Conteúdo disponível em: https://www.ifgoiano.edu.br/periodicos/multiscience

Multi-Science Journal

Website do periódico:
https://www.ifgoiano.edu.br/periodicos/index.php/multiscience

Artigo de Revisão

TÉCNICAS DE EXTRAÇÃO DE CONHECIMENTO POR MEIO DE DADOS FALTANTES E MINERAÇÃO DE DADOS SOBRE AS VÍTIMAS DO CÉSIO-137:

Danyllo Sudário Cardoso[1], Hugo Pereira Leite Filho[2], Rafael Souto[3]

[1] Universidade Estadual de Goiás Campus Ceres
[2] Universidade Estadual de Goiás Campus Henrique Santillo
[3]Centro de Excelência em Ensino, Pesquisa e Projetos Leide das Neves

*Autor para correspondência: danyllo.sudario@hotmail.com

INFO ARTIGO

Histórico do artigo
Recebido: 18/03/2016
Aceito: 27/03/2017

Palavras chaves:
dosimetria citogenética, imputação múltipla, modelos mais acurados, mineração de dados, regressão linear.

RESUMO

Lidar com dados massivos sem perda ou distorção de resultados requer a aplicação de técnicas aprimoradas de mineração de dados (Witten, Frank, & Hall, 2011). O problema não tratado dos dados faltantes distorce a realidade gerando modelos tendenciosos (Haukoos & Newgard, 2007). Aqui é exposto um ensaio sobre o problema de dados faltantes em pesquisas clínicas mediante técnicas de mineração de dados, métodos estatísticos de regressão linear e múltipla imputação. Foram analisados relatórios de dosimetria citogenética, dos acidentados com o Césio-137 em Goiânia, divididos quanto ao índice de dermatites apresentadas em: Grupo I e Grupo II. O Grupo I, apresentou percentagem de dados faltantes de quase vinte e oito por cento, já o Grupo II, apresentou falta de dados de cerca de sessenta e dois por cento, havendo assim nos dois casos, a degradação da amostra. Para os dois grupos foram aplicados métodos de regressão linear pré- e pós-imputação. O estudo exposto neste trabalho, mostra que a preocupação de pesquisadores, quanto à coleta de dados (Haukoos & Newgard, 2007), é realmente relevante. A imputação múltipla revela-se uma excelente escolha para o tratamento de dados faltantes, culminando na realização de modelos mais acurados, dirimindo deste modo, problemas de degradação da amostra.

Keywords:
Reuse
productive chain
biofuels.

ABSTRACT

Dealing with massive data without loss or distortion of results requires the application of improved techniques of Data Mining (Witten, Frank, & Hall, 2011). The untreated problem of missing data distorts reality generating biased models (Haukoos & Newgard, 2007). Here is exposed an essay on the missing data problem in clinical research by means of Data Mining techniques, statistical methods of linear regression and multiple imputation. Reports of cytogenetic dosimetry of the accident victims with Cesium-137 in Goiânia have been examined, in which the data was distributed by dermatitis index in: I-Group e II-Group. The I-Group presented proportion of missing data of about twenty-eight percent, and the II-Group presented lack of data about sixty-two percent, and thus in both cases, degradation of the sample. For both groups were applied linear regression methods, pre- and post-imputation. The study exposed in this work shows that the worry of researchers concerning the collection of data (Haukoos & Newgard, 2007), is relevant. Multiple imputation reveals itself an excellent choice for the treatment of missing data, culminating in the realization of more accurate models, settling thus the sample degradation problems.

1. Introdução Geral

Devido a uma grande massa de dados espalhados pelo mundo e nas nossas vidas, segundo Witten, Frank, *&* Hall (2011), torna-se imprescindível a necessidade de lidarmos com esses dados de maneira adequada.

Em pesquisas clínicas, autores da atualidade vêm buscando lidar com um problema que tem substancial influência nas informações geradas (Haukoos & Newgard, 2007) a partir da mineração de dados (Galvão, 2007) para extração de conhecimento, problema este referenciado na literatura como "*missing data*" (MD).

Aqui serão utilizados métodos estatísticos para evidenciar a tendência citada por Haukoos *&* Newgard (2007), imposta a modelos estatísticos que utilizam métodos simples de modelagem e executam a deleção por lista.

Feita uma pesquisa bibliográfica, verificou-se que poucos autores realmente lidam explicitamente com a falta de dados em suas pesquisas (Osborne, 2013), portanto aqui será executada uma análise comparativa de modelos a partir de métodos de análise de caso completo (SAS Institute Inc., 2013), contrapondo-se a modelos mais aprimorados mediante conjuntos de dados completados como descritos nos trabalhos de Taugourdeau e sua equipe (2014) via imputação múltipla.

Os resultados aqui obtidos se deram pela análise de modelos de regressão linear e modelos de regressão linear mediante múltipla imputação, sobre dados de uma base de dados dos acidentados com o Césio-137 classificados nos Grupo I e II em Fuini, Souto, Amaral, *&* Amaral (2013) totalizando 99 observações.

A tentativa é a de evidenciar a possível tendência imposta aos resultados modelados pela aplicação de métodos estatísticos simples, diante da presença de dados faltantes (Haukoos & Newgard, 2007).

2 - Knowledge Discovery in databases (KDD) e data mining

Aqui serão utilizados conceitos de KDD e técnicas de *Data Mining*, mediante a utilização de um conjunto de dados referente à dosimetria de radiação nos pacientes dos Grupos I e II do acidente radioativo com o Césio-137 em Goiânia, Fuini, Souto, Amaral, *&* Amaral (2013), o qual possui variáveis com dados incompletos, referenciados na literatura largamente como dados faltantes e que tem sido um problema generalizado como concordam Baneshi *&* Talei (2010) e Brown *&* Kros (2003).

