Handling Missing Data: Traditional Techniques Versus Machine Learning

Andrei Cojocaru
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
a.cojocaru@student.utwente.nl

ABSTRACT

Missing data is a serious problem in data science and in other fields that rely on statistical inference. Improper handling of missing data often leads to biased or invalid conclusions. Given this risk, existing research compares many techniques for the practical analysis of a dataset with missing data, all of varying levels of quality. This paper examines and documents the methodological flaws that affect many of these studies, presenting a comparison based on more realistic assumptions. Traditional statistical techniques are compared to machine learning algorithms, and the strengths and weaknesses of each category are described based on the observed results, offering some prescriptions for the right time to apply machine learning to missing data problems.

Keywords

Data analysis, missing data imputation, machine learning, multiple imputation, random forest, expectation maximization, k-nearest neighbors, perceptron, bayesian ridge regression

1. INTRODUCTION

At its very core, data science is about drawing conclusions about a population from a given sample (inference) or making predictions (building predictive models) based on datasets. Statistical inference, in particular, is a process that is not unique to data science, but recurs throughout a wide range of different disciplines, from education and family studies to nursing and psychology, even impacting genetics.

A frequent and often poorly understood problem in drawing conclusions from datasets is that of missing data. Due to a variety of causes such as measurement errors or participant non-response, datasets in practical analysis often have a number of missing data points. The researchers' approach to handling these missing data points has the potential to severely bias the results of the research, even rendering them entirely invalid. Within data science itself, improper handling of missing data interferes not only with inference, but with predictive model building as well. An algorithm trained on biased data will create biased results.

31st Twente Student Conference on IT July 5th, 2019, Enschede, The Netherlands.

This risk has motivated a significant amount of research regarding the optimal way to handle missing data.

The techniques for handling missing data used in practical analysis vary widely, from ad-hoc methods such as mean substitution, to more sophisticated ones such as multiple imputation and expectation maximization. Over the years, there have been a wide variety of studies that compare these techniques. The purpose of this paper is to address the flaws of these comparative studies, and offer a more realistic and practical comparison between these techniques and machine learning algorithms, the latter of which have rarely been considered in the missing data literature.

1.1 Research Questions

RQ1 What are the most frequently applied imputation methods in missing data handling?

RQ2 What prior work has been done with respect to comparing imputation techniques, and how were the comparisons carried out?

> **RQ2.1** What are the flaws in prior comparative studies?

RQ3 How does the performance of traditional imputation techniques compare to that of machine learning techniques applied for missing data imputation?

> **RQ3.1** Which missing data mechanism is most relevant within practical data analysis?

1.2 Summary

The problem of missing data adversely affects the performance of data science solutions. Similarly, it damages the scientific quality of empirical data-centric studies in a wide variety of disciplines. Consequently, many researchers of different fields have attempted to document and compare the various techniques for handling missing data (**Section 3**). Taken collectively, these works form a multi-disciplinary methodology for missing data handling in science.

Many of the aforementioned comparative works suffer from a consistent pattern of methodological flaws and unrealistic assumptions, such as using only one dataset. These flaws are explained and documented herein, paper by paper (**Section 3.1, Table 3**).

A more realistic methodology for technique comparison is designed (**Section 4**) in order to conduct an experiment with more generalizable results. Moreover, machine learning algorithms are also compared with the traditional missing data handling techniques.

The resulting R-Squared and F1 scores for each method are documented graphically (**Section 5** and **Appendix**)

Table 1. Complete age and height data matrix

Person	Height (meters)	Age
1	1.90m	18
2	1.75m	16
3	1.85m	23
4	1.93m	17

Table 2. Incomplete age and height data matrix

Person	Height (meters)	Age
1	1.90m	N.A.
2	1.75m	16
3	1.85m	23
4	1.93m	N.A.

and interpreted, explaining the most suitable dataset types for machine learning imputation methods according to their relative performance (**Section 6**).

