An Assessment of Open Data Sets Completeness

Abdulrazzak Ali[1]
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh
Melaka, Malaysia

Siti A. Asmai[3]
Optimization, Modeling, Analysis,
Simulation and Scheduling (OptiMASS) Research Group
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh, Melaka, Malaysia

Nurul A. Emran[2]
Computational Intelligence Technologies (CIT) Research Group
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh
Melaka, Malaysia

Amelia R. Ismail[4]
Department of Computer Science, Kulliyyah of ICT
International Islamic University Malaysia
P.O. Box 10, 50728
Kuala Lumpur

***Abstract*—The rapid growth of open data sources is driven by free-of-charge contents and ease of accessibility. While it is convenient for public data consumers to use data sets extracted from open data sources, the decision to use these data sets should be based on data sets' quality. Several data quality dimensions such as completeness, accuracy, and timeliness are common requirements to make data fit for use. More importantly, in many cases, high-quality data sets are desirable in ensuring reliable outcomes of reports and analytics. Even though many open data sources provide data quality guidelines, the responsibility to ensure data of high quality requires commitment from data contributors. In this paper, an initial investigation on the quality of open data sets in terms of completeness dimension was conducted. In particular, the results of the missing values in 20 open data sets measurement were extracted from the open data sources. The analysis covered all the missing values representations which are not limited to nulls or blank spaces. The results exhibited a range of missing values ratios that indicated the level of the data sets completeness. The limited coverage of this analysis does not hinder understanding of the current level of data completeness of open data sets. The findings may motivate open data providers to design initiatives that will empower data quality policy and guidelines for data contributors. In addition, this analysis may assist public data users to decide on the acceptability of open data sets by applying the simple methods proposed in this paper or performing data cleaning actions to improve the completeness of the data sets concerned.**

Keywords*—Data completeness; missing values; open data; open data sources; data collection

I. Introduction

Data completeness is an essential dimension in data quality like accuracy and timeliness. Data completeness plays a major role to guarantee the completeness of query answers [1], [2] and ensure reliable analysis [3], [4]. In the context of software quality, completeness is also an important attribute that determines the quality of Software Requirement Specification (SRS) [5]. Several types of data completeness exist [6], [7]. Given a data set with a set of attributes, the most common case of data completeness are as follows: (1) all attributes' values are missing for a record (missing record/tuple), (2) some of the values of the attributes are missing for a record (missing values) [7], [8]. The first case represents the total loss of information where the attributes' values for the whole record are missing. The second case however, represents some level (ratio) of the incompleteness of the attributes of a record. For example, assume that we have a simple data set which is supposed to have ten records of students' information. This dataset consists of 6 attributes, namely, $StudentId, Name, Sex, Level, Class, Grade$ as shown in Table I.

TABLE I. Students Information

No	StudentId	Name	Sex	Level	Class	Grade
1	B11	John	M	5	A	B
2	B12	Mona	F	4	A	A
3	B13	Marta	F	4	C	B
4	B15	Helen	F		B	
5	B16	Mark	M			
6	B17		F		A	D
7	B18		M	4		C
8	B19	Sozan	F	5	B	
9	B20	Ahmed	M	5	B	B

Table I illustrates a set of student records with the completeness problem. All records are uniquely identified by an identification attribute, $StudentId$. A missing record can be represented by the absence of the student's record with id 'B14' from the data set. Records 4^{th} to 8^{th} are examples of records with missing values. In this example, missing values are detected through the blank spaces in the table. Missing values are not necessarily represented by nulls or blank spaces as illustrated in Section II-B. Thus, the effort to detect missing values requires a proper understanding of its representation. The ability to detect missing values is the pre-requisite for any missing values effort. In many data-intensive applications, the recoverability of missing values is the key to avoid failure in query answering [1]. The execution of queries on any data set containing missing values may result in unrealistic and high cost [9], [10], [11].

