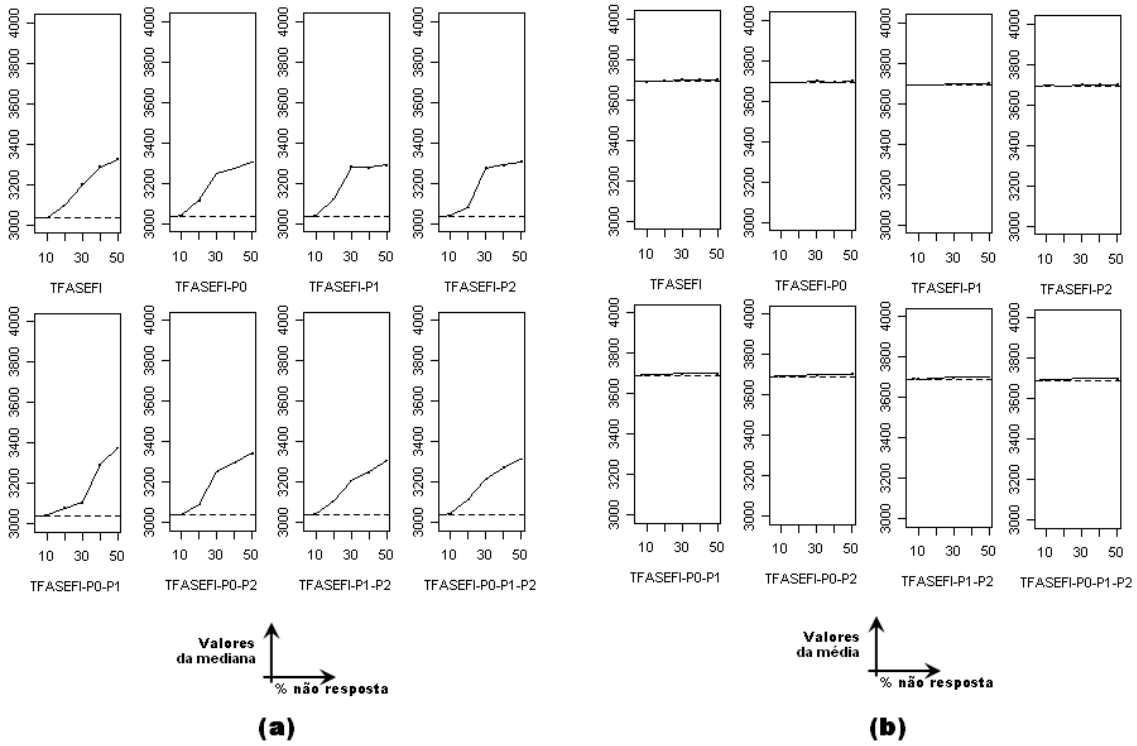


Figura 4.20: Resultados na consistência estatística a partir de 500 simulações da rede crime–papéis–tempo para imputação na variável de tempo em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

podem ser visualizados nas Figuras 4.20 (a) e 4.20 (b). A exemplo do ocorrido na rede da seção anterior, a mediana sofreu maior influência da imputação pelo método. O acréscimo no valor da mediana é sempre ascendente na direção da média e sempre cresce a medida que se aumenta o percentual de não resposta simulado. Com a média, indiferente do percentual de não resposta e da perturbação aplicada aos dados, após a imputação, esta quantidade permanece inalterada, com poucas variações.

4.6.3 Rede réu–crime–papéis–tempo

Esta é a rede mais complexa das três aplicadas aos dados de homicídios em Campinas. Reúne variáveis dos grupos de características de réu, do crime, do processo e do tempo. Especificamente para esta rede, o tempo a ser trabalhado é aquele decorrido entre o registro da ocorrência e a abertura do inquérito policial, denominado de TPOL1. O grafo correspondente à rede Bayesiana após seu ajuste a partir do *deal* (BØTTCHER; DETHLEFSEN, 2003) pode ser visto na Figura 4.21 a seguir.

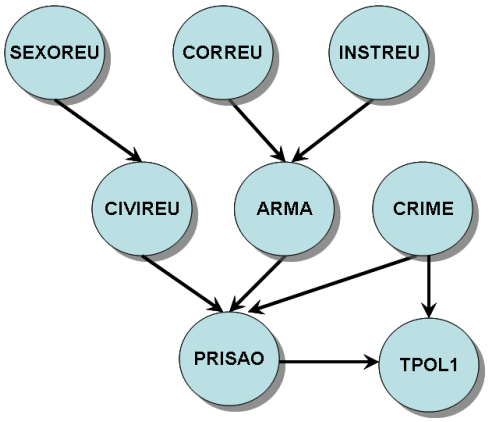


Figura 4.21: Grafo da rede ajustada para variáveis de réu, crime, papéis e tempo para imputação aos dados de homicídio de Campinas

A parte discreta desta rede é a mesma avaliada na Seção 3.6.1. O acréscimo da variável TPOL1 não provocou alterações nos relacionamentos de independência condicional anteriormente detectadas. Na Figura 4.21 podem–se identificar os subconjuntos de variáveis a serem utilizados na imputação dos itens faltantes simulados na base. O subconjunto P_0 compreende as variáveis sexo (SEXOREU), cor (CORREU) e grau de instrução (INSTRUREU) do réu, além da variável tipo de crime (CRIME). As variáveis estado civil do réu (CIVIREU) e tipo de arma utilizada na ocorrência (ARMA) formam o grupo de variáveis do subconjunto P_1. Em P_2 está a variável prisão durante o processo (PRISAO) e P_3 compreende a variável de tempo TPOL1. Novamente, as avaliações das consistências serão conduzidas para a variável contínua TPOL1.

A avaliação da consistência da base de dados foi realizada com base no coeficiente de correlação entre o valor real observado e o seu correspondente imputado. O comportamento desta medida de consistência pode ser visualizado nos gráficos da Figura 4.22 para as diversas perturbações aplicadas. O que se nota é a correlação muito fraca entre os valores de X_j e $\widetilde{X}_j$, independente do percentual de não resposta simulado e do tipo da perturbação.

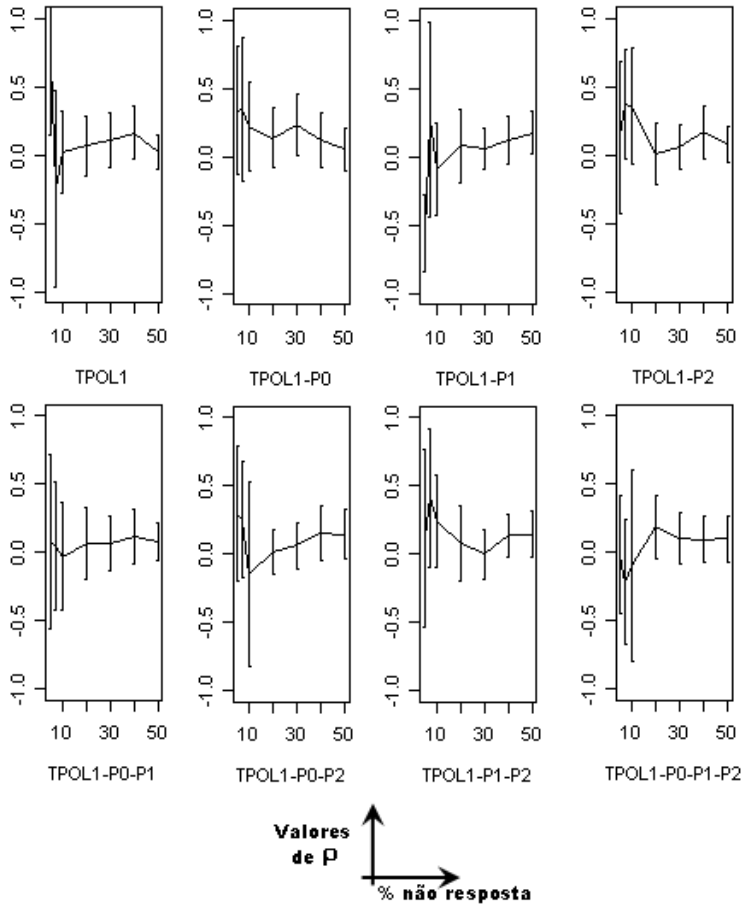


Figura 4.22: Resultados do coeficiente de correlação na consistência da base de dados a partir de 500 simulações da rede réu–crime–papéis–tempo, para imputação na variável de tempo em diferentes percentuais de não resposta

Na consistência estrutural, ocorre o mesmo fato das duas redes das seções anteriores, uma não dependência entre o percentual de estruturas após a imputação na mesma classe de equivalência da estrutura original da rede e o percentual de não resposta ou blocos de variáveis imputadas. A comparação entre, por exemplo, 28% de estruturas obtidas na imputação em TPOL1–P_0–P_1–P_2 com 30% de não resposta simulada é menor que os 62% de resultado da mesma perturbação em 50% de não resposta. Apesar de a medida de consistência estrutural para as redes construídas a partir dos dados do Censo Demográfico ter apresentado melhor desempenho em menores percentuais de não resposta, este fato

não se observou nos dados de homicídios de Campinas. Nesta rede como das demais desta base de dados, os resultados também se apresentaram ora altos, ora baixos, independente do percentual de não resposta simulado e do tipo de perturbação aplicada.

Na rede réu–crime–papéis–tempo, a avaliação da consistência estatística na imputação da variável TPOL1 foi observada para os valores da mediana e da média. Os comportamentos destas quantidades oriundos das 500 simulações de imputação da rede em questão podem ser visualizados nas Figuras 4.23 (a) e 4.23 (b) respectivamente. Assim como nos resultados obtidos para as outras duas redes, a mediana sofreu maior influência da imputação com relação ao percentual de não resposta e ao tipo de perturbação.

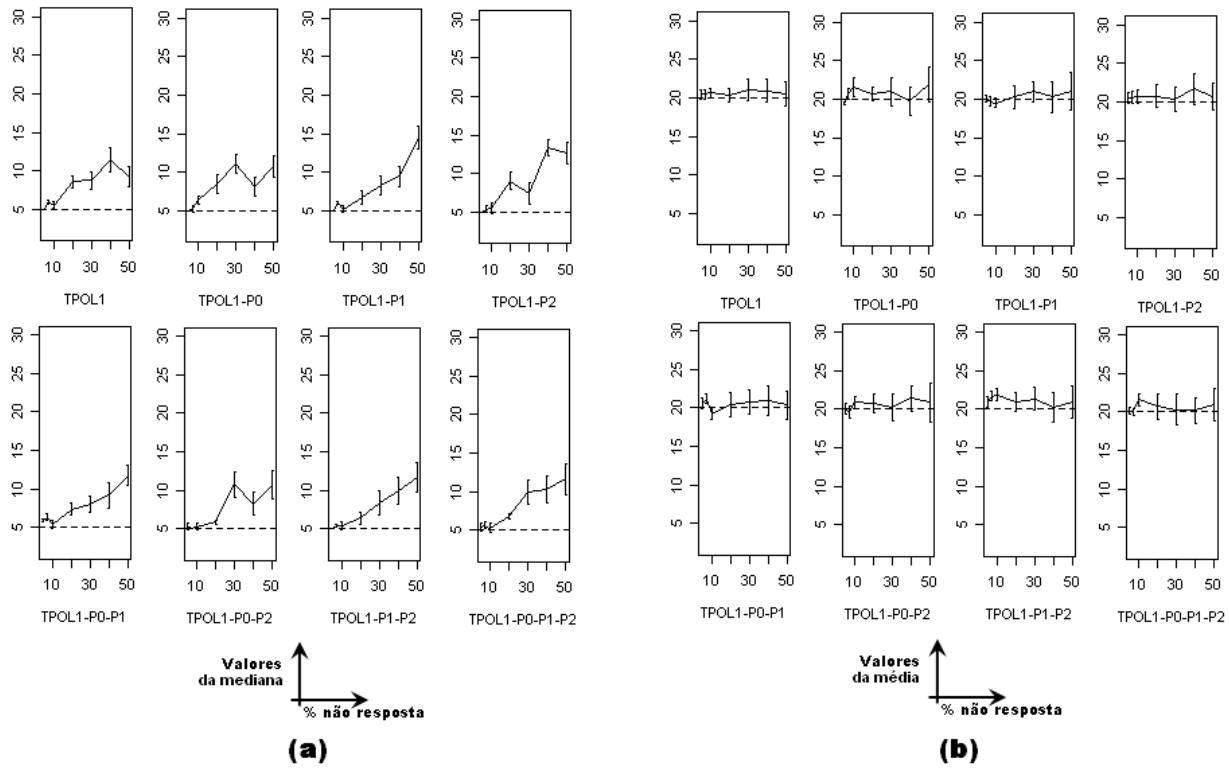


Figura 4.23: Resultados na consistência estatística a partir de 500 simulações da rede réu–crime–papéis–tempo para imputação na variável de tempo em diferentes percentuais de não resposta (a linha tracejada refere–se à quantidade calculada a partir da base original) (a) valores da mediana (b) valores da média

4.7 Conclusões e futuros direcionamentos

Sendo estes os primeiros resultados extraídos da imputação em variáveis contínuas a partir de redes Bayesianas, cabem algumas observações que passam desde o algoritmo proposto até os resultados das avaliações em função dos dados utilizados.

Em duas bases de dados diferentes, de características distintas, uma com poucos dados e outra em que é possível construir relacionamentos com um pouco mais de segurança, as medidas de consistência estatística apresentaram conclusões semelhantes em redes construídas a partir das duas bases de dados. Embora todas as redes tenham sido construídas a partir de variáveis discretas e contínuas, a avaliação se deu apenas no nó contínuo. Ressalta–se que os resultados oriundos de perturbação em outros blocos, além daquele que contivesse o nó contínuo, podem influenciar os desempenhos atingidos por este. Este fato ocorre em todas as conclusões e sabe–se que deve ser considerado em todos os itens citados a seguir.

Nas avaliações das consistências apresentadas na Seção 4.4, combinam–se em maior grau as idéias de Chambers (2000), por não haverem até então resultados na literatura da aplicação de redes Bayesianas mistas para imputação em variáveis aleatórias contínuas. As conclusões são apresentadas em sequência.

- o algoritmo para imputação;

  Na verdade, o algoritmo proposto para imputação em nós contínuos é o mesmo que o aplicado para redes discretas, certificando–se de que variáveis discretas não terão variáveis contínuas como pais, pela limitação do pacote computacional utilizado. Outra observação está na ordenação das variáveis dentro de cada subconjunto formado pela estrutura da rede de entrada. A convenção empregada neste texto foi a de que as variáveis contínuas seriam as últimas a serem imputadas em cada bloco. Não se sabe se este fato levaria a alguma influência sobre os resultados, uma vez que a construção da estrutura depende da ordenação em que se encontram os atributos.

- na consistência estrutural, os resultados são mais instáveis para bases pequenas;

  O que ocorreu nas redes sobre os dados de homicídios em Campinas induziu a essa conclusão. Em algumas situações de perturbação, uma maior proporção de redes na mesma classe de equivalência da rede original se deu para maiores percentuais de não resposta. Não se atribuiu este fato à complexidade da rede pois este comportamento foi recorrente em todas as estruturas. Já para os dados do Censo Demográfico, a estrutura $\mathcal{G}$ equivalente à rede de três nós (trabalhada na Seção 4.5.1) deu origem a 100% de estruturas $\widetilde{\mathcal{G}}$ em sua classe de equivalência. As demais estruturas tratadas com os dados do Censo Demográfico não apresentaram desempenho inferior a 70%, e parecem conduzir a piores valores nesta medida de consistência para maiores percentuais de não resposta. Intuitivamente, a consistência estrutural

é dependente da complexidade da estrutura da rede, quando a base de dados tiver um número muito grande de observações. Esta intuição é um reflexo da afirmação de Neapolitan (2004) de que a classe de equivalência de uma dada estrutura pode ser obtida para $n \to \infty$ em determinadas classes de modelos em redes discretas.

- na consistência da base de dados, os resultados do coeficiente de correlação entre o real observado e o imputado exprimem que a rede Bayesiana para imputação em variáveis aleatórias contínuas não é um bom método para manter os valores individuais da base;

  Esta afirmação teria mais alguns elementos que certamente justificariam melhor os resultados obtidos. Um deles é a estrutura da rede Bayesiana original e o outro, o percentual de não resposta nos dados. O primeiro por causa dos parâmetros da rede, que são responsáveis pelos elementos imputados e o segundo pelo percentual de valores que pode ser imputado conforme um parâmetro muito distante da sua distribuição.

  Como estes dois elementos não sobressaíram a ponto de se extrair comentários mais precisos a este respeito, justifica–se esta sentença comparando–se os valores de ρ nas redes do Censo Demográfico e dos dados de homicídios em Campinas. Nos percentuais de não resposta e tipos de perturbação simulados para as redes nos dados de homicídios, a correlação deu–se quase nula, e em alguns casos, a ocorrência de correlação negativa exprimiu que os itens individuais imputados seguiram a direção oposta do real observado. Nos dados do Censo Demográfico, atingiu–se uma correlação de fraca a moderada (próxima de 40%) em certas perturbações, tendo seus resultados em sua maioria interpretados como associação fraca entre o imputado e o real observado.

  Em geral, os métodos de imputação não buscam atender a essa característica de preservar os dados individuais, mas atingir uma correlação moderada em uma variável de renda com uma rede Bayesiana mista é um resultado que anima aos estudos neste tópico.

- ao se aumentar o percentual de não resposta, o valor de ξ_{X_j} construído a partir de (4.4.7) tende a zero;

  Antes de se optar pelo coeficiente de correlação proposto por Chambers (2000) para avaliar imputações em variáveis quantitativas, estudou–se um conjunto de outras possibilidades que iam desde a soma de diferenças simples entre x_{ij} e $\widetilde{x}_{ij}$, até a soma dos módulos das diferenças entre o real observado e o imputado. Dentre

essas outras possibilidades surgiu o valor de $\xi_{X_j} = \frac{\sum_{i=1}^{n^*}(x_{ij}-\widetilde{x}_{ij})^2}{n^*}$ que foi calculado em todas as redes, mas só foi apresentado para as aplicações dos dados do Censo Demográfico. Avaliando esta quantidade após a simulação, por exemplo na Figura 4.12 (a), observa–se um acentuado decaimento e tende a se estabilizar próximo de zero à medida que se aumenta o percentual de não resposta. Embora tenhamos visto que a correlação entre X_j e $\widetilde{X}_j$ é fraca (e algumas vezes, nula), a característica compensatória entre os valores imputados permite manter a regularidade em estimativas subsequentes, como ocorre com a média.

- a distribuição da variável quantitativa imputada permanece inalterada até um percentual de não resposta em torno de 10%.

  Essa afirmação independe do número de unidades da base original, pois foi um comportamento observado na consistência estatística em todas as estruturas consideradas. A média da variável permaneceu estável e sempre em torno do valor da média calculado com os dados originais. Os valores da mediana calculados após a imputação apresentaram comportamento ascendente, sempre na direção da média. Quanto a isso, reporta–se à suposição de normalidade condicional do nó contínuo em seus pais. Apesar da transformação logaritmo efetuada na variável renda nos dados do Censo Demográfico, a consistência foi verificada em sua unidade original, e isso demostrou o quão forte está a suposição da distribuição no nó.

  À medida que mais valores da distribuição suposta são inseridos na base de dados, mais a distribuição resultante vai se aproximando da distribuição condicional Gaussiana suposta no nó. Como aspecto complementar, observou–se que a mediana é uma quantidade sensível à esta suposição.

Embora estes resultados sejam preliminares para redes mistas e específicos para determinadas situações, reforça–se que sua importância está em sua recorrência e por isso encontram–se aqui registrados. Não obstante os itens citados, alguns outros tópicos seriam importantes de serem mencionados, como o padrão de referência limitante dos tempos nos dados de homicídios de Campinas. Nas simulações conduzidas na Seção 4.6, os parâmetros da rede original foram estimados dos dados disponíveis, mas se tivessem sido utilizados os padrões do Código do Processo Penal, a variável TMP, por exemplo, deveria ter um tempo máximo de trinta dias enquanto que o tempo mediano observado foi de quarenta dias. Esse fato pode ocorrer em quaisquer outras bases de dados em que os parâmetros de entrada sejam definidos por especialistas. Quando se trata de avaliar os parâmetros

dentro da configuração dos pais do nó contínuo, pode–se amenizar um pouco os efeitos dos padrões de referência.

Buscando uma analogia do método de imputação a partir das redes Bayesianas com o desenvolvido por Pessoa e Santos (2004) e Pessoa, Moreira e Santos (2004) nos dados do Censo Demográfico, a proposição do uso de árvores de regressão (BREIMAN et al., 1984) teve o intuito de que em cada nó a probabilidade de não resposta fosse uniforme, onde seria então utilizado um método hot–deck para imputação a partir de um doador da renda. Na rede Bayesiana, a configuração dos pais delimitam classes de imputação onde, para cada uma destas, a distribuição da variável a imputar (no caso, LOG(RENDA)) fosse normal para o pacote utilizado nas aplicações. Não se verificou o comportamento da não resposta dentro de cada configuração de pais, mas este seria um bom exercício de trabalho futuro. A simulação da não resposta do tipo MCAR foi feita para a variável como um todo, o que não garante ter permanecido com a mesma característica após a divisão dentro dos nós.

5 *Avaliação da consistência estatística em redes Bayesianas mistas a partir de imputação múltipla*

Para uma avaliação mais profunda do impacto do método de imputação em quantidades de interesse será empregada a imputação múltipla (RUBIN, 1987) de forma preliminar. Conforme Rubin (1996), a imputação múltipla combina resultados de análises em múltiplas bases de dados imputados.

Este capítulo traz duas aplicações da imputação múltipla para mensurar a variabilidade associada às redes Bayesianas como método de imputação[1]. Uma delas avalia os parâmetros de um modelo de regressão linear normal nos pais dos nós contínuos na rede, na tentativa de se dispor de informações relevantes com respeito à influência da imputação por redes Bayesianas. Na Seção 5.2 avaliam–se as médias associadas às variáveis contínuas das redes mistas. A outra aplicação refere–se aos parâmetros do modelo de riscos proporcionais (COX, 1972) quando tratada a imputação a partir de redes Bayesianas sob um contexto em que se permite o aproveitamento da informação parcial da variável resposta (Seção 5.3).

Na Seção 2.5 apresentou–se de forma resumida os principais tópicos de imputação múltipla que serão utilizados. Para a condução das simulações foram escolhidas duas redes mistas que foram trabalhadas no capítulo anterior: a rede domicílio–pessoa–renda (Seção 4.5.3) e a rede réu–crime–papéis–tempo (Seção 4.6.3).

[1]Os resultados preliminares apresentados neste capítulo ajudam a compor um artigo que, juntamente a um tratamento mais detalhado da técnica, será submetido para publicação ao *Journal of Statistical Computations and Simulation*.

5.1 Introdução

Imputação múltipla (RUBIN, 1987) é uma técnica na qual cada valor faltante é substituído por $m > 1$ valores simulados (SCHAFER, 1997a). Os m conjuntos de imputações refletem a incerteza sobre os verdadeiros valores dos itens perdidos. Depois que as imputações múltiplas são criadas, tem–se m versões da base de dados completa e cada uma delas pode ser analisada a partir de uma técnica padrão de análise estatística para dados completos. Os resultados das m análises são então combinados para produzir um valor inferencial simples (por exemplo, um p–valor ou um intervalo de confiança), que inclui a incerteza devido a não resposta.

As fases de imputação e análise podem ser desenvolvidas em ocasiões diferentes e por diferentes pessoas. Este é o caso, por exemplo, dos institutos oficiais de estatística que têm as bases de dados de uso público sobre alguma pesquisa como um de seus produtos. Dessa maneira, o produtor do dado e o pesquisador ou usuário tornam–se diferentes. De acordo com Schafer (1997a) esta separação temporal entre as fases da imputação múltipla é uma das vantagens de sua inferência porque os aspectos do tratamento da não resposta passam a ser de inteira responsabilidade da primeira fase, e não há necessidade de nenhum procedimento especial para considerar os itens faltantes quando se executar a fase de análise.

Rubin (1996) cita que a imputação múltipla é a técnica que deve ser escolhida no caso do tratamento da não resposta quando o usuário final e o produtor da base de dados são entidades distintas. Esta afirmação se deve ao fato de que o usuário final em geral não possui conhecimento específico sobre os elementos teóricos necessários ao tratamento da não resposta. Além disso, apenas uma única base de dados imputada desconsideraria a incerteza que estaria associada aos dados faltantes. Segundo Schafer (1997a), um conjunto de m imputações pode efetivamente resolver os problemas devido aos dados faltantes para um grande número de futuras análises.

