Torres Munguía, Juan Armando

Article

Comparison of imputation methods for handling missing categorical data with univariate pattern

Revista de Métodos Cuantitativos para la Economía y la Empresa

Provided in Cooperation with:
Universidad Pablo de Olavide, Sevilla

Suggested Citation: Torres Munguía, Juan Armando (2014) : Comparison of imputation methods for handling missing categorical data with univariate pattern, Revista de Métodos Cuantitativos para la Economía y la Empresa, ISSN 1886-516X, Universidad Pablo de Olavide, Sevilla, Vol. 17, pp. 101-120

REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (17). Páginas 101–120.
Junio de 2014. ISSN: 1886-516X. D.L: SE-2927-06.
URL: http://www.upo.es/RevMetCuant/art.php?id=91

Comparison of Imputation Methods for Handling Missing Categorical Data with Univariate Pattern

Torres Munguía, Juan Armando
Maestría en Estadística Aplicada
Instituto Tecnológico y de Estudios Superiores de Monterrey (México)
Correo electrónico: ja.torresmunguia@gmail.com

ABSTRACT

This paper examines the sample proportions estimates in the presence of univariate missing categorical data. A database about smoking habits (2011 National Addiction Survey of Mexico) was used to create simulated yet realistic datasets at rates 5% and 15% of missingness, each for MCAR, MAR and MNAR mechanisms. Then the performance of six methods for addressing missingness is evaluated: listwise, mode imputation, random imputation, hot-deck, imputation by polytomous regression and random forests. Results showed that the most effective methods for dealing with missing categorical data in most of the scenarios assessed in this paper were hot-deck and polytomous regression approaches.

Keywords: imputation methods; hot-deck; polytomous regression; random forests; smoking habits; missing categorical data.
JEL classification: C18; C80; C83.
MSC2010: 6207; 62P20; 62P25.

Artículo recibido el 9 de agosto de 2013 y aceptado el 2 de junio de 2014.

Una comparación de métodos de imputación de variables categóricas con patrón univariado

RESUMEN

El presente estudio examina la estimación de proporciones muestrales en la presencia de valores faltantes en una variable categórica. Se utiliza una encuesta de consumo de tabaco (Encuesta Nacional de Adicciones de México 2011) para crear bases de datos simuladas pero reales con 5% y 15% de valores perdidos para cada mecanismo de no respuesta MCAR, MAR y MNAR. Se evalúa el desempeño de seis métodos para tratar la falta de respuesta: *listwise*, imputación de moda, imputación aleatoria, *hot-deck*, imputación por regresión politómica y árboles de clasificación. Los resultados de las simulaciones indican que los métodos más efectivos para el tratamiento de la no respuesta en variables categóricas, bajo los escenarios simulados, son *hot-deck* y la regresión politómica.

Palabras clave: métodos de imputacin; *hot-deck*; regresión politómica; árboles de clasificación; hábitos de consumo de tabaco; valores perdidos en variables categóricas.
Clasificación JEL: C18; C80; C83.
MSC2010: 6207; 62P20; 62P25.

1. INTRODUCTION

Researchers have often to deal with the problem of missing data in surveys and census. One of the most important risks run when conducting studies with missing data is to reach incorrect estimates and results. Given this potential issue that can arise from the presence of missingness, a variety of alternatives for addressing missing data have been developed. One of the most common methods is imputation. Imputing means replacing each missing case with a plausible value (single imputation) or a vector of plausible values (multiple imputation) (Rubin, 1996). Nevertheless, the aim of imputation is not to fill in all missing cases, but to "preserve the characteristics of their distribution and relationships between different variables" (Barceló, 2008).

The new lines of research in imputation are devoted to analyze the impact of these methods on estimates, bias and results. However, it is important to point out that most of these are focused on continuous data imputation (Barceló, 2008; Burton *et al.*, 2007; Ghosh-Dastidar and Schafer, 2003; Follmann *et al.*, 1992) and only a limited number of studies have examined the effect of imputation on categorical data (Eisemann *et al.*, 2011, Bacallao and Bacallao, 2010; Farhangfar *et al.*, 2008; Gimotty and Brown, 1990; Little and Schluchter, 1985).

Therefore, the primary goal of the current study was to compare the performance between different explicitly categorical imputation approaches. To achieve that goal, taking the 2011 National Addiction Survey on smoking habits in Mexico, six datasets were randomly simulated at rates 5% and 15% of missingness, imposing the MCAR, MAR and MNAR mechanisms in the original database. A common method to dealing with this is to perform a complete case analysis (listwise). Nonetheless, this may lead to biased estimates if, for instance, entire smoker subgroups are excluded. Then, the effect on sample proportions in smoking status (current smokers, former smokers and never smokers) of five methods for imputing missingness were evaluated in different scenarios generated by varying missingness mechanism and the proportion of missing cases. Hence, the current simulation study also attempts to provide a framework to compare the performance of different approaches for handling missing data with different missing data mechanisms, as the true value is known.

To satisfy the aim of this paper, section 1 contains an overview to the database used in this paper, the 2011 National Addiction Survey of Mexico. In section 2 the key terms used in discussing missingness in the literature are presented. Section 3 briefly describes the methods for handling missing data used in this paper, the complete case analysis (listwise), imputation of the mode, random imputation, the hot-deck method, imputation by polytomous regression and random forests. In section 4 a summary of the missingness simulations is presented. Section 5 provides the results

of the study and compares the performance of the methods for handling missing data used in this paper. Section 6 summarizes the results and main findings of this study.

2. ABOUT THE NATIONAL ADDICTIONS SURVEY: TOBACCO CONSUMPTION IN MEXICO

The National Addictions Survey (ENA) is a probabilistic, randomized and multistage household survey conducted by the National Institute of Public Health in Mexico (INSP) and the National Institute of Psychiatry "Ramón de la Fuente Muñiz". The data are representative at a national level and also for eight regions of the country. The regions are North Central (Coahuila, Chihuahua y Durango), Northwestern (Baja California, Baja California Sur, Sonora y Sinaloa), Northeastern (Nuevo León, Tamaulipas y San Luis Potosí), Western (Zacatecas, Aguascalientes, Jalisco, Colima y Nayarit), Central (Puebla, Tlaxcala, Morelos, Estado de México, Hidalgo, Querétaro y Guanajuato), Mexico City (Distrito Federal), South Central (Veracruz, Oaxaca, Guerrero y Michoacán) and South (Yucatán, Quintana Roo, Campeche, Chiapas y Tabasco).