Missing values recovery becomes more challenging especially in the case where key attributes are surrogate keys (keys that are not real and unnatural like student Id) and the natural candidate keys are missing. Candidate keys are usually useful in recovering the values of other attributes that are missing, based on their functional-dependency property. (The role of functional dependency can also be seen in storage space optimization (refer to [12]).

In the next section, the problem of missing values will be elaborated in terms of the causes and representation aspects. Section III presents the methodology used to conduct the analysis. Section IV consists of the analysis results and discussion. Finally, Section V concludes this paper.

II. Background of Missing Values Problem

Missing data or missing value can be defined as a value that is not stored or not exist in a dataset [13], [14], [15]. It is either referred as blank, unknown or null (in database world). The problem of missing values is fairly common and can occur at various stages of data processing, ranging from data collection to data storage phase [16]. The presence of a small ratio of missing values can greatly affect the results that can be derived from most databases including electronic medical records [17], [18]. Hence, missing values are a crucial problem in many decision-making systems as precise decision depends on the completeness of the information at hand [19], [20], [21].

The impact of missing values can also be seen in large data sets [22]. Many statistical methods have difficulties in dealing with the missing data especially in assigning an arbitrary value to the missing data. The fact that missing values is a common (but yet unsolved) problem has motivated many researchers in improving the existing system ability to work on incomplete data sets. (See an example of clinical support systems in [23].)

In general, handling the problem of missing values need to follow the following steps [17]:

1) Determine the reasons for missing data.
2) Determine the representation of the missing value.
3) Analysis of the percentage of missing values.
4) Determine the proper method for handling missing values.

A. Reasons of Missing Data

Several studies reported the causes of missing values in databases as follows:

- Lack of data constraints: No restrictions are imposed on the user to enter all data. This leaves some fields missing [24].
- Insufficient users' experience: Low user experience in dealing with data entry systems and lack of knowledge in the correct mechanism of data entry may cause the loss of data. For example, users might leave the date and time fields blank if they are not sure of the format.
- Merge multiple data sources of different schema: Data repositories such as data warehouse (DW) data sets are merged from several sources. Database schema differences among the contributing sources will cause data sets originated from the source with lack of attributes to be missing values within the merged data sets [2].
- Respondents' answering behavior: In survey data, missing values are often caused by reasons like respondents refuse to answer the survey or they do not understand the questions in the questionnaires [25], [26].
- Error in data collection tools: Research data is also prone to missing values problem due to an error in data collection tool (such as sensors) or human researcher's fault. So, failure software and hardware are significant examples that cause the problem of missing data [2].

Most statistical programs work to remove the missing values automatically from the original data sets. This approach leads to the lack of sufficient data to complete an analysis and thus may give misleading results [27], [28].

Kalkan (2018) stated that although there is a direct correlation between the rate of missing values and the quality of statistical analysis, there is no acceptable proportion of missing values in the data set for the correct statistical conclusion [29], [26]. However, Schafer (1999) argued that the ratio of 5% or less of missing values is inconsequential [30]. Bennett (2001) stated that if the amount of missing values is greater than 10%, the results of the statistical analysis will be biased [31].

B. Missing Values Representation

Data completeness studies have been conducted since 1970 where missing information in the database community was the crux of the problem. The problem of missing values representation was overcome within the relational tables. Completeness studies on distinguishing the null types are triggered by the desire to ascertain the existence of the completeness problem. If nulls are present in the 'non-existence' case, then the presence is treated as legitimate unlike in the 'unknown' case. In short, while the presence of 'non-existence' nulls shows no completeness problems, the presence of "unknown" nulls shows the contrary [32][21].

The @ symbol [33], ! and 'x', 'y' and 'z' have often represented nulls [34]. ANSI/SPARC interim report listed 14 manifestations of nulls. However, the two common types of null used are the unknown nulls in which the values are missing because of the unknown status, and the non-existence nulls in which the values are missing because the attributes relation are not applicable. For example, if someone is not a vehicle owner in London, the attribute 'vehicle owned' is considered null.