Na fase de imputação, imputações repetidas são obtidas da distribuição preditiva a priori dos valores faltantes sob um modelo especificado (RUBIN, 1987) (RUBIN, 1996), ou seja, a partir de um particular modelo Bayesiano para os dados e o mecanismo da não resposta. Este modelo é representado graficamente pela estrutura $\mathcal{G}$ da rede Bayesiana e seus parâmetros indicam as probabilidades de ocorrência de uma variável aleatória em cada combinação de classes de seus pais, no caso de atributos discretos, ou da sua média mais provável, no caso de atributo contínuo.

Conforme definido pelo algoritmo proposto neste texto, a combinação do conhecimento de especialistas com a estrutura gráfica que codifica a independência condicional entre as variáveis, direcionam a imputação sob os itens faltantes.

Na fase de análise, uma técnica convencional da análise estatística é aplicada às m bases imputadas e a fase de combinação utiliza–se da Equação (2.5.1) para a obtenção da estimativa de θ de interesse. A variância de $\widehat{\theta}$ será estimada por (2.5.4), uma vez que são calculadas as variâncias dentro de cada base e entre as imputações.

De acordo com Rubin (1987), não é necessário um valor muito grande de m conforme demonstrado na Seção 2.5. Portanto, ao avaliar o resultado esperado do erro padrão na fração de dados faltantes de $\lambda = 50\%$, decidiu–se utilizar $m = 20$ nas aplicações conduzidas na sequência.

Todas as simulações executadas deram–se a partir dos pacotes computacionais *deal* (BØTTCHER; DETHLEFSEN, 2003) para o ajuste das redes Bayesianas de entrada no algoritmo de imputação e da *mitools* (LUMLEY, 2004) que foi o dispositivo utilizado para analisar e combinar os resultados dos m bancos de dados imputados.

5.2 Modelos de regressão linear

Tipicamente os usuários finais de bases de dados têm disponíveis várias rotinas que permitem obter regressão linear simples ou múltipla, regressão logística, análise fatorial, modelos de riscos proporcionais e outras, com a característica principal de excluir da análise as unidades que contenham pelo menos uma não resposta ao item. Isso é o que ocorre em vários pacotes computacionais presentes no mercado.

A escolha de uma avaliação dos parâmetros de um modelo de regressão linear através da imputação múltipla agrega a influência do método de imputação e a visualização do que ocorre simultaneamente com os parâmetros da rede Bayesiana, responsáveis pela imputação nas variáveis contínuas quando sujeitos a diferentes percentuais de não resposta.

O estudo refere–se a um modelo do tipo:

$$\widetilde{X}_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i; \qquad i = 1, \ldots, n,$$

que será realizado de três formas:

1. avaliando a média da variável contínua imputada, através de $\overline{\mu}(\widetilde{X}) = \widehat{\beta}_0$, supondo

$e_i \sim N(0, \sigma^2)$;

2. avaliando os parâmetros de imputação nas configurações dos pais discretos da variável contínua a ser imputada, a partir de $\widetilde{X}_{ij} = \widehat{\beta}_{0,X_j|pa_D(X_j)}$, supondo $e_{ij} \sim N(0, \sigma^2_{pa_D(X_j)})$; e

3. avaliando a contribuição dos pais da variável contínua imputada, através de $\widetilde{X}_i = \widehat{\beta}_j X_{ij|pa_D(\widetilde{X})}$, supondo $e_i \sim N(0, \sigma^2_{\widetilde{X}|pa_D(\widetilde{X})})$.

Como as redes tratadas neste capítulo apresentam apenas pais discretos, a avaliação das demais situações ficam sugeridas como trabalhos futuros.

Para obter os $m = 20$ valores imputados para cada item faltante executou–se o algoritmo proposto no Quadro 5 e as perturbações são simuladas apenas nas variáveis contínuas.

5.2.1 Média e parâmetros da rede domicílio–pessoa–renda

A rede domicílio–pessoa–renda foi trabalhada para a imputação da renda em dados de domicílios do Censo Demográfico 2000. A estrutura desta rede pode ser vista na Figura 5.1 (a) que identifica o seu grafo e detalha a parte modelada nesta seção. Na Figura 5.1 (b) mostra–se o comportamento da distribuição da variável renda antes e depois da transformação logaritmo.

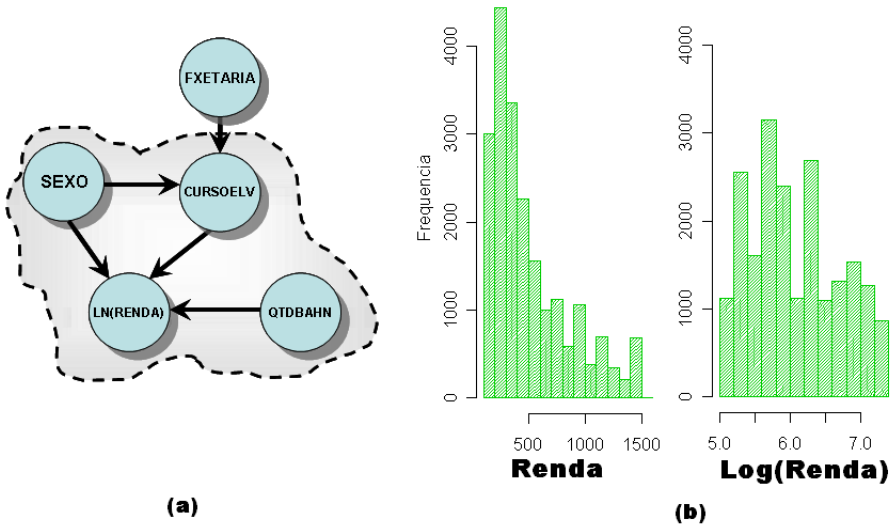


Figura 5.1: (a) Grafo da rede domicílio–pessoa–renda para imputação da variável renda nos dados do Censo Demográfico. Em destaque estão as variáveis avaliadas por imputação múltipla (b) Comportamento da variável renda e de sua transformada nos dados do Censo Demográfico

As imputações múltiplas próprias são execuções independentes de $\mathbf{X}_{per}$ da distribuição a priori dos dados faltantes, $P(\mathbf{X}_{per}|\mathbf{X}_{obs})$, onde $\mathbf{X}_{per}$ é o conjunto de variáveis que contêm itens faltantes e $\mathbf{X}_{obs}$ é o conjunto de variáveis com valores observados. Conforme Rubin (1996) e Schafer (1997a) obtêm–se valores plausíveis para $\mathbf{X}_{per}$ a partir de *Markov Chain Monte Carlo* (MCMC) (TANNER, 1993) (GILKS; RICHARDSON; SPIEGELHALTER, 1996) ou *data–augmentation* (TANNER; WONG, 1987), tendo como parâmetros iniciais os estimadores de máxima verossimilhança de θ e finalizando sua execução após, por exemplo, dez passos de iteração da cadeia. O valor final de $\mathbf{X}_{per}$ em cada cadeia seria o seu valor imputado.

No caso da rede Bayesiana, $\widehat{\theta}_{X_j|pa(X_j)}$, fornecido por especialistas ou estimados por base de teste, é o vetor de parâmetros iniciais do processo. A partir destes são simuladas as m bases de dados. No caso de redes em que os parâmetros fossem atualizados após cada imputação (ou "ativas"), poder–se–ia avaliar a convergência destes no processo.

O interesse primeiramente está em $\overline{\mu}(\widetilde{X}) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i$, onde $\widetilde{X}$ é a variável renda. Relembra–se que a média dos dados observados é de R$ 532,00 reais. O cálculo da média em cada percentual de não resposta é feito após a transformação inversa do item, na unidade original do atributo. A Tabela 14 traz os resultados da inferência por imputação múltipla para a média da variável renda após $m = 20$ imputações da base, a partir da rede Bayesiana identificada na estrutura $\mathcal{G}$ da Figura 5.1 (a). O que se observa é uma relativa estabilidade no valor estimado da média à medida que se aumenta o percentual de não resposta, fato este já identificado antes na avaliação da média na consistência estatística conduzida na Seção 4.5.3 com base em quinhentas imputações na simulação aplicada aos dados. Os valores de ζ são grandes, sugerindo que as estimativas da variância do total são estáveis, principalmente se pensar que estão baseadas em vinte imputações apenas.

A Tabela 14 também mostra duas medidas de diagnóstico: o aumento relativo na estimativa da variância devido a não resposta ($\widehat{r}$) e a fração estimada de informação faltante que afeta o valor da estimativa do parâmetro ($\widehat{\lambda}$). Os baixos valores de $\widehat{r}$ e $\widehat{\lambda}$ mostram a baixa influência do método de imputação na estimativa da variância. O aumento relativo na estimativa da variância devido a não resposta é obtido por:

$$\widehat{r} = \frac{(1 + m^{-1})Var_e(\widehat{\theta})}{Var_d(\widehat{\theta})}, \tag{5.2.1}$$

e quando não há informação perdida sobre $\widehat{\theta}$, os valores de $Var_e(\widehat{\theta})$ e $\widehat{r}$ são nulos (RUBIN, 1987). Quando o $\widehat{r}$ calculado é muito pequeno, o número de graus de liberdade será grande, o que dá uma indicação da normalidade na distribuição de teste sobre $\widehat{\theta}$ (SCHAFER,

Tabela 14: Inferências com base em imputação múltipla ($m = 20$) para a média da renda imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico. Perturbações aplicadas apenas na variável RENDA

Percentual de não resposta na variável	$\widehat{\theta} = \overline{\mu}(\widetilde{X})$	$\sqrt{Var_t(\widehat{\theta})}$	ζ	intervalo 95% confiança	$\widehat{r}$	$\widehat{\lambda}$
1%	532,427	352,412	$2{,}424 \times 10^{14}$	[0 ; 1.223,155]	$2{,}800 \times 10^{-07}$	$8{,}251 \times 10^{-15}$
3%	532,189	352,187	$1{,}132 \times 10^{13}$	[0 ; 1.222,476]	$1{,}296 \times 10^{-06}$	$1{,}767 \times 10^{-13}$
5%	532,786	351,671	$2{,}505 \times 10^{12}$	[0 ; 1.222,061]	$2{,}754 \times 10^{-06}$	$7{,}984 \times 10^{-13}$
7%	533,251	352,131	$3{,}159 \times 10^{12}$	[0 ; 1.222,427]	$2{,}454 \times 10^{-06}$	$6{,}331 \times 10^{-13}$
10%	533,946	351,179	$5{,}223 \times 10^{11}$	[0 ; 1.222,256]	$6{,}031 \times 10^{-06}$	$3{,}829 \times 10^{-12}$
20%	534,509	347,130	$1{,}232 \times 10^{11}$	[0 ; 1.214,884]	$1{,}242 \times 10^{-05}$	$1{,}623 \times 10^{-11}$
30%	535,057	346,574	$1{,}976 \times 10^{11}$	[0 ; 1.214,342]	$9{,}806 \times 10^{-06}$	$1{,}012 \times 10^{-11}$
40%	536,981	346,848	$1{,}348 \times 10^{11}$	[0 ; 1.216,803]	$1{,}187 \times 10^{-05}$	$1{,}484 \times 10^{-11}$
50%	537,512	346,093	$4{,}401 \times 10^{09}$	[0 ; 1.215,855]	$2{,}078 \times 10^{-05}$	$4{,}544 \times 10^{-11}$

1997a). Outra estatística útil está na fração estimada de informação faltante que afeta o valor da estimativa do parâmetro, ou $\widehat{\lambda}$, que é obtida por:

$$\widehat{\lambda} = \frac{(r+2)/(\zeta+3)}{(r+1)}. \tag{5.2.2}$$

Apesar de os percentuais de não resposta serem de até 50%, a fração estimada de dados faltantes calculada na Tabela 14 abaixo de 0,1%, mostra que as correlações com as demais variáveis da rede fornecem informação adequada sobre a renda. A Figura 5.2 (a) mostra as estimativas da renda derivadas a partir da imputação múltipla após a imputação da não resposta pelo método da rede Bayesiana. No gráfico pode–se observar melhor o comportamento da variância total da estimativa obtida pela combinação entre a variância dentro da base e a estimação entre as bases, respectivamente calculadas pelas Equações (2.5.9) e (2.5.10). Já no gráfico da Figura 5.2 (b), verifica–se a evolução de $\sqrt{Var_t(\widehat{\theta})}$ e $Var_e(\widehat{\theta})$ no cálculo das estimativas da média da variável renda por imputação múltipla em diversos percentuais de não resposta. O comportamento de $Var_d(\widehat{\theta})$ assemelha–se à variância total e, como era de se esperar, a variância entre as imputações aumenta à medida que se aumenta o percentual de não resposta simulado. Na Tabela D.1 do Apêndice D, encontra–se uma extensão da Tabela 14 com os valores particionados das variâncias obtidos nesta simulação por imputação múltipla.

A segunda aplicação da imputação múltipla aos dados imputados através da rede Bayesiana avalia os parâmetros de imputação da rede para a variável renda. Tendo como pais os atributos SEXO, CURSOELV e QTDBAHN, a imputação da renda de

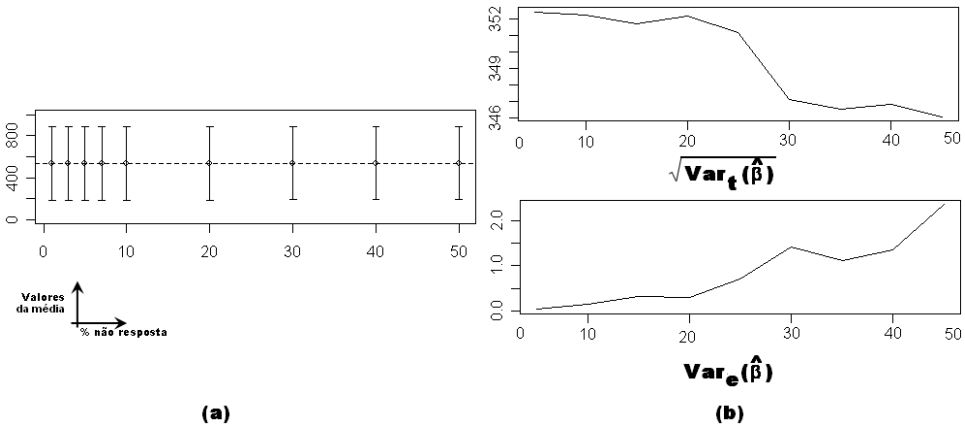


Figura 5.2: (a) Estimativas da média da variável renda por imputação múltipla em variados percentuais de não resposta (b) Evolução da $\sqrt{Var_t(\widehat{\theta})}$ e $Var_e(\widehat{\theta})$ no cálculo das estimativas da média da variável renda por imputação múltipla em variados percentuais de não resposta

Tabela 15: Parâmetros da rede Bayesiana mista domicílio–pessoa–renda avaliados a partir de imputação múltipla

Variáveis			**Parâmetros**	
SEXO	QTDBAHN	CURSOELV	$\widehat{\beta}_{0,X_j\|pa_D(X_j)}$	$\sigma^2_{pa_D(X_j)}$
1 – Masculino	0	1 – nenhum	5,589	0,200
		2 – básico/fundamental	5,651	0,193
		3 – médio/ 2º grau	5,780	0,266
		4 – superior/ pós graduação	6,234	0,475

acordo com as relações dadas pela Figura 5.1(a) respeita a probabilidade condicional P(LN(RENDA)|SEXO,CURSOELV,QTDBAHN). Como a variável contínua modelada no nó renda recebeu a transformação logaritmo antes da imputação, os valores considerados como parâmetros foram $\log(\widetilde{X}_{ij}) = \widehat{\beta}_{0,X_j|pa_D(X_j)}$, supondo $e_{ij} \sim N(0, \sigma^2_{pa_D(X_j)})$. Foram avaliados os valores na sua unidade transformada.

A combinação entre as categorias dos pais de LOG(RENDA) resultam em trinta e dois parâmetros diferentes. Condicionado ao observado ou imputado nas variáveis SEXO, CURSOELV e QTDBAHN, um valor para LOG(RENDA) é gerado conforme média e variância especificadas. O item é imputado após a transformação inversa equivalente. A Tabela D.2 do Apêndice D quantifica os valores de $\widehat{\beta}_{0,X_j|pa_D(X_j)}$ e de $\sigma^2_{pa_D(X_j)}$ em todos os cruzamentos das classes nos pais do atributo renda. O uso da imputação múltipla dar–se–á na parte da Tabela D.2 reproduzida na Tabela 15. A inferência para os demais parâmetros seguiriam os mesmos procedimentos.

De acordo com a Equação (2.5.8), a estimativa dos parâmetros é obtida com base

Tabela 16: Estimativas com base em imputação múltipla ($m = 20$) para os parâmetros da rede e sua variância total no nó contínuo da variável LOG(RENDA) imputada (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	**Estimativas de** $\widehat{\boldsymbol{\beta}}_{\mathbf{0},\boldsymbol{X}_{j\|pa_D(X_j)}}$				$\boldsymbol{Var_t}(\widehat{\boldsymbol{\beta}}_{\mathbf{0},\boldsymbol{X}_{j\|pa_D(X_j)}})$			
	$\widehat{\beta}_{0,101}$	$\widehat{\beta}_{0,102}$	$\widehat{\beta}_{0,103}$	$\widehat{\beta}_{0,104}$	$\widehat{\sigma}^2_{101}$	$\widehat{\sigma}^2_{102}$	$\widehat{\sigma}^2_{103}$	$\widehat{\sigma}^2_{104}$
1%	5,589	5,650	5,782	6,241	0,217	0,199	0,265	0,472
3%	5,591	5,648	5,763	6,266	0,201	0,190	0,262	0,465
5%	5,602	5,792	5,834	6,307	0,195	0,183	0,254	0,460
7%	5,618	5,817	5,889	6,351	0,192	0,177	0,251	0,458
10%	5,631	5,943	5,910	6,419	0,183	0,164	0,243	0,451
20%	5,707	5,999	5,985	6,500	0,181	0,160	0,243	0,448
30%	5,795	6,071	6,102	6,598	0,165	0,151	0,237	0,435
40%	5,900	6,132	6,117	6,637	0,157	0,148	0,229	0,430
50%	6,004	6,216	6,268	6,669	0,144	0,146	0,220	0,427

na média simples das imputações e sua variância total, estimada por (2.5.11), é uma combinação entre $Var_d(\widehat{\theta})$ e $Var_e(\widehat{\theta})$. A Tabela 16 apresenta os resultados da inferência por imputação múltipla para os parâmetros da rede na variável contínua LOG(RENDA) após $m = 20$ imputações a partir da rede Bayesiana identificada na estrutura $\mathcal{G}$ da Figura 5.1 (a). Para efetuar essa avaliação considerou–se a estrutura fixa da rede, de modo que $\widetilde{\mathcal{G}} = \mathcal{G}$ em todas as imputações. Esse procedimento serviu para garantir que todos os parâmetros da rede pudessem ser estimados em $\widehat{\theta}$.

Na Tabela 16, $\widehat{\beta}_{0,101}$ equivale ao parâmetro de imputação na variável transformada para o subconjunto de valores pertencentes ao sexo masculino (SEXO = 1), que não dispõem de banheiro no domicílio (QTDBAHN = 0) e que não têm instrução (CURSOELV = 1). O mesmo ocorre aos demais parâmetros, alternando apenas a classe da variável CURSOELV no terceiro dígito subscrito em $\widehat{\beta}_0$.

Mais uma vez, combinar as estimativas das m bases de dados, é uma aplicação direta das equações apresentadas na Seção 2.5. O que se observa na Figura 5.3 (a) do comportamento das estimativas quando se aumenta o percentual de não resposta é que os parâmetros da rede para esse subconjunto permanece inalterado ao longo das perturbações aplicadas. O comportamento dos parâmetros estimados pode ser acompanhado na Figura 5.3 (a). A Figura 5.3 (b) traz a variância total das estimativas de $\widehat{\beta}_{0,X_j|pa_D(X_j)}$ e da componente da variância que quantifica a variabilidade dependente do método de imputação, $Var_e(\widehat{\theta})$. Como esperado, a parcela que representa a contribuição da rede Bayesiana para imputação nos parâmetros da rede aumenta à medida que se aumenta o percentual de não

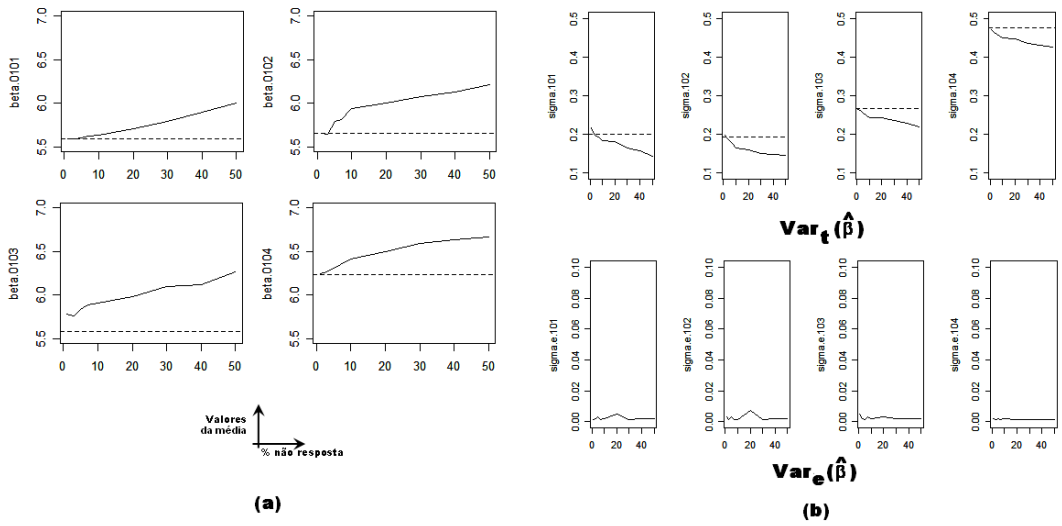


Figura 5.3: (a) Estimativas dos parâmetros de imputação da variável LOG(RENDA) por imputação múltipla em variados percentuais de não resposta (b) Evolução da $Var_t(\widehat{\theta})$ e $Var_e(\widehat{\theta})$ no cálculo das estimativas dos parâmetros da rede por imputação múltipla em variados percentuais de não resposta

resposta simulada nos dados e maior quantidade de itens passa a depender da distribuição no nó contínuo.

Nesta aplicação também foram calculados os valores de $\widehat{r}$, ou o aumento relativo na estimativa da variância devido a não resposta e o $\widehat{\lambda}$, que mede a fração de informação faltante estimada. Nos parâmetros da rede, embora os percentuais de não resposta simulados estejam entre 1% e 50%, os valores de $\widehat{\lambda}$ estimados a partir de imputação múltipla não ultrapassaram 31%, indicando que a dependência entre a variável LOG(RENDA) e seus pais fornecem informação adequada sobre os itens imputados. As quantidades calculadas de $\widehat{r}$ apontam um aumento relativo na estimativa da variância entre 0,001 e 0,030, ou seja, entre 0,1% e 3% de aumento na variância pode ser creditada à não resposta. Os valores destas simulações podem ser visualizadas na Tabela D.3 no Apêndice D.

