UNIVERSITI PUTRA MALAYSIA

COLD DECK MISSING VALUE IMPUTATION WITH A TRUST-BASED SELECTION METHOD OF MULTIPLE WEB DONORS

MOHD IZHAM BIN MOHD JAYA

FSKTM 2018 79

COLD DECK MISSING VALUE IMPUTATION WITH A TRUST-BASED SELECTION METHOD OF MULTIPLE WEB DONORS

By

MOHD IZHAM BIN MOHD JAYA

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

December 2018

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

COLD DECK MISSING VALUE IMPUTATION WITH A TRUST-BASED SELECTION METHOD OF MULTIPLE WEB DONORS

By

MOHD IZHAM BIN MOHD JAYA

December 2018

Chair : Assoc. Prof. Fatimah binti Sidi, PhD
Faculty : Computer Science and Information Technology

Missing value is a common problem in any dataset and its occurrence decreases data completeness as data values are missing. Moreover, the problem reduces data quality and negatively impacted the result of data analysis. Existing cold deck imputation coped with this problem by selecting a replacement value from a pool of donors identified in other data sources during the imputation process. In comparison to other imputation methods, existing cold deck imputation has less risk on model misspecification and preserves data distribution in the dataset.

Nevertheless, the limitation of the existing cold deck imputation is the chances in finding trusted plausible donor is narrow due to a usage of single data source in each imputation process. The availability of various web data sources today alleviates this limitation. However, as values from multiple web data sources are commonly conflicted to each other, adopting existing cold deck imputation with multiple web donors is not a practical solution as trust score on each of the conflicted values is not measured. Thus, it is difficult to select the most plausible value during imputation process. This research concentrates on improving data completeness by imputing missing values using a trust based cold deck imputation.

Trust Based Cold Deck Missing Values Imputation with Multiple Web Donor is presented in this research. The proposed method takes advantage of multiple web donors from web data sources in order to provide higher chances in finding the most plausible values to impute missing values. The plausible values are selected based on the trust score computation's novelty which is measured by accuracy score and reliability score of the web donor.

The performance of the proposed method is evaluated by running a prediction model on the imputed dataset. A number of experiments are carried out to quantify the accuracy of the prediction model, Root Mean Squared Error (RMSE), and the F-Measure. The results demonstrate that the proposed method improves the performance of existing cold deck imputation. Additionally, the results are then compared with other imputation methods which are K-Nearest Neighbor (KNN), Mean Imputation (AVG), Case Deletion (IGN), Predictive Mean Matching (PMM) and MissForest. The results showed that the RMSE, prediction accuracy and F-Measure is improved when the prediction model is trained with datasets imputed using the proposed method. This research contributed to the improvement of data quality especially to the information system (IS) and database field where good data quality benefited the data analysis performance.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

IMPUTASI NILAI HILANG DEK SEJUK DENGAN KAEDAH PEMILIHAN BERASASKAN KEPERCAYAAN UNTUK BERBILANG PENYUMBANG WEB

Oleh

MOHD IZHAM BIN MOHD JAYA

Disember 2018

Pengerusi : Professor Madya Fatimah binti Sidi, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Nilai hilang adalah masalah yang biasa ditemui pada kebanyakan set data dan kehadirannya akan menyebabkan ketidaksempurnaan data di dalam set data meningkat. Tambahan pula, masalah nilai hilang juga menyebabkan kemerosotan kualiti data dan memberi impak negatif kepada hasil analisis data. Imputasi dek sejuk yang sedia ada berupaya untuk mengatasi masalah ini dengan memilih nilai pengganti dari kolam penyumbang yang dikenalpasti daripada sumber data yang lain. Berbanding dengan kaedah imputasi yang lain, imputasi dek sejuk yang sedia ada mempunyai risiko yang lebih rendah terhadap kesalahan spesifikasi model dan memelihara distribusi data di dalam set data.

Walaupun begitu, peluang untuk menjumpai penyumbang munasabah yang boleh dipercayai adalah kecil dalam kaedah imputasi dek sejuk yang sedia ada kerana hanya sumber data tunggal yang digunakan di dalam setiap proses imputasi. Hari ini, dengan kebolehsediaan pelbagai sumber data web, halangan ini dapat diatasi. Walau bagaimanapun, nilai yang diperoleh dari berbilang sumber data web biasanya bercanggah di antara satu sama lain. Penggunaan kaedah imputasi dek sejuk yang sedia ada adalah penyelesaian yang tidak praktikal kerana skor kepercayaan untuk setiap nilai yang bercanggah tidak dinilai. Oleh itu, adalah sukar untuk memilih nilai yang paling munasabah dan paling boleh dipercayai semasa proses imputasi. Kajian ini menumpukan kepada pembaikan kesempurnaan data dengan imputasi terhadap nilai hilang menggunakan imputasi dek sejuk berasaskan kepercayaan.

Kaedah imputasi nilai hilang dek sejuk berasaskan kepercayaan untuk berbilang penyumbang web dipersembahkan di dalam kajian ini. Kaedah yang dicadangkan ini memanfaatkan berbilang penyumbang web dari sumber data web untuk memberikan peluang yang lebih tinggi dalam mencari nilai yang paling

munasabah dan paling boleh dipercayai untuk imputasi nilai hilang. Nilai yang paling munasabah dan paling boleh dipercayai adalah dipilih berdasarkan kepada skor kepercayaan yang diukur melalui skor ketepatan data dan skor keutuhan penyumbang web.

Prestasi kaedah yang dicadangkan adalah dinilai melalui model ramalan yang dilarikan dengan set data yang diimputasi. Beberapa eksperimen telah dijalankan untuk menyatakan peratusan ketepatan model ramalan, Ralat Punca Min Kuasa Dua (RPMKD), dan nilai-F. Keputusan eksperimen menunjukkan kaedah yang dicadangkan dapat memperbaiki prestasi kaedah imputasi dek sejuk sedia ada. Keputusan eksperimen untuk pendekatan yang dicadangkan juga dibandingkan dengan kaedah imputasi yang lain seperti Jiran-K yang Terdekat (KNN), imputasi purata (AVG), penghapusan kes (IGN), penyesuaian purata yang dijangka (PMM), dan MissForest. Secara umumnya, prestasi model ramalan telah ditingkatkan apabila dilatih menggunakan set data yang telah diimputasi menggunakan kaedah yang dicadangkan. Penyelidikan ini menyumbang kepada penambahbaikan kualiti data terutamanya dalam bidang Sistem Informasi (SI) dan pangkalan data di mana kualiti data yang baik memberi manfaat kepada prestasi analisis data.