The aim of the ENA is to estimate the prevalence of consumption of tobacco, alcohol and illegal drugs in the Mexican population aged from 12 to 65 years old. In the latest ENA, held in 2011, the respondents answered a computerized version of the questionnaire. A total of 16,249 persons were interviewed, of which 3,849 were adolescents (12-17 years), and 12,400 were adults (18-65). Of the sample, females represented the 55.44%. By region, 19.73% of the respondents came from the North Central region, 13.10% were from the South, 12.84% from the Western, 12.39% were from the Northwestern, 12.13% from the South Central, while 11.53%, 9.46% and 8.82% came from the Central, Northeastern and Mexico City regions, respectively.

About tobacco consumption, most of the participants were classified as never smokers (57.52%). 22.91% of all respondents were former smokers and 19.57% were current smokers.

The questions used to determine the smoking status were "During your lifetime, have you ever smoked tobacco even once?" and "When did you last smoke a cigarette?" Those who answered "No" to the first question were classified as never smokers. If the respondent answered "Yes" to the first question, then the person was questioned about the last time they smoked. Those who smoked a year before the baseline interview were classified as former smokers. The current smokers are the respondents who reported having smoked during the last twelve months.

Among people whose status were current smokers, 86.67% were adults, 66.19% were men, 36.98% were single and 36.16% were married. The 91.16% of the former smokers were adults, 54.76% were men, 47.1% were married and 27.65% were single. 66.88% of the classified as never smokers were adults, 66.85% were women, 47.43% were single and 34.57% were married.

3. MISSING DATA PATTERNS AND MECHANISMS

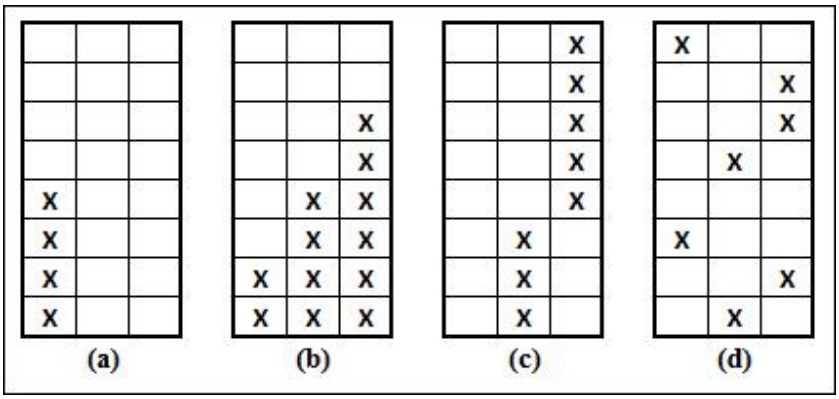
Missing data represents a common problem for statisticians and researchers working with surveys and census. Although a variety of imputation methods have been developed to handling missing data, each method suffers from several limitations and may not perform reasonably well under some circumstances. One reason for this is that most of these techniques make assumptions about how the missing values are distributed within the data set. To decide how to handle missing data it is worthwhile to know the underlying missingness pattern and mechanism, i.e., to identify which values are missing and why.

In the standard taxonomy, three types of missing data patterns can be distinguished. If there is only one variable with missingness, then the pattern is called univariate. When the multivariate pattern is observed, means that the nonresponse occurs in more than one variable.

When observations are ordered and item k is missing, the pattern is said to be monotone if all $k + 1, \ldots, n$ cases are also missing. If this pattern is not monotone, then it is called general or non-monotone.

If any non-missing data point can be reached from any other non-missing data point through a sequence of horizontal or vertical moves, then the missing data pattern is said to be connected (van Buuren, 2012).

Figure 1. Missing data patterns: (a) univariate pattern, (b) monotone pattern, (c) connected pattern, (d) random pattern



Source: Own elaboration based on van Buuren (2012)

One major problem with nonresponse is the missing data mechanism. Little and Rubin (1987) introduced a useful classification of the mechanisms that lead data being missing. They defined three different assumptions: Missing Completely at Random (MCAR), Missing at Random

(MAR), and Missing Not at Random (MNAR). According to Song *et al.* (2005), "the missingness mechanism concerns whether the missingness is related to the study variables or not."

Many data sets can be arranged in a rectangular or matrix form. Let X be an $n \times p$ matrix of partially observed sample data, where the rows, n, correspond to the sample size, and the columns, p, are the number of variables that have been measured. Symbol X_j^{obs} indicates an observed case of the variable j, and X_j^{mis} denotes a missing value in the column j, $j = 1, \dots, p$.

To describe missingness accurately it is important to consider it as a probabilistic phenomenon (Rubin, 1976). When the probability of being missing is unrelated to any of the variables in the dataset (neither on the variable subject to nonresponse nor on any other variable) and the missingness is a random sample from the observed data, then the data are said to be MCAR.

$$P\left[X_j^{mis} \middle| X_1, \dots, X_p\right] = P\left[X_j^{mis}\right]$$

This implies that the probability of being missing is the same for all the units (Rubin, 1987). MCAR is the easiest scenario to face up, however missing data are very rarely MCAR. To determine whether MCAR assumption is satisfied, Little (1988) has provided a statistical test based on a chi-square distribution. The null hypothesis is that the data are MCAR, so a non-significant test supports the assumption of MCAR mechanism.

The MAR assumption is weaker, more common and more realistic than the MCAR mechanism. If the MAR assumption is not rejected, then the probability of nonresponse depends only on the available information but not on the missing values (Durrant, 2005):

$$P\left[X_j^{mis} \middle| X_1, \dots, X_p\right] = P\left[X_j^{mis} \middle| X_1, \dots, X_{j-1}, X_{j+1}, X_p\right]$$

A common way to test if MAR assumptions are held is by modeling missingness as a binary (dichotomous) response regression, such as logistic or probit models, where the response variable equals 1 for missing values and 0 for observed.

When data are MCAR or MAR, then the missing data mechanism can be considered as ignorable.