2. BACKGROUND

A complete understanding of the research problem requires some familiarity with missing data theory. This section aims to introduce all immediately relevant concepts in a concise way, while also providing a starting point for further reading.

Firstly, it is necessary to establish a clear definition of how data is generally structured in a practical analysis.

2.1 Data Matrices

A data matrix is defined as a set of observations (rows) and a set of variables (columns). For each of these variables, any given observation will have either a value representing the realization of the variable (e.g. a value of 36 for an age variable), or a missing value, denoted N.A. A data matrix with at least one missing value is referred to as incomplete.

It is instructive to consider the notion of hypothetically complete data, as outlined by Enders 2010 [5, p.9]. Suppose there is a simple data matrix which records the age and height of a number of individuals. An incomplete data matrix arises, for example, when some of the individuals refuse to provide their age (**Table 2**). The complete data refers to the ages that would have been obtained had all the participants agreed to provide them (**Table 1**).

Using this concept, one can view each incomplete data matrix as a result generated from a complete data matrix through some generative process. For instance, all participants in the prior example had ages, but some chose not to provide them – thus, the incomplete data matrix was generated from the complete data matrix containing all the participants' actual ages. This generative process is referred to as a missing data mechanism.

2.2 Missing Data Mechanisms

Literature in missing data theory recognizes three major missing data mechanisms. The following are explanations of these mechanisms using the perspective of generative processes from a complete data matrix.

Useful notions, such as the diagrams and formal notation, are adapted from Enders (2010) [5].

Let R be an auxiliary indicator variable that is encoded to 1 when the value of a particular variable (in this case, age) is missing, and 0 otherwise. Let $\boldsymbol{\phi}$ denote a set of parameters that govern the relationship between R and the data. Furthermore, let $\boldsymbol{Y_{obs}}$ refer to the observed parts of the data, and $\boldsymbol{Y_{mis}}$ to the missing parts of the data.

Using these notions, the following mechanisms can be described formally.

2.2.1 *Missing Completely at Random (MCAR)*

An incomplete data matrix is said to follow the MCAR mechanism if the missingness of each variable has no correlation to the values of other variables or to its own real value in the hypothetically complete data (e.g. people who are taller don't have a tendency to hide their age, and people of older/younger ages aren't systematically more likely to hide their age). Essentially, a sufficient condition for MCAR is if the observations with missing values are picked at random from the complete data matrix, with equal probability.

Formally, the distribution of MCAR data can be defined as follows:

$$p(R|\phi)$$

This is illustrated by **Figure 1**.

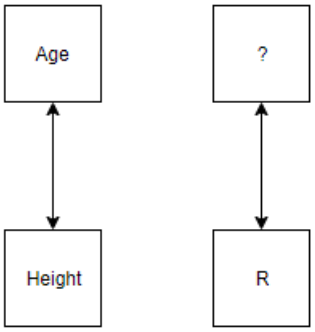

Figure 1. MCAR mechanism in a simple dataset (? corresponds to variables not present in the dataset)

An example of a generative process that takes a complete data matrix and creates an MCAR incomplete data matrix can be described in terms of a uniform probability distribution. Consider the case in **Table 1**: If each observation has the same probability of having a missing age, the resulting incomplete data matrix will be MCAR, since this satisfies both prior conditions.

2.2.2 *Missing at Random (MAR)*

The MAR mechanism implies that the missingness of a given variable is correlated to the values of at least one other variable. For instance, if, in the prior example, participants who are taller are less likely to report their age, this would imply a MAR mechanism for the data.

Formally, the distribution of MAR data can be defined as follows:

$$p(R|Y_{obs}, \phi)$$

This is illustrated by **Figure 2**.

The generative process for this would be similar to the previous, with non-uniform probabilities: a higher value for a given participant's height would imply a higher probability that their age is missing.