ACKNOWLEDGEMENTS

First and foremost, I'm thankful to Allah for all His blessings and His merciful in giving me strength and patience in completing this research. My deepest gratitude to my supervisor, Associate Professor Dr. Fatimah Sidi for her endless supports, prayers and guidance. I am also thankful to my supervisory committee, Associate Professor Dr. Lilly Suriani Affendey and Associate Professor Dr. Marzanah A. Jabar for the insightful advices and comments towards the accomplishment of this research. My special thanks to Dr. Iskandar Ishak for his commitments and priceless support throughout this research. I am indebted to my sponsor, Jabatan Perkhidmatan Awam Malaysia for giving me the opportunity to conduct this research.

I dedicated this research to my beloved wife, Zulinda Zulkifli, my three little heroes, Izz Ammar Daniel, Izz Elman Ali and Izz Emir Adrien, my parents and family. There is no words can described every loves, patience, strengths, dreams and hope that you gave me during this journey.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Fatimah binti Sidi, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Lilly Suriani binti Affendey, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Marzanah binti A. Jabar, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ________________________ Date: ___________________

Name and Matric No.: Mohd Izham bin Mohd Jaya, (GS40595)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABT	Abbott Laboratories
ANN	Artificial Neural Network (ANN)
AVG	Mean Imputation
BIS	Business Intelligent System
CEG	Constellation Energy Group Inc.
CSV	Comma-separated values
EM	Expectation-maximization
ERP	Enterprise Resource Planning
FASB	Financial Accounting Standards Board
FN	False Negative
FP	False Positive
FRO	Financial Report Ontology
GPC	Genuine Parts Company
GRG	Gray Relational Grade
HEOM	Euclidean-overlap metric
IGN	List-wise Deletion
KNN	K-Nearest Neighbour
kNN-FWPD	KNN Feature Weighted Penalty Based Dissimilarity
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multivariate Imputation by Chained Equation
MTB	M&T Bank Corp
NMAR	Not Missing at Random
OFFDM	Ontology-based framework for financial decision-making
OWL	Web Ontology Language
PCA	Principle Component Analysis
PMM	Predictive Mean Matching
RCSE	Relative Change in Stock Earning
RKNN	Reduced Relational Grade KNN
RMSE	Root Mean Squared Error
RRG	Reduced Relational Grade
SEC	U.S. Securities and Exchange Commission

SMOTE	Synthetic Minority Oversampling Technique
TN	True Negative
TP	True Positive
US GAAP	Generally Accepted Accounting Principles
USB	U.S Bancorp
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 Background

Missing values is a common problem found in dataset from any field of research. Liu et al. (2016a) defined missing values as the absence of data values in a dataset in which the data records have the undesirable null values. A data value in a dataset can be missing due to several reasons such as non-response items in the interview and survey, equipment malfunction, human error, and faulty data transmission. The occurrence of missing values in a dataset need to be managed using appropriate methods to estimate the approximate values to replace the missing values. The inability to manage missing values in a dataset could reduce the analysis performance. For example, in predictive modelling, Rahman & Islam (2016), Rubright et al. (2014) and Roth & Switzer (1995) stressed that the occurrence of missing values in a dataset can caused biased result in the prediction model and threaten its prediction accuracy. The same problem occurs in classification algorithms such as neural networks. Liu et al. (2016b) and Zhu et al. (2011) discussed that bias caused by missing values occurring in the training dataset could impact the quality of learned pattern and decrease the classification performance.

Missing values are also associated with data quality and measured by its dimension of data completeness. Data quality is defined as a state in which data are free from defect and 'fit for use' (Lee & Strong, 2003; Levitin & Redman, 1998; Strong et al., 1997; Wang & Strong, 1996; Wang, 1998). As missing values occur in the dataset, the dataset is no longer free from defects. Even worst, it may cause severe problems to the organization that own the data (Haug et al. Strong et al., 1997). For example, the organization required to put more effort to rectify missing values in the customer address as wrong address in product delivery can caused severe impact to the business. Such example showed the increment in the organization's operational cost due to poor data quality i.e. missing values. Furthermore, poor data quality within the organization gave negative influence towards user perception, experience, trust and believability of the specific application usage such as Enterprise Resource Planning (ERP) and Business Intelligent System (BIS). ERP and BIS applications are important for the organization as they strengthen the organization operations and support the decision making process. Hartl & Jacob (2016) and Popovič et al. (2012) discussed the barriers created between specific application usage and user acceptance as data quality decreases in the organization.

As mentioned earlier, the occurrence of missing values in a dataset is measured by the dimension of data completeness. Data completeness is measured as the number of data values that exists against the total number of data values (Liu et al., 2016a; Wechsler & Even, 2012; Batini & Scannapieca, 2006). Data is considered as complete when all necessary values pertaining to the data exist and contain no undesirable null values (Jayawardene et al., 2013; Bovee et al., 2003; Kahn et al., 2002; Wand & Wang, 1996). Previous research in data completeness proposed various methods to solve the missing values problem. These methods can be categorized into two main categories which are case deletion and imputation. The imputation methods comprise two main categories named multiple imputation and single imputation. Single imputation methods can be further classified into three main categories which are model-based methods, machine leaning-based methods and data driven methods.

Cold deck imputation method belongs to data driven methods and is able to produce almost the same imputation accuracy as multiple imputation but with lower computational cost (Garciarena & Santana, 2017). Unlike multiple imputation methods, cold deck imputation method do not require multiple times of imputation process, which can be computationally expensive. Furthermore, model misspecification problem is less likely to occur in cold deck imputation compared to model based imputation. The only problem with cold deck imputation is the chance to find the most suitable value to replace the missing value is small due to the limited number of possible donor. The number of possible donor can be increased by gathering web donor from web data sources.