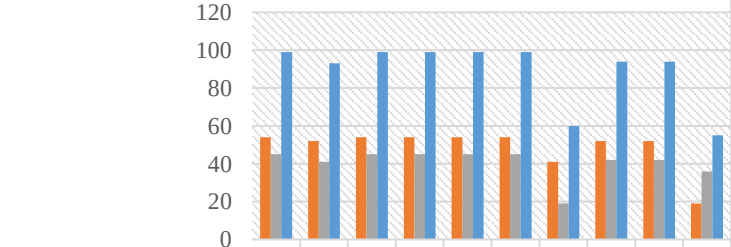
Na fase de mineração de dados, será dado foco no tratamento de MD encontrado no conjunto de dados referentes a relatórios de dosimetria (veja Gráfico 1 e Gráfico 2) mencionadas acima e partindo da modelagem através de um algoritmo estatístico simples de regressão linear dada, que usa método de deleção para lidar com dados faltantes como consta no capítulo "*The* REG *Procedure*" do "SAS/STAT® 13.1 *User's Guide*" que utiliza a equação $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, considerando Y como a variável dependente, x_i variável regressora, e, β_0 e β_1, parâmetros desconhecidos a serem estimados, e ϵ_i o termo de erro para a i-ésima observação (SAS Institute Inc., 2013).

3 - O problema dos dados faltantes

É comum encontrar, em bases de dados, informações incompletas, por isso é de suma importância a fase de limpeza e integração dos dados (Galvão, 2007), para promover uma higienização dos dados, e diminuir ruídos. Uma maior incidência de dados faltantes, pode influenciar a modelagem de resultados.

Haukoos *&* Newgard (2007), defendem que pesquisadores de áreas clínicas, devem buscar utilizar-se de toda a quantidade de dados que puderem lançar mãos, para reduzir os erros sistemáticos

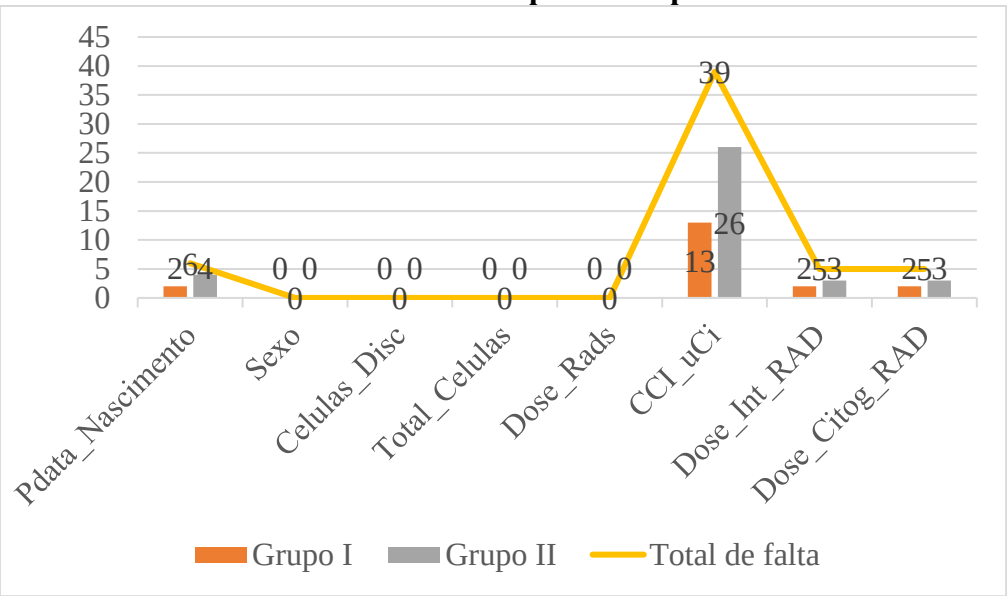
Gráfico 1 – Dados Higienizados do Relatório de Dosimetria Citogenética Grupo I e Grupo II



	Group-Obs	Pdata_Nascimento	Sexo	Celulas_Disc	Total_Celulas	Dose_Rads	CCI_uCi	Dose_Int_RAD	Dose_Citog_RAD	Dados Faltantes
Grupo I	54	52	54	54	54	54	41	52	52	19
Grupo II	45	41	45	45	45	45	19	42	42	36
Total da Amostra	99	93	99	99	99	99	60	94	94	55

Fonte: Relatório de Dosimetria Fundação Leide das Neves Ferreira – FUNLEIDE de 21/01/1991.

Gráfico 2 – Falta por Variável nos Dados do Relatório de Dosimetria Citogénetica Grupo I e Grupo II



Fonte: Relatório de Dosimetria Fundação Leide das Neves Ferreira – FUNLEIDE de 21/01/1991.

(tratados pelos autores como tendências/viés) e apurar a mais válida estimativa risco/benefício.

No seu trabalho dividido em duas partes, a dupla evidencia a diferença entre MD, como sendo aqueles que nunca foram conhecidos (Osborne, 2013), e dados censurados, que podem estar presentes ou não em um conjunto de dados, e que foram intencionalmente removidos pelo investigador (Haukoos & Newgard, 2007).

Osborne (2013), classifica o conjunto de MD em: dados faltantes legítimos, os que cuja ausência é realmente apropriada, não tendo correlação com o elemento de estudo, ou porque o entrevistado não enxerga alguma relação e, nesse caso, a falta de dados por si pode trazer informação; e dados faltantes ilegítimos, são os que podem resultar de problemas logísticos que interrompam a coleta de dados, ou abstenção do entrevistado em responder dados importantes.

Dada a frequência de MD em estudos clínicos (Haukoos & Newgard, 2007), pode ser potencialmente problemático ignorá-los, o que é apoiado pelos resultados obtidos por Baneshi *&* Talei (2010) em pesquisa pela instituição iraniana Centro de Pesquisas do Câncer, em que se verificou que a exclusão de casos com MD culminou em subestimação da verdade na taxa de sobrevivência global para os casos analisados.

Dados faltantes podem ser classificados nas categorias, Ausência Completamente Aleatória – MCAR; Ausência Aleatória – MAR; e Ausência não Aleatória – MNAR; como discorrem Osborne (2013), Haukoos *&* Newgard (2007), Donders, Heijden, Stijnen, *&* Moons (2006) e Taugourdeau, Villerd, Plantureux, Huguenin-Elie, *&* Amiaud (2014).

Contudo, métodos estatísticos mais simples podem conferir substancial tendência aos resultados (Haukoos & Newgard, 2007), uma metodologia que vem obtendo resultados satisfatórios é a Imputação Múltipla (MI) como evidenciado nos trabalhos de Taugourdeau, Villerd, Plantureux, Huguenin-Elie, *&* Amiaud (2014), a segunda parte da pesquisa de Newgard e Haukoos (2007), e da equipe de Donders (2006).