C. Methods for Handling Missing Values

In literature, several ways are adopted to handle missing values. In a customer database of a shopping centre, for example, some customers' data such as the age data might be missing. This situation can be handled in one of the following ways [8]:

1) Ignore records that contain missing values: Records which have the missing values are separated from the analysis [35]. For example, special software for statistical analysis is utilized in the analysis task which will run multiple times and the results are maintained, ignoring the missing

values [36]. In general, ignoring missing values is inefficient unless the record has very little missing values. This method may affect the overall results of the planned analysis if there are many missing values in a large number of records in a dataset.

2) Manual completion of missing values: This method can be considered a waste of time and effort [37], [38]. It may also be impossible to be adopted especially in the case of a large amount of data with a large number of missing values. The data sets may be used only if a small number of values are missing [39].
3) Use a global or uniform constant to replace missing values: In this case, all the missing values in a field can be replaced by a constant and uniform value or a label such as “unknown", but in this case when performing the analysis and data mining, the exploration programs will believe that this common label has a particularly important meaning. The number of relatively large missing values will indicate a poor analysis[40].
4) Use one of the values of central tendency metrics instead of missing values: This method is used to fill the numeric type missing values with the measures of central tendency such as the mean or the arithmetic mean [41]. For example, if we have a customer database of a shopping centre and the missing value is the customer’s age, the mean age is used to replace the missing values in the ‘Age’ box. This method will enhance that value in the database and increase the unwanted repetition of a large number of customers which may affect the results of analysis and exploration.
5) Use the most likely value by predicting the missing values: This complex method is acquired through specialized exploration techniques such as the decision tree, aiming to predict missing values by exploring existing and available data of the results of the analysis [42].

Osborne (2013) pointed that, despite obvious distortion that missing values can cause, the number of researchers that deals explicitly with this problem is limited. In a survey conducted with his students in prestigious journals of the American Psychological Association, 38.89% of the authors reported that some data are missing in the data sets that have been used in their articles. Nevertheless, it is uncertain whether the remaining authors (61%) were failed to report their missing data or they completed their data in their articles [22]. In the case where missing values are recovered, the question of whether they effectively deal with the lost data remains unanswered.

In the context of time-series data, researchers in [43] reported that the level of missing values acceptability varies. Some data sets can contain missing values from 5%-50% while others allow up to 80% percent of missing values. As the level of missing values acceptability is high, the results of the analysis drawn from the data sets is incomplete.

Kim *et al.* (2019) estimated that the missing values in the precipitation data of the Korea Meteorological Agency will be up to 16% from year 2015–2016, and about 19% for weather data in 2017 [44]. This estimation drives the Korean government to plan for data imputation strategy as the missing values can affect power generation prediction performance.

III. Methodology

In order to understand the problem of missing values in open data sets, 20 data sets were extracted from two open data sources: Center for Intelligent Learning and Intelligent system (UCI Machine Learning Repository) and data.gov.uk. UCI provided over 350 databases that were used for the automated learning of the experimental analysis, while data.gov.uk, stores data of the government agencies, public bodies and local authorities in the United Kingdom (UK). The data sets consist of information about the government works, research, applications and services.

The selected data sets cover several domains such as education, healthcare, agriculture, and communities. In this paper, the types of completeness concerning the missing values were analyzed. We measure the ratio of missing values in each data set (on attribute level) and the coverage of affected attributes.