Cold deck imputation using possible web donors from web data source is proposed in Du & Zhou (2012). The proposed imputation method has been compared to existing missing values imputation methods which are mean imputation, deletion and K-Nearest Neighbor (KNN). The result proved that using web donor to replace missing value produced higher accuracy in the prediction model compared to the existing imputation methods. In the evaluation process, missing values were imputed using the proposed imputation method and the completed dataset is then used to build a prediction model. The prediction accuracy, root mean squared error (RMSE) and F-Measure are then compared to evaluate the performance of each imputation methods.

Even though the proposed method in Du & Zhou (2012) produced the highest accuracy in the prediction model, the implementation of the proposed method is only restricted to a single web data source at one time imputation. In the web data source, there is no assurance that the provided data value is correct as in most cases, the data values from multiple web data sources are conflicting even though they referred to the same data item (Dong et al., 2015). Thus, multiple web data sources should be allowed in cold deck imputation to give more chance in getting the most suitable web donor. Another problem rise when multiple web data sources are used, a method to measure and determine the most suitable web donor is absent in the method proposed by Du & Zhou (2012).

Looking further, the method used to select the most suitable web donor should be able to determine the amount of trust for each available web donor from multiple web data sources and rank the web donors according to their trust score. The web donor with the highest trust score then can be used to impute the missing values in the dataset. The trust score also enables users to evaluate the accuracy and the reliability of each web donor before the imputation taking place. This is important in order to provide believability to users and to give more trust towards the imputed dataset. The limitations mentioned above motivate us to conduct this research. The main goal of this research is to improve data completeness where the number of trusted web donor used to replace the missing values in the dataset is increase compared to the existing cold deck imputation method.

1.2 Problem Statement

The data completeness problem happens due to several factors such as human errors, equipment malfunction, manual data entry process, and incorrect measurement (Deb & Liew, 2016; Tsai & Chang, 2016). In previous research, various missing values imputation methods have been proposed and these imputation methods can be arranged according to their complexity and performance. Methods such as case deletion and mean imputation are less complex and easy to use but perform poorly in terms of bias and imputation accuracy (Cox et al., 2014). However, complex imputation methods such as multiple imputation and machine learning based methods give more imputation accuracy and reduce bias but required high computational resources due to multiple imputation and iteration during the imputation (Nakai & Ke, 2011). Same problems happened in model based imputation which required suitable model specification to allow it imputes accurately (Andridge & Little, 2010). A more promising imputation method is the hot deck imputation which gives the same prediction accuracy as the multiple imputation method but with less computational cost (Garciarena & Santana, 2017). However, as the donor comes from the same dataset, the chance to have a more suitable donor to replace the missing values is limited especially in a small size dataset.

The chances to have a more suitable donor can be increased by looking for the possible donor to replace the missing values from other data sources, particularly, the web data sources. However, the success of this approach is dependent on the level of trust that a user has towards the web donor's value and the web data sources itself (Wang et al., 2017). Replacing missing values with untrusted data will not just increase the risk of wrong decision and false analysis, but also ruin the organizational operation in the long run.

Web data sources contain large amount of data that can be used to replace missing value. For example, Yahoo!Financial, and Google Finance stored a huge collection of financial data to replace missing values in financial datasets. However, as each web data source adopted a different data schema, problems such as conceptual inaccuracy and terminological ambiguity limit the ability to make used of these data in missing value imputation. Du & Zhou (2012), adopted an ontology mapping approach to resolve conceptual inaccuracy and terminological ambiguity problems from web data sources and make the matching between identified web donor and missing values during the imputation become possible. However, the approach is only limited to a value from a single web donor for each missing value replacement and ignored the variation of web donor values especially when more than one value are available to replace the missing value. Thus, limits the chances in finding the most suitable value to replace the missing value.

There are various sources of web donor on the web with unknown accuracy and reliability and thus, the web donor values cannot be fully trusted. Therefore, replacing missing values with web donor values may lead to inaccurate imputation (Wang et al., 2017). In fact, web donor from multiple web data sources can have different data values even though it referring to the same data item. Importantly, the approach failed to answer critical questions such as "How much I can trust the imputed data?" and "Which data from which data source is more trusted?" It is known that data from web data sources are usually conflicting with each other (Dong et al., 2015). Thus, answer to the questions raised before is important to increase believability to the analysis derived from the imputed dataset.

Trusted imputed dataset is highly depending on the selection of trusted data to replace missing value. Chu et al. (2015) and Batini & Scannapieca (2006) discussed that trusted data can only be derived from a trusted data source. As example, if data derived from 'Source A' is more trusted than data derived from 'Source B', then replacing missing values with values from 'Source A' will make the imputed dataset more trustworthy compared to replacing the missing values with data from 'Source B'. As the selection of trusted data is important, ranking of trust score between possible web donors from multiple web data sources will further help users to determine the most trusted data. Thus, the selection of trusted web donor can be done before the imputation process.

Trust level for each possible web donor needed to be assessed before imputation process and required metrics to measure the expected characteristics of trust, namely: accuracy and reliability (Dong et al., 2015; Li et al., 2014a; Kitchens et al., 2014; Asmare & McCann, 2014; Li et al., 2012; Batini et al., 2009; Batini & Scannapieca, 2006). Accuracy measures the correctness of web donor's value when compared to their value of reference in the dataset. On the other hand, reliability is a measure that assess the extent of claimed values in web donor's data source that is correct and trusted. As it is impossible to know the accurate value of that missing data, a metric to assess accuracy and reliability based on the available observed data in the dataset is needed (Li et al., 2016; Dong et al., 2015; Asmare & McCann, 2014; Li et al., 2014a; Li et al., 2014b). Web donor which is provided by a web data source with the highest accuracy and reliability score is given the highest trust score and regarded as more trusted to replace the missing value.

Therefore, this research is essential to investigate and propose a new method that answers the following questions:

1. How to measure trust for each possible web donor if more than one web data sources is used in cold deck imputation?

2. How to measure reliability and accuracy for each possible web donor based on the observed data values in the dataset?

3. How to determine the most trusted web donor in cold deck imputation if more than one web data sources is used in order to improve data completeness?