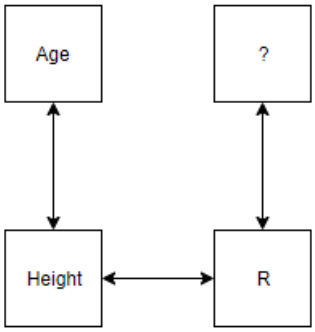

Figure 2. MAR mechanism in a simple dataset (? corresponds to variables not present in the dataset)

2.2.3 *Missing Not at Random (MNAR)*

MNAR implies that the probability of missingness of a given variable's value is correlated to that value (i.e. its actual value in the hypothetically complete data). Considering the example from **Table 1** once more, if older participants are less likely to report their age, this would imply a MNAR mechanism.

The MNAR mechanism can be formally defined thusly:

$$p(R|Y_{obs}, Y_{mis}, \phi)$$

This is illustrated by **Figure 3**

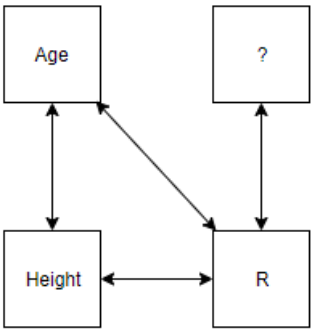

Figure 3. MNAR mechanism in a simple dataset (? corresponds to variables not present in the dataset)

The generative process for MNAR data is similar to that for MAR: a higher value for a given participant's actual age would imply a higher probability that this age is not reported.

2.2.4 *Practical Considerations*

Practically speaking, MCAR does not arise frequently in real datasets. MAR tends to be a safer assumption, with some MNAR cases being convertible to MAR [5, p. 14-16]. This makes missing data handling methods that perform well given a MAR mechanism quite useful. Consequently, this also suggests that those which only perform well with the MCAR mechanism are of limited use in practical analysis.

2.3 Further Reading

Enders 2010 [5] provides an excellent, in-depth treatment of missing data theory along with traditional and modern missing data handling techniques and their efficacy. Schafer & Graham (2002) [17] provide a more formal statistical coverage of the same, with some discussion regarding techniques for MNAR data. Rubin (1976) [16] offers a technical, theoretical foundation for missing data theory and missing data analysis.

A precise description of the missing data handling methods is beyond the scope of this paper. Enders (2010) [5] covers both modern (e.g. multiple imputation, maximum likelihood) and ad-hoc (e.g. mean, median, mode) imputation techniques.

3. RELATED WORK

There have been numerous papers in the domain of missing data handling regarding the various techniques and their efficacy. Schafer and Graham (2002) [17] identify bias, variance, and mean square error of the estimated quantity as performance criteria for these techniques. They also primarily recommend the usage of multiple imputation and maximum likelihood methods. Furthermore, as a practical consideration, they add that, according to Collins, Schafer, & Kam (2001) [4], assumptions of MAR when the data is actually MNAR often have only a minor impact in practice.

Johnson and Young (2011) [9] concur that Multiple Imputation (MI) and Maximum Likelihood (ML) methods are the primary choice for modern missing data analysis, also citing a consensus in the literature ([1], [7] and [17]) regarding the superiority of MI and ML to alternatives.

In the context of a longitudinal study on stress, Musil et al. (2002) [13] conduct a comparison of ad-hoc techniques (e.g. listwise deletion, mean substitution) and the expectation maximization (EM) technique, finding that EM is superior.

Similarly, in the context of longitudinal studies in psychology, Peeters et al. (2015) [15] conduct a comparison of traditional ad-hoc techniques versus multiple imputation using bias and the width of confidence intervals as performance metrics, finding multiple imputation to be the most effective method.

Schmitt (2015) [18] compares mean imputation, K-nearest neighbors (KNN), and multiple imputation with some more exotic techniques, such as fuzzy K-means. Methodologically, they use multiple datasets derived from practical data collection, assuming a Missing Completely at Random mechanism.

More recently, Duy Le (2018) [11] compares multiple imputation, KNN, EM imputation, and mean imputation on two datasets. Likewise, this work assumes a Missing Completely at Random mechanism.

