IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 27, 2018, accepted October 8, 2018, date of publication October 22, 2018, date of current version November 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877269

A Survey on Data Imputation Techniques: Water Distribution System as a Use Case

MUHAMMAD S. OSMAN[1], ADNAN M. ABU-MAHFOUZ[2,3], (Senior Member, IEEE), AND PHILIP R. PAGE[1]

[1]Built Environment, Council for Scientific and Industrial Research, Pretoria 0184, South Africa
[2]Modelling and Digital Science, Council for Scientific and Industrial Research, Pretoria 0184, South Africa
[3]Department of Electrical, Electronic and Computer Engineering, Tshwane University of Technology, Pretoria 0183, South Africa

Corresponding author: Muhammad S. Osman (mosman@csir.co.za)

ABSTRACT The presence of missing data is problematic in most quantitative research studies. Water distribution systems (WDSs) are not immune to this problem. In fact, missing data is an inherent feature of a WDS. There are various techniques and methods to address missing data ranging from simply deleting the data to using complex algorithms to impute missing data. This paper reviews the different imputation options available from traditional methods (such as deletion and single imputation) to more modern and advanced methods (such as multiple imputation, model-based procedures, and machine learning techniques). The concept, application, and qualitative advantages and disadvantages of these methods are discussed. In addition, a novel approach for selecting an applicable technique is presented. The approach is a ''top-down bottom-up'' two-prong approach for the selection of a data analysis and missing data technique. The bottom-up approach facilitates the top-down selection of a suitable technique by analyzing the data and narrowing down the selection options. As a use case, this paper also reviews techniques that are used to impute missing data in WDSs.

INDEX TERMS Data imputation, deletion, machine-learning methods, missing data, model based procedures, multiple imputation, single imputation, water distribution systems.

I. INTRODUCTION

Most scientific and research domains whether they be medical, biological, psychological or climatic science [1]–[5] observe missing data which can be problematic. Data imputation is a key strategy that is used to reconstruct or substitute missing data. Various techniques are mentioned in the literature to impute data to find the most probable answer for missing values in a dataset. These techniques range from traditional methods (such as deletion and single imputation) to more modern and advanced methods such as multiple imputation, model-based procedures and machine learning techniques.

García-Laencina *et al.* [6] analysed and compared various pattern classification techniques to handle missing data. They presented a top-down pattern classification flowchart, which categorised the various missing data approaches into four groups. They emphasized machine-based solutions and highlighted the advantages and disadvantages thereof. Subsequently, Nishanth and Ravi [7] proposed a machine learning technique (probabilistic neural network) which produced efficient results when compared to mean, K-Nearest Neighbour (K-NN), Hot Deck (HD) and a decision tree technique. Gómez-Carracedo *et al.* [8] studied air quality data and found that multiple imputation produced more variable results when compared to single imputation methods. Galán *et al.* [9] used genetic algorithms to impute missing data in the knowledge and skills domain. Wang and Chaib-draa [10] used an online Bayesian framework incorporating Gaussian Process Regression for surface temperature analysis. The authors concluded that their proposed technique outperforms other Gaussian process techniques such as sparse pseudo-input Gaussian process (SPGP) and sparse spectrum Gaussian process (SSGP). Finch [11] compared the performance of three techniques for imputing missing data for surveys and questionnaires. Multiple Imputation for continuous data (MI), multiple imputation for categorical data (MIC) and stochastic regression imputation (SRI) were compared. It was found that MI or SRI produced less bias than MIC and hence was preferred to MIC. Earlier, Blend and Marwala [12] compared an auto-associative

neural network (AANN), a neuro-fuzzy (NF) system and a hybrid AANN/NF system in their analysis of Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS) data. It was found that the AANN outperformed the NF system by an average of approximately 6%, while the hybrid method was approximately 16% more accurate than the standalone AANN or NF systems. However; the hybrid system was 50% less computationally efficient. Dauwels *et al.* [13] presented an innovative tensor-based imputation method based on canonical polyadic (CP) decomposition which they compared to mean imputation, regression imputation and K-NN. Their proposed method was assessed with medical questionnaires and the results showed that the imputation accuracy improved. Tensor based imputation methods are also widely used methods in traffic information systems and road sciences and is well documented in literature [14]–[16]. Tensor decomposition techniques are also used in psychology, chemometrics, signal processing, bioinformatics, neuroscience, web mining and computer vision [17]

This paper presents a ''top-down bottom-up'' two-prong approach for data analysis and missing data technique selection as shown in the flowchart in Section II. The bottom-up approach facilitates the top-down selection of a suitable technique by analyzing the data and narrowing down the options. In addition, this paper also reviews the various imputation methods in both the traditional and modern categories and qualitatively compares them against each other. Furthermore, the underpinning concept, application, advantages and disadvantages are highlighted. Lastly, data imputation in a WDS is discussed as a use case.

II. CATEGORICAL HANDLING OF MISSING DATA

Before categorically classifying data, it is essential to understand the different types of missing data and their mechanisms [4]. Missing data can generally be attributed to one of three missing data mechanisms [4], [18]: Data that is missing completely at random (MCAR), data that is missing at random (MAR) and/or data that is missing not at random (MNAR). Data that are MCAR and MAR are sometimes referred to as ignorable missing data whereas MNAR data is referred to as non-ignorable missing data [19].

- **Missing Completely At Random (MCAR):** For this mechanism there is an independent relationship between the missing value and other variables in the dataset. Typical examples of MCAR include: customer information (such as gender or contact numbers) missing from the database, when a tube containing a blood sample is accidently dropped and breaks [1] or when questionnaires are unintentionally lost. Human error due to manual data entry procedures in water distribution networks, incorrect water reading measurements, instrumentation error, changes in experimental design etc. are some of the possible reasons for data to be deemed MCAR. The direct result is that the data are completely missing i.e. the probability that an observation is not related to any other variable [1]. Statistically, the MCAR mechanism can be expressed as [20]:

$$f\,(M|Y,\phi) = f\,(M\,|\,\phi)\ \textit{for all}\ Y,\phi \tag{1}$$

  where Y *and* M denote a vector of observed data values and a vector of missingness indicators respectively. ϕ is an unknown parameter and the function f denotes the conditional probability distribution.
- **Missing At Random (MAR):** For this mechanism there is a dependent relationship between the missing value and other variables in the dataset but the probability that a value is missing depends on observed values of other variables and not on the other missing values of the target variable. An example of MAR is when the income level of a client is missing but it can be estimated from other variables like the client's profession, experience and qualification. The MAR mechanism can be formally expressed as [19] and [21]:

$$f\,(M|Y,\phi) = f\,(M\,|\,Y_{obs},\phi)\ \textit{for all}\ Y_{mis},\phi \tag{2}$$

  where Y_{obs} and Y_{mis} are the observed and missing components of target variable Y. The unknown parameter ϕ can be estimated by relating Y_{obs} with other explanatory variables.
- **Missing Not At Random (MNAR)**: For this mechanism there is a direct dependant relationship between the values being missing and the nature of the variable. For instance, if citizens of a country opt not participate in a survey, then MNAR occurs. Mathematically, MNAR can be expressed as [20]:

$$f\,(M,Y|\theta,\phi) = f(Y|\theta)f\,(M\,|\,Y,\phi) \tag{3}$$

  where θ is a parameter of the distribution of Y that is estimated from the observed data and ϕ is a parameter that characterises the distribution of the missingness pattern.