1.3 Research Objectives

The main objective of this research is to improve data completeness in a dataset by imputing the missing values with trusted data values from multiple web data sources. The objective is further described as follows:

1. To propose a new method to measure trust for each web donor in cold deck missing value imputation based on the accuracy and the reliability of the web donor from multiple web data sources with the aim to resolve conflicted web donor values and to determine a trusted web donor.
2. To propose a new cold deck imputation method on improving data completeness by imputing missing values using a trusted web donor from multiple web data sources.

1.4 Scope of the Research

The scope of this research work is defined in the following points:

- The type of data that is considered in this research is structured data, in this case it is limited to numerical data type. The structured dataset used in this research comprises of tables with rows and columns.

- This research focuses on column completeness which measures the availability of each attribute value in the dataset and more related to missing values occurrence. Due to this, schema completeness and population completeness are out of the scope of this research.

- This research works on finding the most trusted values to impute missing value in data completeness dimension. Various expected characteristics influenced trust such as accuracy, reliability, believability, and reputation. Unlike accuracy and reliability, reputation is not inferred directly from the data and depended on user's personal preferences and judgement. In this research, reputation is regarded as the ranking of possible web donors based on their reliability and accuracy scores. Additionally, believability can also be achieved when user expectation is met. For example, if the information of data source reliability and data source accuracy is provided, user can compare his expectation and decide to believe the data source if his expectation is met.
- In the literature study, other expected characteristics that influence trust such as credibility, verifiability, relevancy, objectivity, licensing and provenance have also been found and elaborated in Table 2.6 of Section

2.5. But, only a few literature that associated these characteristics with trust. Furthermore, these characteristics are large topic by itself and in some cases, characteristics such as licensing and provenance are not described in some web data sources. As for that, this research focuses only on accuracy and reliability as important characteristics to describe trust.

- Despite the various categories of imputation methods as discusses in Chapter 2, this research focuses only on cold deck imputation in data driven imputation method category. The proposed trust score measurement method requires a comparison of the claimed values from web data source and the corresponding values in the dataset in order to determine a trusted web donor. This approach helps to reduce the dependencies of the imputation method performance to multiple imputation process and model specification problem as occurred in multiple imputation and model-based imputation methods.

- The nature of dataset that is considered by this research is limited to a dataset where variables with non-missing values that are related to the variable with missing value are available. In which, data values for the variables with non-missing values and the corresponding claimed values from the web data source are compared and used to measure accuracy score, reliability score, and trust score.

1.5 Organization of the Thesis

The first chapter of this thesis is an introductory chapter which discusses the problem statement, objectives and the scope of research. The rest of this thesis is organized as follows:

Chapter 2 reviews the fundamental concepts of data quality, data completeness, data accuracy, data reliability and missing values. It also reviews relevant works proposed by previous researchers in missing values imputation. The missing values imputation methods are classified based on their imputation mechanism, namely: case deletion, multiple imputation and single imputation. The features of these imputation methods are presented in term of their strengths and weaknesses towards new research opportunity. The chapter also illuminates on the notion of trust and its related expected characteristic that are relevant to this research.

Chapter 3 describes the research methodology used in this research which includes discussions on different phases of this research. The performance metrics, experiments setup and the dataset that is used in this research are presented as well.

Chapter 4 presents in detail the proposed trust measurement method based on the accuracy and the reliability of multiple web donor as defined in the first objective of this research. This chapter also presents and discusses the measurement method used to measure reliability and accuracy of each web donor.

Chapter 5 elucidates the new cold deck imputation method with multiple web donor and incorporated the trust measurement method in order to achieve the second objective of this research.

Chapter 6 presents the results of the experiments to evaluate the performance of the proposed cold deck imputation methods and its comparison with existing imputation methods. This chapter also discusses the results with respect to the number of web donors and percentage of missing values in the dataset.

Chapter 7 concludes the research by providing a summary of the contributions and recommendation for future research.

REFERENCES

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*, *67*(4), pp.1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

AghaKouchak, A., Habib, E., & Bárdossy, A. (2010). Modeling Radar Rainfall Estimation Uncertainties: Random Error Model. *Journal of Hydrologic Engineering*, *15*(4), pp.265–274. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000185

Akande, O., Li, F., & Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American Statistician*, *71*(2), pp.162–170. https://doi.org/10.1080/00031305.2016.1277158

Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, *78*(1), pp.40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Asmare, E., & McCann, J. A. (2014). Lightweight Sensing Uncertainty Metric—Incorporating Accuracy and Trust. *IEEE Sensors Journal*, *14*(12), pp.4264–4272. https://doi.org/10.1109/JSEN.2014.2354594

Ayad, S., & Cherfi, S. S. (2012). Domain Knowledge Based Quality for Business Process Models. In *Proceedings of the 17th International Conference on Information Quality (ICIQ-12)* (pp. 70–84). ICIQ. Retrieved from http://mitiq.mit.edu/ICIQ/2012/2012 ICIQ CDproceedings final.pdf

Baillie, C., Edwards, P., & Pignotti, E. (2015). QUAL: A Provenance-Aware Quality Model. *Journal of Data and Information Quality*, *5*(3), pp.1–22. https://doi.org/10.1145/2700413

Baio, G., & Leurent, B. (2016). An Introduction to Handling Missing Data in Health Economic Evaluations. In *Care at the End of Life* (Vol. 39, pp. 73–85). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-28267-1_6

Balachandran, P. V., Young, J., Lookman, T., & Rondinelli, J. M. (2017). Learning from data to design functional materials without inversion symmetry. *Nature Communications*, *8*, 14282. https://doi.org/10.1038/ncomms14282

Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, *31*, pp.150–163. https://doi.org/10.2307/2631512

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), pp.1–52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieca, M. (2006). *Data Quality. Concepts, Methodologies and Techniques*. (M. J. Carey & S. Ceri, Eds.). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, *87*, pp.251–260.

