Ain Shams Engineering Journal 15 (2024) 102717

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

A comparative analysis of missing data imputation techniques on sedimentation data

Wing Son Loh [a], Lloyd Ling [b], Ren Jie Chin [b,*], Sai Hin Lai [c,d], Kar Kuan Loo [a], Choon Sen Seah [e]

[a] Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[b] Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[c] Department of Civil Engineering, Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[d] UNIMAS Water Centre (UWC), Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[e] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

ARTICLE INFO

Keywords:
Missing data
Imputation techniques
artificial neural network (ANN)
Fine sediment
Sedimentation

ABSTRACT

Sediment data pertains to various hydrological variables with complex sediment hydrodynamics such as sedimentation rates which are often incompletely presented. Thus, the availability of sedimentation data is of utmost necessity for data accessibility. A comparative analysis on the missing fine sediment data imputation performance was made based on four different techniques, namely the k-Nearest Neighbourhood (k-NN), Support Vector Regression (SVR), Multiple Regression (MR), and Artificial Neural Network (ANN), under the single imputation (SI) and multiple imputation (MI) regimes. Across different missing data proportions (10%-50%), the ANN demonstrated optimal results with consistent performance metrics recorded over both SI and MI regimes. For the highest missing data proportion (50%), the ANN presented the best imputation performance with a reported root mean squared error (RMSE) 0.000882, mean absolute error (MAE) 0.000595, coefficient of determination (R^2) 71%, and Kling-Gupta Efficiency (KGE) 72%. The imputation performance ranking is as follows: ANN, SVR, MR, and k-NN.

1. Introduction

1.1. Background and problem Statement

The transport mechanism of sediment particles constitutes a critical aspect of the hydrological cycle, influencing the sustainability of the aquatic ecosystems, balance of water quality and quantity, maintaining the aquatic habitat conditions, and the overall ecosystem preservation. Throughout the recent years, the intensified anthropogenic activities stemming from urbanisation, timber extraction, and agriculture have introduced heavy sediment loads into the locations of dams, rivers and oceans, carrying detrimental impacts to both the environment as well as the economy [1,2]. The motion of fine sediment particles during the settling process in water bodies wields substantial influence towards siltation rates [3]. Additionally, it is common that real data derived from the hydrological studies typically encounters data incompleteness issue such as instrumental failures or budget constraints [4]. Thus, the pivotal role of missing data imputation techniques must not be trivialized in the context of sedimentation data. In fact, the existence of missing data presents an obstacle in deciphering the complex sediment hydrodynamics such as sedimentation rate of fine sediments in water [5]. Furthermore, lacking of a complete series of data and / or using inaccurate data values for analysis would produce misleading results and eventually lead to invalid research studies and decisions being made. In order to properly handle missing data without sacrificing the data reliability and validity, appropriate imputation techniques must be considered. In this regard, different types of imputation techniques were analyzed and compared in this study, ranged from basic methods, to complex and algorithm based modeling techniques. The missing data imputation process was carried out on the missing sedimentation database based on four stipulated missing proportion, 10 %, 20 %, 30 %, 40 %, and 50 %. The proportion of missing data is a dominant factor in the studies of missing data imputation as the availability of the complete oservations from the data set reduces [6]. Past literatures had suggested that a common range between 10 % and 50 % of missing proportion was adapted in missing data related studies [6,]. Based on the rule of thumb, the underlying assertion regarding the missing proportion in this study is that the missing data imputation procedure is not cost effective and is

* Corresponding author.
E-mail address: chinrj@utar.edu.my (R.J. Chin).

https://doi.org/10.1016/j.asej.2024.102717
Received 30 September 2023; Received in revised form 13 February 2024; Accepted 18 February 2024
Available online 11 March 2024

considered to be insignificant whenever the missing proportion is below 5 % [7]. On the contrary, excessive missing data has extremely high potential of introducing bias to the analysis as a result of an imbalanced data set [8]. As a consequence of a biased analysis which was extracted from analyzing the remaining available data from a largely incomplete data set, biased estimated parameters with high error fluctuation will be produced. Besides, the characterization of the data depends heavily on the completeness of the data set and thus there will be a high likelihood that the data that were missing carried significant properties and influential information from the original complete data set. Such results hold utterly deficient statistical power which would hinder the computed statistical analyses [8,9].

1.2. Missing data mechanisms

Over the past decades, it had been widely recognized that issues invited by the presence of missing data is a pervasive concern within a multitude of hydrological databases. Such examples encompasses of missing observations from precipitation data [10], riverflow data [7], rainfall and runoff data [10], water quality index data [8,12], and sediment load data [5]. There were handful of factors that contributes to the presence of missing sedimentation data. For instance the discrepancy in calibration readings [10], ramification of defective sensor components and failure of in-situ measuring instruments [13,14], occurrence of unexpected catastrophic disasters like landslides and flash floods due to excessive downpour of stormwater [15], and error-prone manual data entry processes [16].