5.2.2 Regressão linear a partir da rede domicílio–pessoa–renda

Avaliam–se os parâmetros da regressão de $\widetilde{X}$ sob seus pais SEXO, CURSOELV e QTDBAHN. Conforme observado na Figura 5.1 (b) a não normalidade da variável RENDA afetaria a suposição de condicionais Gaussianas nos nós contínuos. Por isso, procedeu–se a avaliação a partir da variável transformada LOG(RENDA). Apesar de diminuir a assimetria, essa transformação não resolve o problema da normalidade condicional e a utilização desta variável novamente está relacionada com a indisponibilidade

Tabela 17: Ajuste do modelo de regressão linear tendo LN(RENDA) como variável dependente e SEXO, CURSOELV e QTDBAHN como preditivas para os dados observados do Censo Demográfico

Variável	**Estimativa** $(\widehat{\boldsymbol{\beta}})$	$\boldsymbol{Var_t(\widehat{\beta})}$	$\boldsymbol{t}$	**p–valor**
SEXO1 $(\widehat{\beta}_1)$	5,553	2,231	21,418	$< 2 \times 10^{-16}$
SEXO2 $(\widehat{\beta}_2)$	5,423	2,249	18,085	$< 2 \times 10^{-16}$
CURSOELV1 $(\widehat{\beta}_3)$	4,150	2,064	7,541	$4,850 \times 10^{-14}$
CURSOELV2 $(\widehat{\beta}_4)$	4,862	2,081	14,845	$< 2 \times 10^{-16}$
CURSOELV3 $(\widehat{\beta}_5)$	5,951	2,142	38,736	$< 2 \times 10^{-16}$
QTDBAHN2 $(\widehat{\beta}_6)$	4,709	2,159	10,790	$< 2 \times 10^{-16}$
QTDBAHN3 $(\widehat{\beta}_7)$	5,824	2,205	29,766	$< 2 \times 10^{-16}$
QTDBAHN4 $(\widehat{\beta}_8)$	6,076	2,295	31,232	$< 2 \times 10^{-16}$

de outra na base de dados do Censo Demográfico para o ajuste da rede mista. Lembra–se que o corte efetuado também minimiza o efeito da acentuada assimetria da variável.

Dessa maneira, a aplicação da imputação múltipla será no modelo:

$$\begin{aligned} \widetilde{Z}_i &= \text{LOG(RENDA)} = (\beta_1 + \beta_2)\text{SEXO}_i \\ &+ (\beta_3 + \beta_4 + \beta_5)\text{CURSOELV}_i \\ &+ (\beta_6 + \beta_7 + \beta_8)\text{QTDBAHN}_i + e_i, \qquad i = 1, \ldots, n. \end{aligned} \tag{5.2.3}$$

Os valores estimados dos $\beta^{'}$s encontram–se na Tabela 17. Cabe observar que as quantidades na Tabela 17 não possuem interpretação direta. A interpretabilidade dos parâmetros é dada após a transformação inversa dos mesmos. Não foram verificados efeitos de outras ordens ou interações significativas entre os atributos explicativos considerados e a variável renda transformada.

Para cada uma das $m = 20$ bases de dados imputados, foram obtidas as estimativas de máxima verossimilhança para os elementos de β e então combinados os resultados através das equações para inferência por imputação múltipla apresentadas na Seção 2.5. O comportamento dos valores de $\widehat{\lambda}$ aparece na Figura 5.4. Em todos os parâmetros da regressão, as quantidades estimadas de não resposta que os afetam são inferiores àquelas geradas, que estão representadas pelas linhas tracejadas em cada gráfico da Figura 5.4.

O padrão de comparação dos valores simulados dos parâmetros e de suas variâncias está no modelo ajustado a partir dos dados originais. Os resultados apresentados na Figura 5.5 mostram que o aumento relativo na estimativa da variância foi constante em todos os percentuais de não resposta simulados. As demais estimativas figuram nas Tabelas D.4 a

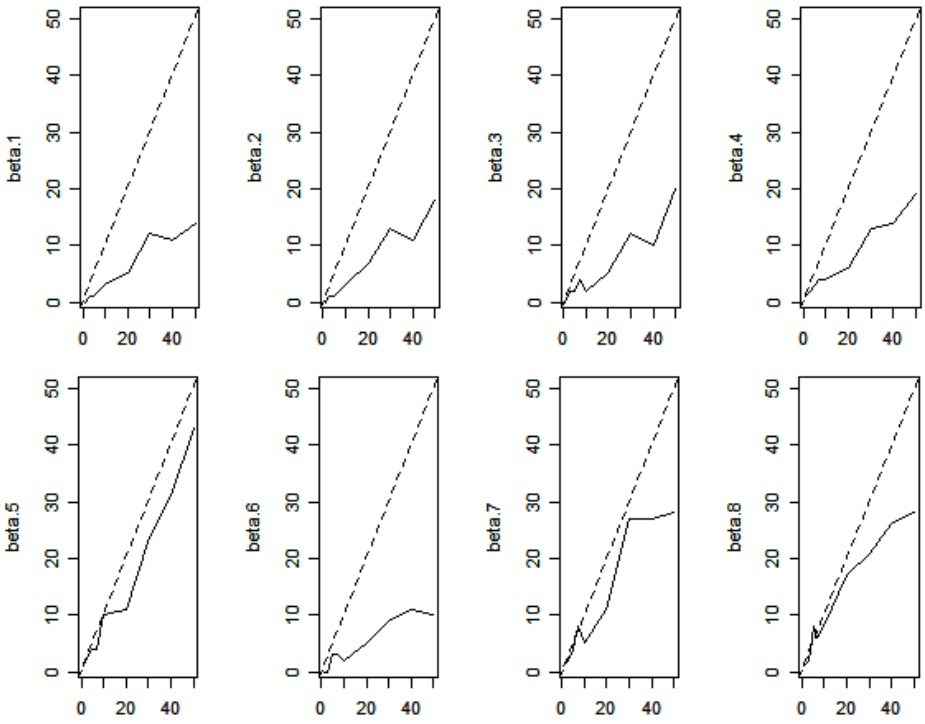


Figura 5.4: Estimativas dos valores de $\widehat{\lambda}$ *versus* os valores simulados de percentual de não resposta nos dados do Censo Demográfico

D.7, do Apêndice D.

5.3 Modelos de sobrevivência na rede réu–crime–papéis–tempo

Um dos objetivos em analisar os dados de tempo na base de homicídios em Campinas estava em obter atributos influentes nas variadas fases do sistema de justiça e concatenar as variáveis de tempo num fluxo contínuo de informações. O objetivo em tratar o tempo de forma contínua está em considerar unidades que abandonam o sistema por qualquer motivo (VARGAS, 2006). Um exemplo deste tipo de informação incompleta está no caso em que o réu vem a falecer na fase de investigação do crime, sendo o seu processo arquivado sem o decurso das demais fases do sistema. Outro exemplo, este típico de casos de crime de estupro, está naquelas unidades em que a vítima retira a acusação em alguma etapa do processo (VARGAS, 2004).

Apesar de não se encontrarem completas, estas unidades contêm uma informação com respeito à variável "tempo até a ocorrência de algum evento" que, se for desconsiderada, pode trazer vícios para as estimativas da análise em questão. Nestes casos não

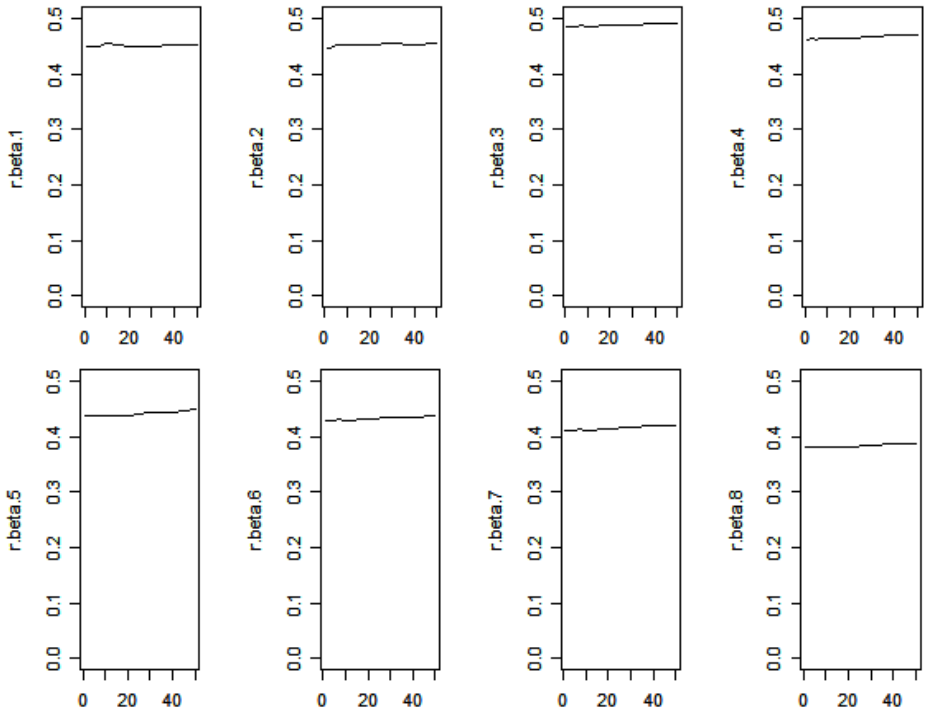


Figura 5.5: Estimativas dos valores de $\widehat{r}$ *versus* os valores simulados de percentual de não resposta nos dados do Censo Demográfico

convém imputar o tempo restante, pois o dado parcial traz elementos de conhecimento susceptível ao fato analisado.

Uma técnica estatística que permite avaliar dados sobre tempo até a ocorrência de algum evento de interesse, incluindo aquelas unidades cuja informação está parcial, é a análise de sobrevivência (COLLETT, 1994) (KLEIN; MOESCHBERGER, 1997). Utilizada em larga escala em dados de pesquisa clínicas, onde as amostras em geral são muito pequenas, a análise de sobrevivência trata os dados incorporando a informação parcial através das funções de sobrevivência e de risco (BLAVATSKY, 2002). Na classe dos modelos, Cox (1972) estima os parâmetros através da função de verossimilhança parcial e estabelece uma classe de modelos semi–paramétricos, também chamada de modelos de Cox ou de modelos de riscos proporcionais (COX, 1972). Estes modelos foram aplicados aos dados de homicídios em Campinas (VARGAS, 2004) (VARGAS, 2006) em um trabalho que trata da análise do fluxo de tempo contínuo do sistema de Justiça. O modelo a ser tratado tem a forma:

$$h(t) = h_0(t) exp\Big\{ \sum_j \beta_j X_j \Big\}, \tag{5.3.4}$$

Tabela 18: Ajuste do modelo de riscos proporcionais tendo TPOL1 como variável resposta e SEXO, CRIME e PRISAO como preditivas para os dados observados de homicídios em Campinas

Variável	**Estimativa** $(\widehat{\beta})$	$\boldsymbol{exp(\widehat{\beta})}$	**d.p.**$(\widehat{\beta})$	z	**p–valor**
SEXO $(\widehat{\beta}_1)$	-0,977	0,376	0,499	-1,960	0,050
CRIME $(\widehat{\beta}_2)$	0,443	1,557	0,231	1,920	0,055
PRISAO $(\widehat{\beta}_3)$	0,830	2,294	0,233	3,570	$< 0{,}001$

onde $h(t)$ é a chamada função de risco que é modelada em função da variável de tempo e das covariáveis de interesse e $h_0(t)$ é a função de risco básica que engloba todos os componentes não paramétricos do modelo (KLEIN; MOESCHBERGER, 1997). Esta função em geral não é estimada e deixa de ser um parâmetro de perturbação quando do uso da função de verossimilhança parcial para estimar os valores de β (COLLETT, 1994).

A variável de tempo modelada através da análise de sobrevivência é TPOL1, que representa o tempo decorrido em dias entre o registro da ocorrência e a abertura do inquérito policial. Nos dados disponíveis, não ocorre a censura da observação e a não resposta é gerada aos percentuais de 5%, 7%, 10%, 20%, 30%, 40% e 50%. Após $m = 20$ imputações da base seguindo ao grafo da Figura 4.21, uma análise destas m bases imputadas foi realizada de acordo com o descrito na Seção 2.5.

O ajuste do modelo de riscos proporcionais para a variável TPOL1 é apresentado na Tabela 18, que traz os valores dos parâmetros para a análise sobre as variáveis presentes no grafo da Figura 4.21 no Capítulo 4.

Uma das interpretações dos parâmetros na Equação (5.3.4) é o de verificar o quanto as covariáveis aceleram ou desaceleram a função de risco. Como exemplo, o valor de $exp(\widehat{\beta}_3) = 2,294$ na Tabela 18, refere–se ao risco associado à ocorrência de abertura do inquérito policial para réus que já se encontram presos é quase 2,3 vezes maior que o risco de réus que estão soltos durante o processo. Essa conclusão é feita quando as demais covariáveis no modelo (SEXO e CRIME) são mantidas fixas.

O interesse nesta seção está em aplicar a imputação múltipla para avaliar o comportamento da variância das estimativas dos parâmetros. Nota–se que os p–valores associados aos atributos SEXO e CRIME encontram–se no limite da região de rejeição estabelecido em 5% pelos pesquisadores mais tradicionais. A não resposta do tipo MCAR, gerada para esta base de dados e submetida a imputação pela rede Bayesiana da Figura 4.21 permite observar alguns aspectos que ocorreriam a outras situações de modelagem. Os resultados das simulações encontram–se na Tabela 19 para os valores de $\widehat{\beta}$, para o seu

intervalo de confiança e para os valores de $\widehat{r}$ e $\widehat{\lambda}$.

Tabela 19: Resultados com base em imputação múltipla ($m = 20$) para intervalos de confiança e medidas de diagnóstico para o modelo de riscos proporcionais tendo TPOL1 como variável resposta e SEXO, CRIME e PRISAO como preditivas para os dados observados de homicídios em Campinas

Percentual de	**Intervalo de confiança para $\widehat{\boldsymbol{\beta}}$**			$\widehat{\boldsymbol{r}}$			$\widehat{\boldsymbol{\lambda}}$		
não resposta	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
na variável	SEXO	CRIME	PRISAO	SEXO	CRIME	PRISAO	SEXO	CRIME	PRISAO
5%	[-1,836; 0,135]	[0,099; 1,056]	[0,351; 1,297]	0,057	0,046	0,039	3	6	4
7%	[-1,906; 0,055]	[0,003; 0,930]	[0,303; 1,242]	0,001	0,045	0,050	2	3	5
10%	[-1,911; 0,051]	[-0,047; 0,895]	[0,308; 1,249]	0,005	0,040	0,050	1	4	3
20%	[-1,786; 0,176]	[0,287; 1,297]	[0,451; 1,452]	0,075	0,099	0,125	3	11	9
30%	[-1,994; -0,012]	[0,039; 1,005]	[0,320; 1,349]	0,087	0,077	0,179	8	7	16
40%	[-1,726; 0,275]	[0,171; 1,242]	[0,259; 1,312]	0,094	0,189	0,190	6	17	16
50%	[-1,693; 0,340]	[0,457; 1,633]	[0,316; 1,444]	0,095	0,210	0,214	12	16	14

Os intervalos de confiança para os valores de $\widehat{\beta}$ calculados após a aplicação da imputação múltipla encontram–se na Tabela 19. Se o zero estiver contido no intervalo, não existem indícios de que exista um valor significativo que quantifique a influência da variável equivalente na função de risco modelada. O que ocorre com $\widehat{\beta}_1$ (que teve p–valor de 0,050 no modelo ajustado a partir dos dados reais observados), é que apenas no percentual de não resposta simulado de 30% este parâmetro passa a ter significância no modelo. O inverso ocorre com $\widehat{\beta}_2$ (cujo p–valor foi de 0,055 no mesmo modelo), apenas em um caso não se pode afirmar que exista influência da variável CRIME no risco de abertura do inquérito policial. Aparentemente não existe influência da taxa de não resposta na definição da área de rejeição do parâmetro. A mesma tabela apresenta as medidas de diagnóstico, $\widehat{r}$ e $\widehat{\lambda}$, para a inferência a partir da imputação múltipla. Apesar do número pequeno de observações nessa base de dados, e de percentuais de não resposta simulados que vão de 5% a 50%, a maior estimativa esteve em torno dos 16%, indicando que a imputação a partir da rede Bayesiana fornece a informação necessária sobre a variável resposta a partir de seus pais.

5.4 Discussões e futuros direcionamentos

Neste capítulo foram apresentadas de forma preliminar as técnicas de imputação múltipla para estimação de quantidades relacionadas às variáveis contínuas em duas redes já trabalhadas no capítulo anterior. Tomando por base os parâmetros da rede para imputação nos dois casos, realizou–se um conjunto de $m = 20$ imputações e os resultados foram combinados para se destacar da necessidade de considerar a distribuição da não resposta na estimação após a imputação.

Reitera–se ser esta seção apenas uma aplicação da imputação múltipla para o caso de m bases imputadas a partir da rede Bayesiana para uma análise de dados. Sendo assim, cabem apenas algumas discussões com respeito aos resultados alcançados destas aplicações:

- o resultado obtido a partir da imputação múltipla para a avaliação da média da variável RENDA foi semelhante ao obtido na consistência estatística apresentada no Capítulo 4. O ganho está na parte computacional, tendo em vista que aqui foram considerados apenas $m = 20$ bases de dados imputadas enquanto que no anterior as estimativas foram construídas com base em 500 imputações;

- os parâmetros de imputação da variável LOG(RENDA), ou seja, suas médias dentro das combinações de classes de seus pais, são afetadas pela não resposta, e conseqüentemente pelo método de imputação;

- assumiu–se que os modelos construídos com os dados reais observados era o padrão de comparação para aqueles ajustados nas bases imputadas. Em geral, este procedimento não fornece muita segurança pois não se tem certeza de que o modelo em questão seja o mais adequado aos dados.

- na regressão linear, a dependência condicional entre as variáveis representada pela rede Bayesiana auxilia na estimativa dos parâmetros de maneira a reduzir a taxa de não resposta estimada que influencia a sua obtenção. Neste contexto, o aumento relativo na variância das estimativas não parece depender do percentual de não resposta simulada;

- na análise de sobrevivência construída com a base de dados de homicídios em Campinas, a pouca quantidade de observações e os resultados não muito promissores na construção de uma medida de consistência estrutural não influenciaram na definição dos limites da região de rejeição dos parâmetros.

Riggelsen (2006) traz um estudo sobre o aprendizado dos parâmetros de uma rede Bayesiana a partir de uma base de dados incompletos usando *importance sampling*. Embora seja possível encontrar uma avaliação dos parâmetros da rede por imputação múltipla, esta se dá para redes discretas e o processo de imputação é decorrente do interesse às estimativas nos nós. Dessa forma, a base de dados completa aparece como resultado secundário e não é realizada uma avaliação de seu método especificamente para a imputação.

Di Zio et al. (2004) citam como trabalho aberto a aplicação de imputação múltipla a imputações oriundas da aplicação de redes Bayesianas como método. Apesar de não estabelecer os elementos técnicos para o seu emprego, a utilização da imputação múltipla neste capítulo se mostrou o início de um longo trabalho futuro neste contexto. Fay (1992) diz que a imputação múltipla produz resultados aceitáveis para estimar média geral e para médias em classes de imputação quando os dados se encontram em amostra aleatória simples. Mas isso não é válido para obter a mesma quantidade para subdomínios que são omitidos no modelo para imputação. Esta observação estimula a questionar quando se torna eficaz a estimação via imputação múltipla se o interesse estiver em estimar parâmetros de variáveis imputadas a partir das redes Bayesianas sob certas situações. Por exemplo, naquelas em que se deseja estimar quantidades em tabelas de contingência onde os nós sejam condicionalmente independentes. Mais além, está o estudo da importância de se quantificar a variabilidade nestas situações.

Em casos como estes em que não se considera os modelos verdadeiros para os dados, fica como trabalho futuro a construção de distribuições associadas aos parâmetros a partir de *bootstrap* (EFRON; TIBSHIRANI, 1993) como padrão de comparação.

6 *Síntese dos resultados e comentários*

Neste capítulo é apresentado um resumo dos resultados que foram obtidos ao longo deste trabalho. O objetivo principal é fornecer uma reflexão sobre as aplicações de redes Bayesianas como método de imputação a partir das avaliações conduzidas no texto. Este está aquém de um estudo conclusivo, tendo em vista que as bases de dados utilizadas são muito específicas e não representam todas as situações possíveis para a análise das mesmas. Conforme Chambers (2000), não existe um único método de imputação que seja melhor que os demais. Um exemplo disso está no principal foco do projeto EUREDIT, que era identificar o método a melhor se adequar a cada tipo diferente de base de dados.

Tal qual encadeadas as simulações ao longo do texto, descrevem–se as discussões e citam–se a seguir os aspectos computacionais para obtê–las.

6.1 Redes Bayesianas para imputação

Por ser um método de imputação recente, poucos resultados conclusivos estão disponíveis para fornecer informações sobre sua teoria e aplicação. Thibaudeau e Winkler (2002) apenas citam do uso de redes Bayesianas para imputação com uma grande rede obtida a partir de variáveis demográficas categorizadas (com cerca de 951 parâmetros) e alertam que, quando existirem restrições de crítica ou edição, a não resposta imputada será sempre do tipo não ignorável. Isso ocorreria também devido a limitação provocada pela combinação entre as classes dos pais da variável a ser imputada. Estes autores citam ainda que usaram como *software* para o ajuste da estrutura e a estimação dos parâmetros, o *WinMine Toolkit* da *Microsoft Research*[1] (2001), mas não existem informações sobre o algoritmo utilizado e tampouco estabelecem critérios para a construção da rede.

[1]Disponível em: http:\\research.microsoft.com\ ∼dmax\WinMine\tooldoc.htm.

Hruschka–Jr. (2003) propõe dois algoritmos para imputação com base no algoritmo K2 (COOPER; HERSKOVITS, 1992) para a construção da rede. Ambos os algoritmos partem de uma base completa para a estimação dos parâmetros, o que implica em se excluir da base as unidades que contenham pelo menos um item perdido antes de se iniciar o processo. Essa característica também descarta a sua utilização em variáveis que apresentem zeros estruturais. Duas outras limitações estão no mecanismo de não resposta e na variável ser do tipo categorizada.

Di Zio et al. (2004) aparecem como o trabalho mais recente, também aplicado apenas a atributos discretos. A definição de seu algoritmo para imputação estabelece o critério da confiabilidade para ordenar as variáveis e com isso obter a rede Bayesiana que guiará a imputação. O *software* utilizado para a construção da rede é o *Hugin Tools* (MADSEN et al., 2003) que se utiliza do algoritmo PC (SPIRTES; GLYMOUR, SCHEINES, 1993) para o aprendizado.

Este trabalho apresenta–se como alternativa às formas de se definir o método para imputação com o uso de redes Bayesianas. Inicialmente pela maneira com que se estabelece a rede Bayesiana de entrada, que agrega o conhecimento de especialistas como informação a priori do relacionamento de independências condicionais entre as variáveis. Depois, pela possibilidade de se imputar variáveis qualitativas e quantitativas a partir de um mesmo elemento representado pela estrutura da rede Bayesiana mista. O ajuste da rede é conduzido a partir do *deal* (BØTTCHER; DETHLEFSEN, 2003) que segue o procedimento *Master Prior* (BØTTCHER, 2004) para a escolha da melhor estrutura que representa a combinação entre os dados e o conhecimento a priori do(s) especialista(s).

Para avaliar a aplicação da rede Bayesiana para imputação foram empregados os três tipos de consistências sugeridos por Di Zio et al. (2004): microdados, lógica e estatística, e proposta a consistência estrutural, que trata da manutenção da estrutura $\mathcal{G}$ da rede quando construída a partir de uma base de dados após o processo de imputação.

Antes de tecer algumas observações, cabem alguns comentários que certamente influenciaram os resultados obtidos ao longo das simulações neste texto:

- Assumiu–se que a perda de informação era do tipo ignorável e que o mecanismo de não resposta era do tipo MCAR nas variáveis e dentro das combinações de pais de cada nó contínuo. É fato que na variável renda, trabalhada com os dados do Censo Demográfico, existem indícios de que esta suposição não é atendida. Além disso, como foi utilizado o *deal* para o ajuste da rede, havia a suposição de normalidade

condicionada aos pais. Por isso a transformação logaritmo a minimizar o efeito da não normalidade no nó para essa variável.

- Considerar variáveis auxiliares para imputação de outras que são suas dependentes pode ocasionar melhoria na consistência estatística. Segundo as avaliações de Albieri (1989), os métodos que não consideram variáveis auxiliares no processo produzem vícios que podem ser elevados e conduzem a interpretações errôneas.

- Foram tomados como informação a priori sobre a construção das redes Bayesianas para imputação, o conhecimento de técnicos e gestores das duas bases de dados trabalhadas. Fundamentalmente, os especialistas contribuíram para estabelecer as relações de independência condicional entre as variáveis. Selecionaram–se os atributos para cada rede após a combinação de experiências próprias e o relato de análises práticas dos conhecedores das duas áreas abordadas.

- Os parâmetros relacionados a cada nó foram estimados a partir das bases de dados disponíveis, com eventuais ajustes estabelecidos pelos especialistas citados no item anterior.