Based on these results, multiple imputation, expectation maximization, and maximum likelihood can be identified as the most common standard non-machine learning imputation techniques. Likewise, ad-hoc techniques such as mean imputation are also commonly applied.

Overall, the literature contains a plethora of comparisons of ad-hoc techniques with modern ones, but the subject of comparing the performance of traditional techniques versus that of those based on machine learning has never been thoroughly explored.

3.1 Other Comparative Works and Their Limitations

Each comparative study mentioned herein suffers from at least one of the following methodological limitations:

1. The study uses purely randomized (simulated) data
    - This creates an epistemic problem: it cannot be guaranteed that the fully simulated data recreates every relevant property of real data from a practical domain. Thus, the techniques may perform differently in a realistic situation.
2. The study uses only one dataset, or a few datasets from the same source and domain
    - This is once again an epistemic problem. It is feasible that certain datasets originating from certain domains have unique individual properties that make them materially different from those in other domains, in a way that is significant to the results of imputation techniques.
3. The study does not account for the size difference in datasets
    - This ignores an important aspect of performance: how well certain techniques scale, in both accuracy and speed, with increasingly larger datasets.
4. The study does not consider machine learning techniques for comparison versus traditional techniques
    - It could be the case that, given enough data with a strong enough predictive relation to the missing values, machine learning techniques are superior to traditional imputation.
5. The study assumes an unrealistic (i.e. Missing Completely at Random) missing data mechanism
    - Most missing data patterns in practical analysis conform to at least the Missing at Random mechanism. Thus, such a comparison is not entirely realistic given this assumption.

Table 3 lists notable comparative studies and their particular flaws.

Table 3. Flaws of Similar Comparative Studies

Paper	F1	F2	F3	F4	F5
Schafer 2002 [17]	✗	✗	✗	✗	
Musil 2002 [13]		✗	✗	✗	
Buhi 2008 [3]	✗	✗	✗	✗	
Johnson 2011 [9]		✗	✗	✗	
Graham 2012 [6]		✗	✗	✗	
Peeters 2015 [15]		✗	✗	✗	
Schmitt 2015 [18]					✗
Duy Le 2018 [11]	✗	✗			✗

3.2 Techniques Compared

Table 4 describes which techniques this paper evaluates that are also common in one or more similar works. It is important to note that the techniques listed are not *all* of the techniques assessed in the related works, but the ones that are also assessed here.

4. METHODOLOGY

The experiment was carried out in the Python programming language, within a Jupyter Notebook [10] environment. The Pandas library was used to easily represent and work with data matrices, the impyute library was used for traditional statistical imputation techniques, and scikit-learn [14] was used to apply machine learning techniques.

Table 4. Techniques Examined in Similar Comparative Studies

Paper	Mean	Median	Mode	EM	MI
Schafer 2002 [17]	✗				✗
Musil 2002 [13]	✗			✗	
Buhi 2008 [3]	✗				✗
Johnson 2011 [9]	✗				✗
Graham 2012 [6]	✗				✗
Peeters 2015 [15]					✗
Schmitt 2015 [18]	✗				✗
Duy Le 2018 [11]	✗			✗	✗
This Paper	✗	✗	✗	✗	✗

4.1 Dataset Selection Criteria

For selecting datasets to use in the experiment, several criteria were chosen:

- The size of the dataset file should be lower than 5 megabytes, since the traditional techniques are too slow past this size
- Each dataset should have at least one numerical and one categorical variable
- The data should not require expert knowledge to understand
- The dataset should be in CSV format for ease of use
- The dataset's variables should be clearly named and described

4.2 Selected Datasets

Based on the above criteria, several datasets of diverse originating domains were selected, to account for the potential differences in technique performance resulting from the individual characteristics of each dataset. For a clearer picture, the datasets were also selected so as to be of ascending size.