4 - Categorias de dados faltantes

Na literatura, como descrito em Taugourdeau, Villerd, Plantureux, Huguenin-Elie, *&* Amiaud (2014), os dados faltantes são categorizados em três diferentes distribuições como segue:

MCAR: aqui cabe a explicação, simplificada de Allison (2001) que dá como exemplo uma variável Y, que contém dados faltantes; é dito então que se a probabilidade de dados faltantes em Y é a de dados não relatados para valores de Y ou para os valores de quaisquer outras variáveis no conjunto de dados, então Y está em MCAR; assim sendo, MCAR permite a possibilidade de haver relação da falta de dados em Y e a falta de dados em alguma variável X;

MNAR: Haukoos *&* Newgard (2007), explicam a presença do mecanismo MNAR, como quando o padrão de censura está relacionado a variáveis que não foram coletadas ao conjunto observado de Y; ou então mais relacionadas ao conjunto de dados faltantes de Y mais do que ao conjunto de dados observados de Y; também referem-se a MNAR como "*nonignorable*" (Allison, 2001);

MAR: finalmente tomando mais uma vez a descrição de Allison (2001), que considera ser esta uma suposição consideravelmente fraca, sobre o mecanismo MAR explica que, a probabilidade de um conjunto de dados em Y, estar em MAR é a probabilidade de falta de dados em Y não estar relacionada a valores de Y, depois de controlada para as outras variáveis na análise; ou a probabilidade de dados faltantes em Y, dados Y e X, é igual a probabilidade de dados faltantes em Y dado X;

Allison incorpora o elucidado por $P(Y\ missing|Y,X) = P(Y\ missing|X)$;

Das três categorias anteriormente citadas a menos plausível para Hawkoos & Newgard (2007), seria MCAR, posto que a existência dessa suposição pode implicar que sujeitos censurados são requeridos para ser uma amostra aleatória da população em estudo, e sujeitos sem dados censurados são requeridos para ser uma amostra aleatória da população fonte. Os autores assumem que o mecanismo MAR é o mais presumível dado ser o menos restritivo, e também mais defensável que a suposição de MCAR. MAR é o pressuposto requerido para a maioria das formas de imputação, inclusive MI.

Apesar dessa dispendiosa classificação, Osborne (2013) ressalta que as melhores práticas em manuseio de MD apresentam ser igualmente efetivas sem levar em conta se os dados estão em MCAR, MNAR ou MAR.

5 - Métodos de imputação múltipla

Newgard & Haukoos (2007) falam que a base geral de MI está no uso de valores observados para gerar valores plausíveis, para cada valor censurado baseando-se em correlações existentes entre variáveis, com a pretensão de representar uma faixa plausível de valores que se aproximem dos valores faltantes.

Nunes, Klück, & Fachel (2010) introduzem seu artigo dando breve histórico sobre o método de imputação múltipla, que foi proposto na década de 1980 por Donald Rubin a fim de solucionar problemas de não resposta em pesquisas. Porém, a princípio, a técnica sugerida por Rubin não pôde ser computacionalmente implementada devido aos recursos tecnológicos da época.

Allison (2000) em um artigo seu, cita o trabalho elaborado por Rubin, no qual descreve os passos para o processo de MI:

(a) imputam-se valores faltantes usando um modelo que incorpora variação aleatória;

(b) procede-se isso M vezes, obtendo M conjuntos de dados completos;

(c) realiza-se a análise desejada em cada conjunto de dados usando métodos padrões para dados completos;

(d) calcula-se a média dos parâmetros estimados das amostras dos modelos para produzir uma estimativa única;

(e) e, por fim, calculam-se os erros padrões pela combinação da média dos quadrados dos erros padrões das estimativas M com o cálculo da variância das estimativas de parâmetros M através da amostra usando a equação $\sqrt{\frac{1}{M}\sum_k\left(S_k^{\,2}\right) + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k\left(b_k-\bar{b}\right)^2}$.

Dessa forma, MI permite que cada conjunto de dados completos imputados sejam analisados, e combinados seus resultados para produzir estimativas e intervalos de confiança que acabam por incorporar os dados faltantes, detalham Taugourdeau, Villerd, Plantureux, Huguenin-Elie, & Amiaud (2014).

A equipe de Nunes (2010) sustenta que a principal vantagem da imputação múltipla em relação à imputação é a de considerar a variabilidade entre as imputações nos resultados.

6 - Softwares para imputação múltipla

Com o avanço dos estudos dos métodos de imputação no âmbito computacional, algumas soluções de software começaram a surgir dentre as quais Nunes, Klück, & Fachel (2009), relaciona: *Statistical Analysis Software* – SAS, S-Plus, SOLAS, NORM, BMPD e MICE, este último de domínio público e baseado na linguagem R.

Aqui não serão abordadas as particularidades de cada uma dessas aplicações estatísticas, mas, para fim de se esclarecer a motivação da escolha de uma destas soluções, segue-se alguma explanação acerca da aplicação elegida.

7 -SAS/STAT software

Trata-se de um ambiente que conta com a implementação de vários algoritmos estatísticos, incluindo-se métodos para imputação múltipla e análise bayesiana, dados pela contribuição de várias pessoas (SAS Institute Inc., 2013) nas pesquisas, testes, consultorias e revisão de documentação.

Ao momento da escrita deste trabalho, a versão mais atual é a 9.2; trata-se de uma aplicação comercial porém existe uma edição universitária (SAS® University Edition) que é disponibilizada gratuitamente para fins acadêmicos e é executada em um ambiente emulado Linux Red Hat via Virtual Box 4.3.10 e/ou VMware Player 6.0 ou superiores, e disponível para download no sítio <http://www.sas.com/en_us/software/university-edition/download-software.html>.

A aplicação oferece uma interface facilitadora que executa no cliente através da porta 10080, e permite implementar alguns procedimentos pelo mecanismo de arrastar e soltar, além de facilitadores que permitem salvar bibliotecas, sessões e programas para reaproveitamento de resultados. É baseada na linguagem estruturada SAS *programing language*.

