Alternative ways to handle missing values problem: A case study in earthquake dataset

Muhammad Fakhruddin*1, Muhamad Syazali1, and Kenny Candra Pradana2

1) Department of Mathematics, Faculty of Military Mathematics and Natural Sciences, Indonesia Defense University, IPSC Area, Sentul, Bogor 16810, Indonesia
2) Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas islam negeri raden intan lampung, Lampung 35131, Indonesia

*Corresponding author: muhammad.fakhruddin@idu.ac.id

Abstract. Dataset is a basic foundation that is often used in understanding a problem. It provides information for researchers to get solutions to the problem. In the data retrieval process, some errors may occur and cause the data to be incomplete for any reason. It was a problem in how to recover the missing values in a dataset. The first step is to look at the characteristics of the data. In this paper, we proposed three alternative ways to obtain the missing values of the dataset. In this case, we used the earthquake dataset that has special properties. We then present the results to see the performance of the proposed methods. The results show a good agreement for the missing data. This is a preliminary result of our research related to missing data in the earthquake dataset. This study has some limitations such as if the missing values occur in a large enough data block, the methods need to be improved.

Keywords: Incomplete data, missing value, earthquake.

1. Introduction

Dataset is a basic foundation that is often used in understanding a problem and can be used to obtain certain specific or new information [1]. It provides information for researchers to get solutions to a problem. In the data retrieval process, some errors may occur and cause the data to be incomplete for any reason. It was a problem in how to recover the missing values in a dataset. The need for data completeness from the observation results is necessary for further analysis [1]. Some problems of missing data methods such as complete case (CC) analysis and last-observation-carried-forward (LOCF) [2].

Over the last decade, there has been extensive exploration using both simulated and empirical data sets for the effects of missing data. Missing values are frequently encountered while collecting data [3]. Missing data are cases where the data entry is incomplete or missing for any reason on the observed variables in the database [4]. Missing data are ubiquitous and may seriously compromise inferences from randomized clinical trials, especially if missing data are not handled appropriately [5].
is relatively common in all types of research [6]. Meanwhile, the problem of missing data is exacerbated, especially in longitudinal data due to the measurement is repeated sampling [7]. This missing data has potentially led to bias and loss of precision [8]. The missing values and patterns may provide rich information about target labels [9]. Moreover, missing data becomes especially problematic when they are non-randomly distributed [10].

The presence of missing values causes the believed sample size to be smaller than expected which greatly affects the study results [3]. Missing values can result from a huge variety of events [11]. In summary, missing values can be generated randomly or not randomly. Missing data can be caused by the data generating process or due to random loss for other reasons [12]. Many studies do not explicitly report how they handle missing data [13], some implicit methods used in statistical software. Most software applications now implement one or more sophisticated missing data handling routines [14]. A commonly used approach is to replace missing data with imputation methods. However, many other methods are used as well.

Nowadays, there have been tremendous developments in clinical data analysis [15]. Earthquake is the most common type of disaster in Indonesia. Tsunami hazard and risk assessments have become an important issue in tsunami-prone regions [16]. Indonesia is prone to severe earthquakes and volcanic eruptions because Indonesia is located at the confluence of four very active tectonic plates of the world. They are the Eurasian plate, Pacific plate, Indo-Australian plate, and one microplate [17, 18]. Especially in Eastern Indonesia, it is located in a very complex tectonic area [19].

The series of earthquakes do not occur randomly but follows a spatial or geographic pattern that triggers an earthquake [20]. Earthquake prediction is widely known to be a difficult problem [21]. Thus, the earthquake prediction system provides benefits in early warning systems to minimize the risk of losses due to earthquakes. Unfortunately, most systems are designed for moderate and large earthquakes so that low magnitude earthquakes cannot be recorded because they are masked by seismic noise [22]. This is what causes the dataset to become missing data.