The following datasets were selected:

- Howell demographic data ([12], retrieved from `https://github.com/rmcelreath/rethinking/blob/master/data/Howell1.csv`
- UCI heart disease data ([8], retrieved from `https://www.kaggle.com/ronitf/heart-disease-uci`
- 1985 to 2016 suicide rates (retrieved from `https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016`)
- King County house sales (retrieved from `https://www.kaggle.com/harlfoxem/housesalesprediction`)

To account for performance differences based on differing variable types, one numerical and one categorical variable was selected for separate testing from each dataset.

4.3 MAR Simulation

Using the idea of generative processes as outlined in the background, two incomplete data matrices (having a numerical variable and a categorical variable with missing values, respectively) were generated using a MAR mechanism from the complete matrix of each dataset. The detailed process of generating a MAR incomplete data matrix was conducted in the following steps:

Firstly, since a MAR missing data mechanism is desired, transitive MNAR relations should be avoided. To illustrate: Suppose that a given dataset contains an income, weight, and height variable. Suppose furthermore that the height variable is the target for simulating missingness, and the values of the weight variable are used to determine the probability that the corresponding values of the height variable are missing. Since there is a strong positive correlation between weight and height, then, transitively, there will be a strong correlation between the value of height and the probability that this value will be missing after the simulation. This is the very definition of a MNAR mechanism, hence, this situation is undesirable.

For this reason, a synthetic float-type variable was introduced to the data matrix. Each row of the data matrix was assigned a random (seed: "thesis") number drawn from a standard normal distribution, corresponding to the realization of the synthetic variable. Since this variable could not be correlated with any of the existing variables in the matrix, assigning it a positive correlation to the probability of missingness of the target variable will create a MAR mechanism without incidentally creating a transitive MNAR one, as well.

Heuristically, based on the number of observations in each selected dataset, 1.2% and 15% were assigned as the minimum and maximum probabilities of missingness, respectively. The maximum generated value of the synthetic variable was tied to the maximum probability of missingness, and the minimum generated value was tied to the minimum probability of missingness.

Let $\boldsymbol{S_{max}}$ represent the maximum of the synthetic values, $\boldsymbol{S_{min}}$ the minimum, $\boldsymbol{P_{max}}$ the maximum probability of missingness, and $\boldsymbol{P_{min}}$ the minimum probability of missingness.

Viewing the probability of missingness as a linear function of the synthetic value, the following function with unknown slope and intercept can be defined:

$$P_{miss}(s) = a + bs$$

where s represents a given synthetic value.

Taking the minimum and maximum probabilities along with the minimum and maximum synthetic values as two input-output pairs within this function, the slope and intercept can be inferred by solving the following system of equations:

$$P_{min} = a + bS_{min}$$
$$P_{max} = a + bS_{max}$$

Since both P and S min/max values are known, this is a solvable system which gives the intercept and slope for the desired function, which maps the generated synthetic values to probabilities of missingness. Given this mapping, a corresponding probability was computed for each synthetic value in a given data matrix row. A real number from 0 to 1 was drawn from a uniform distribution for each of these probabilities, and the value of the target variable was omitted if the drawn number was below the probability.

4.4 Dataset Preprocessing

Each dataset's variables were scaled using scikit-learn's StandardScaler object, to ensure that all machine learning algorithms run as smoothly and correctly as possible. Categorical variables were encoded to integers using scikit-learn's LabelEncoder.

4.5 Train/Test Split

Since the experiment focused in part on applying machine learning algorithms, the various datasets had to be split into training data and test data. For each dataset, the rows containing missing data served as the test set, and all the remaining rows served as the training set. Since the missingness was artificially generated, the actual values of the missing data were kept to evaluate the predictions of each algorithm.

4.6 Evaluated Techniques

On the traditional technique side, the mean, median, mode, expectation maximization, and multiple imputation techniques were applied. Random imputation, defined as simply picking a random value from the non-missing set, was also used as a benchmark. Maximum Likelihood imputation was omitted from the experiment due to time constraints and lack of an already available implementation in the Python language.