In general, missing data can be classified into two groups: traditional data analysis and modern data analysis. Fig. 1 provides a selection diagram for handling missing data.

Fig. 1 describes a two-prong angle for assisting in selecting an appropriate technique to analyse and impute the missing data. This two-prong angle consists of a top-down approach, which classifies the various available techniques into the traditional and modern types. The traditional types are further subdivided into deletion and single imputation techniques. The modern types are subdivided into multiple imputation, model based techniques and machine based learning techniques. Each of these methods is discussed in detail in the next sections. The bottom-up approach simplifies the top-down selection for a suitable technique by narrowing down the options. This is achieved by numerically analyzing the data and determining the governing mechanism pertaining to its missingness (MCAR, MAR or MNAR) and the percentage of missing data. The percentage missing data is case dependant and a detailed analysis considering factors such as logical

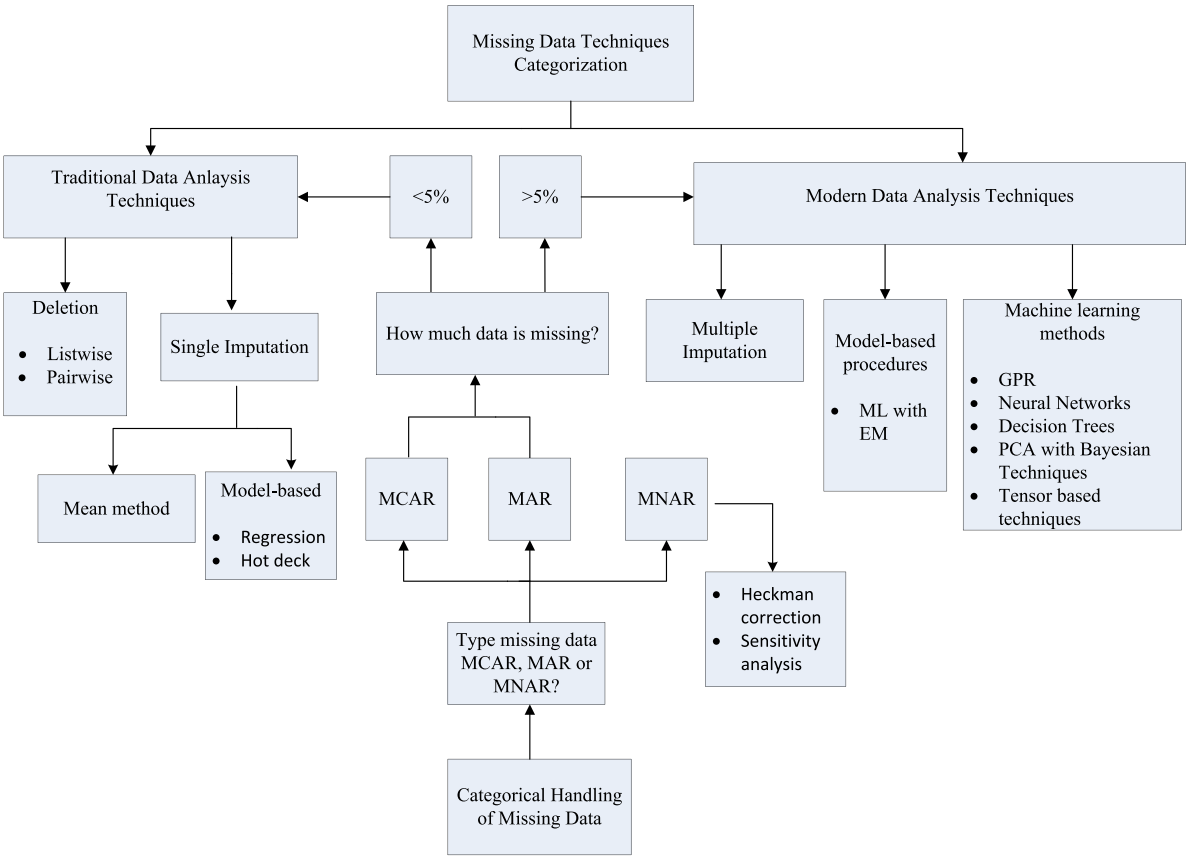

FIGURE 1. Selection diagram to assist in selecting an appropriate handling techniques (derived from [4]).

and structural interdependencies, distribution variability and missing data patterns will be required to determine the correct imputation technique. In some statistical software (such as SPSS), 5% is used as the distinguishing point [22]. For example, if less than 5% of data is missing and the MCAR or MAR mechanisms are applicable, the missing data can potentially be omitted or a suitable single imputation technique can be used to fill in the missing data. On the other hand for the same scenario (i.e. MCAR or MAR missingness) if >5% of the data is missing, advanced techniques can be used to fill the missing data. If the missing data is MNAR and the missingness is due to selection bias, correction factors such as the Heckman correction can be used [23].

III. TRADITIONAL DATA ANALYSIS TECHNIQUES

Traditional data analysis techniques comprise of a number of techniques that can be used to handle missing data. It is a useful tool when a small percentage of the data (<5%) is missing [24]. The most common traditional techniques are deletion (listwise and pairwise) and single imputation. Single imputation is a process that involves analyzing the data together with other variables with the intention of finding the most likely value that can be placed in the data. Single imputation does not involve rigorous computation and provides the dataset with a specific number in place of the missing data. There are several types of single imputation techniques available.

A. DELETION

Deletion techniques are the easiest to execute and are the default choice for many statistical software packages. As the name suggests, the technique simply deletes the cases that contain missing data. There are two common deletion techniques: listwise deletion and pairwise deletion. Listwise deletion, also called complete-case analysis, involves the omission of all the data from an analysis/scenario that contains missing values. Listwise deletion assumes the MCAR mechanism to classify the data. The disadvantage of this technique is that it may introduce serious bias especially when there are a large number of missing values and if the original data set is too small [25]. Pairwise deletion (also referred to as available case analysis), on the other hand, is a more selective method, which determines the extent of missing data on a case by case basis. The cases with high levels of missing data are deleted. This deletion technique tries to minimise data loss and is effective when the overall sample size is small or when the number of missing data observations are large [25].