Essa foi a opção escolhida para a demonstração de resultados em conjuntos de dados com dados faltantes no Relatório de Dosimetria dos acidentados com o Césio-137 do Grupo I e Grupo II, devido a facilidades oferecidas, fácil disponibilidade da aplicação e descomplicada configuração.

8 - Procedimentos usados na modelagem estatística

Na criação de modelos, foram utilizados os procedimentos PROC REG (The REG Procedure, 2013), PROC MI (The MI Procedure, 2013) e PROC MIANALYZE (The MIANALYZE Procedure, 2013), todos disponíveis no SAS/STAT® 13.1 *User's Guide* do SAS *INSTITUTE INC.*

9 - Procedimento PROC REG

Como consta no Guia do Usuário SAS/STAT® 13.1 (The REG Procedure, 2013), o objetivo geral deste procedimento é a regressão linear, porém a aplicação conta com outros procedimentos que proveem aplicações mais especializadas.

Aqui decidimos pela utilização deste procedimento devido ao fato de poder assim evidenciar

a modelagem estatística sobre análise de caso completo também "*listwise deletion*" referidas nos trabalhos de Haukoos & Newgard (2007), que segundo os autores realiza a exclusão de dados censurados para a variável regressora ou variável de interesse, limitando a análise às observações com dados completos.

Devido à redução da amostra com a abordagem acima reduz-se a precisão dos dados estimados, concordam Haukoos & Newgard (2007) e Goeij *et al.* (2013).

O Guia do Usuário SAS/STAT® relaciona as capacidades providas em PROC REG. Dentre as quais se destacam: valores preditos, residuais, residuais estudados, limites de confiança, estatísticas de influência.

Os resultados serão analisados sobre os parâmetros estimados evidenciados no modelo conjuntamente com a análise do R^2 ou coeficiente de determinação, que é um modelo de regressão que mede a proporção da variabilidade na resposta que é apresentada pelas variáveis regressoras, conforme esclarece o *User's Guide* do SAS/STAT 13.1 (2013).

A fórmula do modelo R^2 é dada por $R^2 = 1 - SSE/SST$, na qual SSE é o erro residual das somas dos quadrados, e SST é o total da soma dos quadrados corrigido pela média.

PROC REG também dispõem o resultado do R^2 ajustado, que leva em conta o número de parâmetros no modelo (SAS Institute Inc., 2013), dado por $ADJRSQ = 1 - (n-i)/(n-p)(1-R^2)$, em que n é o número de observações usadas para encaixar o modelo, p é o número de parâmetros no modelo, incluindo-se os interceptados e i é 1 se o modelo inclui algum termo interceptado, caso contrário i é 0.

O procedimento origina ainda outros resultados em índices numéricos e plotagens, mas nos ateremos à análise dos parâmetros estimados e do R^2 pré- e pós-imputações; sendo que os modelos aqui apresentados apresentam alguns dados aglutinados dos modelos completos gerados.

10 - Procedimento PROC MI

O PROC MI é um procedimento de múltipla imputação que cria conjuntos de dados multiplamente imputados para dados multivariados p-dimensionais incompletos SAS *INSTITUTE INC.*, *The* MI *Procedure* (2013). Ele usa métodos que incorporam variabilidade apropriada através de m imputações, o método de escolha de imputação depende do padrão de completa falta nos dados e o tipo de variável imputada.

O Guia SAS esclarece que, para conjuntos de dados com padrões de dados faltantes arbitrários, pode-se usar um método Markov Chain Monte Carlo (MCMC), assumindo-se normalidade multivariada, ou um método de especificação "*fully conditional specification*" (FCS).

Schafer (1999) detalha que MCMC é uma coleção de métodos de simulação de extração de distribuições não padronizadas via cadeias de Markov, usada para simulação de parâmetros para a criação de um grande número de (tipicamente dependentes) parâmetros extraídos aleatoriamente de uma distribuição Bayesiana a *posteriori* sob modelos paramétricos complicados.

Nos trabalhos de van Buuren (2007), FCS é especificado como uma alternativa semiparamétrica e flexível que especifica modelos multivariados através de uma série de modelos condicionais, um para cada variável incompleta. O método provê muita flexibilidade e é fácil de aplicar, mas suas propriedades estatísticas são difíceis de estabelecer.

O método padrão usado em PROC MI é o MCMC através da "MCMC *statement*", pelo seu valor padrão detalhado em SAS *INSTITUTE INC.*, *The* MI *Procedure* (2013).

Depois dos m conjuntos de dados completos serem analisados o PROC MIANALYZE, pode ser usado para gerar inferências estatísticas válidas sobre esses parâmetros pela combinação dos resultados das m análises (SAS Institute Inc., 2013).

A eficiência relativa de um estimador m de imputações é alta para casos com poucos dados faltantes SAS *INSTITUTE INC.*, *apud* Rubin (2013).

11 - Procedimento PROC MIANALYSE

O procedimento combina os resultados das análises das imputações e gera inferências estatísticas válidas, em *The* MIANALYZE *Procedure*, (SAS Institute Inc., 2013), as análises das imputações são obtidas usando procedimentos SAS padrão para conjuntos de dados completos, como o PROC REG. Não importa qual análise de dados completos é usada, o processo de combinação de resultados a partir de diferentes conjuntos de dados imputados é essencialmente o mesmo.

O PROC MIANALYZE faz a leitura dos parâmetros estimados e erros padrões associados ou matriz de covariância, que são computados pelo procedimento estatístico padrão para cada conjunto de dados imputados, como consta no guia de referência mencionado.

12 - Resultados e discursão

Foram modelados, para fins de estudo, um modelo de regressão linear usando análise de casos completos para cada conjunto de dados; posteriormente, aos conjuntos foram aplicados os métodos de imputação múltipla e regressão linear, a fim de aferir com acurácia a verdade estimada aplicando-se em um modelo com conjunto de dados completos.

Por fim, combinaram-se os resultados das análises das imputações geradas para análise das inferências estatísticas, de acordo com Spratt *et al.* (2010), como se segue:

ANÁLISE DOS MODELOS DE REGRESSÃO LINEAR

A Tabela 1 mostra dados modelados a partir do relatório de dosimetria citogenética dos pacientes do Grupo I e Grupo II.