For categorical variables, the following machine learning techniques were applied:

- Random Forest Classification
- Logistic Regression
- Perceptron Classification
- K-Nearest Neighbors Classification

For numerical variables, the following machine learning techniques were applied:

- Random Forest Regression
- Linear Regression
- Bayesian Ridge Regression
- K-Nearest Neighbors Regression

4.7 Implementation & Evaluation

The traditional ad-hoc techniques were applied using the impyute library. Multiple imputation was applied from the fancyimpute library, specifically with the IterativeImputer algorithm.

All methods applied to categorical variables were assessed using scikit-learn's classification_report function, which computes the precision and recall per category. All methods applied to numerical variables were assessed using scikit-learn's r2_score (R-squared calculator) function.

These techniques were executed twice (once for categorical, once for numerical) per dataset. The models were fit to a training set (all the non-missing data) and evaluated on a test set (all the missing data).

The resulting performance metrics were compiled into .txt format reports on a per-dataset basis and saved to disk for later analysis.

4.8 Adaptive Rounding

The Multiple Imputation and Expectation Maximization techniques generate real (non-integer) numbers for both numerical and categorical variables. Thus, their results are useless without rounding in the categorical case.

For this purpose, multiple rounding schemes can be used. The first and most basic rounding scheme is naive rounding, which is essentially rounding to the nearest integer.

Enders (2010) cites multiple sources indicating that this naive rounding scheme inadvertently results in biased conclusions, particularly for binary categorical variables. He cites Bernaards (2007) [2], who proposes the following adaptive rounding scheme:

Let $\hat{\mu}_{UR}$ be the mean of the imputed values (not yet rounded). Then,

$$c = \hat{\mu}_{UR} - \phi^{-1}(\hat{\mu}_{UR})\sqrt{\hat{\mu}_{UR}(1-\hat{\mu}_{UR})}$$

where c is the threshold for rounding and ϕ represents the standard normal density function. Then, iterating through each imputed value, the following rule is applied: If the value is above the threshold, it is encoded to 1. Otherwise, it is encoded to 0.

Adaptive rounding is applied for each categorical variable imputation with the Multiple Imputation and Expectation Maximization techniques.

4.9 Hyperparameter Tuning

Both the Random Forest and K-nearest neighbors (KNN) algorithms have a number of complex parameters which impact their performance significantly. The optimal values for these parameters depend on the dataset the algorithms are being applied on – thus, these optimums must be estimated on a per-dataset basis. This process is referred to as hyperparameter tuning.

In the context of this experiment, hyperparameter tuning is conducted via the RandomizedSearchCV algorithm from scikit-learn, with 3-fold cross validation. The values used for the hyperparameter search algorithm for the Random Forest and KNN algorithms are listed in **tables 5 and 6**, respectively.

Table 5. Random Forest Hyperparameter Values

Parameter	Values
n_estimators	5, 10, 20, 40, 60, 100
max_features	auto, sqrt
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrap	True, False

Table 6. KNN Hyperparameter Values

Parameter	Values
n_neighbors	All integers in [2, 12]
weights	uniform, distance
algorithm	auto, ball_tree, kd_tree, brute
leaf_size	10, 20, 30, 40, 50, ..., 100
p	1, 2, 3, 4, 5

5. RESULTS

The performance scores of each technique are compared graphically using color-coded bar plots (blue for traditional techniques, red for machine learning). Each individual bar plot corresponds to technique performance on a variable in a certain dataset.

The two most significant plots on any given dataset-variable combination are displayed in **Figure 4**, illustrating the contrast between the relative performance of machine learning classification techniques on the smallest (UCI) versus the largest (suicide) dataset in the study. Likewise, the comparatively poor increase of traditional technique performance is also visible.

The full results are illustrated in the Appendix (**Figure 5**). Multiple imputation is omitted from the housing price graph since its R-squared is negative (-4.9).