The steps conducted are as follows:

- Download data sets from the open data source: Data sets were downloaded from the UCI and data.gov.uk. Table II shows the details of the selected data sets covered by the analysis. The total number of records for all data sets is around three million records.
- Convert data sets into Excel spreadsheet format: The data sets were originally recorded in several formats, such as a text file (.txt) and Excel file (.Xls, .Csv). The data sets in the text files format (.txt) were converted to Excel format for standardization and processing ease. The data sets are categorized into four categories, namely, Medical, Educational, Security and Miscellaneous.
- Perform missing values detection: In order to detect missing values in the data sets, we refer the representation of missing values presented earlier in Section B. In addition to nulls (or blank cells), symbols “?" and “unknown" are detected for missing values in the data sets under study.
- Measure missing values ratio: To describe the formula used to measure missing values, simple ratio method (refer to [45]) which is usually applied to measure completeness is used. The following are descriptions of the notations:
  Suppose that:

  D is the data set under measure,
  A is the set of attributes in D, where $A = \{a_1, a_2, a_3, \ldots, a_n\}$, where n is the number of attributes,
  R is the set of records in D, where $R = \{r_1, r_2, r_3, \ldots, r_m\}$,
  $|V_1|$ is the number of values that are supposed to be in a_1.
  $|V_1^{'}|$ is the number of missing values in a_1
  As $|V_1| = |R|$, the ratio of missing values (in percentage) for a_1 is calculated as:

$$\frac{|V_1^{'}|}{|V_1|} \times 100 = \frac{|V_1^{'}|}{|R|} \times 100 \qquad (1)$$

  The ratio of missing values for D is calculated as:

$$\frac{\sum_{a=1}^{n} |V_a|}{|R|} \times 100 \qquad (2)$$

TABLE II. DATA SETS UNDER STUDY

No	Type	Dataset Name	Attributes	Instances
1	**Medical**	Arrhythmia	**279**	**452**
2		Diabetes 130-US hospitals	**55**	**101767**
3		Cervical cancer (Risk Factors)	**36**	**858**
4		Details of GPs, GP Practices, Nurses and Pharmacies from Organisation Data Service	**20**	**85280**
5		GPS_Details_Nuers_Pharmaces	**19**	**85280**
6		GP Prescribing Data	**17**	**473116**
7		Sickness Absence Rates in the NHS	**10**	**85839**
8		Numbers of Patients Registered at a GP Practice	**9**	**326328**
9		Recorded Dementia Diagnoses	**7**	**608281**
10	**Educational**	Leeds schools all information	**54**	**264**
11		School Locations	**37**	**1555**
12		Schools List	**14**	**101**
13		Public libraries in England	**7**	**3079**
14	**Security**	Communities and Crime normalized Data Set	**147**	**2215**
15		Communities and Crime	**128**	**1994**
16		Street Level Crime Data	**10**	**68177**
17	**Miscellaneous**	Plants	**70**	**34781**
18		BasicCompanyData-2017-12-01-part1	**55**	**849999**
19		Indicator data	**16**	**33930**
20		Traffic Commissioners goods and public service vehicle operator licence records	**13**	**37097**

TABLE III. THE RESULTS OF MISSING VALUES MEASUREMENT

Type	Dataset Name	Number of Affected Attribute	Affected Attribute (%)	Missing Values (%)
Medical	Arrhythmia	**5/279**	**1.79**	0.32
	Diabetes 130-US hospitals	**7/55**	**12.73**	3.79
	Cervical cancer (Risk Factors)	**26/36**	**72.22**	11.73
	Details of GPs, GP Practices, Nurses and Pharmacies from Organisation Data Service	**11/20**	**55**	14.76
	GPS_Details_Nuers_Pharmaces	**10/19**	**52.63**	14.83
	GP Prescribing Data	**2/17**	**11.76**	0.61
	Sickness Absence Rates in the NHS	**3/10**	**30**	0.39
	Numbers of Patients Registered at a GP Practice	**3/9**	**33.33**	11.18
	Recorded Dementia Diagnoses	**1/7**	**14.29**	0.01
Educational	Leeds schools all information	**38/54**	**70.37**	46.75
	School Locations	**14/37**	**37.84**	16.46
	Schools List	**10/14**	**71.43**	10.04
	Public libraries in England	**6/7**	**85.71**	3.35
Security	Communities and Crime normalized Data Set	**41/147**	**27.89**	13.7
	Communities and Crime	**97/128**	**75.78**	15.36
	Street Level Crime Data	**5/10**	**50**	5.59
Miscellaneous	Plants	**68/70**	**97.14**	86.16
	BasicCompanyData-2017-12-01-part1	**41/55**	**74.55**	51.03
	Indicator data	**2/16**	**12.5**	0.34
	Traffic Commissioners goods and public service vehicle operator licence records	**2/13**	**15.38**	0.03