Similar to listwise deletion, pairwise deletion also assumes the MCAR mechanism.

Despite deletion techniques being a simple and easy-to-use, this traditional method is not a popular choice amongst researchers and has been branded ''amongst the worst methods for practical use'' as quoted by [4].

B. SINGLE IMPUTATION TECHNIQUES

1) MEAN IMPUTATION

This imputation technique involves replacing the missing value with the arithmetic mean ($\bar{x}$) of all the other cases [26]:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

The advantage of this method is that it is fairly simple to use. The disadvantage of this technique is that the results may be distorted due to unevenness in the sample distribution. This technique is used for small data samples and is usually used in surveys. It is usually not used for rigorous multivariate data systems.

Mean /median imputation is a tempting technique but is not recommended by statisticians for the abovementioned reasons. Gómez-Carracedo *et al.* [8] considered a ''modified median'' approach to improve the disadvantages of the traditional mean method.

2) HOT-DECK AND COLD-DECK IMPUTATION

Hot-Deck (HD) imputation is a statistical method that is a popular choice and is the most widely used data imputation method in survey research [19]. Principally, the hot deck technique involves finding a similar or closely matched dataset [6], [27]. The mean of this similar match is then computed and the upper and lower bounds are determined. The advantage of this technique is that it is simple but it can be computationally inefficient. A further shortcoming of this technique is that the missing data estimate is based on a single dataset resulting in underestimates of the standard errors and variability is underestimated [28].

There are numerous different HD configurations such as the Last Observation Carried Forward (LOCF) method. In this method, the missing value is imputed from the last observation in the dataset. This method makes the unrealistic assumption that there is no change at all since the last measured observation [29]. Other hot-deck imputation approaches are the distance function approach and the pattern matching approach. The distance function approach, also called the nearest neighbour approach, imputes the missing value with the smallest squared distance to the case with the missing value. The matching pattern method is more common. In this technique, the sample is separated into separate similar groups and the imputed value for the missing case is randomly drawn from values in the same group [30].

Cold-Deck imputation is principally similar to hot-deck imputation. However, the difference is that the data source must be different from the current data set [6], [31].

3) REGRESSION IMPUTATION

Regression imputation is a statistical tool used to estimate the relationship between an input and an output or between a data point and its associated variable. It is usually presented as a function in the form $y = f(x)$ where y is the output described as a function of an input x. The least-squares fit method is a form of the commonly known linear regression. In its simplest case it is termed simple linear regression and has the form of $y = f(x) = mx + c$, where m is the gradient and c describes the intercept. The function $f(x)$ may also have other forms such as multiple linear, quadratic, cubic as well as non-polynomial.

A drawback of regression imputation is that bias is introduced as the technique fails to account for data variability [4]. In other words, the problem is that the imputed data does not have an error term included in the estimation. There is therefore a fit perfectly along the regression line without any residual variance. The chosen regression model predicts the most likely value of missing data but does not supply any information pertaining to the uncertainty related to the imputed value.

Stochastic regression imputation is a modified version of regression imputation, which attempts to correct the absence of an error term. As described above in regression imputation a regression equation is generated to predict the missing values. However, an error term is generated and added to introduce variance to the missing value.

IV. MODERN DATA ANALYSIS TECHNIQUES

The traditional methods work well for small amounts of missing data. When there is a considerable (>5%) amount of missing data, more sophisticated state of the art techniques and models are required [24].

A. MULTIPLE IMPUTATION (MI)

Multiple imputation seeks to solve the above mentioned problem. Multiple imputation like single imputation is a statistical technique for analyzing incomplete data sets which have some missing values. However; multiple imputation involves three phases [32]: imputation, analysis and pooling. Fig. 2 [33] visually illustrates these steps.

Imputation phase: The missing data values from incomplete data sets are filled in m times ($m = 5$ in Fig. 2). This step results is m complete different data sets. Typically 20 data sets is a good rule of thumb [34]. For this phase single imputation techniques can be used to impute one or more sets. Various algorithms have also been proposed, Baraldi and Enders [4] mentions that the data augmentation procedure is arguably the most widely-used approach. A two-step iterative algorithm is used in this approach. The first step is an imputation step (I-step), which is identical to stochastic regression imputation, which is used to estimate the missing data. Thereafter, the posterior step (P-step) is carried out. In this step Bayesian estimation principles [35], [36] are used to generate new estimates of the means and the covariances. Conceptually,

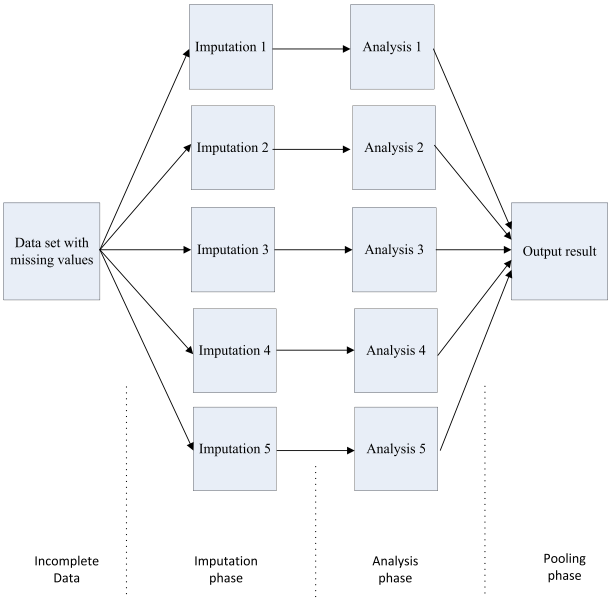

FIGURE 2. A schematic illustration of multiple imputation [33] with m = 5.

the means and the covariances in the P-step differ from the I-step. Using these updated values, the process is repeated several times and multiple copies of the data set are obtained each containing a unique set of estimates of the missing values.

Analysis phase: Each of the m complete data sets are analysed using standard imputation procedures that would have been used in complete data sets [4]. The standard error is also computed during this phase. This step results in m analyses with m sets of standard errors.

Pooling: Combine or integrate the m analyses and standard errors into a final result. The estimates and standard errors of these analyses are usually averaged into a single set of values.