Bhattacherjee, A. (2002). Individual Trust in Online Firms: Scale Development and Initial Test. *Journal of Management Information Systems*, *19*(1), pp.211–241. https://doi.org/10.1080/07421222.2002.11045715

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, *18*(1), pp.51–74. https://doi.org/10.1002/int.10074

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), pp.211–252. Retrieved from https://www.jstor.org/stable/2984418?seq=1

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), pp.5–32. https://doi.org/10.1023/A:1010933404324

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. *Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pp.641–644. Retrieved from http://www.lrec-conf.org/lrec2004/article.php3?id_article=20#disambiguation

Brick, J., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*(3), pp.215–238. https://doi.org/10.1177/096228029600500302

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), pp.219–242. https://doi.org/10.18637/jss.v045.i03

Casey, R. J., Gao, F., Kirschenheiter, M. T., Li, S., & Pandit, S. (2016). Do Compustat Financial Statement Data Articulate? *Journal of Financial Reporting*, *1*(1), pp.37–59. https://doi.org/10.2308/jfir-51329

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, pp.321–357. https://doi.org/10.1613/jair.953

Cheah, Y.-W., & Plale, B. (2014). Provenance Quality Assessment Methodology and Framework. *Journal of Data and Information Quality*, *5*(3), pp.1–20. https://doi.org/10.1145/2665069

Chen, T. T., & Takaishi, T. (2014). Box-Cox transformation of firm size data in statistical analysis. *Journal of Physics: Conference Series*, *490*(1), 012182. https://doi.org/10.1088/1742-6596/490/1/012182

Chen, J. V., Rungruengsamrit, D., Rajkumar, T. M., & Yen, D. C. (2013). Success of electronic commerce Web sites: A comparative study in two countries. *Information & Management*, *50*(6), pp.344–355. https://doi.org/10.1016/j.im.2013.02.007

Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., Papotti, P., Tang, N., & Ye, Y. (2015). KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15* (pp. 1247–1261). New York, New York, USA: ACM Press. https://doi.org/10.1145/2723372.2749431

Çokluk, Ö., & Kayri, M. (2011). The effects of methods of imputation for missing values on the validity and reliability of scales. *Educational Sciences: Theory & Practice*, *11*(1), pp.303–309.

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with Missing Data in Higher Education Research: A Primer and Real-World Example. *The Review of Higher Education*, *37*(3), pp. 377–402. https://doi.org/10.1353/rhe.2014.0026

Cugnata, F., & Salini, S. (2017). Comparison of alternative imputation methods for ordinal data. *Communications in Statistics - Simulation and Computation*, *46*(1), pp.315–330. https://doi.org/10.1080/03610918.2014.963611

Curé, O. (2012). Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies. *Journal of Data and Information Quality*, *4*(1), pp.1–21. https://doi.org/10.1145/2378016.2378019

Datta, S., Misra, D., & Das, S. (2016). A feature weighted penalty based dissimilarity measure for k-nearest neighbor classification with missing features. *Pattern Recognition Letters*, *80*, pp.231–237. https://doi.org/10.1016/j.patrec.2016.06.023

Deb, R., & Liew, A. W.-C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences*, pp.*339*, 274–289. https://doi.org/10.1016/j.ins.2016.01.018

Dellschaft, K., & Staab, S. (2008). Strategies for the Evaluation of Ontology Learning. In P. Buitelaar & P. Cimiano (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (Vol. 167, pp. 253–272). IOS Press. Retrieved from https://s3.amazonaws.com/academia.edu.documents/3253068/A_Methodology_for_Ontology_Learning.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1536117437&Signature=ppPrZN0h4Gcx1FcNygD6H%2FxOylw%3D&response-content-disposition=inline%3B filename%3DA_Methodolo

Delmotte, F., & Borne, P. (1998). Modeling of reliability with possibility theory. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *28*(1), pp.78–88. https://doi.org/10.1109/3468.650324

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), pp.1–38. Retrieved from http://www.jstor.org/stable/2984875

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*, *6*(1), 21689. https://doi.org/10.1038/srep21689

Destercke, S., Buche, P., & Charnomordic, B. (2013). Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), pp.92–105. https://doi.org/10.1109/TKDE.2011.179

Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., … Zhang, W. (2015). Knowledge-based Trust: Estimating the Trusworthiness of Web Sources. *Proceedings of the VLDB Endowment*, *8*(9), pp.938–949. https://doi.org/10.14778/2777598.2777603

Du, J., & Zhou, L. (2012). Improving financial data quality using ontologies. *Decision Support Systems*, *54*(1), pp.76–86. https://doi.org/10.1016/j.dss.2012.04.016

Enders, C. K. (2010). *Applied missing data analysis*. *The Guilford Press*. The Guilford Press.

Fernandez, M., Gomez-Perez, A., & Juristo, N. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In *Proc. AAAI Spring Symposium Series* (pp. 33–40). Menlo Park, California: AAAI Press. Retrieved from http://oa.upm.es/5484/1/METHONTOLOGY_.pdf

Gao, L., Bruenig, M., & Hunter, J. (2013). Semantic-based Detection of Segment Outliers and Unusual Events for Wireless Sensor Networks. In *Proceedings of the 18th International Conference on Information Quality (ICIQ-13)* (pp. 102–119). Retrieved from https://drive.google.com/file/d/0B81NXHLVoIS3T0VsZGlHcW9HUG8/view?usp=sharing

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, *19*(2), pp.263–282. https://doi.org/10.1007/s00521-009-0295-6

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, pp.52–65. https://doi.org/10.1016/j.eswa.2017.07.026

Geisler, S., Weber, S., & Quix, C. (2011). Ontology-Based Data Quality Framework for Data Stream Applications. In *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)* (pp. 145–159). ICIQ. Retrieved from http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 2011/Papers/02_03_ICIQ2011.pdf

Ghapor, A. A., Zubairi, Y. Z., & Imon, A. H. M. R. (2017). Missing Value Estimation Methods for Data in Linear Functional Relationship Model. *Sains Malaysiana*, *46*(02), pp.317–326. https://doi.org/10.17576/jsm-2017-4602-17

Ghasemi, A., & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), pp.486–489. https://doi.org/10.5812/ijem.3505

Gopal, K., Abdul, R. M. F., & Adam, M. B. (2017). Box-cox transformation of monthly Malaysian gold price range. *Malaysian Journal of Mathematical Sciences*, *11*(S2), pp.107–118. Retrieved from http://einspem.upm.edu.my/journal/fullpaper/vol11sapril/9. kathibun.pdf

Govidan, K., & Mohapatra, P. (2012). Trust Computations and Trust Dynamics in Mobile Adhoc Networks: A Survey. *IEEE Communications Surveys & Tutorials*, *14(2)*, pp.279-298. https://doi.org/10.1109/SURV.2011.042711.00083

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), pp.199–220. https://doi.org/http://dx.doi.org/10.1006/knac.1993.1008

Guo, B. H. W., & Goh, Y. M. (2017). Ontology for Design of Active Fall Protection Systems. *Automation in Construction*, *82*, pp.138–153. https://doi.org/10.1016/j.autcon.2017.02.009

Hartl, K., & Jacob, O. (2016). The Role of Data Quality in Business Intelligence – An empirical study in German medium-sized and large companies. In *Proceedings of the International Conference on Information Quality* (*ICIQ) 2016* (p. 4:1-4:10).