IV. RESULTS, ANALYSIS AND DISCUSSIONS

As shown in Table III high ratio of missing values (more than 40%) was found in three data sets. The first data set is ‘Plant’, where this dataset is extracted from the USDA plants database. It contains all plants (species and genera) in the database and the states of the USA and Canada where the plants exist. This dataset exhibits the highest ratio of missing values as compared to other data sets which are about 86.16% of missing values. From 70 attributes, 97% of it consists of missing values.

The second data set is ‘BasicCompanyData’ which consists of the first part of basic company data of live companies registered in the UK. The ratio of missing values for this data sets is also high (51.3%) that affects 74.55 of its attributes.

The third data set is ‘Leeds schools all information’, with 46.75% of missing values involving 70.37% of its attributes. Nine data sets exhibit low ratio in missing values (less than 10%) while other data sets show between 11-20%. Fig. 1 illustrates that the completeness ratio (in percentage) for data sets from ‘Medical’ and ‘Security’ category are considerably high (80% and above).

High ratio of missing values in some data sets may due to several reasons. The merging operation that requires data from several sources (with different database schemas) to be integrated is a common cause of missing values. The immediate consequence of this scenario is attributes that do not originally exist in their contributing source will be populated with nulls.

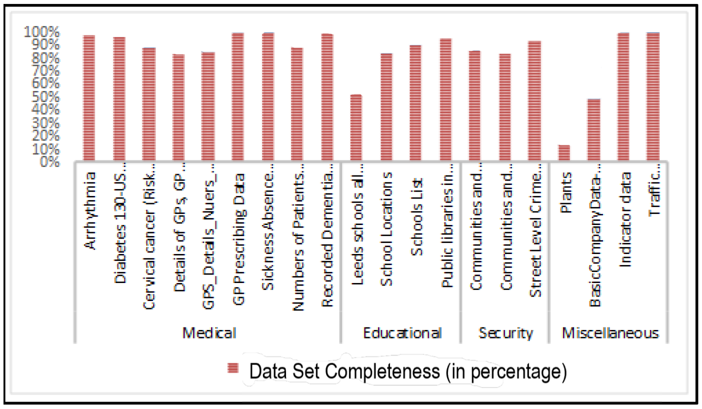

Fig. 1. The ratio of data sets completeness

Secondly, the use of nulls as the default values might be the reason for the missing values in the data sets. For Plants data sets in particular, the missing values might be due to the nature of Plants' species with unknown properties. In this case, missing values are mostly legitimate during data collection.

Another reason of missing values might be caused by lack of enforcement to complete the data sets, which is closely related to data governance and policy.

V. Conclusions

In conclusion, we presented the results of assessing data completeness problem in open data sets. The assessment results involving twenty open data sets show varying missing values ratios that perhaps can be explained by the nature of the data set, data collection policy and enforcement (which is set by the contributing sources).

The findings support our hypothesis on the varying completeness of open data sets that may require further action by data consumers and open data source providers. The findings reported in this paper may motivate further research on dealing with missing values involving open data sets.

Even though most statistical methods will easily calculate the presence of missing data, future work could focus on examining the appropriateness of the methods used and investigate the mechanism that may affect the validity of the results.

Acknowledgment

We would like to thank and acknowledge financial assistance received from the Universiti Teknikal Malaysia Melaka (UTeM) in publishing this work.