When MCAR nor MAR assumptions are not satisfied, data can be classified as MNAR. Contrary to MCAR, when the MNAR assumption holds, then it means that the probability of nonresponse is related to the missing (unobservable) values. One of the implications of MNAR is that a missing cases have a different distribution than the observed, even when they otherwise have the same characteristics. As a consequence, since the value of the missing cases depends on information not available, they cannot be predicted unbiasedly.

The MNAR mechanism of missingness is non-random and cannot be considered as ignorable.

To make these ideas concrete to the variable of this study (smoking habits), I give some examples. The data could be considered MCAR, if the decision to answer or not answer a question about smoking habits is unrelated to the respondent's smoking habits or to their marital status, gender or age. If married participants were more likely to omit reporting smoking habits than single respondents, then the MAR assumption holds, because missingness would be related with marital status. When someone fails to report smoking habits and their decision to report or not report depends on their smoking habit, then the data are MNAR. For example, when current smoker respondents are less likely to answer the questions about their smoking habits than never smoker people, then the missingness is not ignorable.

4. STRATEGIES FOR ADDRESSING MISSING DATA

There are several methods to handling missingness of categorical data in surveys. In this paper is evaluated the performance of six of these methods: complete case analysis, imputation of the mode, random imputation, the hot-deck method, imputation by polytomous regression and random forests.

4.1. Complete case analysis

The complete case analysis, also known as listwise deletion, eliminates all observations with missing values in at least one variable. Because of convenience, this is the most widely applied approach of handling missing data. Actually, this is the default method applied in many statistical packages (SAS, SPSS, Stata, R). In some cases, when the missing data are MCAR, the listwise deletion can provide better estimations than other methods, because the observations with missing data are a random sample of the full sample (Farhangfar *et al.* 2008; and, Matsubara *et al.*, 2008). But since the subsample produced by the complete case analysis will always have fewer cases, the standard errors and significance levels are often larger relative to all available data. In the other hand, if the missing-data mechanism is MAR or MNAR, this technique can introduce bias and result in a considerable efficiency loss, as shown by Desai *et al.* (2011), Little and Rubin (2002) and Schafer and Graham (2002).

4.2. Imputation methods

Rather than removing the non-observed cases, there are approaches that retain all the data, replacing (imputing) each missing observation with a plausible value. However, the aim of imputation is not to fill in all missing cases, but to "preserve the characteristics of their distribution and relationships between different variables" (Barceló, 2008).

Consider X, the $n \times p$ partially observed matrix, as an approximation to the true sample data Y. Thus Y is a fully observed $n \times p$ matrix. The process of imputation is the set of procedures

applied to the partially observed matrix, X, with the aim to find a fully recorded matrix Y^*, that is an approximation to the completely observed matrix Y.

Keeping the full sample size can be advantageous for bias and precision, nevertheless, using imputation methods without carefully bearing in mind the assumptions required for the valid application of each method, can yield to misleading results. In the following lines are described the imputation methods applied in this paper.

4.2.1. Mode imputation

One of the easiest ways in the case of categorical data is to fill in each missing value with the sample mode. This is a common practice; nonetheless, the major disadvantage of mode imputation is that it creates spikes in the distribution by concentrating all the imputed values in the mode, as a consequence, the variance is reduced artificially (Kalton and Kish, 1981). This is a single imputation method, since only one value is used to replace each missing observation.

4.2.2. Random imputation

Let R be a $n \times p$ matrix of data with univariate pattern, and let m be the number of non-observed cases in the only one variable with missingness j. The random imputation consists of taking a simple random sample of size m from the $n - m$ non-missing values in the partially observed variable j, and returns these as imputations, obtaining an imputed fully recorded vector j^*.

One of the advantages of this method is that it does not produce impossible values, nevertheless one important drawback is that random imputation can introduce an additional amount of variability due to the random selection of residuals (Chauvet, Deville and Haziza, 2011).

4.2.3. The hot-deck method

Hot-deck imputation implicates replacing missing cases on incomplete records (recipient) using values from complete observations of the same data set (donors) that matches the case that is missing. When two or more observations are similar to the non-respondent with respect to characteristics observed, then the method uses the expected value of the scores. Hot- deck imputation is appropriate when dealing with categorical data and is usually non-parametric. This approach appears to be reasonable; however, it presents two main disadvantages: it assumes perfect correlation between the variables, disregarding variability, and the more variables uses the less likely to find a match (Andridge and Little, 2010; Durrant, 2005).

4.2.4. Imputation by polytomous regression

Imputation by polytomous regression is applied when the dependent variable is a categorical variable with more than two categories. A general expression for the conditional probability is:

$$P[Y = k|X_i] = \frac{e^{\theta(k|X_i)}}{e^{\theta(1|X_i)+\cdots+\theta(K|X_i)}}$$

Assuming $Y = 1, \ldots, K$, the log odds ratio between categories k and K (base category) is defined as $\theta(k|X_i) = \log\frac{P[Y=k|X_i]}{P[Y=K|X_i]}$, $k = 1, \ldots, K$. The model assumes $\theta(k|X_i) = \beta_{k0} + \beta_{k1}x_{i1} + \cdots + \beta_{kp}x_{ip}$, where $\beta_{k0}, \ldots, \beta_{kp}$ are the regression parameters and p is the number of variables in the model.

In general, the method consists of the following steps, first it fits a categorical response as a polytomous model, then computes the predicted categories, and finally adds appropriate noise to predictions (Hosmer and Lemeshow, 1989; Souverein *et al.*, 2006; and, Silva-Ramírez *et al.*, 2011).

4.2.5. Random forests

The random forests imputation method is a machine learning method and it is one of the newest techniques for imputing. In this approach regression trees are constructed iteratively from bootstrap. The set of trees grown constitutes a forest. Each tree "votes" and this vote is used to classify each instance based on the majority (mode) vote over all trees (Stekhoven and Bühlmann, 2012; Pantanowitz and Marwala, 2008; Segal, 2004; Rieger *et al.*, 2010).

Random forest offers advantages in terms of dealing with mixed-type data, it is relatively robust to outliers and noise, and does not have the same assumptions of normality, linearity, homoscedasticity (Hill, 2012; Rieger *et al.*, 2010)

The hot-deck, imputation by polytomous regression and the random forests approaches are considered as multiple imputation methods, since a vector of plausible values is used to replace each missing observation.