The following equations can be used to calculate the standard errors as alluded to in the pooling phase:

$$W = \frac{\sum (SE_t)^2}{m} \tag{5}$$

$$B = \frac{\sum \left(\hat{\theta}_t - \bar{\theta}\right)^2}{m-1} \tag{6}$$

$$SE = \sqrt{W + B + \frac{B}{m}} \tag{7}$$

In the above equations, SE is for standard error, t refers to a particular imputed dataset, m the total number of imputed datasets. W is the arithmetic average, B the variability of the estimates across the dataset, $\bar{\theta}$ and $\hat{\theta}_t$ are the average parameter estimate and the parameter estimate for a particular dataset respectively.

The greatest disadvantage of multiple imputation is that it is complex in nature. Its complexity not only involves running the analyses, but also combining the results and using the data correctly. On the other hand, multiple imputation introduces the variability in order to find a range of possible responses.

B. MODEL–BASED PROCEDURES

1) EXPECTATION–MAXIMIZATION (EM) ALGORITHM

An EM algorithm is an iterative method used to find maximum likelihood estimates (MLE) of variables in statistical models. It is one of the most used and versatile techniques because there are different EM algorithms for different applications [24]. The EM iteration alternates between performing an expectation (E) step and a maximization (M) step. In the E-step the missing data is firstly estimated from the observed data and the current estimate model parameters. In the M-step, the likelihood function is maximized under the assumption that the missing data are known [6], [37]. Depending on the application, each version of the EM algorithm produces a different solution for the raw data that is entered. The maximum likelihood estimate (MLE) function is mostly coupled to the EM algorithm. Mathematically, the EM algorithm can be expressed as [6]:

$$L(\theta, X, Z) = p(\theta, X, Z) = \sum\nolimits_{j=1}^{J} p(\theta, Z|\theta) \tag{8}$$

where X is the known observed data, Z the missing values, θ the vectors for the unknown parameters, L the likelihood function and p the probability function. The E –step in the EM algorithm can expressed as [6]:

$$Q(\theta|\theta_s) = E\left[L_c(\theta|X, Z)|X, \theta_s\right] \tag{9}$$

and the M-step is shown as [6]:

$$\theta_{s+1} = arg\, max Q(\theta \mid \theta_s) \tag{10}$$

where Q is the iterative procedure used in the E-step and M-step, whereas L_c is defined as the 'complete-data' log likelihood and is defined by the following mathematical function [6]:

$$L_c(\theta|X, Z) = \sum\nolimits_{n=1}^{N} \sum\nolimits_{j=1}^{J} z_{nj} logp\left(x_n|z_n, \theta_j\right) P\left(z_n, \theta_j\right) \tag{11}$$

For the E-step, data is entered individually and if a value is present, the sums, sums of squares, and sums of cross products are augmented. If the value is missing, the current best guess for that value is used instead. The best guess, with all other variables in the model used as predictors, is based on regression-based single imputation [24]. In the M-step, a similar approach is taken for the iteration. The parameters (variances, covariances, and means) are calculated based on the current values of the sums, sums of squares, and sums of cross-products. Depending on the covariance matrix at this iteration, new regression equations are calculated for each variable. These regression equations are then used to re-iterate the best guess for missing values during the E-step of the next iteration. The process continues until the covariance matrix stop changing or until the change is considered negligible and EM is said to have converged [24].

For MLE, the parameter estimates (means, variances, and covariances) from the EM algorithm are excellent. However, a downside to the EM algorithm is that it does not provide standard errors as an automatic part of the process.

C. MACHINE LEARNING METHODS

Machine learning methods are sophisticated and modern procedures that are derived from the study of pattern recognition and computational learning theory [38]. It involves the creation and construction of algorithms that make predictions on data that is missing. Machine learning generally consists of a predictive model that estimates the missing data based on information available in the dataset [39]. Some of the most common machine learning techniques are mentioned below.

1) GAUSSIAN PROCESS REGRESSION

As described in the subsequent section, regression analysis is a statistical tool used to estimate the relationship between an input and an output. Gaussian Process Regression (GPR) on the other hand is a more refined approach than the conventional regression methods described above. It is a powerful technique allowing complex data to be analyzed and described [40], [41].

The GPR mathematical regression process is expressed as covariance matrices K and K_* [42].

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{pmatrix} \quad (12)$$

$$K_* = [k(x_*,x_1)\, k(x_*,x_2) \ldots .k(x_*,x_n)] \quad (13)$$

The key assumption is that the data can be expressed as a multivariate Gaussian distribution [42]:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \quad (14)$$

The point of interest and the variance can be calculated from [42]:

$$K_{**} = k(x_*,x_*) \quad (15)$$

The co-variance function (k) can be determined from the popular squared exponential [40]:

$$k = \sigma_f^2 exp\left[\frac{-(d)^2}{2l^2}\right] \quad (16)$$

where $d = |x - x'|$ is the absolute distance between input samples, l is the length parameter and σ_f is the maximum allowable covariance [40].

According to [41] and [42] the Matèrn function is a better alternative and provides better flexibility:

$$k_m = \frac{h^2\,(2)^{1-\nu}\, K_\nu \sqrt{2\nu d}}{\gamma(\nu)\, w}\left(\frac{\sqrt{2\nu d}}{w}\right)^{\nu} \quad (17)$$

where K_ν is the modified Bessel function, h and w are the height and width respectively, γ the Gamma function defined by [43]. ν is the smoothing parameter and d the absolute distance as defined above.

Depending on the complexity of the missing data, [40] points out two sophisticated covariance functions:

$$k = \sigma_{f1}^2 exp\left[\frac{-(d)^2}{2l_1^2}\right] + \sigma_{f2}^2 exp\left[\frac{-(d)^2}{2l_2^2}\right] + \sigma_n^2\delta \quad (18)$$

$$k = \sigma_{f1}^2 exp\left[\frac{-(d)^2}{2l_1^2}\right] + exp\left\{-2sin^2[\phi\pi d]\right\} + \sigma_n^2\delta \quad (19)$$

with δ defined as the Kronecker delta function and ϕ is the frequency function [40].

2) PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is a statistical data analysis technique that aims to reduce the dimensionality of a data set, which consists of a large number of variables that are interrelated [44], [45]. However, the technique looks to retain the maximum amount of variance present in the data set. This is achieved by using an orthogonal transformation to derive a new set of variables, which are termed the principal components. These are uncorrelated variables from a set of observations of possibly correlated variables. The principal components are ordered so that the first few retain most of the variation present in all of the original variables [45].

Ilin and Raiko [44] have extensively reviewed the use of PCA in the presence of missing values and discussed various approaches to it. They clearly stipulated that the simplicity of PCA is lost the moment missing values appear. Biasness is introduced and the covariance matrix of the data becomes difficult and thus the solution by eigen-decomposition cannot be derived directly [44]. Furthermore, in the PCA algorithm convergence to a unique solution cannot always be assured even for the simplest models. The presence of missing values also creates the potential of overfitting and thus there is a need for some form of regularization. Regularization can be performed better using probabilistic models.