Haug, A., Zachariassen, F., & Liempd, D. van. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, *4*(2), pp.168–193. https://doi.org/10.3926/jiem.2011.v4n2.p168-193

He, Y., & Pi, D. C. (2016). Improving KNN method based on reduced relational grade for microarray missing values imputation. *IAENG International Journal of Computer Science*, *43*(3), pp.89–95. Retrieved from http://www.iaeng.org/IJCS/issues_v43/issue_3/IJCS_43_3_11.pdf

Heinrich, B., & Klier, M. (2015). Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. *Decision Support Systems*, *72*, pp.82–96. https://doi.org/10.1016/j.dss.2015.02.009

Huang, Y. M., Hung, C. M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, *7*(4), pp.720–747. https://doi.org/10.1016/j.nonrwa.2005.04.006

Hwang, R.C., Cheng, K. F., & Lee, C.F. (2009). On multiple-class prediction of issuer credit ratings. *Applied Stochastic Models in Business and Industry*, *25*(5), pp.535–550. https://doi.org/10.1002/asmb.735

Jayalakshmi, T., & Santhakumaran, A. (2010). A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. In *2010 International Conference on Data Storage and Data Engineering* (pp. 159–163). IEEE. https://doi.org/10.1109/DSDE.2010.58

Jayawardene, V., Sadiq, S., & Indulska, M. (2013). An Analysis of Data Quality Dimensions. *ITEE Technical Report No. 2013-01*, *01*, 1–32. Retrieved from http://espace.library.uq.edu.au/view/UQ:312314/n2013-01_TechnicalReport_Jayawardene.pdf

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), pp.105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks. *Journal of Business Finance & Accounting*, *44*(1–2), pp.3–34. https://doi.org/10.1111/jbfa.12218

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), pp.2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, *10*(3), pp.215–236. https://doi.org/10.1016/0925-2312(95)00039-9

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM*, *45*(4ve), pp.184–192. https://doi.org/10.1145/505999.506007

Kaur, N., & Aggarwal, H. (2017). Evaluation of Information Retrieval Based Ontology Development Editors for Semantic Web. *International Journal of Modern Education and Computer Science*, *9*(7), pp.63–73. https://doi.org/10.5815/ijmecs.2017.07.07

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1). https://doi.org/10.1145/1629175.1629210

Kim, J. W., & Lim, J.-H. (2011). IT investments disclosure, information quality, and factors influencing managers' choices. *Information & Management*, *48*(2–3), pp.114–123. https://doi.org/10.1016/j.im.2011.03.001

Kitchens, B., Harle, C. a., & Li, S. (2014). Quality of health-related online search results. *Decision Support Systems*, *57*(1), pp.454–462. https://doi.org/10.1016/j.dss.2012.10.050

Kombo, A. Y., Mwambi, H., & Molenberghs, G. (2017). Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, *44*(2), pp.270–287. https://doi.org/10.1080/02664763.2016.1168370

Kon, H. B., Madnick, S. E., & Siegel, M. D. (1995). Good Answers from Bad Data : a Data Management Strategy. *Sloan School of Management, Massachusetts Institute of Technology* (No. 3868–95). https://dspace.mit.edu/handle/1721.1/2601

Kulikowski, J. L. (2014). Data Quality Assessment: Problems and Methods. *International Journal of Organizational and Collective Intelligence*, *4*(1), pp.24–36. https://doi.org/10.4018/ijoci.2014010102

Lakshminarayan, K., Harp, S. a., Goldman, R., & Samad, T. (1996). Imputation of Missing Data Using Machine Learning Techniques. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* , pp.140–145.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), pp.259–275. https://doi.org/10.1023/A:1008334909089

Lancaster, H. O., & Seneta, E. (2005). Chi-Square Distribution. In *Encyclopedia of Biostatistics* (Vol. 2). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011815.b2a15018

Lang, K. M., & Little, T. D. (2016). Principled Missing Data Treatments. *Prevention Science*, pp.1–11. https://doi.org/10.1007/s11121-016-0644-5

Lausen, B., Böhmer, M., & Krolak-Schwerdt, S. (2015). Donor Limited Hot Deck Imputation: A Constrained Optimization Problem. *Studies in Classification, Data Analysis, and Knowledge Organization*, *48*, pp.1–22. https://doi.org/10.1007/978-3-662-44983-7

Lee, Y. W., & Strong, D. M. (2003). Knowing-Why About Data Processes and Data Quality. *Journal of Management Information Systems*, *20*(3), pp.13–39. Retrieved from http://web.mit.edu/tdqm/www/tdqmpub/Knowing-why.pdf

Lee, Z. W. Y., Cheung, C. M. K., & Chan, T. K. H. (2015). Massively multiplayer online game addiction: Instrument development and validation. *Information & Management*, *52*(4), pp.413–430. https://doi.org/10.1016/j.im.2015.01.006

Leedy, P. D. and Ormrod, J. E. (2016). Practical Research: Planning and Design, 11th edition. Boston:Pearson Education

Levitin, A. V., & Redman, T. C. (1998). Data as a resource: properties, implications, and prescriptions. *Sloan Management Review*, *40*, pp.89–101. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=2491177

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., … Han, J. (2014a). A Confidence-Aware Approach for Truth Discovery on LongTail Data. *Proceedings of the VLDB Endowment*, *8*(4), pp.425–436. https://doi.org/10.14778/2735496.2735505