References

[1] W. Nutt, S. Razniewski, and G. Vegliach, "Incomplete Databases : Missing Records and Missing Values," in *In International Conference on Database Systems for Advanced Applications*, vol. 7240. Berlin, Heidelberg: Springer, 2012, pp. 298–310.

[2] F.-Z. Hannou, B. Amann, and M.-A. Baazizi, "Explaining Query Answer Completeness and Correctness with Minimal Pattern Covers," *VLDB Endowment*, vol. 12, p. 14, 2019.

[3] N. A. Emran, S. Embury, and P. Missier, "Measuring Population-Based Completeness for Single Nucleotide Polymorphism (SNP) Databases," in *Advanced Approaches to Intelligent Information and Database Systems*, J. Sobecki, V. Boonjing, and S. Chittayasothorn, Eds. Cham: Springer International Publishing, 2014, pp. 173–182.

[4] N. A. Emran, S. Embury, P. Missier, and A. K. Muda, "Measuring Data Completeness for Microbial Genomics Database," in *In Asian Conference on Intelligent Information and Database Systems*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 186–195.

[5] U. Anuar, S. Ahmad, and N. A. Emran, "A Simplified Systematic Literature Review: Improving Software Requirements Specification Quality with Boilerplates," in *9th Malaysian Software Engineering Conference (MySEC)*. IEEE, 2016, pp. 99–105.

[6] N. A. Emran, S. M. Embury, and P. Missier, "Model-driven component generation for families of completeness," in *QDB/MUD*. Auckland, New Zealand: CTIT workshop proceedings series, 2008, pp. 123–132.

[7] N. A. Emran, "Data Completeness Measures," in *In Pattern Analysis, Intelligent Security and the Internet of Things*. Cham: Springer International Publishing, 2015, pp. 117–130.

[8] P. Bhatia, *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press, 2019.

[9] A. A. Alwan, H. Ibrahim, N. I. Udzir, and F. Sidi, "Estimating Missing Values of Skylines in Incomplete Database," in *In Proceedings of the 2th International Conference on Digital Enterprise and Information Systems*, 2013, pp. 220–229.

[10] W. Cheng, X. Jin, J. T. Sun, X. Lin, X. Zhang, and W. Wang, "Searching dimension incomplete databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 725–738, 2014.

[11] J. Cambronero, J. K. Feser, M. J. Smith, and S. Madden, "Query Optimization for Dynamic Imputation," *VLDB Endowment*, vol. 10, no. 11, pp. 1310–1321, 2017.

[12] N. A. Emran, N. Abdullah, and M. N. M. Isa, "Storage Space Optimisation for Green Data Center," *Procedia Engineering*, vol. 53, pp. 483–490, 2013.

[13] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[14] R. L. Vaishnav and K. M. Patel, "Analysis of Various Techniques to Handling Missing Value in Dataset," *International Journal of Innovative and Emerging Research in Engineering*, vol. 2, no. 2, pp. 191–195., 2015.

[15] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, and L. Halawi, *Applying Predictive Analytics: Finding Value in Data*. Springer, 2019.

[16] R. S. Somasundaram and R. Nedunchezhian, "Missing Value Imputation using Iterative Refined Mean Substitution," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 4, pp. 306–313, 2012.

[17] C. Salgado, C. Azevedo, H. Proença, and S. Vieira, "Missing Data," in *Secondary Analysis of Electronic Health Records*. Springer, 2016, ch. Missing Da, pp. 143–162.

[18] F. Leza and N. Emran, "Data Accessibility Model Using QR Code for Lifetime Healthcare Records," *World Applied Sciences Journal (Innovation Challenges in Multidiciplinary Research & Practice)*, vol. 30, pp. 395–402, 2014.

[19] J. W. Graham, "Missing Data Analysis: Making It Work in the Real World," *Annual Review of Psychology*, vol. 60, no. 1, pp. 549–576, 2009.