Ilin and Raiko [44] introduced a novel algorithm that includes PCA and variation Bayesian learning. Their approach attempted to address the overfitting problem but introduced uncertainty.

PCA is mostly used as a tool in exploratory data mining analysis and can be done by eigenvalue decomposition of a matrix [46], [47]. It is also closely related to factor analysis. PCA has also been used in canonical correlation analysis (CCA). The orthogonal feature of PCA optimally describes the variance in a single dataset while CCA defines coordinate systems that optimally describe the cross-covariance between two datasets [48].

3) K-NEAREST NEIGHBOUR (KNN) ALGORITHM

The KNN algorithm is amongst the simplest of all machine-learning algorithms [39]. This technique is a popular hot deck method, in which the missing data is substituted with similar data from the nearest neighbour. The nearest, most similar, neighbour ''donors'' are found and classified by minimizing a distance function [49]. The distance function is one of the key aspects of the KNN method. Mathematically, it can be

accurately defined by the heterogeneous Euclidean overlap metric (HEOM) as described by [49]:

$$D(x_a, x_b) = \sqrt{\sum_{i=1}^{n} D_i(x_{ia}, x_{ib})^2} \quad (20)$$

where $D_i(x_{ia}, x_{ib})$ is the distance between the input vectors x_{ia} and x_{ib} on the i-th attribute. Batista and Monard [49] compared the KNN method with two other decision tree machine algorithms and found that KNN was suitable for large missing data sets. They also found it to outperform the other two algorithms. Troyanskaya *et al.* [2] compared KNN with mean imputation in the genetic research domain and found it to be a far better imputation technique. knn does have one disadvantage in that it looks for similar cases only and this can be associated with a high cost [6].

4) DECISION TREES (DT)

Decision tree machine learning is a technique that uses a decision tree as a predictive model to map data and observations (represented in the branches of a tree). The aim is to arrive at conclusions about the target value (represented in the tree leaves) [50], [51]. The decision tree technique is one of the most widely used supervised learning methods.

The major advantage to the use of decision trees is that the data can be visualized and it also allows for the data structure to be easily understood. Typically, the aim is to find the optimal decision tree by minimizing the standard error [52].

There are three well-known decision tree methods namely Iterative Dichotomiser (ID3), C4.5, and CN2 [6], [50], [53]. ID3 is a basic top–down decision tree algorithm that has proved to be a popular and effective method. C4.5 is an extension of ID3 and uses a probabilistic approach to handle missing values. The CN2 is an algorithm that uses a simple imputation method to treat missing data. Every missing value is filled in with its attribute most common known value, before computing the entropy measure. Entropy is defined as the probabilistic measure of uncertainty that exists in a data sample [54].

5) NEURAL NETWORKS (NN)

A neural network (NN) is a learning algorithm that mimics the structure and functional aspects of biological neural networks. It is structured in terms of interconnected groups of ''artificial neurons'', processing information using a connected computational approach [55].

Most neural networks are non-linear modern statistical tools that are usually used to model complex relationships between inputs and outputs and to find patterns in data. They have various variations and derivatives as described below.

Feedforward and Feedback Neural Networks: A feedforward neutral network is a NN that is synonymous with feedforward control loops and systems. Feedforward systems are not error-based, instead they are knowledge-based system where the knowledge of the process is used to infer the probable missing values. It consists of neurons, which are organized in layers. Each neuron in a layer is linked with all the other neurons in the previous layer as shown in Fig. 3 [56]. Data enters the network at the input point and passes through layer by layer until an output is achieved. The Probabilistic Neural Network (PNN) [57] is an example of a feed–forward neural network configuration. PNN is an implementation of a statistical algorithm in which the neurons are organized into a four layer feed forward network consisting of an input layer, a pattern layer, a summation layer and an output layer.

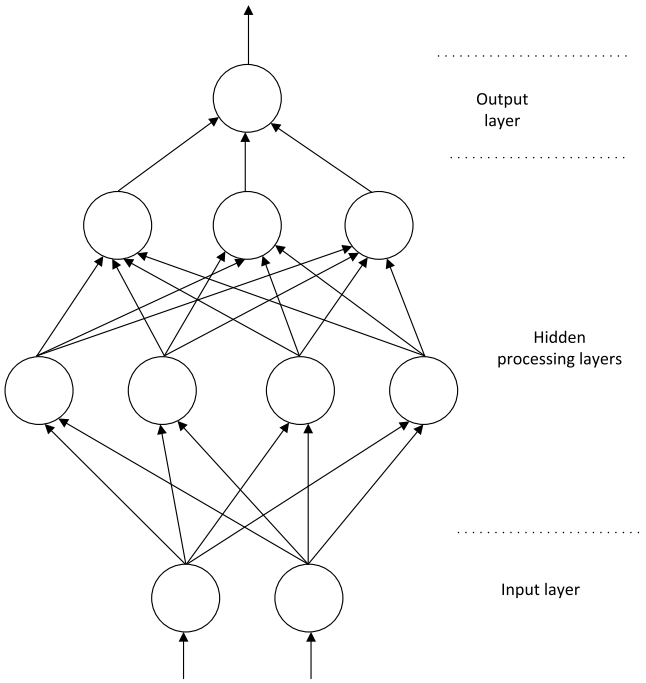

FIGURE 3. Schematic layout of the feedforward network. Deduced from [56].

Another form of neural network is the feedback neural network, an example of which is the recurrent neural network (RNN). It is an architecture that is similar to PNN however, unlike PNN, it operates on feedback connections [6]. The missing values are calculated using single imputation. These values are iteratively updated using feedback connections similar to feedback control systems. Feedback control systems and loops are error based where the aim is to minimise the error. In the last iteration, the sum of a set of recurrent links from the preceding steps (hidden and missing) is iterated.

A third type of NN is an auto-associative neural network (AANN). In this NN type, the network first looks at and learns from complete cases before replicating for the case where missing data is present. Unlike PNN and RNN, there is no summation step; in fact the missing values are replaced by the network outputs [58], [59].

The main disadvantage with NN is that it may require many neurons that can result in multiple combinations.

6) TENSOR BASED TECHNIQUES

Tensors are multi-dimensional arrays that can be used to represent and store multi-dimensional data [60]. A tensor can

TABLE 1. Advantages and disadvantages of some of the traditional data imputation techniques.