While it is vital to develop a reliable and technically sound approach to impute the missing sedimentation data, the missing mechanisms must be understood to ensure the imputation techniques appropriately address the underlying association between the studied variables as well as the probability of the observed data that is missing [17,18]. Generally, the types of missing data mechanisms can be broken down into three principal categories, which are the missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [19].

First and foremost, the MCAR mechanism suggests the scenario of an almost zero or absolute absence of a relationship between the dependency of the variables observed and the likelihood of the unobserved data being missing [15]. In other words, missing data classified under the MCAR mechanism assumes that the original data value is fully independent of the missingness, which is completely random. Cases such as missing recorded data due to the inappropriate use of measuring tools, impaired laboratory equipments, non-responsive data transmissions and overlooked value caused by human related errors are clear examples from the MCAR mechanism [20]. MAR instead interprets the missingness to be related to the observable complete data values, but is unrelated to the unobserved missing data values. Hence, it can be said that MAR claims that non-available missing data as a result of disregarded records follow a random stochastic manner which is predictable from the data pattern discoverable from the observed data [21].

Last but not least, the MNAR missing mechanism states that the missingness of the unobserved data is directly associated with the other missing unobserved data values. This means that the likelihood of the data point being missing with the observed data supplied, has full dependence of the remaining unobserved missing value, and completely independent of the observed complete data set. In this regard, the MNAR mechanism is known to be the most challenging missing mechanism to address [20,22].

By defining a general set of data matrix, $\boldsymbol{D}$ that consists both the observable and missing data variables denoted by $\vec{D}_O$ and $\vec{D}_M$ respectively, the interconnected relationship between the different variables based on the data missingness could be visualized in Fig. 1, where Q represents the cause of the missingness that is unrelated to the $\vec{D}_M$, and R represents the resulting missingness.

More specifically, the likelihood of the sample observation, $\underline{\theta}$, associated with the missing data patterns that are described by the three distinct missing mechanisms can be expressed in accordance with the mathematical equations for MCAR (Eq (1), MAR (Eq (2), and MNAR (Eq (3) [19].

$$\Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)\right) \forall \boldsymbol{D}_{ij} \tag{1}$$

where $\underline{\theta}$ is independent of $\vec{D}_O$ and $\vec{D}_M$.

$$\Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)\right) \forall \boldsymbol{D}_{ij} \tag{2}$$

where $\underline{\theta}$ is independent of $\vec{D}_M$.

$$\Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\vec{D}_O, \vec{D}_M\right)\right) \forall \boldsymbol{D}_{ij} \tag{3}$$

where $\underline{\theta}$ is dependent of $\vec{D}_M$.

1.3. Missing data imputation techniques

In the past decade, there were a large number of studies carried out to perform missing data imputation accross various fields such as applications in financial data [23], biological gene expressions [24], educational production functions [25], ground electromagnetism from the magnetic data acquisition system [20], drill cutting settling rate predictotion [26], and more. Nevertheless, missing data imputation is also actively being researched in the context of missing hydrological databases as mentioned previously. The nature of imputation techniques could be generally grouped into two variations, namely the theoretical based imputation technique, and the empirical based (i.e. function modelling) imputation technique [10,27]. In most cases, the theoretical based approach requires fundamental theories derived from the domain knowledge of the specific field. Such approaches are usually supported by a list of theoretical assumptions which are required to be satisfied.

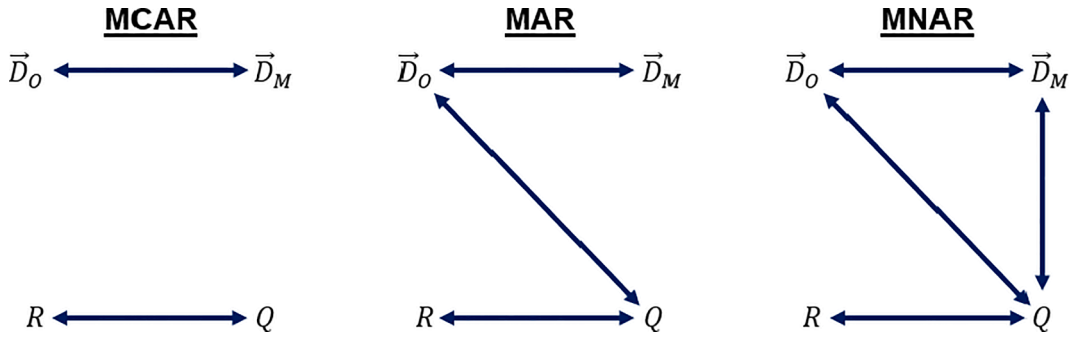

Fig. 1. Missing mechanism relationship illustration.