- Em cada subconjunto de nós na rede, quando ocorria a perturbação em dois ou mais atributos, imputaram–se primeiro as variáveis do tipo discreto e depois as contínuas respeitando a sua ordem de entrada no algoritmo.

- A variável contínua RENDA foi imputada respeitando os limites estabelecidos no corte inicial da base de dados (ou seja, entre mais de um e dez salários mínimos). De acordo com Schafer (1997a) não existem problemas em se imputar segundo restrições da própria variável, mas este fato deve ser observado para se tecer algum comentário sobre as medidas de consistência avaliadas no Capítulo 4.

Das aplicações dos algoritmos citados nos Quadros 4, para redes discretas e 5 para redes mistas, observamos as três medidas de consistência propostas em Di Zio et al. (2004): da base, lógica e estatística, e propomos a medida de consistência estrutural. Temos resumidamente que:

– na medida de consistência da base de dados para variáveis discretas, o percentual de não resposta não influenciou os resultados obtidos, sendo este valor afetado pelo número de categorias da variável considerada. Para as variáveis contínuas, nas redes ajustadas para os dados do Censo Demográfico, a correlação foi classificada como baixa entre o valor real observado e o imputado. Em alguns casos, a correlação entre

os valores real e imputado foi classificada como moderada, mas não ultrapassando uma estimativa de 0,40. Nos dados de homicídios em Campinas, o valor de $\rho(X_i, \widetilde{X_i})$ foi nulo ou negativo;

– na medida de consistência estrutural, esta não depende do tipo de estrutura (se completamente discreta ou se mista), sendo em geral mais fácil de ser obtida quando o número de nós e arestas é pequeno. De acordo com Neapolitan (2004), a classe de equivalência de uma dada estrutura pode ser obtida para $n \to \infty$ em determinadas classes de modelos. A estrutura da rede Bayesiana mostrou–se sensível quando a imputação é conduzida sob um percentual alto de não resposta em base de dados pequenas;

– para a consistência lógica, nas redes discretas esta sempre foi mantida quando especificadas as restrições nos parâmetros da rede. Nas redes mistas, apesar de ter levado o mesmo nome, essa medida refletiu também a imputação sob os valores com baixas probabilidades, especialmente aqueles situados nas caudas da distribuição no nó. Este fato deu–se por causa dos efeitos indesejados na potencialização dos valores extremos da distribuição da variável a ser imputada. Os resultados apontaram uma proporção de 0,800 dos valores imputados pertencentes a um intervalo delimitado por um desvio da sua média calculada com base nos valores reais observados;

– na consistência estatística em redes discretas, o valor de Δ melhora à medida que se aumenta o percentual de não resposta simulada na base. De acordo com Chambers (2000) esta é a medida de consistência mais fácil de ser atingida com um método de imputação. Observa–se que o valor de Δ é calculado somente sobre os itens faltantes, o que significa que, à medida que se aumenta o valor de n^*, $\widetilde{f}(x_i)$ se aproxima de $f(x_i)$ neste subconjunto dos dados. A extensão para o cálculo da consistência estatística de outras quantidades em redes discretas, como foi o caso da aplicação ao valor V de Cramer, uma medida de associação para dados categorizados. A associação entre duas variáveis após a imputação permitiu verificar que, quanto maior esta quantidade, mais esta é afetada ao se aumentar o percentual de não resposta simulado. Possivelmente, esta mudança no nível de associaÇão pode estar relacionada com a mudança na relação observada entre as variáveis, que é explicitada na estrutura da rede. As relações de dependência entre os nós não são necessariamente as mesmas captadas pelo valor V de Cramer. Neste caso, a imputação a partir de redes Bayesianas como método tende a diminuir a associação existente entre as variáveis à medida que se aumenta o percentual de não resposta;

– na consistência estatística em redes mistas, avaliando os valores da média e mediana da variável contínua imputada, verificou–se que a média do atributo não sofre o efeito da não resposta (ou de forma equivalente, do método de imputação). Já a mediana recebe um acréscimo positivo na direção da média ao se aumentar o percentual de não resposta simulada. Esse comportamento foi observado tanto para as redes construídas com os dados do Censo Demográfico, quanto para os dados de homicídios em Campinas. Até um percentual de não resposta em torno de 10%, a distribuição da variável quantitativa permanece inalterada. A avaliação que se tece acerca deste fenômeno reside na inserção de observações geradas conforme uma distribuição normal condicionada, o que faz com que a mediana se aproxime cada vez mais da média;

– a avaliação das partes discreta e contínua podem ser feitas separadamente devido à própria característica de fatoração da função de verossimilhança (BØTTCHER, 2004). Cabe destacar que, pela maneira como foram definidas as redes, o desempenho obtido pela imputação na parte discreta poderia afetar a execução na parte contínua. A recíproca não seria verdadeira por não terem sido avaliados nós discretos com pais contínuos;

– os resultados encontrados foram semelhantes para os desempenhos em redes discretas e mistas distintas, e para as diferentes bases de dados.

Em linhas gerais, a imputação em atributos discretos a partir de redes Bayesianas funcionaram de forma satisfatória para as bases de dados examinadas, exceto para o objetivo de preservar os dados reais observados. Para a imputação de variáveis contínuas com base em redes Bayesianas mistas, os resultados dependerão em larga escala da suposição que é feita no nó. No caso da variável RENDA, mesmo após a transformação logaritmo para amenizar a sua assimetria, a irregularidade da sua distribuição comprometeu a estimação da mediana após a imputação. Apesar da distribuição de TPOL1 não possuir característica de simetria, a informação de conhecedores da área induziram à não transformação da variável em sua origem. Esse aspecto permitiu avaliar o mesmo comportamento na consistência estatística, independente da forma da distribuição da variável.

A proposição da consistência estrutural permitiu a indagação de recorrentes aspectos observados. Em um deles, questionou–se nas bases do Censo Demográfico o fato de a consistência estrutural ser maior para menores percentuais de não resposta enquanto que para a consistência estatística, o melhor desempenho está nos mais altos percentuais de não resposta nas variáveis perturbadas. A intuição leva a crer que, enquanto $n \to \infty$ e

$t_{NR} \rightarrow 0$, a estrutura apresenta–se mais freqüente em sua classe de equivalência, sendo t_{NR} a taxa de não resposta. Estes não são os únicos itens a afetar esta medida de consistência, sendo a complexidade da rede um fator extremamente importante a considerar.

Restou ainda o tratamento de quantidades associadas à proposição da consistência estrutural, como é o caso das medidas baseadas na função escore. Pela dificuldade no tratamento desta função e de suas propriedades, estas não foram avaliadas aqui, ressaltando que estudos mais aprofundados seriam interessantes neste sentido.

6.2 Imputação múltipla em redes Bayesianas para imputação

Apesar de pouco conclusivos, os experimentos com imputação múltipla em dados imputados a partir de redes Bayesianas fornecem informações de que as variâncias das estimativas em geral necessitam de correção para a não resposta. De acordo com Schafer (1997a), alguma medida de diagnóstico deve ser apresentada juntamente à análise. Neste texto foram calculados os valores de $\widehat{r}$, o aumento relativo na estimativa da variância, e $\widehat{\lambda}$, a fração estimada de informação faltante que afeta o valor da estimativa do parâmetro. Estas quantidades são apresentadas em alguns trabalhos que avaliam inferências realizadas com a presença de dados imputados em diversas situações (SCHAFER, 1997a) (SCHAFER, 1997b) (SCHAFER; OLSEN, 1998). Uma unanimidade entre os pesquisadores da área está no pensamento de que o usuário da base de dados deve ter a condição de calcular uma medida de diagnóstico e decidir se é satisfatório considerar a análise de interesse com o método de imputação utilizado.

Nas aplicações efetuadas no Capítulo 5, para a base do Censo Demográfico, as baixas estimativas de $\widehat{r}$ e $\widehat{\lambda}$ nos diferentes percentuais de não resposta para o cálculo da média da variável permitem conferir pouca influência do método de imputação às inferências consideradas. Para os valores de $\widehat{\lambda}$ no cálculo dos parâmetros da rede e em $\widehat{\lambda}$ e $\widehat{r}$ para os parâmetros do modelo de regressão linear, estes acompanharam respectivamente os percentuais de não resposta simulados e os valores das variâncias. Mas nestes casos, a partição delimitada pelo cruzamento das categorias dos pais da variável contínua diminui o número de elementos nas classes de imputação. Na base de dados de homicídios de Campinas, onde o número de unidades é pequeno, é observado um acréscimo no valor de $\widehat{r}$ ao se aumentar o percentual de perturbação.

O que é interessante registrar é que o resultado obtido a partir da imputação

múltipla para a avaliação da média da variável RENDA foi semelhante ao obtido na consistência estatística apresentada no Capítulo 4. O ganho deu–se sobretudo na parte computacional, tendo em vista que em imputação múltipla foram avaliadas apenas $m = 20$ bases de dados imputadas. Nas simulações, as estimativas foram construídas com base em 500 imputações. Um outro resultado a discutir está nos parâmetros de imputação da variável LOG(RENDA), ou seja, suas médias dentro das combinações de classes de seus pais são afetadas pela não resposta e conseqüentemente pelo método de imputação.

Na análise de sobrevivência construída com a base de dados de homicídios em Campinas, a pouca quantidade de observações e os resultados não muito promissores na construção de uma medida de consistência estrutural, não influenciaram na definição dos limites da região de rejeição dos parâmetros.

6.3 Aspectos computacionais

Nesta seção busca–se apresentar alguns aspectos genéricos da execução, desempenho e tempo de processamento dos procedimentos utilizados para a obtenção dos resultados ao longo do texto. Não se objetiva tratar da complexidade dos algoritmos ou da estratégia de melhor utilização dos recursos computacionais e de tempo, embora se saiba que estes são pontos determinantes para a escolha de um método de imputação, sobretudo nas aplicações que envolvem grandes bases de dados.

As simulações em imputação a partir de redes Bayesianas e imputação múltipla foram particionadas em fases e como foram desenvolvidas em diferentes recursos, tiveram diferentes desempenhos condicionados ao *software*, equipamento e etapas do processo. Descrevem–se a seguir estas fases:

- **Ajuste da rede de entrada** – esta etapa tem como objetivo estimar a estrutura e os parâmetros a partir do *deal* (BØTTCHER; DETHLEFSEN, 2003) no *software R*. Foram combinados os conhecimentos de especialistas para a composição de uma estrutura que refletisse as relações de independência condicional entre as variáveis e a inclusão dos zeros estruturais aos parâmetros da rede. Esta foi uma fase demorada devido ao algoritmo de busca da melhor estrutura sobre os dados e restrições especificadas implementado no *deal* (o procedimento de *Master Prior* especificado detalhadamente em (BØTTCHER, 2004)). Este processo depende fundamentalmente do número de variáveis e de unidades contidas na base de dados.

Tabela 20: Recursos computacionais utilizados nas simulações das redes discretas e mistas para imputação e em imputação múltipla

Rótulo	**Configuração**	**Sistema Operacional**	*Software* **R − versão**	*library* *deal* **− versão**
(1)	PC Intel Pentium 4, 3.4 GHz, 1 GB de RAM	Windows XP Professional Edition 2002	2.3.1	1.2–27
(2)	PC Intel Xeon, 2.8 GHz, 4 MB de cache L2, 4 Gb de RAM	Red Hat Enterprise Linux	2.4.0	1.2–28
(3)	PC Intel Pentium 4, 1.6 GHz, 654 MB de RAM	Windows 2000	2.3.1	1.2–27
(4)	PC Intel Pentium 4, 2.0 GHz, 1 GB de RAM	Windows XP Professional Edition 2002	2.3.1	1.2–27
(5)	Notebook Intel Pentium M, 1.7 GHz, 512 MB de RAM	Windows XP Home Edition 2002	2.3.1	1.2–27

Na base do Censo Demográfico com 15.225 unidades e cinco variáveis discretas, o tempo médio da obtenção da rede foi de quarenta e cinco segundos. Para os dados de homicídios em Campinas com 93 unidades e sete variáveis, este tempo médio foi de menos de vinte segundos.

- **Geração da não resposta** − foi desenvolvida uma função em R para gerar não respostas do tipo MCAR aos dados. Essa função teve como parâmetros de entrada a variável ou vetor de variáveis a serem perturbadas, os percentuais de não resposta associados a cada variável e o número de observações total da base. A idéia desta função estava em gerar perturbações de forma independente para cada variável a partir do percentual de interesse. Para cada posição, era gerado um número aleatório u de acordo a uma variável aleatória uniforme em $(0, 1)$ e feito o seguinte procedimento:

$$\text{Se } (u < \%\text{ não resposta}), \text{ então}$$
$$\text{posição}[X_j] \leftarrow \textit{missing},$$

até que o percentual de não resposta fosse atingido na variável X_j. Para a sua execução, a linha de comando

posicao←geramiss(Base,aimputar,percent)

retornava um vetor para cada variável X_j, com as posições equivalentes aos elementos da base que deveriam ser imputados. Esta função teve desempenho condicionado ao número de observações e variáveis perturbadas na base e o seu tempo de execução era bem pequeno. Em média, para a base do Censo Demográfico com 15.225 unidades e um percentual de 50% de não resposta a ser gerado em uma variável discreta, esta função era executada em três segundos.

- **Algoritmo de imputação** – esta é a etapa de aplicação da rede Bayesiana como método de imputação e foi desenvolvida como função no R. Tem como parâmetros de entrada, os valores de θ da rede Bayesiana que servirão como base para a imputação, a divisão das variáveis ordenadas por subconjunto tal qual definido nas Seções 3.3.3 e 4.3.1, a variável (ou vetor de variáveis) a ser imputada e as posições dos itens faltantes resultantes do tópico anterior. O tempo de execução deste algoritmo sofreu mais influência do percentual de não resposta a ser imputado e do número de parâmetros da variável a ser imputada, que do número de observações da base. À entrada desta função estavam o vetor de itens faltantes gerados na fase anterior e os valores de θ definidos no ajuste da rede de entrada. Se a variável a imputar estivesse em qualquer conjunto que não em P_0, era necessário considerar também os itens (observados ou estimados) de seus pais. À saída desta função estava a base de dados imputada.

- **Cálculo das consistências** – o cálculo das consistências da base, estrutural, lógica e estatística deu–se em uma única função, que era executada logo após a imputação da base de dados. Neste ponto é necessário um novo ajuste da rede para a mensuração da consistência estrutural e da consistência estatística. Nestes procedimentos incluíram–se, além dos cálculos das medidas de consistência, os testes para a consistência da base de dados sugeridos por Chambers (2000) que foram descritos na Seção 2.4.1 e o armazenamento dos resultados. Influenciaram no desempenho desta função, o número de elementos da base de dados e o percentual de não resposta simulado nas variáveis.

- **Imputação múltipla** – na fase da imputação múltipla foram aplicadas as funções de geração da não resposta, de imputação após a estabilidade da cadeia e de ajuste dos modelos estudados. Além disso, foi utilizado o pacote *mitools* (LUMLEY, 2004) para a combinação dos resultados.

- **Repetição do processo** – uma vez que foram delimitadas as etapas do processo, todas as funções eram repetidas o número de vezes de interesse. Por exemplo, para

Tabela 21: Tempo médio de processamento (em segundos) em algumas fases da realização de uma simulação, em imputação e imputação múltipla a partir de redes Bayesianas discretas e mistas

Fase	**Rótulo do computador**	**Tempo médio de processamento (em segundos)**
Ajuste da rede de entrada	(1)	37
	(2)	32
	(3)	60
	(4)	35
	(5)	40
Geração da não resposta	(1)	02
	(2)	01
	(3)	03
	(4)	02
	(5)	02
Algoritmo de imputação	(1)	25
	(2)	14
	(3)	29
	(4)	27
	(5)	25
Cálculo das medidas de consistência	(1)	181
	(2)	105
	(3)	213
	(4)	170
	(5)	210
Procedimentos de imputação múltipla	(1)	–
	(2)	22
	(3)	45
	(4)	–
	(5)	–

o cálculo das consistências nos Capítulos 3 e 4, estabeleceu–se um número de 500 imputações para cada percentual de não resposta e em cada combinação de perturbações nas variáveis.

A Tabela 21 apresenta os tempos médios (em segundos) da execução destas fases de acordo com o recurso computacional equivalente. Devido a variedade de estruturas de rede trabalhadas e dos diferentes tamanhos de bases de dados, os tempos de execução de cada um dos procedimentos listados na Tabela 21 foram estimados para uma única amostra. O tempo total estimado em cada procedimento para cada recurso computacional deve ser multiplicado pelo número de vezes que o procedimento foi repetido. Por exemplo, o ajuste da rede de entrada para o recurso computacional rotulado como (1) em 500 repetições do processo em uma estrutura de uma base de dados teve como tempo total estimado de execução 37 segundos $\times 500 = 18.500$ segundos, o que resulta em aproximada-

mente cinco horas de execução. Reitera–se que este tempo médio é estimado sob diferentes tamanhos de amostra e variados tipos de estrutura. Nas redes construídas com os dados de homicídios ocorridos no município de Campinas, com noventa e três ocorrências, estes tempos foram mais reduzidos do que os tempos médios de processamento listados na Tabela 21.

7 *Projetos futuros*

Ao longo deste trabalho foram registrados elementos da existente literatura sobre a aplicação de redes Bayesianas para imputação em variáveis aleatórias discretas, avaliadas as consistências citadas por Di Zio et al. (2004) e sugerida a consistência estrutural como item a ser avaliado para validar a estrutura de uma rede Bayesiana para imputação. Para os resultados em redes discretas foi utilizado aqui um novo algoritmo estendendo–se o pacote computacional *deal* (BØTTCHER; DETHLEFSEN, 2003) como ferramenta para imputação.

Em sequência, procedeu–se a aplicação de redes Bayesianas mistas para imputação em variáveis discretas e contínuas, numa aplicação do algoritmo sugerido para o caso discreto e onde se definiram extensões diretas dos quatro tipos de consistências tratados no caso discreto. Ainda como contribuição, na tentativa de ampliar o escopo da consistência estatística, foram avaliados resultados a partir de imputação múltipla a parâmetros de modelos de regressão.

Por ser ainda uma aplicação muito recente, os registros de trabalhos em aberto são inúmeros e durante a elaboração deste trabalho, foram identificados alguns itens em que cabe atenção:

- **A função escore** – segundo Bøttcher e Dethlefsen (2003), duas redes Bayesianas são equivalentes se a função escore for a mesma para as duas. Mais amplamente, Neapolitan (2004) mostra que o critério do escore Bayesiano é consistente quando o objetivo é encontrar um grafo direcionado acíclico (e sua correspondente classe de equivalência) fidedigno a uma distribuição $P(\mathcal{G}|\boldsymbol{\theta})$. Mas de acordo com o autor esse resultado apenas vale para tamanhos de amostras suficientemente grandes.

  Tendo como ponto de partida estas duas afirmações, deu–se início à proposição da consistência estrutural a partir da combinação entre a estrutura $\mathcal{G}$ da rede

Bayesiana e a sua função escore equivalente. Ainda nos primeiros desenvolvimentos, entendeu–se que a função escore, apesar da sua relação intrínseca com a estrutura da rede, seria melhor tratada como consistência estatística, em um sentido mais amplo do que aquele citado por Chambers (2000) ou Di Zio et al. (2004). Faz–se necessário avaliar as propriedades da função escore para as estruturas de rede discreta e mista em suas famílias de distribuições. Associada à função escore, também é necessária uma avaliação sobre as quantidades sugeridas em $E[\tau|d]$ e $Var[\tau|d]$ na Seção 3.4.2 do Capítulo 3 para os escores calculados na rede original e na rede obtida após a imputação.

- **A distribuição inicial** – o *deal* (BØTTCHER; DETHLEFSEN, 2003), que foi usado para estimar os parâmetros das redes Bayesianas de entrada e para obter as redes após o processo de imputação, supõe distribuição multinomial para variáveis discretas com posteriori de Dirichlet. No caso da rede mista, a distribuição conjunta para os nós da rede é condicional Gaussiana conforme apresentado no Capítulo 4. O aproveitamento do pacote computacional *deal* foi o limitador para o uso destas distribuições, mas sugere–se que outras sejam implementadas e testadas. Existem diversos *softwares* para ajuste de redes Bayesianas e modelos gráficos, e aqueles que tratam de redes mistas consideram nós condicionais Gaussianos. É o caso do *Bayda* (KONTKANEN et al., 1998) e do *Hugin* (MADSEN et al., 2003).

- **Imputação "ativa"** – os parâmetros da rede de entrada permanecem inalterados durante todo o processo de imputação nos algoritmos que são propostos neste texto. Uma sugestão de trabalho futuro está em atualizar os parâmetros nos nós após cada imputação, principalmente para melhorar o desempenho na consistência estatística em variáveis que contenham muitas classes. O aprendizado ativo (ver, por exemplo, (TONG, 2001)) seria aplicável nesta situação. A aplicação seria principalmente para o caso em que a estrutura da rede de entrada fosse atualizada a partir de base de dados onde ocorresse a não resposta.

- **Ajuste da estrutura com dados faltantes** – nos algoritmos de Hruschka–Jr. (2003), as unidades com pelo menos um item faltante são excluídas da base de dados antes do aprendizado da estrutura da rede e da estimação dos parâmetros. Já em Di Zio et al. (2004) a estrutura da rede é afetada pela ordenação das variáveis segundo a sua confiabilidade, ou seja, esta é dependente dos dados faltantes. Fica como trabalho futuro considerar o ajuste da estrutura e a estimação dos parâmetros da rede a partir da base de dados contendo a não resposta. Para isso existem diversos

algoritmos na literatura, como o PC (SPIRTES; GLYMOUR; SCHEINES, 1993) e o SEM (FRIEDMAN, 1998), que obtêm respectivamente a estrutura, e a estrutura e estimação dos parâmetros. Estimar de uma só vez a estrutura e os parâmetros pode ser combinado com a imputação ativa, como no caso considerado por Riggelsen (2006), em que o objetivo final estava em estimar as quantidades associadas aos nós da rede e, para isso, cada item faltante é calculado em etapas intermediárias.

- **Processo de edição (ou crítica) dos dados** – Di Zio e Scanu (2003) e Coppola et al. (2002) sugerem que o processo de crítica dos dados possa ser combinado à estrutura da rede Bayesiana, facilitando também a identificação de *outliers* e pontos influentes. Apesar de até o momento não haverem estudos específicos da aplicação de redes Bayesianas como instrumento facilitador da crítica de dados, a consistência lógica oferece uma indicação de que seu uso seria promissor. Principalmente porque as regras de edição ou a identificação de pontos influentes e extremos seriam facilmente incorporadas à rede $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$. Restaria definir uma forma de se combinar os processos de crítica e imputação às situações mais realísticas das aplicações, como redes mistas, planos amostrais, etc.

  A permanência dos processos de edição e imputação estabelecidos por Fellegi (1975) e Fellegi e Holt (1976) ainda em vários institutos oficiais de estatística demonstra que existe um certo tradicionalismo no que diz respeito a alteração de processos institucionais de grande escala. Com a realização do projeto EUREDIT, atualizações de alguns procedimentos foram reavaliadas e a adaptação da crítica e edição a um método de imputação de rápida visualização facilitaria sua utilização.

- **Aplicação a dados de planos amostrais** – um dos principais motivadores para o uso de métodos de imputação está na não resposta ao item presente em pesquisas amostrais. Por ser cada vez mais crescente a demanda por pesquisas amostrais, o tratamento da não resposta na unidade e no item têm recebido especial atenção de pesquisadores e institutos oficiais de estatística. Sob o contexto de redes Bayesianas, apenas existem duas aplicações a dados de planos amostrais: Ballin, Scanu e Vicard (2005a) e Ballin, Scanu e Vicard (2005b) em artigos que foram apresentados respectivamente no Simpósio Internacional de Metodologia Estatística do Canadá em 2005 e na reunião de 2005 do *Federal Committee on Statistical Methodology*. Nos dois textos os autores ajustam uma rede Bayesiana a partir dos dados (completos) considerando a variável identificadora do plano amostral, o peso, como um nó na rede. O nó *plano amostral* pode ser pai de todas as variáveis ou ser pai de algumas

delas. Não foi abordado o caso em que este seria filho de outras variáveis, o que possivelmente poderia ocorrer em planos complexos com variáveis de estratificação ou conglomeração que causassem uma dependência direta à variável peso.