Analysis Methods	Brief Description	Advantages	Disadvantages
Traditional	- Deletion and single imputation techniques - It can be used with both MCAR and MAR data	- Simple techniques - Deletion is the default setting in many statistical software programs. - These techniques are quick and easy to apply and understand. - No special mathematical or computational methods are required. - Unbiased results can be produced.	- Deletion technique can decrease the data sample size. - Any potential correlations between variables can be affected. - Use of these techniques can result in reduction of statistical effectiveness and statistical power. - Some techniques can lead to biased results as they do not address variability and no information pertaining to the uncertainty of the imputed value is given. - Discards information. - In hot deck imputation, missing data estimates are based on a single dataset - Standard errors and variability is often underestimated - In some cases, assumptions tend to be unrealistic

TABLE 2. Advantages and disadvantages of some of the modern data imputation techniques.

Analysis Methods	Brief Description	Advantages	Disadvantages
Modern	- Multiple imputation, model based methods and machine learning techniques - It can be used with both MCAR and MAR data	- Some of the techniques are simple to understand and are widely used - Many of the techniques are quick and easy to apply and understand. - There are several algorithms and software packages (e.g. for working with tensors) available. - No simulation is required for some of the techniques - Robust and applicable to many research situations - Some methods (e.g. decision trees and NN) allow for visualization of data - Mostly provides unbiased estimates - Preserves sample size and statistical power. - Standard errors can easily be incorporated.	- PCA is not effective for large amount of missing values. It's an iterative process which can be tedious - In some cases parameters are estimated using single imputation techniques that excludes variability - Some methods (EM and GPR) are a medium to high complex process. The EM algorithm undergoes convergence and as such ad-hoc algorithms may be needed. Also computations can be quite complex and difficult requiring complicated mathematical integrations to determine the required expectations and perform the maximization. - They can be intensive both mathematically and computationally. This makes some of the techniques (e.g. multiple imputation) difficult to program - Tensor based techniques do not always have a finite algorithm that can be used to factorize a tensor into rank-1 tensor components.

be first-order in which case it would be a vector. It can also be second-order which would make it a matrix. Tensors of order three or higher are referred to as higher-order tensors [61]. Mathematically, tensors (χ) can be expressed as:

$$\chi \epsilon \mathbb{R}^{D_1 x D_2 x \cdots x D_N} \tag{21}$$

with N defined as the tensor order and D_i is the size of the ith dimension [62]. Furthermore, the tensor size is defined as:

$$size\,(\chi) = D_1 x D_2 x \cdots x D_N \tag{22}$$

Canonical polyadic (CP) decomposition is commonly used to express a tensor as a minimum length linear combination of rank-1 tensors [61]. Using CP decomposition the tensor χ is factorized and defined as [62]:

$$\chi = \upsilon_1 + \upsilon_2 + \cdots + \upsilon_R = \sum_{r=1}^{R} \upsilon_r \tag{23}$$

where R is a positive integer for $r = 1, \ldots, R$ [61].

The Tucker decomposition technique, which is a form of PCA, is also a method that can be used to decompose a higher-order tensor. There are also many other tensor decomposition techniques and methods that are available in the literature. Some of these are: INDSCAL, PARAFAC2, CANDELINC, DEDICOM, and PARATUCK2, as well as nonnegative variants of all of the above [13], [60], [61], [63], [64].

V. COMPARING THE VARIOUS TECHNIQUES

The traditional and modern techniques discussed in the preceding sections are summarised in Table 1 and Table 2 below. The Tables provide a brief description and some high level advantages and disadvantages are highlighted.

Based on the description in Sections III and IV and Tables 1 and 2, deletion appears to be the simplest technique whereas MI, GPR and EM are more complex and challenging. Furthermore; deletion, mean imputation and hot deck imputation are mathematically easier to understand because in the former missingness is ignored by sub-setting the data and drawing inferences for a sub-population, while in the latter methods missing values are simply replaced by summary statistics. The caveat in using these methods is that they can lead to misleading reduction in population variance and bias. MLE, GPR and model-based MI methods are more complex mathematically because they require specification of the likelihood and/or the posterior distribution from which inferences, namely the predicted values and the associated measures of uncertainty, for the missing data, can be drawn. This means distributional assumptions are needed and in the multivariate data setting, specification of joint distributions would be required.

VI. PERFORMANCE EVALUATION OF DATA IMPUTATION TECHNIQUES

Irrespective of the imputation technique chosen, the imputed value should be as close as possible to the true value. Some of the most common ways of measuring performance is to minimize the root mean-square error (RMSE) or the normalized root mean-square error (NRMSE) [26]. Another approach can be from predictive accuracy (PAC) described by the Pearson correlation [6]. The equations describing these three performance-measuring techniques are given below. The closer the PAC is to 1, the better the imputation technique.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(Y_{obs} - Y_{imp}\right)^2}{N}} \tag{24}$$

$$NRMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(\frac{Y_{obs}-\mu}{\varphi} - \frac{Y_{imp}-\mu}{\varphi}\right)^2}{N}} \tag{25}$$

$$PAC = \frac{\sum_{n=1}^{N}\left(Y_{imp,n} - \bar{Y}_{imp}\right)\left(Y_{obs,n} - \bar{Y}_{obs}\right)}{\sqrt{\sum_{n=1}^{N}\left(Y_{imp,n} - \bar{Y}_{imp}\right)^2\left(Y_{obs,n} - \bar{Y}_{obs}\right)^2}} \tag{26}$$

where Y_{obs} is the observed value, Y_{imp} is the imputed value, $Y_{obs,n}$ is the n-th value of Y_{obs}, $Y_{imp,n}$ is the n-th value of Y_{imp}, $\bar{Y}_{imp}$ and $\bar{Y}_{obs}$ are the mean values of Y_{obs} and Y_{imp} respectively.

Wang *et al.* [65] discussed using the Nash-Sutcliffe efficiency (NSE), the mean prediction intervals (MPI) and the prediction interval coverage probability (PICP) to evaluate the accuracy of data imputation methods. They concluded that the MPI and PICP are more reliable.

VII. DATA IMPUTATION IN A WDS

Water distribution systems (WDSs) are no exception and also suffer from problems surfacing from inherent missing data. WDS and water infrastructure are increasingly being saturated with advanced sensing technologies [66], [67] in an effort to collect a growing volume of data aimed at supporting operational and investment decisions [68]. These sensors monitor system characteristics, i.e. flows, pressures and water quality, for example in pipes. The collected data can be analyzed by various techniques and systems to detect abnormal events. These include leakage detection and localisation [69]. The data can also be used to improve the efficiency of the WDS, such as pressure control systems [70]–[72]. For a WDS the presence of missing sensor data arises from various cases such as mishandling of data and samples, low signal-to-noise ratio, measurement errors due to aging instrumentation, and infrastructure as well as sensor non-response [73], [74].