[20] I. Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg, "Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health," *Future Generation Computer Systems*, vol. 96, pp. 297–308, 2019.

[21] R. J. A. Little and R. B., Donald, *Statistical Analysis with Missing Data*, 2nd ed. Wiley, 2019.

[22] J. W. Osborne, "Six: Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness," in *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do before and after Collecting Your Data*. SAGE, 2013.

[23] M. K. Markey, G. D. Tourassi, M. Margolis, and D. M. DeLong, "Impact of missing data in evaluating artificial neural networks trained on complete data," *Computers in Biology and Medicine*, vol. 36, no. 5, pp. 516–525, 2006.

[24] D. Vucevic and W. Yaddow, *Testing the data warehouse practicum: Assuring data content, data structures and quality*. Trafford Publishing, 2012.

[25] M. N. Norazian Ramli, A. S. Yahaya, N. A. Ramli, N. F. Yusof, and M. M. Abdullah, "Roles of imputation methods for filling the missing values: A review," *Advances in Environmental Biology*, vol. 7, no. 12, pp. 3861–3869, 2013.

[26] Ö. K. Kalkan, Y. Kara, and H. Kelecioğlu, "Evaluating Performance of Missing Data Imputation Methods in IRT Analyses," *International Journal of Assessment Tools in Education*, vol. 5, no. 3, pp. 403–416, 2018.

[27] SPSS, "Missing Data : The Hidden Problem," *Ibm Spss*, pp. 1–8, 2009.

[28] J. P. Hoffmann, *Principles of Data Management and Presentation*. Univ of California Press, 2017.

[29] Y. Dong and C. Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, no. 1, pp. 1–17, 2013.

[30] J. L. Schafer, "Multiple imputation: A primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 3–15, 1999.

[31] D. A. Bennett, "How can I deal with missing data in my study?" *Australian and New Zealand Journal of Public Health*, vol. 25, no. 5, pp. 464–469, 2001.

[32] N. A. Emran, "Definition And Analysis Of Population-Based Data Completeness Measurement," Ph.D. dissertation, University of Manchester, 2011.

[33] E. Codd, "Understanding relations (installment #7)," *FDT-Bulletin of ACM SIGMOD*, vol. 7, pp. 23–28, 1975.

[34] I. Tomasz and W. L. Jr., "Incomplete Information in Relational Databases," *Journal of the ACM (JACM)*, vol. 31, no. 4, pp. 761–791, 1984.

[35] A. Anderson, *Statistics for Big Data For Dummies*. John Wiley & Sons, 2015.

[36] J. Honaker and G. King, "What to Do about Missing Values in Time-Series Cross-Section Data," *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, 2010.

[37] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing value imputation based on data clustering," in *In Transactions on computational science*. Berlin, Heidelberg: Springer, 2008, pp. 128–138.

[38] P. Rahman, C. Hebert, and A. Nandi, "ICARUS: Minimizing Human Effort in Iterative Data Completion," *Proceedings of the VLDB Endowment*, vol. 11, no. 13, pp. 2263–2276, 2018.

[39] Z. M. Kumar and R. Manjula, "Regression model approach to predict missing values in the Excel sheet databases," *International Journal of Computer Science & Engineering Technology (IJCSET)*, vol. 3, no. 4, pp. 130–135, 2012.

[40] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, 2011.

[41] A. D. Banasiewicz, *Marketing Database Analytics: Transforming Data for Competitive Advantage*. Routledge, 2013.

[42] J. J. Faraway, *Linear models with R*. Chapman and Hall/CRC, 2016.

[43] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, "A Review of Missing Values Handling Methods on Time-Series Data," in *International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2016.

[44] T. Kim, W. Ko, and J. Kim, "Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting," *Applied Sciences*, vol. 9, no. 1, pp. 1–18, 2019.

[45] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.