O fato é que, como ainda são bem recentes as aplicações de redes Bayesianas a dados de planos amostrais (os artigos que tratam da questão pontuam, juntos, mais de dez itens para trabalhos futuros, inclusive o cálculo da variância de estimadores de interesse), a utilização destas para imputações sob este contexto ainda aguardam tratamento.

- **O problema da ordenação das variáveis** – é fato que a ordem de entrada das variáveis no algoritmo afetam de alguma forma os resultados obtidos. Hruschka–Jr. (2003) identifica essa necessidade e cria um algoritmo que executa um teste de qui–quadrado entre as variáveis para determinar a melhor ordenação entre elas. Neste texto, a ordenação foi indiretamente definida pelo conhecimento de especialistas quando da especificação da rede de entrada do processo. Maiores estudos são necessários neste sentido.

- **A distribuição nas classes da variável** – um trabalho em aberto também está no estudo mais detalhado da influência da distribuição da variável nos resultados das consistências da base, estrutural, lógica e estatística.

Acredita–se que muito exista a ser estudado e desenvolvido na aplicação de redes Bayesianas para imputação e os resultados obtidos neste trabalho mostram que existe grande utilidade prática e desempenho promissor para os contextos de estatísticas oficiais e mineração de dados. Embora cada tópico destes já tenha grande amplitude por si só, observa–se que existe aplicação sob outros contextos onde resultados já consolidados sob outros métodos possam direcionar novas pesquisas em cada um deles.

Referências

AGRESTI, A. *Categorical Data Analysis,* New Jersey: John Wiley & Sons, 2002.

ALBIERI, S. *A Ausência de Respostas em Pesquisas: Uma Aplicação de Métodos de Imputação.* Tese (Mestrado). Rio de Janeiro: IMPA, 1989.

ALLAN, F. E., WISHART, J. A method of estimating yield of missing plot in field experimental work. *Journal of Agricultural Science*, 20, p. 399–406, 1930.

BALLIN, M., SCANU, M., VICARD, P. Model assisted approaches to complex survey sampling from finite populations using Bayesian networks: a tool for integration of different sources. *Proceedings of XXII Statistics Canada International Methodology Symposium,* Ottawa: 2005a.

BALLIN, M., SCANU, M., VICARD, P. Bayesian networks and complex survey sampling from finite populations. *Federal Committee on Statistical Methodology (FCSM) Research Conference,* Virginia: 2005b.

BARROSO, L. P. *Imputação de Dados em Painéis para Populações Finitas.* Tese (Doutorado). São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 1995.

BEAUMONT, J. F. Edit and imputation in surveys: Theory and methods. *Seminário Internacional de Crítica e Imputação.* Rio de Janeiro: IBGE, 2005. disponível em: www.ibge.gov.br\sici.

BETHLEHEM, J., HOFMAN, L. Imputation with Blaise and Manipula. *Apresentado no 5th. International Blaise Users Conference (IBUC).* Lillehammer, Noruega: 1998.

BINNER, J. M., KENDALL, G., CHEN, S. H. *Applications of Artificial Intelligence in Finance and Economics.* Hardbound: Elsevier, 2005.

BLAVATSKY, I. *Inferência em Modelos Marginais de Sobrevivência Multivariada via Bootstrap.* Tese (Mestrado). Minas Gerais: UFMG, 2002.

BØTTCHER, S. G., DETHLEFSEN, C. deal: A package for learning Bayesian networks. *Journal of Statistical Software*, 8(20), 2003.

BØTTCHER, S. G. *Learning Bayesian Networks with Mixed Variables.* Tese (PhD). Dinamarca: Aalborg University, 2004.

BØTTCHER, S. G. Learning Conditional Gaussian Networks. *Technical Report*

R–2005–22, Dinamarca: Aalborg University, 2005.

BOUCKÆRT, R. R. Bayesian network classifiers in Weka. *Working Paper 14/2004*, Nova Zelândia: University of Waikato, Department of Computer Science, 2004.

BREIMAN, L., FRIEDMAN, J. H., STONE, C. J.,OLSHEN, R. A. *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

BUUREN, S. V., OUDSHOORN, C. G. M. Multivariate imputation by chained equations: MICE V1.0 User's manual. *TNO Report PG/VGZ/00.038*, Holanda: 2000.

CHAMBERS, R. Evaluation criteria for statistical editing and imputation. *Technical Report #28,* National Statistics Methodology Series, 2000.

CHARNIAK, E. Bayesian networks without tears. *AI Magazine,* p. 50–630, 1991.

CLEMENTINE USER GUIDE, *A Data Mining Toolkit, 2000.*
disponível em: www.spss.com\clementine.

COCHRAN, W. G. *Sampling Techniques*. New York: Wiley, 1977.

COLLETT, D. *Modelling Survival Data in Medical Research.* London: Chapman & Hall, 1994.

COOPER, G. F., HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, p. 309–347, 1992.

COPPOLA, L., DI ZIO, M., LUZI, O., PONTI, A., SCANU, M. On the use of Bayesian networks in Official Statistics. *XLI Riunione Scientifica, Sessione Spontanea,* Milão: Universitá di Milano–Bicocca, 2002.

COX, D. Regression models and life tables. *Journal of the Royal Statistical Society.* Series B, 34, p. 187–202, 1972.

COX, D. R., HINKLEY, D. V. *Theoretical Statistics*. New York: Wiley, 1974.

CRAMER, H. *Mathematical Methods of Statistics.* New Jersey: Princeton University Press, 1999.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society.* Series B, 39(1), p. 1–38, 1977.

DI ZIO, M., SCANU, M. *Bayesian networks. Methods and experimental results from the Euredit project.* Technical Appendices E, Volume 2, Euredit deliverable D6.1, on CD with Volume 1, 2003.

DI ZIO, M., SCANU, M., COPPOLA, L., LUZI, O., PONTI, A. Bayesian networks for imputation. *Journal of the Royal Statistical Society A*, 167, Part 2, p. 309–322, 2004.

DI ZIO, M., SCANU, M., VICARD, P. Open problems and new perspectives for imputation using Bayesian networks. *Modelli Complesi e Metodi Computazionali Intensivi per la Stima e la Previsione.* Treviso: 2003.

disponível em: www.dst.unive.it\sco2003\.

DIAS, A. J. R. *Tratamento de Dados Ausentes para Análise Fatorial de Indicadores de Saúde.* Tese (Mestrado). Rio de Janeiro: COPPE, UFRJ, 1990.

DIAS, A. J. R., ALBIERI, S. Uso de imputação em pesquisas domiciliares. *VIII Encontro Nacional de Estudos Populacionais.* Anais, Volume 1: Informação Demográfica, Fecundidade, Demografia Histórica, p. 11–26, São Paulo: ABEP, 1992.

DINIZ, C. A. R., LOUZADA–NETO, F. *Data Mining: Uma Introdução.* 14º SINAPE, Caxambu: ABE, 2000.

DURRANT, G. B. Imputation methods for handling item–nonresponse in the social sciences: A methodological review. *Working Paper Series.* UK: National Centre for Research Methods, 2005.

EFRON, B., TIBSHIRANI, R. J. *An Introduction to the Bootstrap.* London: Chapman & Hall, 1993.

FAY, R.E. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association,* p. 227–232, 1992.

FELLEGI, I. P. Automatic edit and imputation of quantitative data. *Bulletin of the International Statistical Institute,* XLVI, p. 249–253, 1975.

FELLEGI, I. P., HOLT, D. A. Systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, p. 17–35, 1976.

FREY, B. J. *Graphical Models for Machine Learning and Digital Communication.* Cambridge: Bradford Book, 1998.

FRIEDMAN, N. The Bayesian Structural EM Algorithm. *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98),* p. 129–138, 1998.
disponível em: citeseer.ist.psu.edu\article\friedman98bayesian.html

GEIGER, D., HECKERMAN, D. Learning Gaussian networks. *Technical Report MSR-TR-94-10,* Redmond: Microsoft Corporation, 1994.

GETOOR, L., TASKAR, B., KOLLER, D. Selectivity estimation using probabilistic models. *in Proc. Association for Computing Machinery Special Interest Group on Management of Data*, p. 461–472, 2001.
disponível em: www.acm.org\sigmod\sigmod01\eproceedings.

GILKS, W. R., RICHARDSON, S., SPIEGELHALTER, D. J. *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall, 1996.

GOODMAN, L. A., KRUSKAL, W. H. Measures of association for cross–classification. *Journal of the American Statistical Association*, 49, 732–764, 1954.

HAIR, J. F., ANDERSON, R. E., TATHAM, R. L., BLACK, W. C. *Multivariate Data Analysis.* New Jersey: Prentice Hall, 1998.

HECKERMAN, D. E. An axiomatic framework for belief updates. *Uncertainty in Artificial Intelligence 2*, New York: 1988.

HECKERMAN, D., GEIGER, D., CHICKERING, D. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning,* 20(3), p. 197–243, 1995.

HERZOG, T. N., RUBIN, D. B. Using multiple imputations to handle non-response in sample surveys. *Incomplete Data in Sample Survey: Theory and Bibliographies, Vol 2.* (Eds. William G. Madow, Ingram Olkin, Donald B. Rubin), New York: Academic Press, p. 209–245, 1983.

HINKLE, D. E., WIERSMA, W., JURS, S. G. Applied Statistics for the Behavioral Sciences. Boston: Houghton Mifflin, 1994.

HRUSCHKA–JR., E. R. *Imputação Bayesiana no Contexto da Mineração de Dados*. Tese (Doutorado). Rio de Janeiro: COPPE–UFRJ, 2003.

HRUSCHKA–JR., E. R., HRUSCHKA, E. R., EBECKEN, N. F. F. Applying Bayesian networks for meteorological data mining. *The Twenty–fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Inteligence,* Cambridge: 2005.

IBGE. *Metodologia do Censo Demográfico 2000.* Rio de Janeiro: IBGE, 2003.

IDE, J. S. *Algoritmos para Inferência Aproximada em Redes Credais com Variáveis Binárias*. Tese (Doutorado). São Paulo: Escola Politécnica, Universidade de São Paulo, 2005.

KALTON, G. *Compensating for Missing Survey Data.* Michigan: Survey Research Center, University of Michigan, 1983.

KALTON, G., KASPRZYK, D. Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section, American Statistical Association,* p. 146–151, 1982.

KLEIN, J. P., MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censured and Truncated Data.* New York: Springer–Verlag, 1997.

KONTKANEN, P., MYLLYMÄKI, P., SILANDER, T., TIRRI, H. BAYDA: Software for Bayesian classification and feature selection. *Knowledge Discovery and Data Mining*, p. 254–258, 1998.

LAIRD, N. M. Missing data in longitudinal studies. *Statistics in Medicine*, 7, p. 305–315, 1988.

LAURITZEN, S. L. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association,* 87 (420), p. 1098–1108, 1992.

LAURITZEN, S. L, JENSEN, F. Stable local computation with conditional Gaussian distributions. *Statistics and Computing,* 11, p. 191–203, 2001.

LITTLE, R. J. A. Comment on analysis of data with missing values. *Statistics in*

Medicine, 7, p. 347–355, 1976.

LITTLE, R. J. A., RUBIN, D. B. Missing data in large data sets. *Statistical Methods and the Improvement of Data Quality,* New York: Academic Press, p. 15–243, 1983.

LITTLE, R. J. A., RUBIN, D. B. *Statistical Analysis with Missing Data.* Second edition, New Jersey: John Wiley & Sons, 2002.

LITTLE, R. J. A., SMITH, P. J. Edit and imputation for quantitative survey data. *Journal of the American Statistical Association,* 82, p. 58–68, 1987.

LUMLEY, T. mitools: Tools for multiple imputation of missing data, *R Package Version 1.0*, Austria: R Foundation for Statistical Computing, 2004.

MADSEN, A. L., LANG, M., KJÆRULFF, U., JENSEN, F. The Hugin tool for learning Bayesian networks. *Proceedings of The Seventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty,* p. 549–605, 2003.

MICROSOFT RESEARCH. *WinMine Toolkit,* 2001.
disponível em: http:\\research.microsoft.com\ ∼dmax\WinMine\tooldoc.htm.

MORRISON, D. *Multivariate Statistical Methods.* New York: McGraw-Hill, 1976.

MURPHY, K. P. Inference and learning in hybrid Bayesian networks. *Technical Report No. UCB/CSD-98-990*, Berkeley: Computer Science Division, University of California, 1998.

NEAPOLITAN, R. E. *Learning Bayesian Networks.* New Jersey: Pearson Prentice Hall, 2004.

OLKIN, I. Introduction and recommendations. *Incomplete Data in Sample Surveys: Report and Case Studies, Vol 1.* (Eds. William G. Madow, Ingram Olkin, Donald B. Rubin), New York: Academic Press, p. 3–14, 1983.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems.* California: Morgan Kaufmann, 1988.

PESSOA, D. G. C., MOREIRA, G. G., SANTOS, A. R. Imputação de rendimentos no questionário da amostra do Censo Demográfico 2000. *Relatório Interno.* Rio de Janeiro: IBGE, 2004.

PESSOA, D. G. C., SANTOS, A. R. Imputação de renda dos responsáveis por domicílios: Conjunto universo do Censo Demográfico 2000. *Relatório Interno.* Rio de Janeiro: IBGE, 2004.

PLATEK, R., GRAY, G. B. Imputation methodology. *Incomplete Data in Sample Surveys: Theory and Bibliographies, Vol 2.* (Eds. William G. Madow, Ingram Olkin, Donald B. Rubin), New York: Academic Press, p. 255–294, 1983.

POLITZ, A. N., SIMMONS, W. R. An attempt to get not–at–homes into the sample without call–backs. *Journal of the American Statistical Association*, 44, p. 9–31, 1949.

PORCARO, R. M. Sistema de informação estatística e sociedade da informação: desafios e perspectivas da economia eletrônica. *Textos para Discussão.* Rio de Janeiro: IBGE, 2003.

RASMUSSEN, L. K. boblo: An expert system based on Bayesian networks to blood group determination of cattle. *Research Report 16.* Tjele: Research Center Foulum, 1995.

RIGGELSEN, C. Learning Bayesian network parameters from incomplete data using importance sampling. *International Journal of Approximate Reasoning*, 42, p. 69–83, 2006.

RUBIN, D. B. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, 1987.

RUBIN, D. B. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, p. 473–489, 1996.

SAHEKI, A. H. *Construção de uma Rede Bayesiana Aplicada ao Diagnóstico de Doenças Cardíacas.* Tese (Mestrado). São Paulo: Escola Politécnica, Universidade de São Paulo, 2005.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J. *Model Assisted Survey Sampling.* New York: Springer-Verlag, 1992.

SCANU, M., DI ZIO, M., VICARD, P. Computational aspects of imputation with Bayesian networks. *A MaPhySto Workshop on Computational Aspects of Graphical Models.* Aalborg, Dinamarca: Aalborg University, 2003.
disponível em www.math.aau.dk\gr\material\scanu.pdf.

SCHAFER, J. L. *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall, 1997a.

SCHAFER, J. L. Imputation of missing covariates under a general linear mixed model. *Technical Report*, Dept. of Statistics, Penn. State University: 1997b.

SCHAFER, J. L. mix: Estimation/multiple imputation for mixed categorical and continuous data, *R Package Version 1.0–4*, 2003.
disponível em www.stat.psu.edu\ ∼jls\misoftwa.html.

SCHAFER, J. L., OLSEN, M. K. Multiple imputation for multivariate missing–data problems: A data analyst's perspective. *Multivariate Behavioral Research,* 33(4), p. 545–571, 1998.

SHIMAKURA, S. E. Interpretação do coeficiente de correlação. *Notas de Aula Online.* Paraná: UFPR, 2006.
disponível em http:\\leg.ufpr.br\ ∼silvia\CE003.

SILVA, P. L. N. *Crítica e Imputação de Dados Quantitativos Utilizando o SAS.* Tese (Mestrado). Rio de Janeiro: IMPA, 1989.

SILVA, M. P. V., SALOMÃO, N. M. R. Interações verbais e não verbais entre mães–crianças portadoras de Síndrome de Down e entre mães–crianças com desenvolvimento normal. *Estudos de Piscologia.* Vol. 7, Número 2, Natal: UFRN,

p. 311–323, 2002.

SILVA–FILHO, A. M. A era da informação. *Revista Espaço Acadêmico,* Ano I, N^o 02, julho de 2001.

SMYTH, P. Belief networks, hidden Markov models and Markov random fields: a unifying view. *Pattern Recognition Letters*, Volume 18, Issues 11–13, p. 1261–1268, 1997.

SPIEGELHALTER, D., LAURITZEN, S. L. Sequential updating of conditional probabilities on directed graphical structures. *Networks,* 20, p. 579–605, 1990.

SPIRTES, P., GLYMOUR, C., SCHEINES, R. *Causation, Prediction and Search.* New York: Springer–Verlag, 1993.

STUART, A. A test for homogeneity of the marginal distributions in a two–way classification. *Biometrika*, 42, p. 412–416, 1955.

TANNER, M. A. *Tools for Statistical Inference, Methods for Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer–Verlag, 1993.

TANNER, M. A., WONG, W. H. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, p. 528–550, 1987.

THIBAUDEAU, Y., WINKLER, W. E. Bayesian networks representation, generalized imputation, and synthetic micro–data satisfying analytic constraints. *Technical Report #2002-09*, Washington: U.S. Bureau of the Census, 2002.

TONG, S. *Active Learning: Theory and Applications.* PhD. Thesis, Palo Alto: Stanford University, 2001.

VARGAS, J. D. *Crimes Sexuais e Sistema de Justiça.* São Paulo: IBCCRIM, 2000.

VARGAS, J. D. *Estupro: Que Justiça? Fluxo do Funcionamento e Análise do Tempo da Justiça Criminal para o Crime de Estupro.* Tese (Doutorado). Rio de Janeiro: IUPERJ, 2004.

VARGAS, J. D., Metodologia de tratamento do tempo e da morosidade processual na Justiça Criminal. *Relatório Final: Concursos Nacionais de Pesquisas Aplicadas em Justiça Criminal e Segurança Pública.* SENASP, Secretaria Nacional de Segurança Pública, 2006.

YATES, F. The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, 1, p. 129–142, 1933.

YUAN, Y. C. Multiple imputation for missing values: concepts and new development. *SAS Data Analysis Papers and Presentations, Report P267–25,* Maryland: 2000.

Apêndice A

Tabelas de resultados das simulações em redes discretas

Tabela A.1: Proporção de imputações corretas na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede com três nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	TIPODOM	CONDDOM	CONDTER	TIPODOM–CONDDOM	TIPODOM–CONDTER	CONDDOM–CONDTER	TIPODOM–CONDDOM-CONDTER
1%	0,952 (0,014)	0,479 (0,036)	0,938 (0,016)	0,694 (0,019)	0,936 (0,011)	0,672 (0,020)	0,765 (0,012)
3%	0,932 (0,010)	0,439 (0,017)	0,933 (0,007)	0,681 (0,012)	0,934 (0,007)	0,694 (0,011)	0,767 (0,010)
5%	0,932 (0,007)	0,440 (0,019)	0,918 (0,007)	0,687 (0,009)	0,928 (0,005)	0,680 (0,009)	0,754 (0,006)
7%	0,939 (0,006)	0,454 (0,012)	0,927 (0,007)	0,695 (0,007)	0,924 (0,004)	0,678 (0,007)	0,757 (0,006)
10%	0,936 (0,005)	0,443 (0,012)	0,918 (0,004)	0,688 (0,007)	0,932 (0,003)	0,670 (0,006)	0,759 (0,006)
20%	0,925 (0,003)	0,450 (0,006)	0,926 (0,004)	0,691 (0,005)	0,930 (0,002)	0,652 (0,005)	0,749 (0,003)
30%	0,933 (0,003)	0,444 (0,006)	0,928 (0,003)	0,690 (0,004)	0,932 (0,002)	0,635 (0,005)	0,732 (0,003)
40%	0,930 (0,002)	0,449 (0,005)	0,925 (0,002)	0,690 (0,003)	0,928 (0,002)	0,620 (0,004)	0,721 (0,002)
50%	0,930 (0,002)	0,448 (0,005)	0,930 (0,002)	0,689 (0,003)	0,929 (0,002)	0,599 (0,003)	0,708 (0,003)

Tabela A.2: Proporção de imputações corretas na avaliação da consistência lógica após a imputação, a partir de 500 imputações de uma rede com três nós, em diferentes percentuais de não resposta quando não usada a restrição de parâmetros (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**	
	CONDTER	CONDTER|CONDDOM
1%	0,935 (0,011)	0,715 (0,020)
3%	0,941 (0,010)	0,703 (0,017)
5%	0,947 (0,007)	0,711 (0,010)
7%	0,952 (0,006)	0,699 (0,007)
10%	0,955 (0,005)	0,697 (0,004)
20%	0,955 (0,003)	0,684 (0,004)
30%	0,958 (0,003)	0,672 (0,004)
40%	0,957 (0,002)	0,650 (0,003)
50%	0,961 (0,002)	0,649 (0,003)

Tabela A.3: Valores de Δ na avaliação da consistência estatística dos parâmetros da rede após a imputação, a partir de 500 imputações de uma rede com três nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	TIPODOM	CONDDOM	CONDTER	TIPODOM–CONDDOM	TIPODOM–CONDTER	CONDDOM–CONDTER	TIPODOM–CONDDOM-CONDTER
1%	0,022 (0,008)	0,071 (0,028)	0,016 (0,007)	0,048 (0,019)	0,019 (0,009)	0,040 (0,017)	0,037 (0,016)
3%	0,012 (0,007)	0,054 (0,016)	0,007 (0,004)	0,027 (0,010)	0,011 (0,007)	0,026 (0,011)	0,023 (0,008)
5%	0,008 (0,005)	0,026 (0,009)	0,007 (0,004)	0,024 (0,010)	0,009 (0,005)	0,019 (0,007)	0,018 (0,007)
7%	0,006 (0,003)	0,036 (0,010)	0,007 (0,004)	0,020 (0,008)	0,009 (0,004)	0,016 (0,006)	0,016 (0,006)
10%	0,006 (0,004)	0,028 (0,008)	0,006 (0,003)	0,016 (0,006)	0,007 (0,003)	0,014 (0,006)	0,012 (0,005)
20%	0,005 (0,003)	0,014 (0,005)	0,004 (0,002)	0,011 (0,003)	0,008 (0,003)	0,012 (0,005)	0,009 (0,004)
30%	0,003 (0,002)	0,011 (0,004)	0,004 (0,002)	0,011 (0,003)	0,003 (0,002)	0,008 (0,004)	0,008 (0,004)
40%	0,003 (0,002)	0,011 (0,004)	0,003 (0,002)	0,007 (0,003)	0,003 (0,001)	0,010 (0,004)	0,007 (0,003)
50%	0,003 (0,002)	0,010 (0,004)	0,003 (0,001)	0,007 (0,003)	0,003 (0,001)	0,007 (0,004)	0,005 (0,002)

Tabela A.4: Avaliação da consistência estatística (medida de associação), após 500 imputações nas variáveis a partir de uma rede com três nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável imputada a partir da rede Bayesiana**		
	TIPODOM–CONDDOM	TIPODOM–CONDTER	CONDDOM–CONDTER
1%	0,098 (0,001)	0,076 (0,001)	0,708 (0,001)
3%	0,092 (0,002)	0,075 (0,001)	0,705 (0,002)
5%	0,089 (0,003)	0,072 (0,002)	0,689 (0,002)
7%	0,090 (0,003)	0,070 (0,003)	0,663 (0,003)
10%	0,087 (0,007)	0,069 (0,003)	0,620 (0,003)
20%	0,086 (0,005)	0,069 (0,003)	0,604 (0,005)
30%	0,084 (0,006)	0,065 (0,005)	0,571 (0,006)
40%	0,079 (0,007)	0,061 (0,007)	0,551 (0,006)
50%	0,077 (0,007)	0,060 (0,008)	0,542 (0,008)