Aqui convém destacar o descarte de conjuntos de dados para variáveis que apresentaram dados faltantes referidos na bibliografia como deleção por lista (Haukoos & Newgard, 2007). Os modelos reportam para o Grupo I um total de 54 (cinquenta e quatro) observações lidas e 15 (quinze) com dados

faltantes, portanto apenas 39 (trinta e nove) foram usadas para a modelagem.

Semelhantemente para o Grupo II foi utilizado o "*listwise deletion*", sendo utilizados de um total de 45 (quarenta e cinco) observações um número expressivamente menor de casos completos, totalizando apenas 17 (dezessete) observações utilizadas, havendo 28 (vinte e oito) observações com dados faltantes.

Destarte, podemos observar que o modelo de regressão para o Grupo I utilizou uma percentagem de 72,22 (setenta e dois vírgula vinte e dois), enquanto que para o Grupo II somente 37,78% (trinta e sete vírgula setenta e oito por cento) dos dados foram utilizados.

ANÁLISE DOS MODELOS DE REGRESSÃO LINEAR SOB MÚLTIPLA IMPUTAÇÃO

Para uma maior compreensão sobre os dados faltantes os modelos gerados pelo procedimento PROC MI modelam o padrão de dados faltantes, esses detalhes também foram retirados dos modelos gerados e explanados a seguir.

Demonstração Dos Padrões De Dados Faltantes

Procederam-se, neste estudo, 5 (cinco) imputações para os conjuntos de dados incompletos analisados. A Tabela 2 mostra o padrão de dados faltantes para os conjuntos em questão.

Nessa tabela, além dos padrões de MD, representados por "X" (presença do dado) e "." (ausência do dado), pode-se também, observar a média e a frequência para cada variável distribuídas por grupos.

Os modelos de padrões de MD demonstram que, tanto no Grupo I quanto no Grupo II, existem dados faltantes para as variáveis "Pdata_Nascimento", "CCI_uCi", "Dose_Int_RAD" e "Dose_Citog_RAD", em que o Grupo I apresenta 5 (cinco) grupos de padrões de falta, sendo:

(a) **Grupo 1**: variáveis não apresentam dados faltantes, representam 72,22% (setenta e dois vírgula vinte e dois por cento) das observações, no total de 39 (trinta e nove) observações;

(b) **Grupo 2**: apresenta falta da variável "CCI_uCi", representam 22,22% (vinte e dois vírgula vinte e dois por cento) das observações, totalizam 12 (doze) observações;

(c) **Grupo 3**: apresenta falta das variáveis "CCI_uCi", "Dose_Int_RAD" e "Dose_Citog_RAD", em 1,85% (um vírgula oitenta e cinco por cento), 1 (uma) observação;

(d) **Grupo 4**: apresenta falta da variável "Pdata_Nascimento", em 1,85% (um vírgula oitenta e cinco por cento), 1 (uma) observação;

(e) **Grupo 5**: apresenta falta das variáveis "Pdata_Nascimento", "Dose_Int_RAD" e "Dose_Citog_RAD", em 1,85% (um vírgula oitenta e cinco por cento), 1 (uma) observação;

E para o Grupo II são evidenciados 7 (sete) grupos de padrões de MD, como se mostra a seguir:

(a) **Grupo 1**: variáveis não apresentam dados faltantes, representam 37,78% (trinta e sete vírgula setenta e oito por cento) das observações, 17 (dezessete) observações;

(b) **Grupo 2**: apresenta falta das variáveis "Dose_Int_RAD" e "Dose_Citog_RAD", representa 2,22% (dois

Tabela 1 – Modelo Regressão Linear Dosimetria Citogenética
Grupo I e Grupo II

The* REG *Procedure
Models: RegLinGI and RegLinGII – Dependent Variable: Sexo

		I-Group		**II-Group**
Number of Observations Read		54		45
Number of Observations Used		39		17
Number of Observations with Missing Values		15		28
Root MSE		0.47923		0.47350
Dependent Mean		0.41026		0.52941
Coeff Var		116.81152		89.43907
R-Square		0.2455		0.4706
Adj R-Sq		0.0751		0.1530
	Parameter Estimates I-Group		Parameter Estimates II-Group	
Variable	**Parameter Estimate**	**Standard Error**	**Parameter Estimate**	**Standard Error**
Intercept	0.69288	0.26236	-0.49104	0.62779
Pdata_Nascimento	-0.00001307	0.00001760	-0.00002513	0.00003172
Celulas_Disc	0.00702	0.00964	0	.
Total_Celulas	-0.00005131	0.00067287	-0.00105	0.00127
Dose_Rads	-0.00408	0.00532	-0.02677	0.03060

CCI_uCi	-0.00014369	0.00055153	0.19165	0.08872
Dose_Int_RAD	-0.00052435	0.01299	-0.05855	0.15342
Dose_Citog_RAD	0.00198	0.00887	0.08926	0.05450

Fonte: Adaptado de SAS *University Edition* 9.2

vírgula vinte e dois por cento) das observações, 1 (uma) ocorrência;

(c) **Grupo 3**: apresenta falta da variável "CCI_uCi", em 48,89% (quarenta e oito vírgula oitenta e nove por cento) das observações, totaliza 22 (vinte e duas) observações;

(d) **Grupo 4**: apresenta falta das variáveis "CCI_uCi", "Dose_Int_RAD" e "Dose_Citog_RAD", em 2,22% (dois vírgula vinte e dois por cento), 1 (uma) observação;

(e) **Grupo 5**: apresenta falta da variável "Pdata_Nascimento", em 2,22% (dois vírgula vinte e dois por cento), 1 (uma) observação;

(f) **Grupo 6**: apresenta falta das variáveis "Pdata_Nascimento" e "CCI_uCi", representa 4,44% (quatro vírgula quarenta e quatro por cento) das observações com 2 (duas) ocorrências;

(g) **Grupo 7**: apresenta falta das variáveis "Pdata_Nascimento", "CCI_uCi", "Dose_Int_RAD" e "Dose_Citog_RAD", em 2,22% (dois vírgula vinte e dois por cento), 1 (uma) observação;

Depois de conhecido e evidenciado o padrão de MD para esses conjuntos de dados, é fácil intuir a proposição de Haukoos *&* Newgard (2007) de resultados tendenciosos, uma vez que são condicionados a um número fracionado da amostra.