Correct data acquisition is becoming increasingly important in the analysis of WDSs. This data provides the distribution system's characteristics, i.e. flow rates, pressures, velocities and sometimes water quality, as well as flow regimes (laminar, turbulent or transition) in the piping network. The presence of missing values in WDSs data due to various reasons severely hampers the use of the data. Poor metering, poor data acquisition and incorrect data storage techniques often cause missing data. Sometimes the data becomes faulty or corrupt and is unusable. Furthermore, missing data poses challenges [7] because: (1) a substantial amount of bias can be introduced into the system, (2) handling and analysis of the data can be more strenuous, and (3) reductions in efficiency enters the analysis.

In WDSs, the continuous flow of information pertaining to the system's characteristics and health is vital. Continuous information flow needs to be maintained for the following five reasons: 1) to ensure the effectiveness and efficiency of the WDS, 2) so that necessary interventions can take place upfront of a potential failure, 3) to ensure continuous supply of water to the end user, 4) to minimize water loss through leakage and pipe rupture, and 5) energy loss reduction via optimisation of WDS functionality. Energy loss reduction can be obtained via pressure minimization and pump optimisation.

Similar to other scientific and technological domains, the presence of missing values causes a setback and needs to be adequately addressed. However, for the water and WDS domain there appears to be limited literature available.

A. THE TWO-PHASE MODEL APPROACH

Quevedo *et al.* [75] employed a two-phase approach to estimate replacement values for invalid and missing data. The first phase is a time series model based on daily aggregated flows that validates the data and the second is a 10-minute flow model (based on a pattern derived from historical data) that is used to replace the values of invalid or missing data.

The advantage of this approach is that it can handle large amounts of data. However, the disadvantage is that the model can only be validated and reconstructed using its own data, and for these large amounts of historical data (from flow meter readings) are required to establish consumer demand patterns.

B. THE COMBINED MODEL APPROACH

Barrela *et al.* [76] presented a new method for imputing missing values. Their method comprised of a combination of forecast and backcast values generated by the TBATS and ARIMA models. The TBATS model is an acronym for trigonometric, Box–Cox transformation, ARMA errors, trend, and seasonal components. Barrela *et al.* [76] carried out extensive tests to evaluate the suitability and robustness of their proposed method. Their results are effective and the method is advantageous for offline data reconstruction, when compared to a simple forecast or backcast approach.

An advantage of this method is that it can be used on both online and offline data. However, Barrela *et al.* [76] noted that the method needs to be tested on a larger data set to better understand its adaptability.

C. THE LEAST SQUARES–KALMAN FILTER APPROACH

Bennis *et al.* [77] worked on improving the least squares method for estimating missing data. They mention the advantage of the standard least-squares method is that it is simple

and offers a reasonable solution. The disadvantage is that it can be biased and can overestimate or underestimate the missing value. They also found that peak flow in a WDS is estimated with poor accuracy when the least squares technique is used. Bennis *et al.* [77] used a combination of the Kalman filtering technique and the least squares method to improve the accuracy. Bennis *et al.* [77] identified the critical switch over point between the performance of the two techniques.

D. THE LEAST SQUARES–KALMAN FILTER APPROACH

The real-time dynamic hydraulic model (DHM) is an example of a non-static hydraulic model (see Fig. 4), which can be used for potable water loss reduction. It can also be used for planning purposes and can also be retrofitted to improve existing water networks [78], [79]. Real-time data will be fed to the dynamic model, which in turn will evaluate the network's current conditions and automatically sends control signals to various network components. This adjusts the WDS performance and makes it more efficient. Such adjustment includes continuing calibration of the model, which obeys well-established relations between its sensitivities of various model state parameters [80]. The DHM consists of three major components: dynamic hydraulic model, smart water network and active network management. Data imputation is an important yet critical in-built feature of the dynamic hydraulic model, which addresses the need for correct real data.

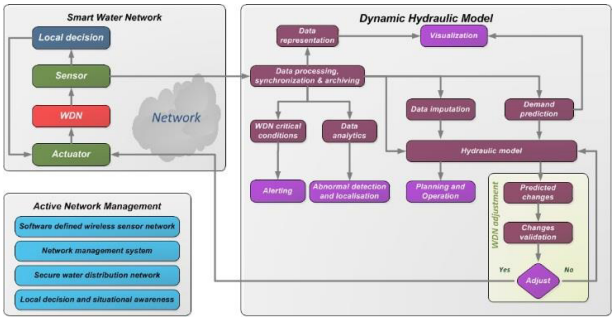

FIGURE 4. Block diagram of the DHM [79].

E. THE VIRTUAL SENSORS CONCEPT

The virtual sensors concept for monitoring water distribution systems is a new approach that is proposed by Goldsmith *et al.* [42] to impute missing data. In this approach, as shown in Fig. 5, permanent and temporary wireless sensors are deployed in the WDS. The permanent sensors provide on-line data on a continuous basis that can be integrated into the hydraulic model component of the DHM. The temporary sensors act as virtual sensors and will be deployed for short periods (7-10 days) at optimum locations in the WDS that are strategically chosen. The virtual sensors will then be removed and their accumulated data will be compared and correlated against the data collected from the permanent sensors [79].

The Virtual Sensor approach utilises the Gaussian Process Regression (described above in section 4.3.1) data imputation technique to correlate the two datasets i.e. the dataset collected from the permanent sensors and the dataset collected from the temporary virtual sensors. This technique collates and combines historical data and spatial correlations between the datasets and predicts missing data using the data provided by the WDSs permanent sensors. In this way, the data inputs to the model are increased without having to increase the sensor count in the WDS.

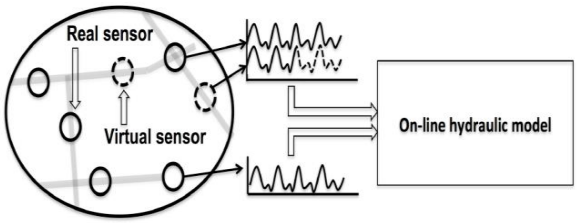

FIGURE 5. The virtual sensor concept [34].

VIII. RESEARCH GAPS AND FUTURE WORK

Most of the imputation techniques available in literature address missing data in most scientific research areas. However, there are some research gaps and future work that have been identified:

- The use of Monte Carlo-Markov chain (MCMC) simulations for error optimisation [12].
- Establishing a distribution pattern before considering random missing value techniques [81].
- Further methodological research into the hot deck method such as improved methods in sourcing donor pools as well as improved methods to assess the trade-off between the quantity and quality of the donor pool [82].
- More work is required to better understand the differences between the MI, MIC and SRI methods. These techniques are fairly similar to each other and should be compared to other statistical methods [11].