Tabela A.5: Proporção de imputações corretas na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede com quatro nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	ABASTEC	TIPODOM–ABASTEC	CONDDOM–ABASTEC	CONDTER–ABASTEC	TIPODOM–CONDDOM-ABASTEC	CONDDOM–CONDTER-ABASTEC	TIPODOM–CONDDOM–CONDTER-ABASTEC
1%	0,908 (0,019)	0,916 (0,016)	0,659 (0,026)	0,913 (0,015)	0,764 (0,022)	0,733 (0,024)	0,783 (0,020)
3%	0,889 (0,011)	0,909 (0,010)	0,680 (0,015)	0,913 (0,010)	0,769 (0,013)	0,748 (0,014)	0,800 (0,012)
5%	0,884 (0,009)	0,902 (0,007)	0,671 (0,012)	0,908 (0,009)	0,760 (0,011)	0,754 (0,011)	0,786 (0,009)
7%	0,878 (0,007)	0,907 (0,006)	0,664 (0,011)	0,905 (0,006)	0,761 (0,008)	0,747 (0,009)	0,792 (0,009)
10%	0,883 (0,006)	0,907 (0,006)	0,666 (0,007)	0,904 (0,005)	0,752 (0,006)	0,745 (0,007)	0,791 (0,007)
20%	0,888 (0,005)	0,912 (0,003)	0,667 (0,006)	0,909 (0,004)	0,752 (0,005)	0,732 (0,005)	0,779 (0,005)
30%	0,887 (0,004)	0,912 (0,003)	0,667 (0,004)	0,907 (0,003)	0,757 (0,005)	0,717 (0,005)	0,769 (0,004)
40%	0,886 (0,003)	0,910 (0,003)	0,669 (0,004)	0,909 (0,003)	0,754 (0,004)	0,705 (0,004)	0,760 (0,004)
50%	0,888 (0,002)	0,908 (0,002)	0,668 (0,004)	0,906 (0,003)	0,754 (0,003)	0,690 (0,004)	0,752 (0,003)

Tabela A.6: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede com quatro nós, em diferentes percentuais de não resposta

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	ABASTEC	TIPODOM–ABASTEC	CONDDOM–ABASTEC	CONDTER–ABASTEC	TIPODOM–CONDDOM-ABASTEC	CONDDOM–CONDTER-ABASTEC	TIPODOM–CONDDOM–CONDTER-ABASTEC
1%	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3%	1,000	1,000	1,000	1,000	1,000	1,000	0,998
5%	1,000	1,000	0,998	0,998	1,000	0,998	0,996
7%	0,998	1,000	0,998	0,998	0,998	0,996	0,998
10%	0,998	0,998	1,000	0,998	0,996	0,998	1,000
20%	0,998	1,000	1,000	0,998	0,994	0,994	0,994
30%	1,000	1,000	0,998	0,996	0,994	0,994	0,992
40%	0,998	0,998	0,998	0,994	0,992	0,992	0,990
50%	0,998	0,998	0,998	0,996	0,994	0,992	0,990

Tabela A.7: Valores de Δ na avaliação da consistência estatística dos parâmetros da rede após a imputação, a partir de 500 imputações de uma rede com quatro nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	ABASTEC	TIPODOM–ABASTEC	CONDDOM–ABASTEC	CONDTER–ABASTEC	TIPODOM–CONDDOM-ABASTEC	CONDDOM–CONDTER-ABASTEC	TIPODOM–CONDDOM–CONDTER-ABASTEC
1%	0,030 (0,017)	0,018 (0,013)	0,043 (0,019)	0,023 (0,012)	0,032 (0,016)	0,062 (0,019)	0,036 (0,016)
3%	0,029 (0,012)	0,010 (0,005)	0,029 (0,011)	0,012 (0,009)	0,021 (0,008)	0,022 (0,009)	0,021 (0,008)
5%	0,011 (0,005)	0,010 (0,004)	0,035 (0,010)	0,009 (0,004)	0,017 (0,008)	0,020 (0,008)	0,018 (0,007)
7%	0,010 (0,005)	0,009 (0,004)	0,026 (0,009)	0,009 (0,005)	0,014 (0,006)	0,017 (0,007)	0,019 (0,006)
10%	0,006 (0,003)	0,007 (0,004)	0,014 (0,006)	0,008 (0,004)	0,012 (0,005)	0,016 (0,006)	0,013 (0,005)
20%	0,008 (0,004)	0,005 (0,003)	0,010 (0,004)	0,007 (0,003)	0,009 (0,004)	0,008 (0,003)	0,008 (0,004)
30%	0,004 (0,002)	0,004 (0,002)	0,008 (0,004)	0,006 (0,003)	0,006 (0,003)	0,008 (0,004)	0,008 (0,003)
40%	0,004 (0,002)	0,004 (0,002)	0,007 (0,003)	0,004 (0,002)	0,006 (0,002)	0,007 (0,003)	0,005 (0,002)
50%	0,003 (0,002)	0,003 (0,001)	0,006 (0,003)	0,003 (0,002)	0,006 (0,002)	0,005 (0,002)	0,005 (0,002)

Tabela A.8: Avaliação da consistência estatística (medida de associação) após 500 imputações nas variáveis, a partir de uma rede com quatro nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável imputada a partir da rede Bayesiana**		
	TIPODOM–ABASTEC	CONDDOM–ABASTEC	CONDTER–ABASTEC
1%	0,053 (0,001)	0,323 (0,001)	0,116 (0,001)
3%	0,050 (0,002)	0,310 (0,002)	0,113 (0,002)
5%	0,051 (0,003)	0,304 (0,003)	0,110 (0,002)
7%	0,048 (0,004)	0,300 (0,003)	0,110 (0,003)
10%	0,045 (0,006)	0,275 (0,004)	0,108 (0,003)
20%	0,043 (0,007)	0,261 (0,005)	0,101 (0,004)
30%	0,042 (0,007)	0,252 (0,006)	0,098 (0,006)
40%	0,040 (0,009)	0,241 (0,007)	0,095 (0,008)
50%	0,039 (0,009)	0,238 (0,007)	0,094 (0,008)

Tabela A.9: Proporção de imputações corretas na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede com cinco nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	TIPOCAN	TIPODOM–TIPOCAN	CONDDOM–TIPOCAN	CONDTER–TIPOCAN	ABASTEC–TIPOCAN	TIPODOM–CONDDOM–TIPOCAN	TIPODOM–CONDDOM–CONDTER-ABASTEC-TIPOCAN
1%	0,940 (0,016)	0,917 (0,018)	0,679 (0,017)	0,910 (0,015)	0,890 (0,018)	0,790 (0,020)	0,787 (0,023)
3%	0,941 (0,013)	0,910 (0,011)	0,681 (0,015)	0,911 (0,015)	0,901 (0,016)	0,791 (0,017)	0,785 (0,021)
5%	0,932 (0,010)	0,905 (0,009)	0,675 (0,010)	0,909 (0,013)	0,899 (0,015)	0,695 (0,014)	0,780 (0,019)
7%	0,932 (0,009)	0,907 (0,008)	0,674 (0,010)	0,905 (0,012)	0,897 (0,010)	0,693 (0,010)	0,777 (0,015)
10%	0,920 (0,007)	0,909 (0,006)	0,662 (0,009)	0,907 (0,011)	0,896 (0,009)	0,692 (0,008)	0,772 (0,010)
20%	0,912 (0,003)	0,910 (0,004)	0,667 (0,008)	0,903 (0,011)	0,903 (0,009)	0,699 (0,006)	0,772 (0,008)
30%	0,913 (0,003)	0,912 (0,004)	0,669 (0,006)	0,899 (0,009)	0,900 (0,006)	0,692 (0,004)	0,770 (0,005)
40%	0,910 (0,002)	0,909 (0,003)	0,665 (0,003)	0,901 (0,006)	0,902 (0,006)	0,690 (0,003)	0,768 (0,004)
50%	0,909 (0,002)	0,908 (0,002)	0,663 (0,003)	0,900 (0,005)	0,905 (0,004)	0,681 (0,003)	0,765 (0,004)

Tabela A.10: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede com cinco nós, em diferentes percentuais de não resposta

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	TIPOCAN	TIPODOM–TIPOCAN	CONDDOM–TIPOCAN	CONDTER–TIPOCAN	ABASTEC–TIPOCAN	TIPODOM–CONDDOM–TIPOCAN	TIPODOM–CONDDOM–CONDTER-ABASTEC-TIPOCAN
1%	0,998	1,000	0,998	0,980	1,000	0,998	1,000
3%	0,960	0,998	1,000	0,996	1,000	0,996	0,960
5%	0,996	0,980	0,960	1,000	0,998	1,000	0,998
7%	0,820	0,860	0,860	0,840	0,900	0,860	0,880
10%	0,840	0,840	0,700	0,820	0,860	0,880	0,800
20%	0,700	0,882	0,820	0,760	0,820	0,780	0,720
30%	0,668	0,720	0,800	0,700	0,760	0,740	0,700
40%	0,668	0,740	0,760	0,680	0,700	0,820	0,660
50%	0,620	0,700	0,720	0,620	0,700	0,668	0,600

Tabela A.11: Valores de Δ na avaliação da consistência estatística dos parâmetros da rede, a partir de 500 imputações de uma rede com cinco nós, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**						
	TIPOCAN	TIPODOM–TIPOCAN	CONDDOM–TIPOCAN	CONDTER–TIPOCAN	ABASTEC–TIPOCAN	TIPODOM–CONDDOM–TIPOCAN	TIPODOM–CONDDOM–CONDTER-ABASTEC-TIPOCAN
1%	0,027 (0,012)	0,019 (0,012)	0,047 (0,020)	0,026 (0,016)	0,023 (0,019)	0,037 (0,015)	0,053 (0,013)
3%	0,020 (0,010)	0,015 (0,011)	0,033 (0,016)	0,018 (0,010)	0,020 (0,011)	0,030 (0,013)	0,031 (0,011)
5%	0,011 (0,007)	0,010 (0,009)	0,025 (0,010)	0,012 (0,008)	0,014 (0,009)	0,024 (0,009)	0,019 (0,008)
7%	0,009 (0,005)	0,010 (0,009)	0,020 (0,010)	0,009 (0,005)	0,012 (0,008)	0,011 (0,007)	0,012 (0,007)
10%	0,009 (0,005)	0,008 (0,007)	0,013 (0,008)	0,008 (0,004)	0,010 (0,007)	0,009 (0,006)	0,008 (0,005)
20%	0,005 (0,004)	0,007 (0,003)	0,011 (0,005)	0,008 (0,004)	0,007 (0,004)	0,006 (0,004)	0,006 (0,005)
30%	0,004 (0,002)	0,004 (0,003)	0,009 (0,003)	0,006 (0,003)	0,006 (0,003)	0,004 (0,004)	0,005 (0,003)
40%	0,003 (0,002)	0,004 (0,002)	0,005 (0,003)	0,004 (0,002)	0,006 (0,003)	0,004 (0,003)	0,003 (0,002)
50%	0,003 (0,002)	0,003 (0,001)	0,005 (0,002)	0,003 (0,001)	0,005 (0,001)	0,002 (0,002)	0,003 (0,002)

Tabela A.12: Avaliação da consistência estatística (medida de associação) após a imputação nas variáveis, a partir de uma rede com cinco nós (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável imputada a partir da rede Bayesiana**			
	TIPODOM–TIPOCAN	CONDDOM–TIPOCAN	CONDTER–TIPOCAN	ABASTEC–TIPOCAN
1%	0,093 (0,001)	0,295 (0,001)	0,121 (0,001)	0,635 (0,001)
3%	0,090 (0,002)	0,292 (0,002)	0,118 (0,002)	0,612 (0,001)
5%	0,084 (0,002)	0,287 (0,002)	0,115 (0,003)	0,587 (0,003)
7%	0,082 (0,002)	0,281 (0,004)	0,107 (0,004)	0,563 (0,004)
10%	0,080 (0,003)	0,273 (0,004)	0,099 (0,004)	0,551 (0,004)
20%	0,076 (0,005)	0,269 (0,005)	0,090 (0,005)	0,545 (0,005)
30%	0,071 (0,007)	0,260 (0,007)	0,083 (0,006)	0,540 (0,008)
40%	0,065 (0,008)	0,254 (0,007)	0,075 (0,008)	0,533 (0,008)
50%	0,062 (0,008)	0,249 (0,009)	0,069 (0,008)	0,512 (0,010)

Tabela A.13: Proporção de imputações corretas na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações na rede réu–prisão dos dados de homicídios em Campinas, em diferentes tipos de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	0,926 (0,181)	0,456 (0,358)	0,707 (0,332)	0,335 (0,252)	0,728 (0,181)	0,616 (0,215)	0,675 (0,278)
7%	0,937 (0,097)	0,359 (0,186)	0,627 (0,196)	0,366 (0,268)	0,711 (0,170)	0,470 (0,227)	0,610 (0,219)
10%	0,926 (0,100)	0,353 (0,157)	0,582 (0,233)	0,326 (0,161)	0,762 (0,147)	0,489 (0,155)	0,605 (0,142)
20%	0,937 (0,063)	0,398 (0,121)	0,510 (0,112)	0,341 (0,111)	0,739 (0,085)	0,448 (0,114)	0,574 (0,098)
30%	0,899 (0,050)	0,342 (0,085)	0,541 (0,089)	0,347 (0,082)	0,738 (0,069)	0,440 (0,088)	0,550 (0,077)
40%	0,870 (0,038)	0,299 (0,076)	0,560 (0,084)	0,346 (0,077)	0,731 (0,059)	0,449 (0,085)	0,574 (0,073)
50%	0,861 (0,033)	0,297 (0,070)	0,390 (0,069)	0,282 (0,067)	0,696 (0,052)	0,399 (0,070)	0,536 (0,059)

Tabela A.14: Proporção de redes na mesma classe de equivalência da rede original, na avaliação da consistência estrutural após a imputação, a partir de 500 imputações da rede réu–prisão e em diferentes percentuais de não resposta nos dados de homicídios em Campinas

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	0,772	0,940	1,000	0,672	0,766	0,888	0,676
7%	0,704	0,738	0,816	0,458	0,708	0,842	0,490
10%	0,764	0,408	0,670	0,474	0,632	0,572	0,522
20%	0,842	0,446	0,640	0,568	0,594	0,480	0,524
30%	0,666	0,412	0,306	0,508	0,608	0,658	0,464
40%	0,612	0,484	0,288	0,418	0,608	0,530	0,448
50%	0,624	0,432	0,452	0,566	0,622	0,374	0,428

Tabela A.15: Valores de Δ na avaliação da consistência estatística dos parâmetros da rede, a partir de 500 imputações da rede réu–prisão dos dados de homicídios em Campinas, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	0,074 (0,181)	0,544 (0,358)	0,610 (0,332)	0,517 (0,188)	0,123 (0,144)	0,396 (0,179)	0,265 (0,237)
7%	0,063 (0,097)	0,477 (0,188)	0,237 (0,173)	0,443 (0,202)	0,164 (0,156)	0,384 (0,227)	0,238 (0,160)
10%	0,073 (0,099)	0,233 (0,118)	0,293 (0,233)	0,308 (0,118)	0,116 (0,115)	0,209 (0,118)	0,258 (0,123)
20%	0,063 (0,063)	0,215 (0,102)	0,231 (0,106)	0,203 (0,089)	0,079 (0,061)	0,153 (0,082)	0,165 (0,084)
30%	0,040 (0,039)	0,182 (0,074)	0,088 (0,065)	0,163 (0,070)	0,079 (0,050)	0,101 (0,057)	0,124 (0,060)
40%	0,035 (0,025)	0,116 (0,052)	0,063 (0,049)	0,131 (0,067)	0,045 (0,037)	0,128 (0,068)	0,107 (0,057)
50%	0,034 (0,024)	0,114 (0,046)	0,054 (0,041)	0,112 (0,051)	0,041 (0,033)	0,078 (0,045)	0,087 (0,046)

Tabela A.16: Proporção de imputações corretas na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações na rede vítima–prisão dos dados de homicídios em Campinas, em diferentes tipos de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	0,981 (0,197)	0,499 (0,392)	0,793 (0,333)	0,525 (0,282)	0,787 (0,232)	0,573 (0,385)	0,613 (0,378)
7%	0,965 (0,111)	0,437 (0,195)	0,754 (0,207)	0,531 (0,219)	0,722 (0,206)	0,519 (0,309)	0,602 (0,221)
10%	0,951 (0,097)	0,364 (0,113)	0,698 (0,185)	0,429 (0,172)	0,768 (0,187)	0,488 (0,281)	0,587 (0,192)
20%	0,947 (0,076)	0,328 (0,099)	0,583 (0,121)	0,435 (0,119)	0,751 (0,113)	0,472 (0,196)	0,530 (0,115)
30%	0,932 (0,053)	0,305 (0,072)	0,567 (0,090)	0,392 (0,081)	0,720 (0,092)	0,416 (0,111)	0,501 (0,090)
40%	0,915 (0,041)	0,291 (0,065)	0,491 (0,082)	0,401 (0,070)	0,693 (0,081)	0,395 (0,091)	0,489 (0,073)
50%	0,898 (0,034)	0,285 (0,054)	0,328 (0,073)	0,400 (0,063)	0,615 (0,079)	0,362 (0,070)	0,473 (0,061)

Tabela A.17: Proporção de redes na mesma classe de equivalência da rede original, na avaliação da consistência estrutural após a imputação, a partir de 500 imputações da rede vítima–prisão e em diferentes percentuais de não resposta nos dados de homicídios em Campinas

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	1,000	0,960	0,898	0,665	0,672	0,780	0,615
7%	0,875	0,628	0,720	0,578	0,432	0,634	0,551
10%	0,713	0,706	0,497	0,632	0,510	0,538	0,437
20%	0,816	0,432	0,635	0,449	0,485	0,662	0,488
30%	0,634	0,527	0,519	0,373	0,396	0,450	0,320
40%	0,561	0,417	0,424	0,320	0,333	0,375	0,288
50%	0,722	0,488	0,376	0,299	0,288	0,452	0,272

Tabela A.18: Valores de Δ na avaliação da consistência estatística dos parâmetros da rede, a partir de 500 imputações da rede vítima–prisão dos dados de homicídios em Campinas, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Blocos de variável(eis) imputada(s) a partir da rede Bayesiana**						
	P_0	P_1	P_2	$P_0 - P_1$	$P_0 - P_2$	$P_1 - P_2$	$P_0 - P_1 - P_2$
5%	0,082 (0,135)	0,437 (0,232)	0,591 (0,352)	0,510 (0,298)	0,120 (0,187)	0,402 (0,153)	0,372 (0,266)
7%	0,073 (0,100)	0,302 (0,177)	0,426 (0,233)	0,403 (0,185)	0,125 (0,113)	0,398 (0,124)	0,238 (0,170)
10%	0,069 (0,091)	0,251 (0,114)	0,265 (0,170)	0,299 (0,114)	0,097 (0,094)	0,243 (0,107)	0,297 (0,113)
20%	0,060 (0,075)	0,198 (0,081)	0,217 (0,111)	0,201 (0,072)	0,082 (0,067)	0,193 (0,085)	0,191 (0,092)
30%	0,052 (0,043)	0,115 (0,073)	0,185 (0,089)	0,119 (0,059)	0,074 (0,050)	0,108 (0,068)	0,114 (0,085)
40%	0,033 (0,026)	0,102 (0,060)	0,090 (0,047)	0,072 (0,043)	0,053 (0,045)	0,097 (0,055)	0,100 (0,069)
50%	0,028 (0,019)	0,094 (0,042)	0,052 (0,034)	0,048 (0,033)	0,038 (0,040)	0,075 (0,049)	0,073 (0,057)

Apêndice B

Exemplo da importância na ordenação das variáveis para a imputação a partir das Redes Bayesianas

Considere três variáveis aleatórias A, B e C na qual se observam os resultados hipotéticos apresentados na Tabela B.1. O objetivo está em imputar itens faltantes que por ventura possam surgir na variável C.

Tabela B.1: Dados de um exemplo hipotético

	Variáveis		
Controle	A	B	C
1	a_1	b_1	c_1
2	a_1	b_1	c_1
3	a_1	b_1	c_2
4	a_2	b_1	c_2
5	a_2	b_2	c_2
6	a_2	b_2	c_2
7	a_1	b_2	c_1
8	a_2	b_1	c_2
9	a_1	b_1	c_1
10	a_2	b_1	c_1

No grafo em (1) na Figura 4.1 reproduzida a seguir na Figura B.1, a imputação seria conduzida sob $P(C|B)$, ou seja, de acordo com:

$$P(C|B) = \begin{cases} P(C = c_1|B = b_1) = \frac{4}{7} \\ P(C = c_2|B = b_1) = \frac{3}{7} \\ P(C = c_1|B = b_2) = \frac{1}{3} \\ P(C = c_2|B = b_2) = \frac{2}{3} \end{cases}$$

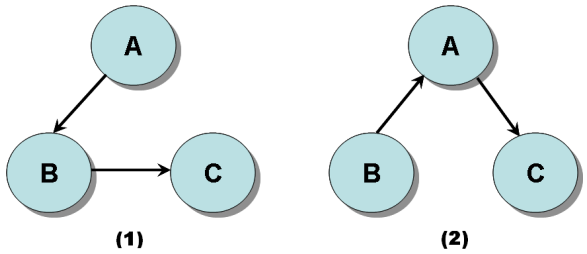


Figura B.1: Reprodução da Figura 4.1 – Grafos do exemplo da importância da ordenação das variáveis para imputação: (1) grafo da rede que descreve o melhor relacionamento entre as variáveis (2) grafo da rede que ordena conforme a confiabilidade das variáveis

Já no grafo em (2) na mesma figura, a imputação seria conduzida sob $P(C|A)$,

ou seja, de acordo com:

$$P(C|A) = \begin{cases} P(C = c_1|A = a_1) = \frac{4}{5} \\ P(C = c_2|A = a_1) = \frac{1}{5} \\ P(C = c_1|A = a_2) = \frac{1}{5} \\ P(C = c_2|A = a_2) = \frac{4}{5} \end{cases}$$

Se calcularmos uma tabela de contingência entre as variáveis B e C após a imputação em C de acordo com as redes definidas pelos grafos das Figuras 4.1 (1) e 4.1 (2), teríamos como valores esperados em B e C o seguinte:

$$\begin{aligned} E[BC|(1)] &= \sum_{i,j=1}^{2} b_i c_j P(B = b_i, C = c_j) \\ &= b_1 c_1 \frac{4}{10} + b_2 c_1 \frac{1}{10} + b_1 c_2 \frac{3}{10} + b_2 c_2 \frac{2}{10} \end{aligned}$$

e

$$\begin{aligned} E[BC|(2)] &= E[B]E[C] \\ &= \left[b_1 \frac{7}{10} + b_2 \frac{3}{10}\right]\left[c_1 \frac{5}{10} + c_2 \frac{5}{10}\right] \\ &= b_1 c_1 \frac{35}{100} + b_2 c_1 \frac{15}{100} + b_1 c_2 \frac{35}{100} + b_2 c_2 \frac{15}{100}. \end{aligned}$$

Portanto, a partir de um exemplo conduzido sob uma estrutura simples, mostra–se que $E[BC|(1)] \neq E[BC|(2)]$ em função da ordenação entre as variáveis para imputação a partir da rede Bayesiana. Para uma estrutura mais complexa ou então para tabelas de mais alta ordem, as diferenças podem ser maiores.