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014b). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14* (pp. 1187–1198). New York, USA: ACM Press. https://doi.org/10.1145/2588555.2610509

Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2012). Truth Finding on the Deep Web: Is the Problem Solved. *Proceedings of the VLDB Endowment*, *6*(2), pp.97–108. https://doi.org/10.14778/2535568.2448943

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., … Han, J. (2016). A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), pp.1–16. https://doi.org/10.1145/2897350.2897352

Little, R. J. A. (1988a). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404). https://doi.org/10.2307/2290157

Little, R. J. A. (1988b). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, *6*(3). https://doi.org/10.2307/1391881

Little, R. J. A., & Rubin, D. B. (1987). Statistical Analysis with Missing Data. *Wiley Series in Probability and Mathematical Statistics* (Vol. 2). Retrieved from http://www.gbv.de/dms/ilmenau/toc/33682193X.PDF

Liu, Y., Li, J., & Zou, Z. (2016a). Determining the Real Data Completeness of a Relational Dataset. *Journal of Computer Science and Technology*, *31*(4), pp.720–740. https://doi.org/10.1007/s11390-016-1659-x

Liu, Z. G., Pan, Q., Dezert, J., & Martin, A. (2016b). Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recognition*, *52*, pp.85–95. https://doi.org/10.1016/j.patcog.2015.10.001

Love, T. (2000) Theoretical Perspectives, Design Research and the PhD Thesis. In *Doctoral Education in Design*, Foundations for the Future, Staffordsihre University Press.

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems* (Vol. 32). https://doi.org/10.1007/s10115-011-0424-2

Martin, N., Poulovassilis, A., & Wang, J. (2014). A Methodology and Architecture Embedding Quality Assessment in Data Integration. *Journal of Data and Information Quality*, *4*(4). https://doi.org/10.1145/2567663

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), pp.709–734. Retrieved from http://www.jstor.org/stable/258792

Mead, R. F., & Kennett, S. R. (2011). Assessment of Information Quality Used in Advice to the Royal Australian Navy. In *Proceedings of the 16th International Conference on Information Quality (ICIQ-11)* (pp. 463–476). MIT Information Quality (MITIQ) Program. Retrieved from http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 2011/Papers/07_02_ICIQ2011.pdf

Moepya, S. O., Akhoury, S. S., Nelwamondo, F. V., & Twala, B. (2016). The role of imputation in detecting fraudulent financial reporting. *International Journal of Innovative Computing, Information and Control*, *12*(1), pp.333–356.

Mohamed, A., Kadir, N. A. N. A., Yap, M.-L., Rahman, S. A., & Arshad, N. H. (2009). Data completeness analysis in the Malaysian Educational Management Information System. *International Journal of Education and Development Using Information and Communication Technology (IJEDICT)*, *5*(2), pp.106–122. Retrieved from http://ijedict.dec.uwi.edu/viewarticle.php?id=477

Mouzhi Ge, & Helfert, M. (2007). A review of information quality research - Develop a research agenda. In *International Conference on Information Quality (ICIQ) 2007* (pp. 76–91). MIT Information Quality (MITIQ) Program. Retrieved from http://digital-library.theiet.org/content/conferences/10.1049/cp_20070800

Nakai, M., & Ke, W. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. *International Journal of Mathematics Analysis*, *5*(1), pp.1–13. Retrieved from http://www.m-hikari.com/ijma/ijma-2011/ijma-1-4-2011/keweimingIJMA1-4-2011.pdf

Paul, C., Mason, W. M., McCaffrey, D., & Fox, S. A. (2008). A cautionary case study of approaches to the treatment of missing data. *Statistical Methods and Applications*, *17*(3), pp.351–372. https://doi.org/10.1007/s10260-007-0090-4

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., … Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), pp.961–970. https://doi.org/10.1111/2041-210X.12232

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4). https://doi.org/10.1145/505248.506010

Popovič, A., Hackney, R., Coelho, P. S., & Jaklič, J. (2012). Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. *Decision Support Systems*, *54*(1), pp.729–739. https://doi.org/10.1016/j.dss.2012.08.017

Rahman, M. G., & Islam, M. Z. (2016). Missing value imputation using a fuzzy clustering-based EM approach. *Knowledge and Information Systems*, *46*(2), pp.389–422. https://doi.org/10.1007/s10115-015-0822-y

Redman, T. C. (1996). *Data quality for the information age*. Boston: Artech House.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), pp.79–82. https://doi.org/10.1145/269012.269025

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). Examining solutions to missing data in longitudinal nursing research. *Journal for Specialists in Pediatric Nursing*, *22*(2), pp.1–12. https://doi.org/10.1111/jspn.12179

Roth, P. L., & Switzer, F. S. (1995). A Monte Carlo analysis of missing data techniques in a HRM setting. *Journal of Management*, *21*(5), pp.1003–1023. https://doi.org/10.1177/014920639502100511

Royston, P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Applied Statistics*, *44*(4. https://doi.org/10.2307/2986146

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. (D. B. Rubin, Ed.), *John Wiley & Sons*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rubright, J. D., Nandakumar, R., & Glutting, J. J. (2014). A Simulation Study of Missing Data with Multiple Missing X's. *Practical Assessment, Research & Evaluation*, *19*(10). Retrieved from http://pareonline.net/getvn.asp?v=19&n=10

Salas-Molina, F., Martin, F. J., Rodríguez-Aguilar, J. A., Serrà, J., & Arcos, J. L. (2017). Empowering cash managers to achieve cost savings by improving predictive accuracy. *International Journal of Forecasting*, *33*(2), pp.403–415. https://doi.org/10.1016/j.ijforecast.2016.11.002

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), pp.147–177. https://doi.org/10.1037//1082-989X.7.2.147

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. *Academy of Management Review*, *32*(2), pp.344–354. https://doi.org/10.5465/AMR.2007.24348410

Sintonen, S., Tarkiainen, A., Cadogan, J. W., Kuivalainen, O., Lee, N., & Sundqvist, S. (2016). Cross-country cross-survey design in international marketing research. *International Marketing Review*, *33*(3), pp.454–482. https://doi.org/10.1108/IMR-11-2014-0348