Análise Comparativa Dos Modelos De Regressão Linear Sob Análise De Caso Completo Versus Modelos De Regressão Linear Sob Múltipla Imputação

Como dito aqui, definiram-se 5 (cinco) imputações para o procedimento PROC MI, que as gerou para uma variável de saída capturada pelo PROC REG, este gerou um modelo de regressão linear para cada um dos conjuntos de dados imputados.

Utilizou-se o modelo gerado da última imputação para, através de uma comparação entre o modelo gerado a partir de conjuntos incompletos e completados por MI, evidenciar o impacto das imputações, Taugourdeau *et al.* (2014).

Serão comparados aqui os parâmetros estimados antes e depois da múltipla imputação e os valores para o R^2, que segundo a bibliografia, é a medida estatística expressada pela razão entre a variação dos valores ajustados e a variação nos valores observados (King, 1996), a qual sido proposta para modelos de regressão geral que utilizam estimação de parâmetros por máxima probabilidade (Freels & Sinha, 2008).

Os conjuntos de dados apresentam alguns dados com valores negativos devido à presença de variáveis do tipo data anteriores a 01 (primeiro) de janeiro de 1960 (mil novecentos e sessenta), como descrito em SAS *INSTITUTE INC.*, (2012) "*Step-by-Step Programming with Base* SAS®".

Como exposto nos trabalhos de Heus (2012), autores consideram que, devido à possibilidade de o conjunto de dados apresentarem valores negativos, o R^2 não é tecnicamente uma proporção da variância, o que limita sua utilidade, mas Heus conclui seu pensamento nas palavras de Preacher *&* Kelley (2011) com a possibilidade de que R^2 pode ter um valor heurístico em certas situações.

Dado o exposto, analisaremos os modelos sob essas duas evidências em análise comparativa, para percebermos o comportamento dos modelos sob os métodos de análise de caso completo e múltipla imputação. Ademais, apesar de não constarem dos modelos, foram gerados ensaios que excluíam a variável de valores negativos em questão, que resultaram em resultados para R^2 muito próximos dos aqui estudados.

Tabela 2 – Padrão de Dados Faltantes Modelo Imputação Múltipla Dosimetria Citogenética Grupo I e Grupo II

The* REG *Procedure

Models: RegLinGI and RegLinGII – Dependent Variable: Sexo

(continua)

	I-Group					II-Group						
Group / **Variable**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Sexo	X	X	X	X	X	X	X	X	X	X	X	X

Tabela 2 – Padrão de Dados Faltantes Modelo Imputação Múltipla Dosimetria Citogenética Grupo I e Grupo II

The* REG *Procedure
Models: RegLinGI and RegLinGII – Dependent Variable: Sexo

(conclusão)

	I-Group					II-Group						
Group / Variable	1	2	3	4	5	1	2	3	4	5	6	7
Pdata_Nascimento	X	X	X	.	.	X	X	X	X	.	.	.
Celulas_Disc	X	X	X	X	X	X	X	X	X	X	X	X
Total_Celulas	X	X	X	X	X	X	X	X	X	X	X	X
Dose_Rads	X	X	X	X	X	X	X	X	X	X	X	X
CCI_uCi	X	.	.	X	X	X	X	.	.	X	.	.
Dose_Int_RAD	X	X	.	X	.	X	.	X	.	X	X	.
Dose_Citog_RAD	X	X	.	X	.	X	.	X	.	X	X	.
Freq	39	12	1	1	1	17	1	22	1	1	2	1
Percent	72.22	22.22	1.85	1.85	1.85	37.78	2.22	48.89	2.22	2.22	4.44	2.22

Fonte: Adaptado de SAS *University Edition* 9.2.

Os parâmetros estimados sob análise de caso completo, para as variáveis com MD dos relatórios de dosimetria citogenética do Grupo I, foram extraídos da Tabela 1, e comparados na Tabela 3, com os parâmetros estimados sob múltipla imputação dados na Tabela 4.

Na Tabela 3, os valores do Grupo I estimados a maior, foram grifados, evidenciando-se que, sob análise de caso completo, houve, no geral, uma subestimação dos parâmetros analisados. Também foram coletados nessa tabela, dados das Tabelas 1 e 4 para comparação dos modelos gerados sob análise de caso completo e múltipla imputação para os relatórios do Grupo II.

Para os dados de dosimetria citogenética do Grupo II, a tendência encontrada no mesmo modelo para o Grupo I se repete e se vê que os parâmetros estimados sob análise de caso completo têm seus valores com viés de subestimação em relação aos dados estimados sob MI.

Analisando-se os valores do R^2 para os modelos, nota-se que, no caso do modelo de regressão linear para o Grupo I, o índice, no caso de análise completa, encontrava-se por volta de 0,25 (zero vírgula vinte e cinco) em MI cai para 0,13 (zero vírgula treze), ou seja, o comportamento da variável regressora não explica o comportamento das demais variáveis.

Tabela 3 – Comparação dos Parâmetros Estimados nos Modelos sob Análise de Caso Completo e Múltipla Imputação Dosimetria Citogenética Grupo I e Grupo II

Parameter Estimates Comparition

(continua)

	I-Group		II-Group	
Variable	**Parameter Estimate Complete Case Analisys**	**Parameter Estimate Multiple Imputation**	**Parameter Estimate Complete Case Analisys**	**Parameter Estimate Multiple Imputation**
Pdata_Nascimento	-0.00001307	**0.00000353**	**-0.49104**	-0.68401

Tabela 3 – Comparação dos Parâmetros Estimados nos Modelos sob Análise de Caso Completo e Múltipla Imputação Dosimetria Citogenética Grupo I e Grupo II

Parameter Estimates Comparition (conclusão)				
	I-Group		II-Group	
Variable	Parameter Estimate Complete Case Analisys	Parameter Estimate Multiple Imputation	Parameter Estimate Complete Case Analisys	Parameter Estimate Multiple Imputation
Celulas_Disc	**0.00702**	0.00512	-0.00002513	**-0.00001657**
Total_Celulas	-0.00005131	**-0.00000501**	0	**2.52948**
Dose_Rads	-0.00408	**-0.00398**	-0.00105	**-0.00087436**
CCI_uCi	-0.00014369	**-0.00013707**	-0.02677	**-0.02475**
Dose_Int_RAD	**-0.00052435**	-0.00231	0.19165	**0.23026**
Dose_Citog_RAD	0.00198	**0.00317**	**-0.05855**	-0.07026

Fonte: SAS *University Edition* 9.2.