Apêndice C

Tabelas de resultados das simulações em redes mistas

Tabela C.1: Valores de ξ_{RENDA} na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista domicílio–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	977,452 (89,328)	794,323 (85,771)	852,766 (81,175)	839,210 (79,492)
3%	295,703 (12,712)	262,118 (14,669)	273,351 (12,466)	278,243 (15,931)
5%	180,098 (8,163)	175,442 (7,730)	172,867 (7,528)	179,081 (7,650)
7%	107,003 (4,596)	116,107 (4,301)	119,028 (4,397)	114,380(5,042)
10%	90,876 (3,013)	88,786 (2,468)	85,206 (2,282)	86,904 (2,739)
20%	40,652 (1,008)	43,961 (0,980)	42,733 (0,916)	42,206 (0,903)
30%	25,954 (0,312)	27,729 (0,402)	27,125 (0,390)	26,899 (0,367)
40%	20,824 (0,233)	21,695 (0,245)	21,142 (0,286)	22,073 (0,278)
50%	14,463 (0,197)	10,792 (0,206)	12,877 (0,214)	12,659 (0,205)

Tabela C.2: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista domicílio–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,263 (0,052)	0,104 (0,060)	0,091 (0,057)	0,242 (0,049)
3%	0,249 (0,041)	0,065 (0,054)	0,154 (0,042)	0,170 (0,039)
5%	0,118 (0,029)	0,228 (0,037)	0,233 (0,039)	0,094 (0,025)
7%	0,197 (0,022)	0,293 (0,024)	0,112 (0,031)	0,155 (0,020)
10%	0,095 (0,017)	0,199 (0,015)	0,282 (0,027)	0,168 (0,012)
20%	0,212 (0,011)	0,128 (0,010)	0,206 (0,010)	0,142 (0,010)
30%	0,245 (0,010)	0,075 (0,008)	0,191 (0,010)	0,296 (0,009)
40%	0,172 (0,008)	0,206 (0,007)	0,110 (0,007)	0,163 (0,008)
50%	0,136 (0,008)	0,188 (0,007)	0,107 (0,007)	0,251 (0,008)

Tabela C.3: Valores de ξ_R na avaliação da consistência lógica após a imputação, a partir de 500 imputações de uma rede mista domicílio–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,972 (0,021)	0,966 (0,029)	0,904 (0,020)	0,900 (0,022)
3%	0,965 (0,014)	0,960 (0,012)	0,899 (0,019)	0,898 (0,017)
5%	0,974 (0,009)	0,979 (0,009)	0,912 (0,010)	0,896 (0,010)
7%	0,972 (0,009)	0,952 (0,008)	0,909 (0,008)	0,896(0,009)
10%	0,961 (0,007)	0,940 (0,008)	0,910 (0,007)	0,872 (0,008)
20%	0,959 (0,007)	0,943 (0,007)	0,911 (0,006)	0,865 (0,008)
30%	0,953 (0,005)	0,935 (0,006)	0,889 (0,006)	0,862 (0,007)
40%	0,946 (0,003)	0,933 (0,005)	0,906 (0,004)	0,859 (0,004)
50%	0,941 (0,002)	0,930 (0,002)	0,903 (0,003)	0,853 (0,003)

Tabela C.4: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista domicílio–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
3%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
5%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
7%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
10%	400,000 (0,000)	402,057 (1,086)	400,000 (0,000)	401,072 (1,035)
20%	400,000 (0,000)	409,891 (2,455)	407,133 (3,878)	400,928 (1,919)
30%	410,157 (4,318)	433,080 (1,651)	442,270 (2,148)	415,771 (1,068)
40%	456,226 (2,915)	464,301 (0,978)	469,158 (1,662)	438,997 (2,720)
50%	470,849 (1,172)	485,039 (1,005)	481,954 (0,876)	465,820 (1,551)

Tabela C.5: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista domicílio–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	532,009 (0,115)	532,115 (0,247)	532,017 (0,315)	532,132 (0,212)
3%	532,287 (0,485)	532,978 (0,119)	532,226 (0,246)	532,095 (0,365)
5%	532,509 (0,689)	533,219 (0,362)	532,970 (0,712)	532,854 (0,410)
7%	534,871 (0,720)	534,082 (0,478)	533,465 (0,688)	532,993 (0,761)
10%	534,996 (0,915)	535,550 (0,925)	534,526 (0,913)	534,139 (0,874)
20%	535,021 (0,982)	535,964 (1,110)	536,118 (0,874)	535,761 (1,082)
30%	536,644 (1,173)	537,261 (0,964)	537,469 (1,138)	536,446 (1,280)
40%	538,175 (1,948)	538,423 (2,127)	538,392 (1,454)	538,678 (1,369)
50%	539,002 (2,064)	538,938 (1,842)	540,487 (0,993)	539,175 (1,644)

Tabela C.6: Valores de ξ_{RENDA} na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	1055,723 (92,662)	986,631 (90,175)	902,225 (90,050)	977,142 (91,674)
3%	274,425 (13,488)	269,130 (14,024)	251,772 (13,951)	270,068 (13,692)
5%	168,326 (7,942)	172,949 (7,671)	171,865 (8,044)	171,676 (7,853)
7%	102,481 (5,027)	115,783 (4,984)	114,269 (4,197)	118,902 (4,701)
10%	91,405 (2,983)	90,881 (3,127)	90,521 (2,836)	92,922 (2,998)
20%	45,742 (1,877)	40,867 (0,952)	41,339 (1,840)	40,801 (1,695)
30%	22,808 (0,365)	25,369 (0,415)	25,972 (0,428)	27,466 (0,479)
40%	20,769 (0,221)	21,040 (0,276)	20,983 (0,219)	20,828 (0,191)
50%	9,804 (0,172)	11,158 (0,114)	11,990 (0,087)	10,434 (0,103)

Tabela C.7: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,249 (0,057)	0,103 (0,053)	0,097 (0,050)	0,234 (0,051)
3%	0,232 (0,031)	0,082 (0,043)	0,182 (0,047)	0,195 (0,044)
5%	0,193 (0,026)	0,315 (0,035)	0,116 (0,032)	0,238 (0,033)
7%	0,187 (0,023)	0,228 (0,022)	0,214 (0,029)	0,111 (0,025)
10%	0,221 (0,021)	0,195 (0,019)	0,073 (0,018)	0,297 (0,017)
20%	0,215 (0,014)	0,207 (0,011)	0,244 (0,015)	0,080 (0,013)
30%	0,098 (0,010)	0,054 (0,010)	0,192 (0,010)	0,153 (0,011)
40%	0,251 (0,009)	0,289 (0,008)	0,211 (0,009)	0,149 (0,009)
50%	0,147 (0,008)	0,116 (0,008)	0,275 (0,009)	0,206 (0,008)

Tabela C.8: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede pessoa–renda, em diferentes percentuais de não resposta

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)$\|P_0$	LN(RENDA)$\|P_1$	LN(RENDA)$\|P_0, P_1$
1%	1,000	1,000	1,000	1,000
3%	1,000	0,980	1,000	1,000
5%	0,900	0,960	0,960	0,980
7%	0,998	0,860	0,960	0,980
10%	0,860	0,880	0,880	0,980
20%	0,920	0,900	0,880	0,900
30%	0,900	0,860	0,960	0,820
40%	0,840	0,840	0,840	0,800
50%	0,900	0,820	0,780	0,920

Tabela C.9: Valores de ξ_R na avaliação da consistência lógica após a imputação, a partir de 500 imputações de uma rede mista pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,815 (0,030)	0,810 (0,025)	0,803 (0,024)	0,800 (0,028)
3%	0,813 (0,010)	0,809 (0,017)	0,801 (0,011)	0,791 (0,016)
5%	0,813 (0,009)	0,805 (0,009)	0,799 (0,010)	0,790 (0,010)
7%	0,809 (0,009)	0,798 (0,008)	0,801 (0,009)	0,784(0,009)
10%	0,808 (0,007)	0,796 (0,006)	0,782 (0,007)	0,781 (0,008)
20%	0,811 (0,006)	0,795 (0,005)	0,784 (0,007)	0,773 (0,006)
30%	0,802 (0,005)	0,795 (0,004)	0,779 (0,004)	0,772 (0,004)
40%	0,800 (0,005)	0,790 (0,004)	0,780 (0,003)	0,760 (0,003)
50%	0,808 (0,003)	0,767 (0,003)	0,778 (0,002)	0,753 (0,002)

Tabela C.10: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
3%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
5%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
7%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
10%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
20%	421,365 (2,492)	408,711 (0,972)	400,000 (0,000)	410,093 (1,893)
30%	438,127 (1,768)	419,135 (1,054)	411,738 (1,132)	419,741 (2,452)
40%	440,498 (2,915)	422,360 (2,688)	430,971 (2,085)	425,162 (2,634)
50%	449,277 (3,139)	437,732 (2,905)	442,817 (2,886)	433,997 (3,042)

Tabela C.11: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	532,004 (0,124)	532,100 (0,244)	532,019 (0,192)	532,108 (0,136)
3%	532,136 (0,230)	532,189 (0,297)	532,254 (0,356)	532,326 (0,277)
5%	532,497 (0,486)	532,504 (0,360)	532,638 (0,393)	532,712 (0,384)
7%	533,048 (0,557)	534,915 (0,417)	532,992 (0,408)	533,185 (0,439)
10%	533,395 (0,743)	533,826 (0,724)	534,507 (0,814)	534,067 (0,761)
20%	534,508 (0,779)	535,991 (0,931)	535,183 (0,932)	535,246 (0,910)
30%	535,246 (1,130)	536,016 (0,996)	536,246 (1,205)	536,325 (1,138)
40%	538,148 (2,016)	542,915 (1,361)	539,172 (1,608)	539,488 (1,722)
50%	537,932 (1,948)	543,366 (1,650)	541,409 (1,704)	541,719 (1,925)

Tabela C.12: Valores de ξ_{RENDA} na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista domicílio–pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	1065,640 (128,201)	939,768 (99,458)	841,906 (81,859)	840,478 (82,462)
3%	300,631 (15,772)	279,233 (13,399)	276,511 (14,479)	294,423 (14,515)
5%	175,693 (7,624)	177,884 (7,471)	163,068 (7,398)	176,597 (7,657)
7%	116,809 (3,658)	116,984 (4,752)	126,284 (3,506)	120,708(4,085)
10%	86,453 (2,556)	82,916 (2,428)	84,317 (2,836)	87,720 (2,047)
20%	41,667 (0,928)	43,791 (0,751)	44,892 (0,870)	44,479 (0,965)
30%	28,198 (0,537)	29,077 (0,498)	29,632 (0,432)	30,619 (0,575)
40%	20,952 (0,272)	22,662 (0,337)	22,825 (0,339)	24,125 (0,305)
50%	17,110 (0,213)	18,378 (0,225)	18,537 (0,206)	19,851 (0,246)

Tabela C.13: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista domicílio–pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,055 (0,059)	0,351 (0,052)	0,208 (0,049)	0,229 (0,056)
3%	0,296 (0,043)	0,197 (0,045)	0,041 (0,037)	0,163 (0,044)
5%	0,352 (0,035)	0,283 (0,033)	0,152 (0,024)	0,310 (0,032)
7%	0,148 (0,027)	0,116 (0,021)	0,196 (0,017)	0,278 (0,029)
10%	0,193 (0,015)	0,322 (0,018)	0,137 (0,012)	0,294 (0,023)
20%	0,094 (0,011)	0,295 (0,011)	0,113 (0,010)	0,181 (0,012)
30%	0,202 (0,009)	0,075 (0,010)	0,279 (0,009)	0,230 (0,010)
40%	0,105 (0,008)	0,114 (0,009)	0,194 (0,008)	0,254 (0,009)
50%	0,224 (0,007)	0,236 (0,008)	0,188 (0,007)	0,101 (0,007)

Tabela C.14: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede domicílio–pessoa–renda, em diferentes percentuais de não resposta

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)$\|P_0$	LN(RENDA)$\|P_1$	LN(RENDA)$\|P_0, P_1$
1%	1,000	0,980	1,000	1,000
3%	1,000	1,000	0,800	0,980
5%	0,980	1,000	0,960	0,980
7%	0,860	0,888	0,980	0,800
10%	0,900	0,900	0,700	0,920
20%	0,800	0,860	0,760	0,840
30%	0,840	0,720	0,880	0,900
40%	0,760	0,820	0,800	0,760
50%	0,820	0,820	0,780	0,760

Tabela C.15: Valores de ξ_R na avaliação da consistência lógica após a imputação, a partir de 500 imputações de uma rede mista domicílio–pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	0,807 (0,020)	0,804 (0,024)	0,812 (0,018)	0,814 (0,021)
3%	0,797 (0,015)	0,792 (0,012)	0,812 (0,011)	0,780 (0,015)
5%	0,798 (0,008)	0,787 (0,010)	0,807 (0,009)	0,800 (0,009)
7%	0,809 (0,009)	0,807 (0,009)	0,793 (0,009)	0,801(0,008)
10%	0,798 (0,006)	0,803 (0,007)	0,810 (0,007)	0,790 (0,006)
20%	0,802 (0,004)	0,787 (0,004)	0,792 (0,005)	0,791 (0,005)
30%	0,798 (0,004)	0,796 (0,005)	0,792 (0,004)	0,791 (0,005)
40%	0,800 (0,003)	0,789 (0,003)	0,788 (0,004)	0,785 (0,003)
50%	0,795 (0,002)	0,788 (0,003)	0,785 (0,004)	0,779 (0,003)

Tabela C.16: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista domicílio–pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
3%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
5%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
7%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
10%	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)	400,000 (0,000)
20%	417,012 (2,849)	411,980 (3,314)	414,672 (3,455)	414,556 (3,622)
30%	433,436 (3,188)	434,892 (3,174)	435,626 (3,596)	439,618 (3,131)
40%	449,416 (1,241)	449,808 (0,673)	449,424 (1,215)	449,540 (1,095)
50%	450,292 (0,831)	452,686 (1,744)	451,526 (1,589)	452,270 (1,579)

Tabela C.17: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista domicílio–pessoa–renda, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**			
	LN(RENDA)	LN(RENDA)–P_0	LN(RENDA)–P_1	LN(RENDA)–P_0–P_1
1%	532,092 (0,174)	532,160 (0,216)	532,338 (0,223)	532,328 (0,229)
3%	532,438 (0,343)	532,100 (0,426)	532,718 (0,368)	531,620 (0,356)
5%	532,530 (0,535)	531,834 (0,471)	532,886 (0,425)	532,782 (0,399)
7%	533,608 (0,586)	533,682 (0,526)	532,698 (0,528)	533,252 (0,608)
10%	533,242 (0,687)	534,014 (0,686)	534,622 (0,602)	532,748 (0,690)
20%	535,236 (0,893)	532,664 (0,773)	533,946 (1,010)	534,372 (0,967)
30%	535,164 (1,192)	536,784 (0,936)	535,660 (0,996)	537,146 (1,209)
40%	537,356 (1,146)	536,564 (1,077)	536,288 (1,251)	537,986 (1,219)
50%	535,370 (1,432)	538,056 (1,501)	535,406 (1,936)	537,984 (1,984)

Tabela C.18: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista vítima–crime–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TMP	TMP–P_0	TMP–P_1	TMP–P_2	TMP–P_0–P_1	TMP–P_0–P_2	TMP–P_1–P_2	TMP–P_0–P_1–P_2
5%	-0,024 (0,368)	0,539 (0,521)	0,166 (0,446)	-0,646 (0,350)	0,357 (0,513)	0,157 (0,606)	0,356 (0,580)	0,202 (0,461)
7%	-0,058 (0,336)	0,094 (0,530)	0,422 (0,511)	0,256 (0,356)	0,080 (0,606)	-0,414 (0,362)	0,165 (0,425)	0,134 (0,483)
10%	0,098 (0,333)	0,214 (0,319)	-0,002 (0,369)	0,210 (0,263)	0,043 (0,621)	0,188 (0,365)	0,179 (0,350)	0,387 (0,422)
20%	-0,049 (0,219)	0,082 (0,235)	0,068 (0,228)	0,100 (0,191)	0,228 (0,241)	0,168 (0,244)	0,260 (0,234)	0,036 (0,238)
30%	0,076 (0,224)	0,250 (0,199)	0,145 (0,237)	0,172 (0,209)	0,147 (0,207)	0,169 (0,225)	0,143 (0,194)	0,273 (0,209)
40%	0,065 (0,133)	0,015 (0,146)	0,066 (0,166)	0,177 (0,151)	0,101 (0,137)	0,048 (0,158)	0,128 (0,147)	0,107 (0,182)
50%	0,186 (0,158)	0,209 (0,136)	0,120 (0,198)	0,197 (0,141)	0,106 (0,146)	0,200 (0,159)	0,045 (0,132)	0,184 (0,204)

Tabela C.19: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede mista vítima–crime–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TMP	TMP–P_0	TMP–P_1	TMP–P_2	TMP–P_0–P_1	TMP–P_0–P_2	TMP–P_1–P_2	TMP–P_0–P_1–P_2
5%	0,820	0,600	0,520	0,380	0,400	0,780	0,300	0,480
7%	0,680	0,520	0,680	0,220	0,560	0,200	0,620	0,700
10%	0,440	0,480	0,600	0,280	0,620	0,460	0,540	0,360
20%	0,800	0,320	0,440	0,400	0,620	0,500	0,400	0,540
30%	0,600	0,400	0,360	0,520	0,540	0,440	0,480	0,400
40%	0,560	0,240	0,500	0,360	0,600	0,520	0,660	0,320
50%	0,380	0,580	0,200	0,700	0,380	0,500	0,660	0,400

Tabela C.20: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista vítima–crime–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TMP	TMP–P_0	TMP–P_1	TMP–P_2	TMP–P_0–P_1	TMP–P_0–P_2	TMP–P_1–P_2	TMP–P_0–P_1–P_2
5%	40,000 (0,000)	40,000 (0,000)	40,872 (0,157)	40,000 (0,000)	40,136 (0,140)	40,000 (0,000)	40,000 (0,000)	40,000 (0,000)
7%	40,133 (0,198)	40,291 (0,115)	40,426 (0,293)	40,000 (0,000)	40,254 (0,131)	40,143 (0,119)	40,096 (0,242)	40,132 (0,101)
10%	41,246 (0,230)	40,981 (0,192)	40,973 (0,181)	40,654 (0,203)	40,763 (0,175)	41,803 (0,241)	40,419 (0,127)	40,903 (0,115)
20%	43,497 (0,832)	42,801 (0,736)	43,106 (0,849)	43,778 (0,942)	44,008 (0,815)	43,662 (0,808)	42,938 (0,762)	43,910 (0,917)
30%	45,860 (1,291)	46,698 (1,321)	45,342 (1,178)	45,980 (1,223)	48,123 (1,367)	46,374 (1,271)	47,243 (1,196)	45,311 (1,273)
40%	48,107 (1,198)	47,367 (1,662)	48,634 (1,247)	50,082 (1,561)	48,948 (1,420)	47,441 (1,638)	49,175 (1,526)	47,708 (1,457)
50%	49,604 (2,201)	49,972 (1,906)	50,311 (1,873)	60,863 (1,734)	49,366 (1,930)	50,476 (1,684)	52,935 (1,871)	49,412 (1,935)

Tabela C.21: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista vítima–crime–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TMP	TMP–P_0	TMP–P_1	TMP–P_2	TMP–P_0–P_1	TMP–P_0–P_2	TMP–P_1–P_2	TMP–P_0–P_1–P_2
5%	202,142 (0,672)	202,725 (0,418)	202,238 (0,135)	202,244 (0,669)	202,821 (0,528)	202,467 (0,192)	201,982 (0,650)	202,769 (0,817)
7%	202,266 (0,781)	202,633 (0,542)	202,545 (0,860)	202,182 (0,932)	203,906 (0,981)	202,205 (0,834)	202,139 (0,667)	202,260 (0,754)
10%	203,154 (0,902)	203,026 (0,828)	202,905 (0,707)	202,766 (0,832)	202,804 (0,661)	202,114 (0,722)	203,518 (0,786)	202,476 (0,843)
20%	202,575 (1,356)	203,403 (1,189)	202,118 (1,045)	203,064 (1,576)	203,224 (1,167)	202,863 (1,902)	203,237 (1,554)	202,686 (1,696)
30%	203,027 (1,638)	204,002 (1,973)	203,826 (1,578)	204,651 (1,637)	203,385 (1,822)	203,186 (1,942)	202,527 (1,979)	203,786 (1,641)
40%	203,030 (1,986)	204,701 (1,743)	203,128 (1,180)	202,917 (1,945)	203,254 (2,031)	203,972 (2,004)	203,014 (1,619)	204,026 (1,802)
50%	204,618 (2,528)	203,154 (2,097)	204,296 (1,923)	203,503 (2,057)	203,279 (2,183)	204,390 (2,020)	205,084 (3,301)	203,693 (2,255)

Tabela C.22: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TFASEFI	TFASEFI–P_0	TFASEFI–P_1	TFASEFI–P_2	TFASEFI–P_0–P_1	TFASEFI–P_0–P_2	TFASEFI–P_1–P_2	TFASEFI–P_0–P_1–P_2
5%	-0,167 (0,416)	0,112 (0,238)	-0,185 (0,513)	-0,308 (0,492)	0,313 (0,495)	-0,617 (0,235)	0,228 (0,658)	0,684 (0,730)
7%	0,493 (0,318)	-0,053 (0,411)	-0,026 (0,704)	-0,092 (0,385)	0,120 (0,994)	0,215 (0,312)	0,384 (0,353)	0,161 (0,601)
10%	-0,225 (0,346)	0,661 (0,281)	0,103 (0,345)	-0,284 (0,407)	0,267 (0,303)	0,317 (0,227)	0,323 (0,494)	-0,010 (0,302)
20%	0,052 (0,228)	0,084 (0,199)	0,142 (0,240)	0,136 (0,235)	-0,028 (0,228)	0,163 (0,201)	0,058 (0,200)	0,237 (0,233)
30%	0,154 (0,196)	0,154 (0,172)	0,002 (0,156)	0,121 (0,180)	0,232 (0,256)	0,084 (0,170)	0,139 (0,190)	0,086 (0,176)
40%	0,174 (0,199)	0,137 (0,183)	0,102 (0,169)	0,179 (0,195)	0,085 (0,139)	0,190 (0,206)	0,016 (0,143)	0,128 (0,186)
50%	0,162 (0,176)	0,102 (0,180)	0,192 (0,163)	0,108 (0,163)	0,154 (0,177)	0,103 (0,150)	0,167 (0,156)	0,154 (0,170)

Tabela C.23: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede mista crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TFASEFI	TFASEFI–P_0	TFASEFI–P_1	TFASEFI–P_2	TFASEFI–P_0–P_1	TFASEFI–P_0–P_2	TFASEFI–P_1–P_2	TFASEFI–P_0–P_1–P_2
5%	0,640	0,800	0,300	0,500	0,600	0,480	0,620	0,700
7%	0,720	0,500	0,540	0,880	0,420	0,300	0,480	0,360
10%	0,480	0,480	0,620	0,200	0,420	0,260	0,500	0,400
20%	0,500	0,520	0,640	0,180	0,360	0,500	0,580	0,620
30%	0,400	0,600	0,300	0,600	0,660	0,420	0,560	0,400
40%	0,640	0,500	0,220	0,420	0,540	0,640	0,180	0,420
50%	0,720	0,320	0,400	0,380	0,520	0,440	0,200	0,500

Tabela C.24: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TFASEFI	TFASEFI–P_0	TFASEFI–P_1	TFASEFI–P_2	TFASEFI–P_0–P_1	TFASEFI–P_0–P_2	TFASEFI–P_1–P_2	TFASEFI–P_0–P_1–P_2
5%	3038,000 (0,000)	3038,020 (0,095)	3038,105 (0,132)	3038,000 (0,000)	3038,000 (0,000)	3038,133 (0,101)	3038,575 (0,182)	3038,273 (0,218)
7%	3038,720 (0,289)	3039,212 (0,968)	3039,571 (0,867)	3038,686 (0,753)	3040,503 (0,665)	3039,299 (0,538)	3039,548 (0,697)	3040,029 (0,651)
10%	3040,231 (0,279)	3045,250 (0,326)	3043,103 (0,488)	3041,274 (0,505)	3045,411 (0,601)	3042,051 (0,520)	3045,227 (0,618)	3045,397 (0,676)
20%	3098,067 (0,660)	3116,119 (0,834)	3125,644 (0,607)	3082,160 (0,733)	3080,857 (0,933)	3089,926 (0,918)	3107,241 (0,903)	3113,603 (0,881)
30%	3202,169 (1,798)	3249,302 (1,837)	3282,235 (1,682)	3277,025 (1,381)	3107,225 (1,616)	3253,414 (1,603)	3209,338 (1,512)	3214,488 (1,377)
40%	3288,703 (1,417)	3275,022 (1,894)	3279,128 (1,722)	3291,408 (1,652)	3293,861 (1,632)	3297,089 (1,646)	3250,452 (1,784)	3272,942 (1,863)
50%	3325,581 (1,770)	3304,810 (1,647)	3290,459 (1,780)	3306,932 (1,810)	3376,472 (1,690)	3343,215 (1,891)	3309,051 (1,972)	3315,132 (1,858)