There have been many studies discussed missing data, including research by K. A. Hallgren et al. (2016) about compares approaches to modeling binary outcomes with missing data in combined study [23]. I. Schwabe et al. (2016) explain about a new approach to handle missing covariate data in twin research [24]. Subsequent research by P. Khosravi et al. (2015) describes how to handle missing data in decision trees: a probabilistic approach [25]. The next research by T. K. De et al. (2020) explains about handling missing data in randomization tests for single-case experiments: A simulation study [26], and research from J. Zhuang et al. (2017) which is about investigating the missing data problem in the Japan Meteorological Agency catalog of the Kumamoto aftershock sequence [27]. In this study, we proposed alternative ways to handle the missing values problem using the earthquake dataset in Lampung, Indonesia, and relate how to recover the missing values in the dataset.
2. Data Descriptions
In this case, we used an earthquake dataset that has special reference because it was taken from the USGS 1960-2019 catalog specifically for Lampung, Indonesia [28]. We used four types of earthquake datasets, namely Latitude, Longitude, Magnitude, and Depth, see Figures 1 and 2. Latitude is a vertical line that measures the angle between a point and the equator. The entire range of latitudes is measured from a point 0 degrees from the equator to 90 degrees at the poles, 90 degrees to the north pole, and -90 degrees to the south pole [29]. Longitude is a horizontal line measuring the angle between a point and the zero point of the earth. The longitude is measured both east and west from the Moon’s Prime Meridian, namely Greenwich in London, United Kingdom which is an internationally accepted point of longitude 0 degrees or 360 degrees [30].

Magnitude is a measure of the strength of an earthquake which describes the amount of seismic energy emitted by the earthquake source and is the result of seismograph observations. Even if calculated from different places, this quantity would be the same. Regional earthquake magnitude frequency distributions obey a negative exponential law (Gutenberg-Richter) [31]. Earthquake Early Warning (EEW) is strongly influenced by the estimated speed of the earthquake magnitude (M), which can provide a signal for countermeasures against earthquake loss [32, 33]. However, advances in data transmission and communication yield high-quality broadband and strong-motion waveforms in near real-time that are fundamental for the rapid determination of earthquake focal mechanisms [34]. Earthquake depth is the depth from the epicenter that occurred, and this parameter is critical for hazard mitigation and a better understanding of their mechanisms. The shallower epicenter of the earthquake poses a greater risk of damage [35].

The estimation accuracy of the four earthquake parameters is important because this is directly related to disaster mitigation efforts that will be carried out on the impact that occurs. Therefore, missing values becomes a problem that should be recovered. In this study, alternative ways of missing values imputation are explained.

3. Methods
In this section, we propose three options for dealing with missing data of the earthquake dataset. For convenience in writing the steps for completion, let define the data as
\[X = \{x_1, ..., x_n\} \subset \mathbb{R}^n, n \in \mathbb{N},\]  
\[X_j = [x_{1j}, x_{2j}, ..., x_{mj}]^T, m \in \mathbb{N} \text{ and } j = 1, ..., n\]  
\[x_j \text{ and } x_{ij} \text{ are an object datum and } i^{th} \text{ attribute value of } j^{th} \text{ object datum, respectively.}\]

In this case, we temporarily ignore the time and place data for the earthquake, but the sequence of events is still considered.

First, we defined \(X^C = \{x_j \in X | x_j \text{ is a complete datum}\}\), \(X^V = \{x_{ij}, 1 \leq i \leq m, 1 \leq j \leq n | \text{the value of } x_{ij} \text{ is available in } X\}\), and \(X^M = \{x_{ij} = ?, 1 \leq i \leq m, 1 \leq j \leq n | \text{the value of } x_{ij} \text{ is missing from } X\}\).
Figure 1. Plots of the earthquake dataset, namely Latitude, Longitude, Depth (km), and Magnitude (Richter).

For clarity, if we have a data $X \in \mathbb{R}^{m \times n}$ where $m = 3$ and $n = 5$, let say

$$X = \begin{bmatrix}
  2 & 8 & 3 \\
  2 & ? & 4 \\
  7 & 1 & 2 \\
  ? & 2 & 6 \\
\end{bmatrix}$$

Then

$$X^C = \begin{bmatrix}
  2 & 7 & 3 \\
  8 & 1 & 4 \\
  3 & 2 & 6 \\
\end{bmatrix}$$

where $X^M = \{x_{22}, x_{14}, x_{24}\}$ and $X^V = \{x_{11}, x_{21}, x_{31}, x_{12}, x_{32}, x_{13}, x_{33}, x_{23}, x_{34}, x_{15}, x_{25}, x_{35}\}$. Thus, the values in $X^V$ and $X$ are the same, otherwise they are missing values. We then used the information in $X^V$ as a guide in recovering the missing values in the dataset.
3.1. First Option

The idea of the first option is to utilize the information around the missing values with the mean approach, which is defined as follows.

\[
    x_{ij}^M = \begin{cases} 
        \frac{(x^V_{i(j-b)} + x^V_{i(j+r)})}{2}, & 1 < j < n, \\
        \frac{(x^V_{i(j+b)} + x^V_{i(j+r)})}{2}, & j = 1 \\
        \frac{(x^V_{i(j-b)} + x^V_{i(j-r)})}{2}, & \text{others}
    \end{cases}
\]

where \( b, r \geq 1 \) such that the available values are the closest values to the missing value in the same object datum.