5. MISSINGNESS SIMULATIONS

Taking the ENA, two datasets at rates 5% and 15%, each for MCAR, MAR and MNAR data were generated as described below. Only the univariate pattern has been considered in this study, the variable with missingness is the one related to the smoking status. These six new versions of the original dataset were used to examine the impact of different methods to handle missingness for the smoking status.

As mentioned before, the original dataset was comprised of 16,249 complete cases. The MCAR datasets were created by randomly adding 5% and 15% of missingness. The MAR datasets were simulated under the specific assumption that adolescents and women were more likely to be missing than the rest of participants. The MNAR datasets were created so that the current smokers were more likely to be missing than never smokers and former smokers.

Table 1. Levels of non-response

Smoking Status	Original dataset	
	N	%
Current Smoker	3180	19.6%
Former Smoker	3722	22.9%
Never Smoker	9347	57.5%
Missing	0	0.0%

Smoking Status	MCAR 5% Little's test chi-sq = 2.180662 p-value = 0.9023		MCAR 15% Little's test chi-sq = 8.161031 p-value = 0.2265		MAR 5% Little's test chi-sq = 3427.891 p-value = 0		MAR 15% Little's test chi-sq = 654.6718 p-value = 0		MNAR 5% Little's test chi-sq = 197.3155 p-value = 0		MNAR 15% Little's test chi-sq = 589.5141 p-value = 0	
	N	%	N	%	N	%	N	%	N	%	N	%
Current Smoker	3042	18.7%	2685	16.5%	3098	19.1%	2763	17.0%	2627	16.2%	1461	9.0%
Former Smoker	3536	21.8%	3186	19.6%	3655	22.5%	3237	19.9%	3541	21.8%	3201	19.7%
Never Smoker	8859	54.5%	7941	48.9%	8684	53.4%	7812	48.1%	9269	57.0%	9150	56.3%
Missing	812	5.0%	2437	15.0%	812	5.0%	2437	15.0%	812	5.0%	2437	15.0%

6. RESULTS

The imputation of the smoking status variable has led to different estimates in the sample proportions of current smokers, former smokers and never smokers. The results of each imputation method are described below. In order to facilitate interpretation of results, they have been represented in tables. To that same end, 95% confidence intervals were produced using the sample mean and variance of the smoking status proportions from each dataset generated.

Under the MCAR mechanism, the easiest assumption to manage, most of the techniques yielded to no statistically significant differences between sample proportions of smoking conditions in each imputed dataset and the proportions of the original dataset. In fact, the only method that consistently produced very different estimates to those from the original dataset was the mode imputation. Concentrating all the imputed values in the mode led to serious underestimation of current and former smokers and an overestimation in never smokers (see Tables 2 and 3)

As the rate of missingness increased from 5% to 15%, the difference between the mode imputation estimates grew even larger from those from the original data.

Figure 2. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MCAR missingness

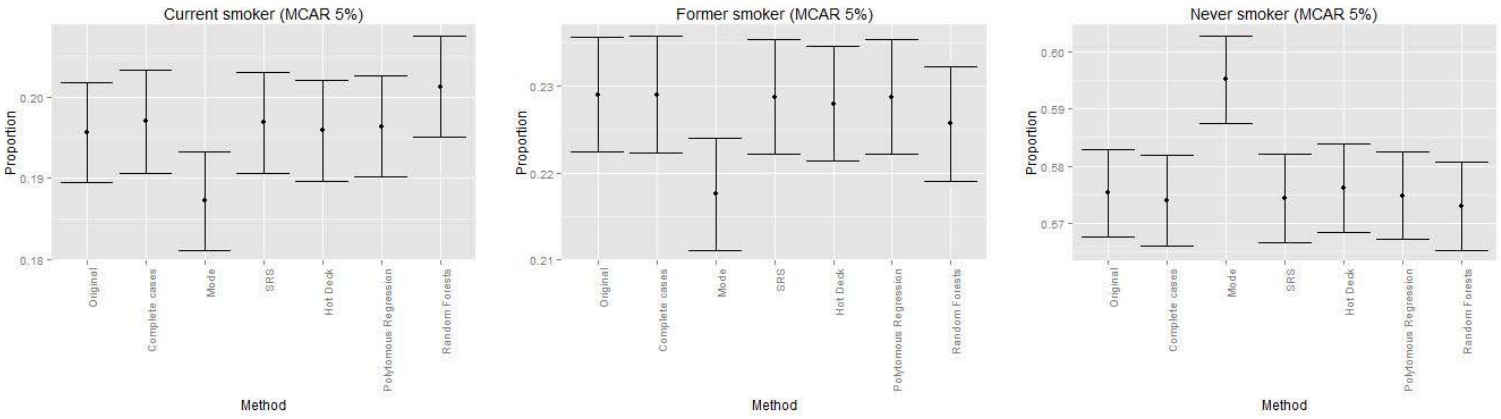


Table 2. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MCAR missingness

Proportion	SE	LCI	UCI
19.57%	0.00311	18.95%	20.19%
22.91%	0.00330	22.25%	23.57%
57.52%	0.00388	56.75%	58.30%

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	18.72%	0.00306	18.11%	19.33%
Former Smoker	21.76%	0.00324	21.11%	22.41%
Never Smoker	59.52%	0.00385	58.75%	60.29%

Proportion	SE	LCI	UCI
19.71%	0.00320	19.07%	20.35%
22.91%	0.00338	22.23%	23.58%
57.39%	0.00398	56.59%	58.18%

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.69%	0.00312	19.06%	20.31%
Former Smoker	22.88%	0.00330	22.22%	23.53%
Never Smoker	57.44%	0.00388	56.66%	58.21%

Proportion	SE	LCI	UCI
19.59%	0.00311	18.97%	20.21%
22.80%	0.00329	22.14%	23.46%
57.61%	0.00388	56.83%	58.39%

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.14%	0.00315	19.51%	20.77%
Former Smoker	22.57%	0.00328	21.91%	23.22%
Never Smoker	57.30%	0.00388	56.52%	58.07%

Proportion	SE	LCI	UCI
19.64%	0.00312	19.02%	20.27%
22.88%	0.00330	22.22%	23.53%
57.48%	0.00388	56.70%	58.26%