Tabela 4 – Parâmetros Estimados Modelo Regressão Linear Dosimetria Citogenética Grupo I e Grupo II Sob Múltipla Imputação

The* REG *Procedure
Model: RegLinGIMI and RegLinGIIMI – Dependent Variable: Sexo - Imputation Number = 5

		I-Group		II-Group
Number of Observations Read		54		45
Number of Observations Used		54		45
Root MSE		0.50168		0.40124
Dependent Mean		0.44444		0.51111
Coeff Var		112.87853		78.50389
R-Square		0.1317		0.4702
Adj R-Sq		-0.0004		0.3700
	Parameter Estimates I-Group		Parameter Estimates II-Group	
Variable	**Parameter Estimate**	**Standard Error**	**Parameter Estimate**	**Standard Error**
Intercept	0.58511	0.23380	-0.68401	0.24078
Pdata_Nascimento	0.00000353	0.00001564	-0.00001657	0.00001396
Celulas_Disc	0.00512	0.00936	2.52948	0.53849
Total_Celulas	-0.00000501	0.00064169	-0.00087436	0.00049779
Dose_Rads	-0.00398	0.00547	-0.02475	0.01691
CCI_uCi	-0.00013707	0.00044964	0.23026	0.04438
Dose_Int_RAD	-0.00231	0.01188	-0.07026	0.10414
Dose_Citog_RAD	0.00317	0.00906	0.09145	0.02767

Fonte: Adaptado de SAS *University Edition* 9.2.

Continuando a aferição do coeficiente de determinação, para o conjunto de dados do Grupo II, nota-se que o índice permanece quase inalterado por volta de 0,47 (zero vírgula quarenta e sete), demonstrando uma correlação mediana para o modelo de regressão.

Análise Da Múltipla Imputação

Por último, foram analisados nos modelos gerados por MI, os valores de informação faltantes, e o aumento relativo na variância dada a não resposta,

mecanismo descrito por SAS *INSTITUTE INC.* (2013) mencionando Rubin, em que a informação faltante, é dada por, $\hat{\lambda} = \left(r + \left(\frac{2}{v_m+3}\right)\right)/(r+1)$, sendo que r é o aumento relativo na variância e v_m é o grau de liberdade; e $r = \left(\left((1+m^{-1}\ )B\right)\right)/\overline{W}$, tendo-se B como a variância entre as imputações e $\overline{W}$ a média da variância da imputação interna de m conjuntos de dados completos. Quando r e B são 0 (zero) não há informação sobre os parâmetros imputados, as estatísticas acima mencionadas aferem como os dados faltantes contribuem para a incerteza sobre esses parâmetros.

Pelos dados dos modelos de análise de MI, demonstrados na Tabela 5, constata-se que, para o Relatório de Dosimetria Citogenética do Grupo I, mesmo apresentando aumento relativo na variância para a maioria das variáveis próximo de 0 (zero), conta com informação faltante para todas as variáveis. A falta de informação é ainda mais acentuada para o Relatório de Dosimetria Citogenética do Grupo II, no qual, os índices de r em sua maioria aproximam-se de ou superam 1 (um).

Tabela 5 – Resultados das Combinações das Análises das Imputações Dosimetria Citogenética Grupo I

MIANALYZE Variance Information						
	Grupo I			Grupo II		
Parameter	**Relative Increase in Variance**	**Fraction Missing Information**	**Relative Efficiency**	**Relative Increase in Variance**	**Fraction Missing Information**	**Relative Efficiency**
Pdata_Nascimento	0.037042	0.036334	0.992786	0.538300	0.386385	0.928266
Celulas_Disc	0.018245	0.018076	0.996398	24.221201	0.971159	0.837358
Total_Celulas	0.017142	0.016992	0.996613	1.271097	0.615528	0.890388
Dose_Rads	0.092322	0.087771	0.982749	0.703379	0.457308	0.916203
CCI_uCi	0.426486	0.328340	0.938379	0.438089	0.334799	0.937242
Dose_Int_RAD	0.281037	0.237513	0.954652	0.178978	0.161415	0.968727
Dose_Citog_RAD	0.056800	0.055111	0.989098	0.767827	0.481076	0.912230

Fonte: Adaptado de SAS *University Edition* 9.2.

CONCLUSÕES DAS ANÁLISES DOS MODELOS

Pelas análises dos modelos aqui apresentados, pode-se concluir que, mesmo sob um baixo índice de dados faltantes, existe substancial perda de informação para um dado conjunto de dados, esse comportamento fica evidente no caso do conjunto de dados para o Grupo I.

Apesar de o Grupo I apresentar um padrão de dados presentes com frequência acima dos 70% (setenta por cento), apresenta viés de subestimação dos parâmetros similar ao apresentado pelo Grupo II.

O valor de R^2 que mostra quão um modelo de regressão se encaixa para uma dada variável, confirma-se para o Grupo II após MI, porém não se confirma para o modelo do Grupo I.

E, apesar da variável regressora escolhida não explicar o comportamento das demais nos modelos do Grupo I, explica medianamente para o Grupo II, ou seja, apresenta independência para o Grupo I e mediana correlação para o Grupo II.

13 - Considerações finais

O ensaio demonstrado neste trabalho evidencia a importância e relevância da posição de alguns pesquisadores, em orientar seus pares quanto ao cuidado necessário em se buscar lançar mãos da maior quantidade de dados possível no intuito de obter a mais válida estimativa risco/benefício.

Os resultados apontam que é justificado o relato na bibliografia quanto ao descuido no manejo de dados faltantes e que esse problema deve ser considerado importante caso conste em pesquisas clínicas.

Os modelos sob múltipla imputação revelam-se uma excelente escolha para o tratamento do problema de falta de dados em muitas situações de não resposta em pesquisas clínicas, através da obtenção de modelos completos, dirimindo-se, assim, o problema da degradação da amostra e da factível perda de informações.