This method assumes that missing data that is not on the edge can be recovered by using the data averaging approach before and after the missing values on the same object datum. Meanwhile, if the missing values are on the edge, then an approach to average the two values before or after the missing values is carried out.
3.2. Second Option

In this second option, we assume that the datum object, which is called the magnitude of the earthquake, is essential, and it rarely has errors in recording. Therefore, this option offers data handling using the closest neighbor between the magnitudes between $x_{j}^C$ and $x_{j}^M$. For more details, it can be explained in the following steps.

(i) Find complete datum $x_j$ which has $|x_{ql} - x_{qj}|$, $1 \leq q \leq m$. $l$ and $q$ is the index of incomplete object datum and attribute value for magnitude, respectively.

(ii) Replace $x_{il} = ?, i \neq q$ with $x_{ij}, i \neq q$.

Thus, the key to this option is that close enough magnitude values are expected to have similar attribute values.

3.3. Third Option

In this final option, we use a norm definition as follows.

$$||U|| = \sqrt{\sum_{s=1}^{m} d_s x_s^2}$$

By using the norm definition, the stages in this recovery option are described as follows.

(i) Find complete datum $x_j$ which has the minimum norm value for incomplete datum using the norm definition (4).

(ii) Replace the missing value, $x_{il} = ?, i \neq q$ with $x_{ij}, i \neq q$. $l$ and $q$ is the index of incomplete object datum and attribute value for magnitude, respectively.

This process requires that there is at least one attribute value on an incomplete datum object.

4. Results and Discussion

Based on the three options proposed, we want to see the performance of each method. However, when we look at the strict assumptions taken and consider the relationship between the data and the time and space they occur, the first option is risky, considering that they are independent of each other. Hence, we have sampled data that is quite close to the assumptions in the first option. The results are shown in Table 4. At the top, the selected missing value data is presented, and below it is the result of the missing value imputation.

The results obtained indicate that this method is quite good at handling the missing values problem. This also applies to the first option if the strict assumptions are met. Further testing can still be done by applying various missing data scenarios, such as missing value blocks, and testing them several times. This has not been done in this preliminary study and can be done in future studies.

Based on the preliminary results of this study, we tried to see the geometric shape of the data in $\mathbb{R}^3$ space, see Figure 3. It can be seen that there are two different groups of earthquake datasets used here. Therefore, the next study can be developed from
Table 1. The comparison table results between selected missing data and its recovery values (missing values are assigned a red block).

| Missing Values Data | |
|---------------------|---------------------|
| Latitude            | -62.814 -56.301 -6.345 -5.822 -6.055 |
| Longitude           | 1,052,078 1,037,242 104,058 104,197 105,016 |
| Depth               | 50.44 40.67 57.1 33 64 |
| Magnitude           | 4.9 4.7 5.1 4.6 5.3 |

| Recovery Values Data | |
|---------------------|---------------------|
| First Option        | -62.309 1,037,242.5 -6.358 33 59 |
| Second Option       | -60,099 1,046,152 -64,064 35 46.9 |
| Third Option        | -55,729 1,045,039 -5,929 55.7 35 |

Figure 3. Plot the dataset in the Third Dimension (3D).

missing value imputation methods that are integrated with the clustering method as in [36] and [37], which combine Fuzzy C-Means clustering.

5. Conclusion
The results presented here are preliminary results from our study, which could provide a bridge for further studies. Earthquake datasets have special characteristics, such as being tied to time and space. Based on the results that have been presented, we can provide some conclusions. The sample result shows a good agreement between the original data and data recovery. However, this certainly needs to be further tested to ensure the performance of these methods. The test can be in the form of a large number of missing values or a block of missing values.
The first option only relies on an average of the closest events. Although this can be used as an alternative approach, it should be noted that it has the potential to be a misconception because the nature of the time and place of events are independent of one another. The key to the second option is that close enough magnitude values are expected to have similar attribute values. The last option process requires that there is at least one attribute value in an incomplete datum object. However, the emergence of quite a large number of missing values will reduce the accuracy of the actual information.