Figure 3. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MCAR missingness

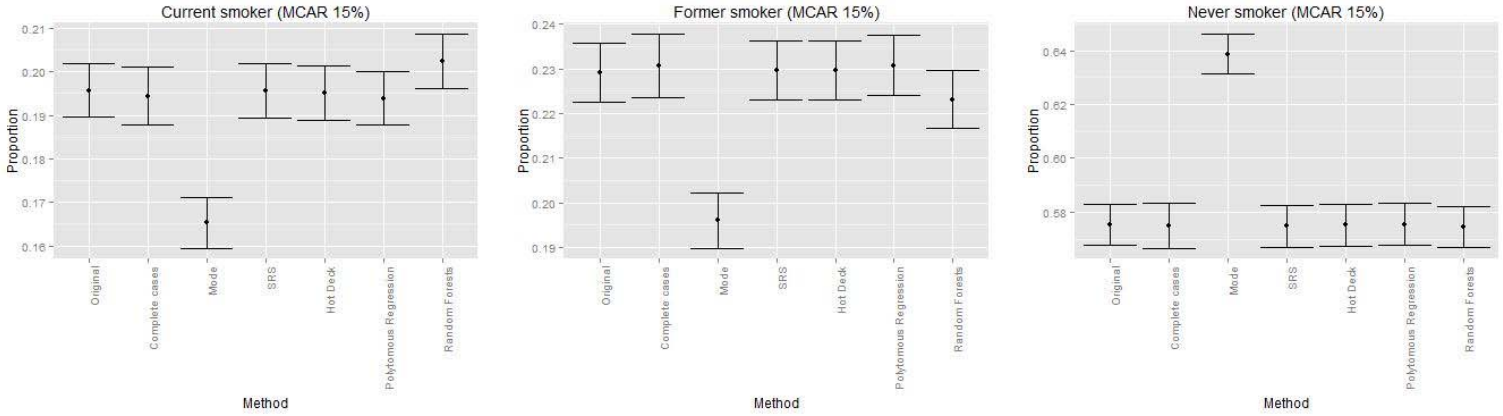


Table 3. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MCAR missingness

Original

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.57%	0.00311	18.95%	20.19%
Former Smoker	22.91%	0.00330	22.25%	23.57%
Never Smoker	57.52%	0.00388	56.75%	58.30%

Mode

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	16.52%	0.00291	15.94%	17.11%
Former Smoker	19.61%	0.00311	18.98%	20.23%
Never Smoker	63.87%	0.00377	63.11%	64.62%

Complete cases

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.44%	0.00337	18.77%	20.11%
Former Smoker	23.07%	0.00358	22.35%	23.78%
Never Smoker	57.49%	0.00421	56.65%	58.33%

SRS

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.56%	0.00311	18.94%	20.18%
Former Smoker	22.96%	0.00330	22.30%	23.62%
Never Smoker	57.48%	0.00388	56.70%	58.26%

Hot Deck

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.52%	0.00311	18.89%	20.14%
Former Smoker	22.97%	0.00330	22.31%	23.63%
Never Smoker	57.52%	0.00388	56.74%	58.29%

Random forest

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.24%	0.00315	19.61%	20.87%
Former Smoker	22.32%	0.00327	21.66%	22.97%
Never Smoker	57.44%	0.00388	56.67%	58.22%

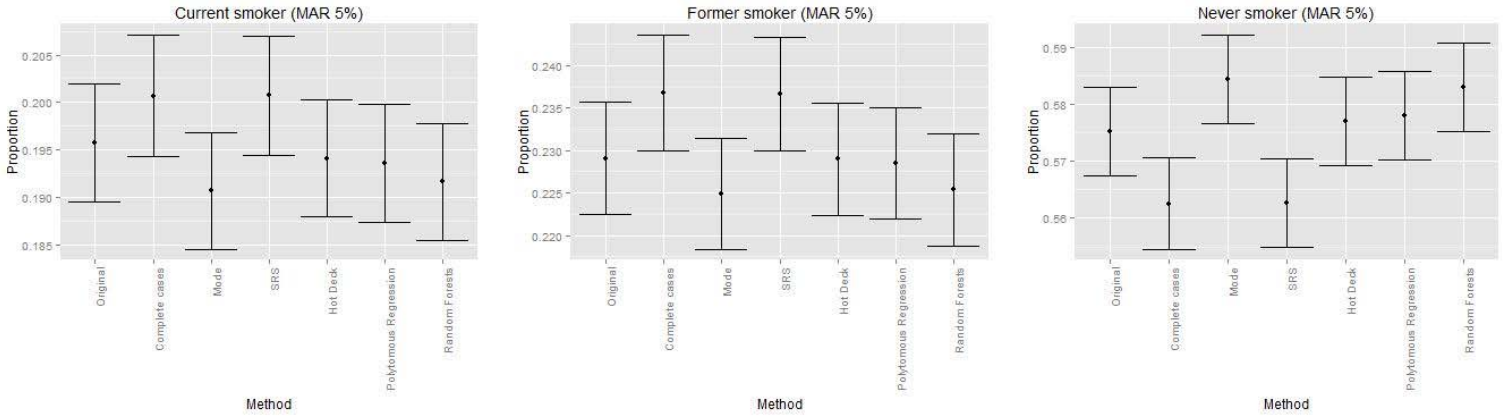
MICE

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.39%	0.00310	18.77%	20.01%
Former Smoker	23.08%	0.00331	22.42%	23.74%
Never Smoker	57.54%	0.00388	56.76%	58.31%

When 5% of the values were missing under the MAR mechanism, three of the methods returned estimates closer to their original values. These were the most sophisticated approaches: hot-deck (p-value = 0.8647), polytomous regression (p-value = 0.7356), and random forests (p-value = 0.133)[1].

At 15% rate of missing cases, no statistically significant difference was found in the estimates between the original dataset and the data handled by the complete cases approach (p-value = 0.07227), and the hot-deck (p-value = 0.7857), the polytomous regression (p-value = 0.4781) and the random imputation (p-value = 0.134) methods (see Tables 4 and 5).

Figure 4. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MAR missingness



[1] The p-values are produced using the outcome of a t-test that compares the proportion of the original sample and the proportion of the imputed dataset.