Em suma: modelos mais realísticos podem ser construídos sobre conjuntos de dados estimados pelos métodos que utilizam a imputação múltipla, dada a aproximação da média real do valor de cada variável imputada pela correção do erro padrão de uma imputação para outra subsequente.

14 - Referências Bibliográficas

Allison, P. D. (February de 2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods & Research, 28*(3), 201-209. doi:10.1177/0049124100028003003

Allison, P. D. (2001). *Missing Data* (7th ed., Vol. 136). (D. Santoyo, Ed.) Thousand Oaks, California, EUA: SAGE Publications, Inc.

Baneshi, M. R., *&* Talei, A. R. (21 de June de 2010). Impact of Imputation of Missing Data on Estimation of Survival Rates: An Example in Breast Cancer. *Iranian Journal of Cancer Prevention, 3*(3), pp. 127-131.

Brown, M. L., *&* Kros, J. F. (2003). *Data Mining* and *the* impact of missing data. *Industrial Management & Data Systems, 103*(8), 611-621. doi:10.1108/02635570310497657

Donders, A. R., Heijden, G. J., Stijnen, T., *&* Moons, K. G. (10 de January de 2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091. doi:10.1016/j.jclinepi.2006.01.014

Freels, S., *&* Sinha, K. (1 de September de 2008). R-squared for general regression models in *the* presence of sampling weights. *Statistics and Probability Letters, 78*(12), 1671–1672. doi:10.1016/j.spl.2008.01.010

Fuini, S. C., Souto, R., Amaral, G. F., *&* Amaral, R. G. (julho de 2013). Qualidade de vida dos indivíduos expostos ao césio-137, em Goiânia, Goiás, Brasil. *Cad. Saúde Pública*, 1301-1310.

Galvão, S. D. (2007). *A Seleção de Atributos e o Aprendizado Supervisionado de Redes Bayesianas no Contexto da Mineração de Dados.* Dissertação, Universidade Federal de São Carlos, São Carlos.

Goeij, M. C., Diepen, M. v., Jager, K. J., Tripepi, G., Zoccali, C., *&* Dekker, F. W. (31 de May de 2013). Multiple imputation: dealing with missing data. *Nephrol Dial Transplant*(28), 2415–2420. doi:10.1093/ndt/gft221

Haukoos, J. S., *&* Newgard, C. D. (July de 2007). Advanced Statistics: Missing Data in Clinical Research - Part 1: An Introduction and Conceptual Framework. (R. J. Lewis, Ed.) *Academic Emergency Medicine, 14*(7), 662-668. doi:10.1197/j.aem.2006.11.037

Heus, P. d. (March de 2012). R squared effect-size measures and overlap between direct and indirect effect in mediation analysis. *Behavior Research Methods, 44*(1), 213-221. doi:10.3758/s13428-011-0141-5

King, G. (1996). Stochastic variation: A comment on Lewis-Beck and Skalaban's "*The* R-Square". *Political Analysis, 6*(1), 1-36. doi:10.1093/pan/6.1.1

Newgard, C. D., *&* Haukoos, J. S. (July de 2007). Advanced Statistics: Missing Data in Clinical Research—Part 2: Multiple Imputation. *Academic Emergency Medicine, 14*(7), 669–678. doi:10.1197/j.aem.2006.11.038

Nunes, L. N., Klück, M. M., *&* Fachel, J. M. (fevereiro de 2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cad. Saúde Pública, 25*(2), 270-278. doi:10.1590/S0102-311X2009000200005

Nunes, L. N., Klück, M. M., *&* Fachel, J. M. (dezembro de 2010). Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. *Rev. Bras. Epidemiol., 13*(4), 595-606. doi:10.1590/S1415-790X2010000400005

Osborne, J. W. (2013). DEALING WITH MISSING DATA OR INCOMPLETE DATA: Debunking *The* Mith of Emptiness. Em J. W. Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data* (p. 296). Los Angeles, California, USA: SAGE Publications, Inc.

Preacher, K. J., *&* Kelley, K. (2011). Effect Size Measures for Mediation Models: Quantitative Strategies for Communicating Indirect Effects. *Psychological Methods, 16*(2), 93–115. doi:10.1037/a0022658

SAS *Institute Inc.* (2012). *Step-by-Step Programming with Base SAS*® (2nd ed.). Cary, North Carolina, United States of America: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). Introduction to Regression *Procedures*. Em S. I. Inc., A. Baxter, *&* E. Huddleston (Eds.), *SAS/STAT® 13.1 User's Guide* (pp. 67-104). Cary: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). *SAS/STAT® 13.1 User's Guide.* (A. Baxter, *&* E. Huddleston, Eds.) Cary, NC, USA: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). *The* MCMC *Procedure*. Em S. I. Inc., A. Baxter, *&* E. Huddleston (Eds.), *SAS/STAT® 13.1 User's Guide* (pp. 4729-4994). Cary: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). *The* MI *Procedure*. Em S. I. Inc., A. Baxter, *&* E. Huddleston (Eds.), *SAS/STAT® 13.1 User's Guide* (pp. 5033-5170). Cary: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). *The* MIANALYZE *Procedure*. Em S. I. Inc., A. Baxter, *&* E. Huddleston (Eds.), *SAS/STAT® 13.1 User's Guide* (pp. 5171-5232). Cary: SAS *Institute Inc.*

SAS *Institute Inc.* (2013). *The* REG *Procedure*. Em S. I. Inc., A. Baxter, *&* E. Huddleston (Eds.), *SAS/STAT® 13.1 User's Guide* (pp. 7019-7206). Cary: SAS *Institute Inc.*

Schafer, J. L. (February de 1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8*(1), 3-15. doi:10.1177/096228029900800102

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., *&* Tilling, K. (8 de July de 2010). Strategies for Multiple Imputation in Longitudinal Studies. *American Journal of Epidemiology, 172*(4), 478–487. doi:10.1093/aje/kwq137

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O., *&* Amiaud, B. (26 de January de 2014). Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecology and Evolution, 4*(7), pp. 944-958. doi:10.1002/ece3.989

van Buuren, S. (June de 2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242. doi:10.1177/0962280206074463

Witten, I. H., Frank, E., & Hall, M. A. (2011). *DATA MINING: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington, MA, USA: Elsevier.