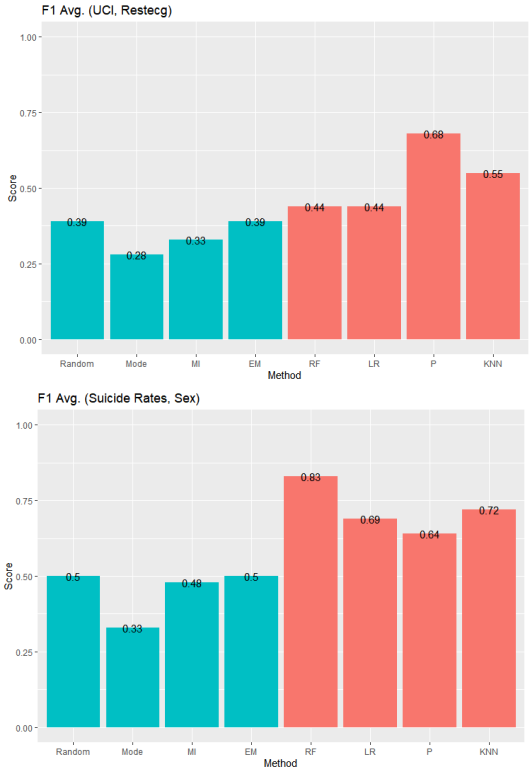

Figure 4. Classification on small vs. large dataset

5.1 Discussion

To begin with, it is clear that the machine learning techniques tend to have a better relative performance within classification than they do within regression. A possible explanation is that regression tends to be more difficult and requires more fine-grained hyperparameter tuning than classification.

An interesting aspect to note is the superior performance of traditional methods relative to machine learning algorithms on the UCI heart disease dataset. This particular dataset is the smallest in this study, with 303 observations. It is important to remember that the traditional techniques used in this study do not have an underlying model that must be fit to a given set of training data. Thus, they are not penalized for not having enough training data, whereas the machine learning-based algorithms are. It is quite likely that the traditional techniques are superior in both result and ease of use when applied to particularly small datasets, generally speaking.

In contrast, the superiority of machine learning algorithms relative to the traditional methods is clearly visible in the context of the suicide rate dataset imputation. This is the largest dataset used in this study. This finding is consistent with the previously proposed idea that traditional techniques are superior on smaller datasets, but do not scale as well on larger ones.

To further support this hypothesis, it is also notable that, on datasets significantly larger than the suicide rate dataset,

the traditional imputation techniques did not finish execution in a reasonable time. However, whether this is the result of the particular implementation used herein or of the underlying theory behind the techniques is a matter for future study.

Another point of interest is the impressive (F1 of 0.99) performance of all machine learning techniques along with mode imputation and expectation maximization on the housing dataset's waterfront variable, in contrast to multiple imputation (F1 of 0.75). The variable in question is highly skewed – evidently, few homes are on a waterfront. 21450 observations have this variable set to 0, and only 163 have it set to 1. This explains the performance metrics observed. Moreover, this suggests that multiple imputation is not a good approach for highly skewed categorical variables.

Finally, it is important to address the performance of multiple imputation on the price variable of the housing dataset. In this case, its R-Squared is negative (-4.90). Upon further investigation, multiple imputation was found to generate some negative house prices, thus explaining the poor score. Based on the reasonable performance of all other techniques, this appears to be a flaw with the method, or its implementation.

There are a few factors negatively influencing the generality of the presented results:

- The MAR mechanism simulation is fairly naive, as it involves generating an otherwise uncorrelated synthetic variable
- Due to time constraints, the datasets are not as numerous as they should be for a truly general study
- The selected datasets or their vertical partitions are of fairly low dimensionality
- The author's relative inexperience with applied machine learning is a significant factor in the machine learning techniques' performance

6. CONCLUSION

The most common state of the art imputation techniques studied are multiple imputation, maximum likelihood imputation, and expectation maximization. Naive methods such as mean imputation are also commonly used. (**RQ1**)

Many studies comparing these techniques have been carried out in the context of various scientific disciplines. The methodology of comparative studies within the field of missing data imputation has a number of recurring flaws, such as the use of only one dataset, the use of fully simulated data, or the assumption of an MCAR mechanism. (**RQ2** and **RQ2.1**).