Table 4. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MAR missingness

Original

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.57%	0.00311	18.95%	20.19%
Former Smoker	22.91%	0.00330	22.25%	23.57%
Never Smoker	57.52%	0.00388	56.75%	58.30%

Mode

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.07%	0.00308	18.45%	19.68%
Former Smoker	22.49%	0.00328	21.84%	23.15%
Never Smoker	58.44%	0.00387	57.67%	59.21%

Complete cases X-squared = 0.2908 p-value = 0.8647

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.07%	0.00322	19.42%	20.71%
Former Smoker	23.68%	0.00342	22.99%	24.36%
Never Smoker	56.25%	0.00399	55.46%	57.05%

SRS

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.08%	0.00314	19.45%	20.70%
Former Smoker	23.66%	0.00333	23.00%	24.33%
Never Smoker	56.26%	0.00389	55.48%	57.04%

Hot Deck

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.41%	0.00310	18.79%	20.03%
Former Smoker	22.90%	0.00330	22.24%	23.56%
Never Smoker	57.69%	0.00388	56.91%	58.46%

Random Forests

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.16%	0.00309	18.55%	19.78%
Former Smoker	22.54%	0.00328	21.88%	23.19%
Never Smoker	58.30%	0.00387	57.53%	59.07%

Polytomous Regression

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.36%	0.00310	18.74%	19.97%
Former Smoker	22.85%	0.00329	22.19%	23.51%
Never Smoker	57.79%	0.00387	57.02%	58.57%

Figure 5. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MAR missingness

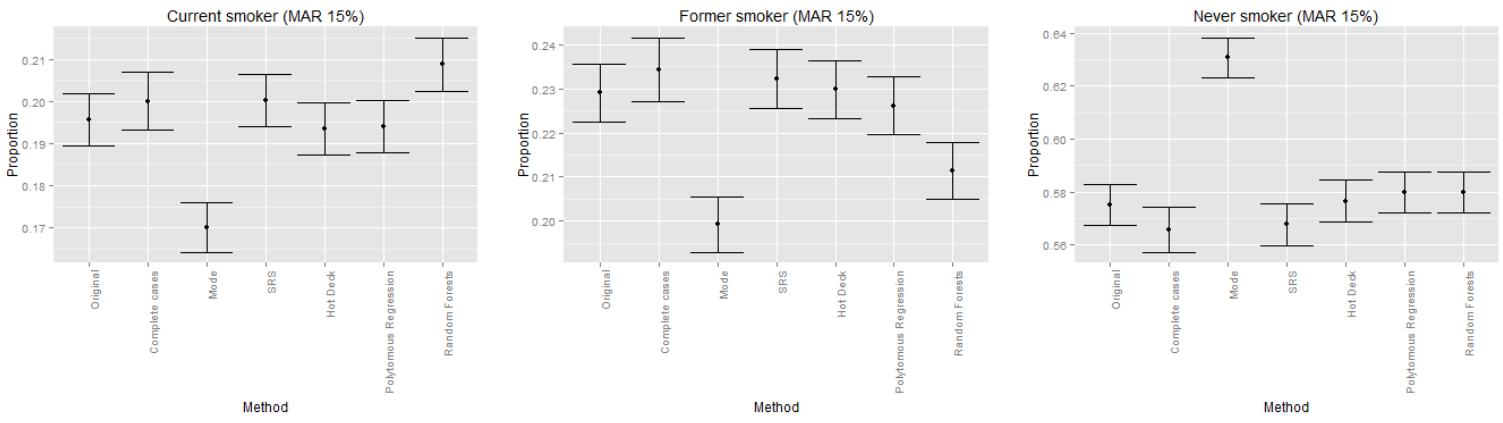


Table 5. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MAR missingness

Original

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.57%	0.00311	18.95%	20.19%
Former Smoker	22.91%	0.00330	22.25%	23.57%
Never Smoker	57.52%	0.00388	56.75%	58.30%

Mode

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	17.00%	0.00295	16.41%	17.59%
Former Smoker	19.92%	0.00313	19.29%	20.55%
Never Smoker	63.07%	0.00379	62.32%	63.83%

Complete cases

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.00%	0.00340	19.32%	20.69%
Former Smoker	23.44%	0.00360	22.72%	24.16%
Never Smoker	56.56%	0.00422	55.72%	57.40%

SRS

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.02%	0.00314	19.39%	20.65%
Former Smoker	23.22%	0.00331	22.56%	23.88%
Never Smoker	56.76%	0.00389	55.98%	57.54%

Hot Deck

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.36%	0.00310	18.74%	19.97%
Former Smoker	22.99%	0.00330	22.33%	23.65%
Never Smoker	57.66%	0.00388	56.88%	58.43%

Random forest

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	20.88%	0.00319	20.24%	21.51%
Former Smoker	21.14%	0.00320	20.50%	21.78%
Never Smoker	57.99%	0.00387	57.21%	58.76%

MICE

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.39%	0.00310	18.77%	20.01%
Former Smoker	22.62%	0.00328	21.96%	23.27%
Never Smoker	57.99%	0.00387	57.22%	58.77%

As was mentioned before, the MNAR assumption is the hardest mechanism to face up. The MNAR, both at 5% and 15% rates, showed the largest differences between the original dataset and the six approaches to handle missing data, all the methods produced biased results even though in some cases the random forest approach performed results very close to the 95% limits (see Tables 6 and 7).