St-Maurice, J., & Burns, C. (2017). An Exploratory Case Study to Understand Primary Care Users and Their Data Quality Tradeoffs. *Journal of Data and Information Quality*, *8*(3–4), pp.1–24. https://doi.org/10.1145/3058750

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), pp.112–118. https://doi.org/10.1093/bioinformatics/btr597

Stockdale, M., & Royal, K. (2016). Missing data as a validity threat for medical and healthcare education research: problems and solutions. *International Journal of Healthcare*, *2*(2), pp.67–72. https://doi.org/10.5430/ijh.v2n2p67

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data Quality in Context. *Communications of the ACM*, *40*(5), pp.103–110. https://doi.org/10.1145/253769.253804

Taft, L. M., Evans, R. S., Shyu, C. R., Egger, M. J., Chawla, N., Mitchell, J. A., … Varner, M. (2009). Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *Journal of Biomedical Informatics*, *42*(2), pp.356–364. https://doi.org/10.1016/j.jbi.2008.09.001

Torgo, L. (2010). *Data Mining with R*. Chapman and Hall/CRC. https://doi.org/10.1201/b10328

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), pp.520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Tsai, C.-F., & Chang, F.-Y. (2016). Combining instance selection for better missing value imputation. *Journal of Systems and Software*, *122*, pp.63–71. https://doi.org/10.1016/j.jss.2016.08.093

Tsai, C.-F., Li, M.-L., & Lin, W.-C. (2018). A class center based approach for missing value imputation. *Knowledge-Based Systems*, *151*, pp.124–135. https://doi.org/10.1016/j.knosys.2018.03.026

Uschold, M., & Gruninger, M. (1996). Ontologies : Principles , Methods and Applications. *Knowledge Engineering Review*, *11*(2), pp.93–136. https://doi.org/10.1.1.111.5903

Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, *112*, pp.111–124. https://doi.org/10.1016/j.dss.2018.06.011

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0178731

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), pp.61–90. https://doi.org/10.1111/stan.12023

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., … Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. https://doi.org/10.1136/bmjopen-2013-002847

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), pp.86–95. https://doi.org/10.1145/240455.240479

Wang, HZ., Qi, ZX., Shi, RX., Li, JZ., Gao, H. (2017). COSSET+ : Crowdsourced Missing Value Imputation Optimized by Knowledge Base. Journal of Computer Science and Technology, 32(5), pp.845-857. /doi.org/10.1007/s11390-017-1768-1

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), pp.58–65. https://doi.org/10.1145/269012.269022

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), pp.5–33. https://doi.org/10.1080/07421222.1996.11518099

Wang, S., & Wang, H. (2007). Mining Data Quality In Completeness. In *ICIQ* (pp. 295–300). Retrieved from http://mitiq.mit.edu/iciq/PDF/MINING DATA QUALITY IN COMPLETENESS.pdf

Wechsler, A., & Even, A. (2012). Assessing Accuracy Degradation Over Time With a Markov-Chain Model. In *Proceedings of the 17th International Conference on Information Quality (ICIQ-12)* (pp. 99–110).

Wilson, D. R., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, *6*, pp.1–34. https://doi.org/10.1613/jair.346

Xu, F., Zhang, H. Y., & Huang, W. (2014). Do CDOs Matter? Assessing the Value of CDO Presence in Firm Performance. In *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)* (pp. 164–170).

Yang, Y., Xu, Z., & Song, D. (2016). Missing value imputation for microRNA expression data by using a GO-based similarity measure. BMC Bioinformatics, 17(Suppl 1):10, pp.109-116. DOI: 10.1186/s12859-015-0853-0

Yajuan, W., Simon, M., Bonde, P., Harris, B. U., Teuteberg, J. J., Kormos, R. L., & Antaki, J. F. (2012). Prognosis of Right Ventricular Failure in Patients With Left Ventricular Assist Device Based on Decision Tree With SMOTE. *IEEE Transactions on Information Technology in Biomedicine*, *16*(3), pp.383–390. https://doi.org/10.1109/TITB.2012.2187458

Yakout, M., Atallah, M. J., & Elmagarmid, A. (2012). Efficient and Practical Approach for Private Record Linkage. *Journal of Data and Information Quality*, *3*(3), pp.1–28. https://doi.org/10.1145/2287714.2287715

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), pp.2141–2155. https://doi.org/10.1080/00949655.2010.520163

YÜCEL, R. M. (2015). R mlmmm Package: Fitting Multivariate Linear Mixed Effects Models with Missing Values. *Turkiye Klinikleri Journal of Biostatistics*, *7*(1), pp.11–24. https://doi.org/10.5336/biostatic.2014-42149

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), pp.63–93. https://doi.org/10.3233/SW-150175

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), pp.2541–2552. https://doi.org/10.1016/j.jss.2012.05.073

Zhou, M., He, Y., Yu, M., & Hsu, C. (2017). A nonparametric multiple imputation approach for missing categorical data. *BMC Medical Research Methodology*, *17*(1). https://doi.org/10.1186/s12874-017-0360-2

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2011). Missing Value Estimation for Mixed-Attribute Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, *23*(1), pp.110–121. https://doi.org/10.1109/TKDE.2010.99

BIODATA OF STUDENT

Mohd Izham bin Mohd Jaya received the Master of Science (Computer Science) from Universiti Sains Malaysia in 2007. He also holds the Bachelor of Computer Science from Universiti Sains Malaysia in 2005.

In September 2014, he enrolled as a full-time student at Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, pursuing a PhD degree in Database system. His current research interest include data quality and missing values imputation.

LIST OF PUBLICATIONS

Mohd Izham Mohd Jaya, Fatimah Sidi, Iskandar Ishak, Lilly Suriani Affendey, Marzanah A. Jabar (2017). A Review of Data Quality Research in Achieving High Data Quality within Organization. Journal of Theoretical & Applied Information Technology, vol. 95 issue 12, pp.2647-2657.

Mohd Izham Mohd Jaya, Fatimah Sidi, Sharmila Mat Yusof, Iskandar Ishak, Lilly Suriani Affendey, Marzanah A. Jabar (2017). Replacing Missing Values using Trustworthy Data Values from Web Data Sources. Journal of Physics: Conference Series, Volume 892.