Tabela C.25: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TFASEFI	TFASEFI–P_0	TFASEFI–P_1	TFASEFI–P_2	TFASEFI–P_0–P_1	TFASEFI–P_0–P_2	TFASEFI–P_1–P_2	TFASEFI–P_0–P_1–P_2
5%	3690,055 (0,826)	3691,382 (0,924)	3690,129 (0,735)	3690,337 (0,855)	3690,106 (0,824)	3690,074 (0,853)	3690,293 (0,822)	3690,091 (0,778)
7%	3691,691 (1,254)	3692,018 (1,356)	3690,920 (1,384)	3692,221 (1,218)	3692,448 (1,404)	3692,173 (1,352)	3692,248 (1,472)	3692,336 (1,378)
10%	3690,483 (1,964)	3691,382 (1,475)	3692,100 (1,584)	3696,045 (1,197)	3692,128 (1,482)	3692,302 (1,157)	3692,088 (1,415)	3692,155 (1,462)
20%	3693,402 (2,574)	3692,207 (1,236)	3692,982 (1,499)	3693,340 (1,254)	3693,410 (1,163)	3692,947 (1,092)	3692,185 (1,292)	3692,236 (1,395)
30%	3700,248 (3,242)	3699,450 (2,813)	3698,140 (2,506)	3699,751 (2,732)	3700,127 (2,041)	3700,348 (2,152)	3700,172 (2,134)	3700,676 (1,851)
40%	3701,108 (2,963)	3693,533 (2,206)	3695,927 (2,146)	3700,124 (2,458)	3701,653 (2,216)	3701,129 (2,158)	3701,101 (2,238)	3701,262 (2,299)
50%	3700,459 (2,196)	3701,089 (2,207)	3700,286 (2,803)	3701,192 (2,606)	3701,029 (2,885)	3702,005 (2,926)	3701,156 (2,790)	3701,271 (2,984)

Tabela C.26: Valores do coeficiente de correlação entre o observado e o imputado na avaliação da consistência da base de dados (microdados) após a imputação, a partir de 500 imputações de uma rede mista réu–crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TPOL1	TPOL1–P_0	TPOL1–P_1	TPOL1–P_2	TPOL1–P_0–P_1	TPOL1–P_0–P_2	TPOL1–P_1–P_2	TPOL1–P_0–P_1–P_2
5%	0,691 (0,544)	0,339 (0,470)	-0,562 (0,281)	0,130 (0,559)	0,077 (0,636)	0,294 (0,493)	0,118 (0,652)	-0,019 (0,430)
7%	-0,245 (0,714)	0,348 (0,527)	0,268 (0,710)	0,371 (0,402)	0,044 (0,467)	0,256 (0,424)	0,411 (0,504)	-0,216 (0,455)
10%	0,023 (0,294)	0,222 (0,324)	-0,085 (0,336)	0,361 (0,423)	-0,031 (0,394)	-0,149 (0,673)	0,238 (0,337)	-0,097 (0,695)
20%	0,067 (0,215)	0,139 (0,217)	0,083 (0,270)	0,008 (0,226)	0,063 (0,260)	0,011 (0,162)	0,077 (0,271)	0,186 (0,231)
30%	0,108 (0,196)	0,234 (0,227)	0,058 (0,151)	0,060 (0,168)	0,067 (0,202)	0,062 (0,169)	0,003 (0,181)	0,103 (0,185)
40%	0,163 (0,193)	0,123 (0,197)	0,120 (0,173)	0,167 (0,190)	0,114 (0,202)	0,158 (0,199)	0,138 (0,154)	0,096 (0,172)
50%	0,027 (0,126)	0,055 (0,152)	0,178 (0,161)	0,082 (0,133)	0,083 (0,136)	0,146 (0,179)	0,147 (0,174)	0,097 (0,169)

Tabela C.27: Proporção de redes na mesma classe de equivalência da rede original na avaliação da consistência estrutural após a imputação, a partir de 500 imputações de uma rede mista réu–crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TPOL1	TPOL1–P_0	TPOL1–P_1	TPOL1–P_2	TPOL1–P_0–P_1	TPOL1–P_0–P_2	TPOL1–P_1–P_2	TPOL1–P_0–P_1–P_2
5%	1,000	0,500	0,580	0,540	1,000	0,660	0,400	0,900
7%	0,980	0,520	0,540	0,660	0,680	0,620	0,800	0,660
10%	0,860	0,520	0,460	0,520	0,960	0,760	0,520	0,400
20%	0,560	0,400	0,420	0,840	0,740	0,540	0,320	0,720
30%	0,420	0,420	0,500	0,520	0,820	0,380	0,440	0,280
40%	0,380	0,620	0,560	0,560	0,560	0,420	0,360	0,540
50%	0,340	0,680	0,440	0,420	0,480	0,480	0,460	0,620

Tabela C.28: Valores da mediana na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista réu–crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TPOL1	TPOL1–P_0	TPOL1–P_1	TPOL1–P_2	TPOL1–P_0–P_1	TPOL1–P_0–P_2	TPOL1–P_1–P_2	TPOL1–P_0–P_1–P_2
5%	5,000 (0,000)	5,000 (0,000)	5,000 (0,000)	5,000 (0,000)	5,960 (0,198)	5,354 (0,460)	5,000 (0,000)	5,432 (0,476)
7%	5,944 (0,227)	5,237 (0,407)	5,960 (0,198)	5,384 (0,467)	6,288 (0,441)	5,000 (0,000)	5,342 (0,161)	5,622 (0,475)
10%	5,554 (0,486)	6,386 (0,463)	5,204 (0,398)	5,559 (0,681)	5,406 (0,480)	5,270 (0,460)	5,412 (0,518)	5,312 (0,574)
20%	8,566 (0,763)	8,436 (1,160)	6,758 (0,837)	9,051 (1,135)	7,378 (0,786)	5,908 (0,260)	6,382 (0,718)	6,786 (0,317)
30%	8,751 (1,191)	11,074 (1,232)	8,274 (1,215)	7,439 (1,363)	8,003 (1,092)	10,787 (1,657)	8,368 (1,611)	9,983 (1,551)
40%	11,471 (1,577)	8,157 (1,206)	9,451 (1,352)	13,368 (1,050)	9,205 (1,663)	8,281 (1,502)	9,932 (1,685)	10,279 (1,759)
50%	9,209 (1,303)	10,719 (1,346)	14,512 (1,469)	12,617 (1,362)	11,757 (1,318)	10,701 (1,807)	11,651 (1,920)	11,619 (2,008)

Tabela C.29: Valores da média na avaliação da consistência estatística após a imputação, a partir de 500 imputações de uma rede mista réu–crime–papéis–tempo, em diferentes percentuais de não resposta (os valores entre parênteses correspondem aos desvios das estimativas)

Percentual de não resposta na variável	**Variável(eis) imputada(s) a partir da rede Bayesiana**							
	TPOL1	TPOL1–P_0	TPOL1–P_1	TPOL1–P_2	TPOL1–P_0–P_1	TPOL1–P_0–P_2	TPOL1–P_1–P_2	TPOL1–P_0–P_1–P_2
5%	20,429 (0,647)	19,439 (0,152)	20,013 (0,436)	20,549 (0,655)	20,638 (0,673)	20,065 (0,669)	20,936 (0,674)	20,111 (0,424)
7%	20,500 (0,609)	20,661 (0,691)	19,702 (0,590)	20,573 (0,775)	21,204 (0,652)	19,579 (0,777)	21,675 (0,602)	19,995 (0,472)
10%	20,614 (0,644)	21,586 (1,209)	19,521 (0,593)	20,683 (0,928)	19,223 (0,796)	20,938 (0,789)	21,881 (0,749)	21,601 (0,762)
20%	20,313 (0,927)	20,625 (0,867)	21,323 (1,486)	20,760 (1,473)	20,344 (1,561)	20,804 (1,220)	20,936 (1,187)	20,704 (1,608)
30%	21,011 (1,454)	20,960 (1,806)	20,981 (1,301)	20,346 (1,502)	20,741 (1,554)	20,237 (1,808)	21,376 (1,427)	20,323 (1,942)
40%	20,923 (1,505)	19,747 (1,774)	20,245 (1,961)	21,688 (1,993)	20,960 (1,860)	21,387 (1,729)	20,243 (1,943)	20,168 (1,670)
50%	20,448 (1,585)	21,848 (2,278)	21,014 (2,436)	20,682 (1,720)	20,332 (1,823)	20,857 (2,525)	20,912 (2,063)	20,942 (2,017)

Apêndice D

Tabelas de resultados das simulações em imputação múltipla para avaliação de redes mistas

Tabela D.1: Inferências com base em imputação múltipla ($m = 20$) para a média da renda imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	$\widehat{\theta} = \overline{\mu}(\widetilde{X})$	$\sqrt{Var_t(\widehat{\theta})}$	$\sqrt{Var_d(\widehat{\theta})}$	$Var_e(\widehat{\theta})$	$\widehat{r}$	$\widehat{\lambda}\%$
1%	532,427	352,412	352,412	0,033	$2{,}800 \times 10^{-07}$	$8{,}251 \times 10^{-15}$
3%	532,189	352,187	352,186	0,153	$1{,}296 \times 10^{-06}$	$1{,}767 \times 10^{-13}$
5%	532,786	351,671	351,670	0,324	$2{,}754 \times 10^{-06}$	$7{,}984 \times 10^{-13}$
7%	533,251	352,131	352,131	0,289	$2{,}454 \times 10^{-06}$	$6{,}331 \times 10^{-13}$
10%	533,946	351,179	351,178	0,707	$6{,}031 \times 10^{-06}$	$3{,}829 \times 10^{-12}$
20%	534,509	347,130	347,128	1,421	$1{,}242 \times 10^{-05}$	$1{,}623 \times 10^{-11}$
30%	535,057	346,574	346,572	1,119	$9{,}806 \times 10^{-06}$	$1{,}012 \times 10^{-11}$
40%	536,981	346,848	346,846	1,357	$1{,}187 \times 10^{-05}$	$1{,}484 \times 10^{-11}$
50%	537,512	346,093	346,090	2,364	$2{,}078 \times 10^{-05}$	$4{,}544 \times 10^{-11}$

Tabela D.2: Parâmetros de imputação da variável LOG(RENDA) para a parte contínua na rede mista domicílio–pessoa–renda nos dados do Censo Demográfico

Variáveis			**Parâmetros**	
SEXO	QTDBAHN	CURSOELV	$\widehat{\beta}_{0,X_j\|pa_D(X_j)}$	$\sigma^2_{pa_D(X_j)}$
1 – Masculino	0	1 – nenhum	5,589	0,200
		2 – básico/fundamental	5,651	0,193
		3 – médio/ 2º grau	5,780	0,266
		4 – superior/ pós graduação	6,234	0,475
	1	1 – nenhum	5,763	0,227
		2 – básico/fundamental	5,869	0,291
		3 – médio/ 2º grau	6,082	0,306
		4 – superior/ pós graduação	6,560	0,290
	2	1 – nenhum	5,938	0,348
		2 – básico/fundamental	6,373	0,352
		3 – médio/ 2º grau	6,439	0,341
		4 – superior/ pós graduação	6,788	0,224
	3 ou mais	1 – nenhum	5,969	0,337
		2 – básico/fundamental	6,519	0,362
		3 – médio/ 2º grau	6,607	0,323
		4 – superior/ pós graduação	6,859	0,178
2 – Feminino	0	1 – nenhum	5,634	0,140
		2 – básico/fundamental	5,484	0,144
		3 – médio/ 2º grau	5,676	0,299
		4 – superior/ pós graduação	5,591	0,105
	1	1 – nenhum	5,792	0,244
		2 – básico/fundamental	5,845	0,291
		3 – médio/ 2º grau	5,928	0,295
		4 – superior/ pós graduação	6,509	0,265
	2	1 – nenhum	6,041	0,371
		2 – básico/fundamental	6,331	0,355
		3 – médio/ 2º grau	6,373	0,319
		4 – superior/ pós graduação	6,696	0,244
	3 ou mais	1 – nenhum	6,307	0,418
		2 – básico/fundamental	6,568	0,317
		3 – médio/ 2º grau	6,640	0,283
		4 – superior/ pós graduação	6,758	0,242

Tabela D.3: Estimativas com base em imputação múltipla ($m = 20$) para os componentes da variância dos parâmetros da rede no nó contínuo e medidas de diagnóstico para inferências na variável renda imputada (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	$Var_d(\widehat{\theta})$				$Var_e(\widehat{\theta})$				$\widehat{r}$				$\widehat{\lambda}$			
	$\widehat{\beta}_{0,101}$	$\widehat{\beta}_{0,102}$	$\widehat{\beta}_{0,103}$	$\widehat{\beta}_{0,104}$	$\widehat{\beta}_{0,101}$	$\widehat{\beta}_{0,102}$	$\widehat{\beta}_{0,103}$	$\widehat{\beta}_{0,104}$	$\widehat{\beta}_{0,101}$	$\widehat{\beta}_{0,102}$	$\widehat{\beta}_{0,103}$	$\widehat{\beta}_{0,104}$	$\widehat{\beta}_{0,101}$	$\widehat{\beta}_{0,102}$	$\widehat{\beta}_{0,103}$	$\widehat{\beta}_{0,104}$
1%	0,216	0,196	0,260	0,470	0,001	0,003	0,005	0,002	0,005	0,016	0,020	0,004	1	1	1	0
3%	0,199	0,189	0,260	0,464	0,002	0,001	0,002	0,001	0,011	0,006	0,008	0,002	1	2	2	1
5%	0,192	0,180	0,253	0,458	0,003	0,003	0,001	0,002	0,016	0,018	0,004	0,005	2	3	3	3
7%	0,191	0,176	0,248	0,457	0,001	0,001	0,003	0,001	0,005	0,006	0,013	0,002	2	7	5	6
10%	0,180	0,163	0,241	0,449	0,002	0,001	0,002	0,002	0,012	0,006	0,009	0,005	1	2	4	7
20%	0,176	0,153	0,240	0,447	0,005	0,007	0,003	0,001	0,030	0,005	0,013	0,002	3	9	9	10
30%	0,164	0,150	0,235	0,434	0,001	0,001	0,002	0,001	0,006	0,007	0,009	0,002	6	11	21	12
40%	0,155	0,146	0,227	0,429	0,002	0,002	0,002	0,001	0,014	0,014	0,009	0,002	5	17	22	19
50%	0,142	0,144	0,218	0,426	0,002	0,002	0,002	0,001	0,015	0,015	0,010	0,002	31	19	16	20

Tabela D.4: Estimativas com base em imputação múltipla ($m = 20$) para os parâmetros de um modelo de regressão linear na variável LOG(RENDA), imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	**Estimativas dos parâmetros**							
	$\widehat{\beta}_1$ **SEXO1**	$\widehat{\beta}_2$ **SEXO2**	$\widehat{\beta}_3$ **CURSOELV2**	$\widehat{\beta}_4$ **CURSOELV3**	$\widehat{\beta}_5$ **CURSOELV4**	$\widehat{\beta}_6$ **QTDBAHN1**	$\widehat{\beta}_7$ **QTDBAHN2**	$\widehat{\beta}_8$ **QTDBAHN3**
1%	5,554	5,423	4,148	4,866	5,950	4,708	5,824	6,076
3%	5,556	5,423	4,135	4,841	5,949	4,720	5,827	6,075
5%	5,552	5,426	4,148	4,859	5,955	4,725	5,824	6,085
7%	5,547	5,412	4,169	4,876	5,951	4,716	5,833	6,081
10%	5,548	5,412	4,192	4,883	5,949	4,725	5,812	6,085
20%	5,572	5,437	4,104	4,842	5,940	4,725	5,822	6,101
30%	5,556	5,436	4,178	4,876	5,941	4,709	5,828	6,089
40%	5,580	5,456	4,071	4,826	5,922	4,721	5,844	6,062
50%	5,579	5,458	4,101	4,841	5,925	4,699	5,831	6,088

Tabela D.5: Estimativas com base em imputação múltipla ($m = 20$) para as variâncias dos parâmetros ($Var_t(\widehat{\theta})$) de um modelo de regressão linear na variável LOG(RENDA), imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	**Estimativas das variâncias**							
	$\widehat{\beta}_1$ **SEXO1**	$\widehat{\beta}_2$ **SEXO2**	$\widehat{\beta}_3$ **CURSOELV2**	$\widehat{\beta}_4$ **CURSOELV3**	$\widehat{\beta}_5$ **CURSOELV4**	$\widehat{\beta}_6$ **QTDBAHN1**	$\widehat{\beta}_7$ **QTDBAHN2**	$\widehat{\beta}_8$ **QTDBAHN3**
1%	2,232	2,530	2,132	2,168	2,299	2,333	2,435	2,641
3%	2,233	2,533	2,138	2,173	2,306	2,330	2,436	2,642
5%	2,233	2,531	2,139	2,180	2,314	2,342	2,450	2,674
7%	2,233	2,532	2,141	2,179	2,310	2,340	2,463	2,661
10%	2,503	2,542	2,138	2,186	2,345	2,339	2,454	2,676
20%	2,240	2,558	2,148	2,188	2,345	2,347	2,478	2,718
30%	2,250	2,578	2,172	2,214	2,402	2,358	2,561	2,729
40%	2,247	2,564	2,159	2,216	2,456	2,364	2,561	2,758
50%	2,253	2,598	2,214	2,238	2,541	2,355	2,561	2,770

Tabela D.6: Partição da variância total das estimativas dos parâmetros com base em imputação múltipla ($Var_d(\widehat{\theta})$ e $Var_e(\widehat{\theta})$) para um modelo de regressão linear na variável RENDA, imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	$\boldsymbol{Var_d(\widehat{\theta})}$								$\boldsymbol{Var_e(\widehat{\theta})}$							
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$
1%	2,231	2,249	2,064	2,165	2,295	2,331	2,431	2,635	0,001	0,001	0,065	0,003	0,004	0,002	0,004	0,006
3%	2,230	2,248	2,063	2,263	2,293	2,329	2,429	2,633	0,003	0,005	0,071	0,009	0,012	0,001	0,007	0,009
5%	2,231	2,249	2,064	2,165	2,294	2,330	2,431	2,635	0,002	0,269	0,071	0,014	0,019	0,011	0,018	0,037
7%	2,228	2,246	2,060	2,158	2,288	2,324	2,424	2,628	0,005	0,272	0,077	0,020	0,021	0,015	0,037	0,031
10%	2,231	2,248	2,063	2,164	2,294	2,330	2,430	2,634	0,259	0,280	0,071	0,021	0,049	0,009	0,023	0,040
20%	2,229	2,246	2,061	2,160	2,289	2,326	2,426	2,630	0,010	0,297	0,083	0,027	0,053	0,020	0,050	0,084
30%	2,223	2,240	2,054	2,146	2,275	2,311	2,411	2,615	0,026	0,322	0,012	0,065	0,121	0,045	0,143	0,109
40%	2,221	2,238	2,052	2,141	2,271	2,307	2,407	2,611	0,025	0,310	0,102	0,071	0,176	0,054	0,147	0,140
50%	2,216	2,234	2,048	2,132	2,261	2,298	2,398	2,602	0,035	0,347	0,158	0,101	0,267	0,054	0,155	0,160

Tabela D.7: Medidas de diagnóstico para inferências com base em imputação múltipla ($\widehat{\lambda}$ e $\widehat{r}$) para um modelo de regressão linear na variável RENDA, imputada por uma rede Bayesiana de características de domicílio–pessoa–renda nos dados do Censo Demográfico (perturbações aplicadas apenas na variável LOG(RENDA))

Percentual de não resposta na variável	$\widehat{\lambda}\%$								$\widehat{r}$							
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$
1%	0	0	0	1	1	0	1	1	0,448	0,445	0,486	0,462	0,436	0,429	0,411	0,380
3%	1	1	2	2	3	0	2	2	0,448	0,445	0,486	0,463	0,436	0,429	0,412	0,380
5%	1	1	2	3	4	3	4	8	0,448	0,451	0,486	0,462	0,436	0,429	0,412	0,380
7%	2	2	4	4	4	3	8	6	0,449	0,451	0,487	0,464	0,438	0,431	0,413	0,381
10%	3	3	2	4	10	2	5	8	0,454	0,451	0,486	0,463	0,437	0,429	0,412	0,380
20%	5	7	5	6	11	5	11	17	0,449	0,452	0,487	0,464	0,438	0,430	0,413	0,382
30%	12	13	12	13	23	9	27	21	0,450	0,454	0,487	0,467	0,442	0,434	0,418	0,384
40%	11	11	10	14	31	11	27	26	0,451	0,453	0,490	0,469	0,444	0,435	0,419	0,386
50%	14	18	20	19	43	10	28	28	0,452	0,455	0,492	0,471	0,448	0,436	0,420	0,387

Tabela D.8: Inferências com base em imputação múltipla ($m = 20$) para coeficientes de modelo de riscos proporcionais na análise do tempo imputado por uma rede Bayesiana de características de réu–crime–papéis–tempo nos dados de homicídios de Campinas (perturbações aplicadas apenas na variável TPOL1)

Percentual de não resposta na variável	$\widehat{\boldsymbol{\theta}}$			d.p.$_t(\widehat{\boldsymbol{\theta}})$			g.l.			Intervalo de confiança para $\widehat{\boldsymbol{\beta}}$		
	$\widehat{\beta}_1$ SEXO	$\widehat{\beta}_2$ CRIME	$\widehat{\beta}_3$ PRISAO	$\widehat{\beta}_1$ SEXO	$\widehat{\beta}_2$ CRIME	$\widehat{\beta}_3$ PRISAO	$\widehat{\beta}_1$ SEXO	$\widehat{\beta}_2$ CRIME	$\widehat{\beta}_3$ PRISAO	$\widehat{\beta}_1$ SEXO	$\widehat{\beta}_2$ CRIME	$\widehat{\beta}_3$ PRISAO
5%	-0,851	0,578	0,824	0,503	0,244	0,241	6.509	6.928	14.543	[-1,836; 0,135]	[0,099; 1,056]	[0,351; 1,297]
7%	-0,926	0,466	0,772	0,500	0,236	0,239	1×10^{07}	13.559	6.475	[-1,906; 0,055]	[0,003; 0,930]	[0,303; 1,242]
10%	-0,930	0,424	0,778	0,501	0,240	0,240	1×10^{07}	14.543	58.074	[-1,911; 0,051]	[-0,047; 0,895]	[0,308; 1,249]
20%	-0,805	0,792	0,951	0,501	0,258	0,255	3.752	2.104	1.502	[-1,786; 0,176]	[0,287; 1,297]	[0,451; 1,452]
30%	-1,003	0,522	0,834	0,506	0,246	0,262	3.102	3.903	580	[-1,994; -0,012]	[0,039; 1,005]	[0,320; 1,349]
40%	-0,725	0,707	0,785	0,510	0,273	0,268	2.642	792	750	[-1,726; 0,275]	[0,171; 1,242]	[0,259; 1,312]
50%	-0,677	1,045	0,880	0,518	0,300	0,288	2.559	617	817	[-1,693; 0,340]	[0,457; 1,633]	[0,316; 1,444]

Tabela D.9: Informação da variância com base em imputação múltipla ($m = 20$) para coeficientes de modelo de riscos proporcionais na análise do tempo imputado por uma rede Bayesiana de características de réu–crime–papéis–tempo nos dados de homicídios de Campinas (perturbações aplicadas apenas na variável TPOL1)

Percentual de não resposta na variável	**Variâncias**									$\widehat{r}$			$\widehat{\lambda}$%		
	$\widehat{\beta_1}$ **(SEXO)**			$\widehat{\beta_2}$ **(CRIME)**			$\widehat{\beta_3}$ **(PRISAO)**								
	$Var_t(\widehat{\theta})$	$Var_d(\widehat{\theta})$	$Var_e(\widehat{\theta})$	$Var_t(\widehat{\theta})$	$Var_d(\widehat{\theta})$	$Var_e(\widehat{\theta})$	$Var_t(\widehat{\theta})$	$Var_d(\widehat{\theta})$	$Var_e(\widehat{\theta})$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
5%	0,253	0,239	0,013	0,060	0,057	0,003	0,058	0,056	0,002	0,057	0,046	0,039	3	6	4
7%	0,250	0,250	0,001	0,056	0,054	0,002	0,057	0,055	0,003	0,001	0,045	0,050	2	3	5
10%	0,251	0,249	0,001	0,058	0,056	0,002	0,058	0,057	0,001	0,005	0,040	0,050	1	4	3
20%	0,251	0,233	0,017	0,066	0,060	0,006	0,065	0,058	0,007	0,075	0,099	0,125	3	11	9
30%	0,256	0,235	0,019	0,061	0,056	0,004	0,069	0,057	0,012	0,087	0,077	0,179	8	7	16
40%	0,261	0,238	0,021	0,074	0,063	0,011	0,072	0,061	0,011	0,094	0,189	0,190	6	17	16
50%	0,269	0,245	0,022	0,090	0,074	0,015	0,083	0,070	0,012	0,095	0,210	0,214	12	16	14