Figure 6. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MNAR missingness

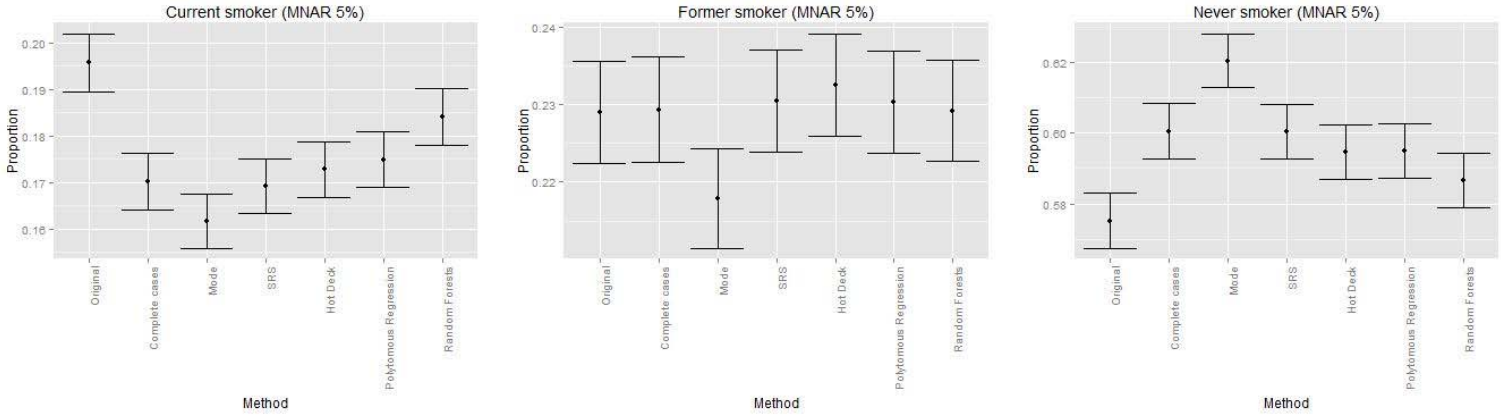


Table 6. Confidence intervals of the sample proportions for the different missing data approaches applied to 5% MNAR missingness

Original

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.57%	0.00311	18.95%	20.19%
Former Smoker	22.91%	0.00330	22.25%	23.57%
Never Smoker	57.52%	0.00388	56.75%	58.30%

Mode

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	16.17%	0.00289	15.59%	16.74%
Former Smoker	21.79%	0.00324	21.14%	22.44%
Never Smoker	62.04%	0.00381	61.28%	62.80%

Complete cases

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	17.02%	0.00302	16.41%	17.62%
Former Smoker	22.94%	0.00338	22.26%	23.62%
Never Smoker	60.04%	0.00394	59.26%	60.83%

SRS

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	16.92%	0.00294	16.33%	17.51%
Former Smoker	23.05%	0.00330	22.39%	23.71%
Never Smoker	60.03%	0.00384	59.26%	60.80%

Hot Deck

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	17.28%	0.00297	16.69%	17.87%
Former Smoker	23.26%	0.00331	22.59%	23.92%
Never Smoker	59.46%	0.00385	58.69%	60.23%

Random forest

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	18.41%	0.00304	17.81%	19.02%
Former Smoker	22.92%	0.00330	22.26%	23.58%
Never Smoker	58.66%	0.00386	57.89%	59.43%

MICE

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	17.49%	0.00298	16.89%	18.09%
Former Smoker	23.03%	0.00330	22.37%	23.69%
Never Smoker	59.48%	0.00385	58.71%	60.25%

Figure 7. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MNAR missingness

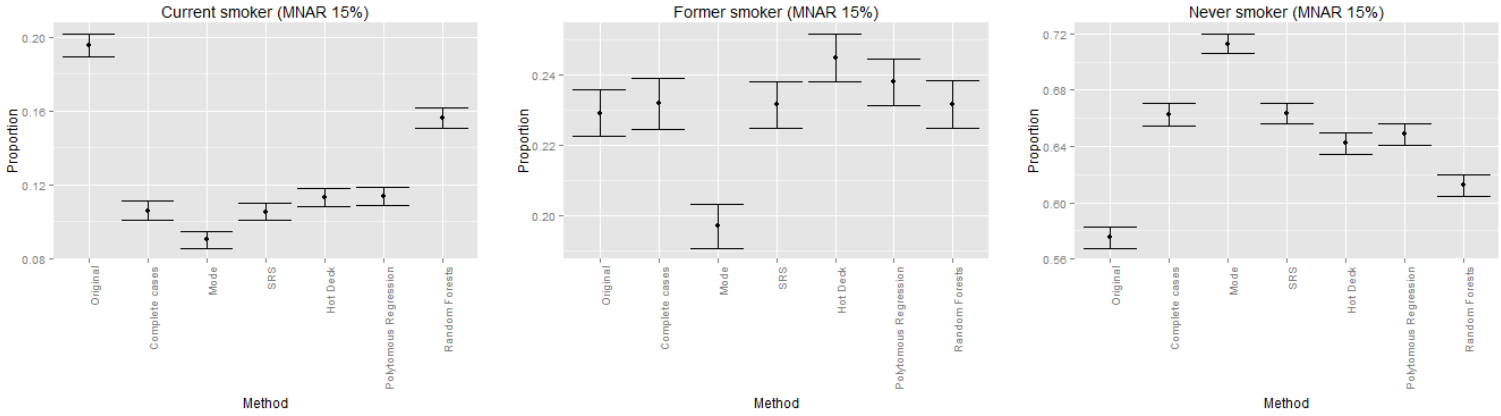


Table 7. Confidence intervals of the sample proportions for the different missing data approaches applied to 15% MNAR missingness

Original

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	19.57%	0.00311	18.95%	20.19%
Former Smoker	22.91%	0.00330	22.25%	23.57%
Never Smoker	57.52%	0.00388	56.75%	58.30%

Mode

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	8.99%	0.00224	8.54%	9.44%
Former Smoker	19.70%	0.00312	19.08%	20.32%
Never Smoker	71.31%	0.00355	70.60%	72.02%

Complete cases

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	10.58%	0.00262	10.05%	11.10%
Former Smoker	23.18%	0.00359	22.46%	23.89%
Never Smoker	66.25%	0.00402	65.44%	67.05%

SRS

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	10.52%	0.00241	10.04%	11.01%
Former Smoker	23.15%	0.00331	22.49%	23.81%
Never Smoker	66.32%	0.00371	65.58%	67.07%

Hot Deck

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	11.29%	0.00248	10.80%	11.79%
Former Smoker	24.49%	0.00337	23.81%	25.16%
Never Smoker	64.22%	0.00376	63.47%	64.97%

Random forest

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	15.61%	0.00285	15.04%	16.18%
Former Smoker	23.16%	0.00331	22.50%	23.82%
Never Smoker	61.23%	0.00382	60.46%	61.99%

MICE

Smoking Status	Proportion	SE	LCI	UCI
Current Smoker	11.34%	0.00249	10.84%	11.84%
Former Smoker	23.79%	0.00334	23.12%	24.46%
Never Smoker	64.87%	0.00375	64.12%	65.61%

7. CONCLUSIONS

Although imputation procedures are often useful, in this paper is noted that no universally best approach to handle missingness exists. Every method suffers from limitations related to the missing data mechanism. Nonetheless, understanding why data are missing can guide the researcher to an appropriate strategy for addressing missingness. The fact that assumptions about non-observed values can affect estimates and results is evident in this paper.