These flaws can be addressed effectively through larger-scale studies with more realistic assumptions. Specifically, MAR mechanisms and real datasets do away with the built-in assumptions of generality and randomness that do not reflect reality (**RQ3.1**). Moreover, researchers would do well to consider the power of machine learning algorithms when comparing such techniques.

Given more realistic assumptions, the results of the comparison indicate that traditional techniques tend to be superior on smaller datasets, whereas machine learning techniques tend to scale better and outperform them as the number of observations in the dataset increases. (**RQ3**) This also makes sense in principle, since machine learning models require training data, and tend to perform better when there is more data. This need not be the case for traditional imputation techniques.

Some limitations of these results are the low number of datasets used, their simplicity, and the fairly naive MAR simulation mechanism. These are experimental constraints that can be relaxed in future studies to paint a better picture of the relative performance of traditional imputation techniques versus machine learning models.

6.1 Future Work

There are numerous experimental constraints to relax in order to achieve more generalizable results. For instance, this study only compares the performance of techniques on a dataset with MAR missingness. Some datasets have a missingness pattern that fits the MCAR mechanism, and some others have an irreducible MNAR mechanism. Comparing these techniques on each possible missing data mechanism for a variety of datasets would paint a much clearer and more informative picture.

Furthermore, multiple libraries could be used in order to account for the differences in technique performance that result from implementation specifics.

On the dataset side, a wider range of dataset sizes and more varied levels of dimensionality would further increase the generality of derived results.

Moreover, this study has not taken into account a number of state of the art non-machine learning imputation techniques, such as model based ones (e.g. FIML) and Bayesian techniques.

Finally, since most imputation-related studies are conducted with the social sciences in mind, there has been little work done in exploring effective imputation in the context of big data. This presents a promising avenue for further research.

7. ACKNOWLEDGEMENTS

The author would like to thank Dr. Doina Bucur for her generous advice on the practical application of machine learning, along with her feedback with respect to this paper.

8. REFERENCES

[1] A. C. Acock. Working with missing values. *Journal of Marriage and family*, 67(4):1012–1028, 2005.

[2] C. A. Bernaards, T. R. Belin, and J. L. Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in medicine*, 26(6):1368–1382, 2007.

[3] E. R. Buhi, P. Goodson, and T. B. Neilands. Out of sight, not out of mind: Strategies for handling missing data. *American journal of health behavior*, 32(1):83–92, 2008.

[4] L. M. Collins, J. L. Schafer, and C.-M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.

[5] C. K. Enders. *Applied Missing Data Analysis*. Guilford Press, 2010.

[6] J. W. Graham, P. E. Cumsille, and A. E. Shevock. Methods for handling missing data. *Handbook of Psychology, Second Edition*, 2, 2012.

[7] D. C. Howell. The analysis of missing data. *Handbook of social science methodology*, pages 208–224, 2008.

[8] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart disease dataset.

[9] D. R. Johnson and R. Young. Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family*, 73(5):926–945, 2011.

[10] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90, 2016.

[11] T. D. Le, R. Beuran, and Y. Tan. Comparison of the most influential missing data imputation algorithms for healthcare. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 247–251. IEEE, 2018.

[12] R. McElreath. *Statistical rethinking: a Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[13] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones. A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, 24(7):815–829, 2002.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] M. Peeters, M. Zondervan-Zwijnenburg, G. Vink, and R. Van de Schoot. How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, 12(4):377–394, 2015.

[16] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

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

[18] P. Schmitt, J. Mandel, and M. Guedj. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6(1):1, 2015.

Figure 5. Appendix (Result Graphs)