References

[1] Pratama I, Pernamasari A E, Ardiyanto I and Indrayani R 2016 A review of missing values handling methods on time-series data 2016 International Conference on Information Technology Systems and Innovation (ICITSI) (IEEE) pp 1–6
[2] Li P and Stuart E A 2019 The American journal of clinical nutrition 109 504–508
[3] Kwak S K and Kim J H 2017 Korean journal of anesthesiology 70 407
[4] Zhu J, Ge Z, Song Z and Gao F 2018 Annual Reviews in Control 46 107–133
[5] Jakobsen J C, Ghud C, Wetterslev J and Winkel P 2017 BMC medical research methodology 17 162
[6] Chhabra G, Vashisht V and Ranjan J 2017 Indian Journal of Science and Technology 10 1–7
[7] Mazen J A, Tong X and Taylor L K 2019 Behavior research methods 51 1321–1335
[8] Hughes R A, Heron J, Sterne J A and Tilling K 2019 International journal of epidemiology 48 1294–1304
[9] Cho Z, Purushotham S, Cho K, Sontag D and Liu Y 2018 Scientific reports 8 1–12
[10] Xi Z, Liu L and Davis C C 2016 Molecular Biology and Evolution 33 838–860
[11] García-Laencina P J, Abreu P H, Abreu M H and Afonoso N 2015 Computers in biology and medicine 59 125–133
[12] Marsman M, Waldorp L and Maris G 2017 Behaviormetrika 44 513–534
[13] Zhang Z 2016 Annals of translational medicine 4
[14] Enders C K 2017 Behaviour research and therapy 98 4–18
[15] Wu X, Khorshidi H A, Aickelin U, Edib Z and Peate M 2019 Health Information Science and Systems 7 19
[16] Muhammad A, Goda K, Alexander N A, Kongko W and Muhari A 2017 Natural Hazards and Earth System Sciences 17 2245
[17] Wang C, Wang X, Xin W, Zhang B, Zhang G and Liu P 2020 Seismological Research Letters 91 733–744
[18] Arlym L, Hermon D, Lanin D, Oktorie O and Putra A 2019 International Journal of Recent Technology and Engineering
[19] Supendi P, Nugraha A, Widiyantoro S, Pesicek J, Thurber C, Abdullah C, Daryono D, Wiyono S, Shiddiqi H and Rosalia S 2020 Geophysical Journal International 221 1845–1855
[20] Shodiq M N, Kusuma D H, Rifqi M G, Barakbah A R and Harsono T 2018 JOIV: International Journal on Informatics Visualization 2 37–43
[21] Kato M 2019 Seismological Research Letters 90 1330–1334
[22] Perol T, Gharbi M and Denolle M 2018 Science Advances 4 e1700578
[23] Hallgren K A, Witkiewitz K, Kranzler H R, Fulk D E, Litten R Z, O’Malley S S, Anton R F and conjunction with the Alcohol Clinical Trials Initiative (ACTIVE) Workgroup I 2016 Alcoholism: Clinical and Experimental Research 40 1548–1557
[24] Schwabe I, Boomsma D I, De Zeeuw E L and Van Den Berg S M 2016 Behavior genetics 46 583–595
[25] Khosravi P, Vergari A, Choi Y, Liang Y and Broeck G V d 2020 arXiv preprint arXiv:2006.16341
[26] De T K, Michiels B, Tanious R and Ongenha P 2020 Behavior Research Methods 1–16
[27] Zhuang J, Ogata Y and Wang T 2017 *Earth, Planets and Space* **69** 36
[28] Abhimanyu H, Raj D K *et al.* 2019 *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)* **53** 182–206
[29] Lowder C, Qiu J and Leamon R 2017 *Solar Physics* **292** 18
[30] Ye H, Guo H, Liu G and Ren Y 2018 *International Journal of Remote Sensing* **39** 5809–5833
[31] Parsons T, Geist E L, Console R and Carluccio R 2018 *Journal of Geophysical Research: Solid Earth* **123** 10–761
[32] Lin J T, Chang W L, Melgar D, Thomas A and Chiu C Y 2019 *Geophysical Journal International* **219** 1148–1162
[33] Noda S and Ellsworth W L 2016 *Geophysical Research Letters* **43** 9053–9060
[34] Tinti E, Scognamiglio L, Michelini A and Cocco M 2016 *Geophysical Research Letters* **43** 10–745
[35] Fang H and van der Hilst R D 2019 *Geophysical Research Letters* **46** 14440–14449
[36] Hathaway R J and Bezdek J C 2001 *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **31** 735–744
[37] Zhang L, Lu W, Liu X, Pedrycz W and Zhong C 2016 *Knowledge-Based Systems* **99** 51–70