Eliminating all observations with missing values in at least one variable (complete case analysis) returned to reasonable estimates under the MCAR assumption.

While one of the easiest ways of filling in the blanks is to replace the missing cases with the mode, this approach showed the worst performance in every rate and under every missingness assumption. The mode imputation produced the largest bias.

While the simple random sample (random) imputation results were characterized by unbiased estimates when the missing data mechanism can be considered as ignorable.

The random forest method led to non-dissimilar results for the MCAR at 5% and 15% rates and for MAR at 5% of missingness.

The most effective methods for dealing with missing data in most of the missing data scenarios assessed in this paper were the hot-deck and the polytomous regression approaches. This finding has key implications because both methods are available in most of the software packages (both free and commercial) such as R, SAS, Stata, SPSS and S-Plus. It is important to remember that missingness under the MAR and MCAR assumptions is linked to the rest of variables (observed), so the methods that performed the best were those that uses values from complete observations of the same dataset.

In addition, another important outcome of this study is that it investigated how the performance of the model was affected by varying amounts of missing data and different missing data mechanisms. In general, with a small number of missing data cases, the various strategies will likely have small impact on estimates. For larger rates of non-observed cases, the effect of the strategies for handling missingness was less efficient. Also, it should be pointed out that MCAR, MAR and MNAR mechanisms led to dissimilar results for a given imputation method. Under the MNAR assumption, both at 5% and 15% rates, none of the methods performed well even though in some cases the random forest approach performed results very close to the 95% limits. This result might be comparable to the findings of Schafer and Graham (2002), who found that some approaches to deal with MAR assumptions can produce unbiased results under MNAR.

Finally, as with any study, there are limitations to the current work that must be considered. First, the simulations were based only under the univariate missingness pattern. Second, the current

study is focused on non-ordered categorical data. Although this type of data is very common in surveys, there are other types of variables that can be considered. Further research should be focused on ordinal, continuous and mixed categorical and continuous data.

REFERENCES

Andridge, R. and Little, R. (2010). A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review, 78 (1), pp. 40–64.

Bacallao, J. and Bacallao, J. (2010). Imputación Múltiple en Variables Categóricas Usando Data Augmentation y Árboles de Clasificación. Investigación Operacional, 31 (2), pp. 133–139.

Barceló, C. (2008). The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the Spanish Survey of Household Finances. (No. 0829). Banco de España.

Burton, A., Billingham, L.J., and Bryan, S. (2007). Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. Clinical Trials, 4 (2), pp. 154–161.

Chauvet, G., Deville, J.C., and Haziza, D. (2011). On balanced random imputation in surveys. Biometrika, 98 (2), pp. 459–471.

Desai, M., Esserman, D.A., Gammon, M.D., and Terry, M.B. (2011). The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects. Epidemiologic Perspectives and Innovations, 8 (1), 5.

Durrant, G.B. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review. NCRM Methods Review Papers. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM/002.

Eisemann, N., Annika, W., and Alexander, K. (2011) Imputation of missing values of tumour stage in population-based cancer registration. BMC Medical Research Methodology, 11.

Farhangfar A, Kurgan L, and Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recognit 41 (12), pp. 3692–3705

Follmann, D., Elliott, P., Suh, I., and Cutler, J. (1992). Variance imputation for overviews of clinical trials with continuous response. Journal of clinical epidemiology, 45 (7), pp. 769–773.

Ghosh-Dastidar, B., and Schafer, J.L. (2003). Multiple edit/multiple imputation for multivariate continuous data. Journal of the American Statistical Association, 98 (464), pp. 807–817.

Gimotty, P.A. and Brown, M.B. (1990). Imputation procedures for categorical data: their effects on the goodness-of-fit chi-square statistic. Communications in Statistics-Simulation and Computation, 19 (2), pp. 681–703.

Hill, J. (2012) Four Techniques for Dealing with Missing Data in Criminal Justice. Paper presented at the annual meeting of the ASC Annual Meeting, Palmer House Hilton, Chicago, IL, Nov 13, 2012.

Hosmer, D.W. and Lemeshow, S. (1989). Introduction to the Logistic Regression Model. Applied Logistic Regression, Second Edition, pp. 1–30.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the survey research methods section (pp. 146–151).

Little, R.J. (1988). A test of missing completely at random for multivariate data with missing values Journal of the American Statistical Association, 83 (404), pp. 1198–1202.

Little, R.J. and Rubin, D.B. (1987). Statistical analysis with missing data (Vol. 539). New York: Wiley.

Little, R.J. and Rubin, D.B. (2002). Statistical analysis with missing values. Wiley, New York.

Little, R.J. and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika, 72 (3), pp. 497–512.

Matsubara, E.T., Prati, R.C., Batista, G.E., and Monard, M.C. (2008). Missing value imputation using a semi-supervised rank aggregation approach. Advances in Artificial Intelligence-SBIA 2008 (pp. 217–226). Springer Berlin Heidelberg.

Panranowitz, A. and Marwala, T. (2009) Missing Data Imputation Through the Use of the Random Forest Algorithm. Advances in Intelligent and Soft Computing Volume 116, pp. 53–62.

Rieger, A., Hothorn, T., and Strobl, C. (2010). Random Forests with Missing Values in the Covariates.

Rubin, D.B. (1976) Inference and missing data. Biometrika, 63, pp. 581–592.

Schafer, J.L. and Graham, J.W. (2002). Missing data: our view of the state of the art. Psychological methods, 7 (2), 147.

Segal, M.R. (2004). Machine learning benchmarks and random forest regression

Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M., and Cubiles-de-la-Vega, M.D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks, 24 (1), pp. 121–129.

Song, Q., Shepperd, M., and Cartwright, M. (2005). A short note on safest default missingness mechanism assumptions. Empirical Software Engineering, 10 (2), pp. 235–243.

Souverein, O.W., Zwinderman, A.H., and Tanck, M.W. (2006). Multiple imputation of missing genotype data for unrelated individuals. Annals of human genetics, 70 (3), pp. 372–381.

Stekhoven, D.J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28 (1), pp. 112–118.

van Buuren, S. (2012). Flexible imputation of missing data. CRC press.