The traditional and modern imputation techniques mentioned in the preceding sections of this paper are rarely used in WDSs. Hence, a glaring research gap in this scientific area is evident concerning how these traditional and modern imputation techniques can be used in WDSs; as well as answering the question on how do these techniques compare, align and correlate to the common WDS imputation techniques (section VII).

Goldsmith *et al.* [42] is a rare case where the machine learning GPR technique is used as a data imputation technique for WDSs. Goldsmith *et al.* [42] points out that understanding and quantifying the implications of using multiple virtual sensors in a WDS will need to be investigated in future work. Furthermore, the techniques highlighted in this paper suggest that the model-based EM-MLE option and the more robust multiple imputation method could potentially be better suited for the virtual sensor technique. It is recommended that these be analysed and explored for potential application in WDSs. The combined model technique and the

two-phase model approach employed in WDSs as described in Section V needs to be tested and researched extensively to better understand these techniques' adaptability and to reduce their reliance on their own data for reconstruction.

IX. CONCLUSION

This paper provides a novel approach to assist in narrowing down and selecting an applicable technique. It also reviews most of the data imputation techniques and methods available and highlights their general real-world applications as well as qualitatively compares their respective advantages and disadvantages.

The various imputation techniques can be grouped into two different types. The first type is the easiest way for dealing with missing data: simply delete or impute using single imputation. The main disadvantages of these methods are the loss of information and statistical power. The second type is to use some of the state of the art techniques. The expectation-maximisation algorithm model-based approach coupled with maximum likelihood estimation is the one which stands out. Machine based learning is comprised of a number of options such as neural networks, principal component analysis, Gaussian processes, multiple imputation, etc. Multiple imputation is rated the most robust and flexible machine learning option but is also highly complex when it comes to computational programming.

As a use case, this paper also looks at some of the techniques used in WDSs. These imputation techniques starkly differ from techniques used in other scientific research areas and it appears that no correlation is evident between the two sets of techniques.

In summary it is important to note that most imputation solutions presented in this paper works well in many situations; but they are case specific. A detailed analysis such as a data distribution analysis or a missing data pattern analysis will be required to determine the correct imputation technique to enhance the accuracy.

REFERENCES

[1] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol*, vol. 59, no. 10, pp. 1087–1091, 2006.

[2] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[3] P. Flyer and J. Hirman, "Missing data in confirmatory clinical trials," *J. Biopharm. Stat.*, vol. 19, no. 10, pp. 969–979, 2009.

[4] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J. School Psychol.*, vol. 48, no. 1, pp. 5–37, 2010.

[5] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[6] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.

[7] K. J. Nishanth and V. Ravi, "Probabilistic neural network based categorical data imputation," *Neurocomputing*, vol. 218, no. 12, pp. 17–25, 2016.

[8] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada, "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets," *Chemometrics Intell. Lab. Syst.*, vol. 134, no. 5, pp. 23–33, 2014.

[9] C. O. Galán, F. S. Lasheras, F. J. de Cos Juez, and A. B. Sánchez, "Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions," *J. Comput. Appl. Math.*, vol. 311, no. 2, pp. 704–717, 2017.

[10] Y. Wang and B. Chaib-Draa, "An online Bayesian filtering framework for Gaussian process regression: Application to global surface temperature analysis," *Expert Syst. Appl.*, vol. 67, no. 1, pp. 285–295, 2017.

[11] W. H. Finch, "Imputation methods for missing categorical questionnaire data: A comparison of approaches," *J. Data Sci.*, vol. 8, no. 3, pp. 361–378, 2010.

[12] D. Blend and T. Marwala. (2008). "Comparison of data imputation techniques and their impact." [Online]. Available: https://arxiv.org/abs/0812.1539

[13] J. Dauwels, L. Garg, A. Earnest, and L. K. Pang, "Tensor factorization for missing data imputation in medical questionnaires," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 2109–2112.

[14] H. Tan *et al.*, "A tensor-based method for missing traffic data completion," *Transp. Res. C, Emerg. Technol.*, vol. 28, pp. 15–27, Mar. 2013.

[15] B. Ran, H. Tan, Y. Wu, and P. J. Jin, "Tensor based missing traffic data completion with spatial-temporal correlation," *Physica A, Stat. Mech. Appl.*, vol. 446, pp. 54–63, Mar. 2016.

[16] H. Tan, Z. Yang, G. Feng, W. Wang, and B. Ran, "Correlation analysis for tensor-based traffic data imputation method," *Procedia Soc. Behav. Sci.*, vol. 96, no. 11, pp. 2611–2620, 2013.

[17] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 24–40, 2011.

[18] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[19] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: Wiley, 1987, p. 381.

[20] F. A. Gordon, S. Negrete-Yankelevich, and J. S. Vinicio, *Ecological Statistics: Contemporary Theory and Application*. London, U.K.: Oxford Univ. Press, 2015, p. 407.

[21] S. Seaman, J. Galati, D. Jackson, and J. Carlin, "What is meant by 'missing at random'?" *Stat. Sci.*, vol. 28, no. 2, pp. 257–268, 2013.

[22] S. Landau and B. S. Everitt, *A Handbook of Statistical Analyses Using SPSS*, vol. 24. Boca Raton, FL, USA: CRC Press, no. 20, 2004, pp. 1–339.

[23] J. R. Cheema, "Some general guidelines for choosing missing data handling methods in educational research?" *J. Mod. Appl. Stat. Methods*, vol. 13, no. 2, pp. 53–75, 2014.

[24] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, no. 1, pp. 549–576, 2009.

[25] Q. Song and M. Shepperd, "A new imputation method for small software project data sets," *J. Syst. Softw.*, vol. 80, no. 1, pp. 51–62, 2007.

[26] R. A. Johnson, *Miller & Freund's Probability and Statistics for Engineers*. Upper Saddle River, NJ, USA: Prentice-Hall, 2004, p. 642.

[27] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. London, U.K.: Chapman & Hall, 1997, p. 444.

[28] P. L. Roth, "Missing data: A conceptual review for applied psychologists," *Pers. Psychol.*, vol. 47, no. 3, pp. 537–560, 1994.

[29] A. M. Wood, I. R. White, and S. G. Thompson, "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals," *Clin. Trials*, vol. 1, no. 4, pp. 368–376, 2004.

[30] S. M. Fox-Wasylyshyn and M. M. El-Masri, "Handling missing data in self-report measures," *Res. Nursing Health*, vol. 28, no. 6, pp. 488–495, 2005.

[31] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 2nd ed. New York, NY, USA: Wiley, 2002, p. 408.

[32] Y. Yuan, "Multiple imputation for missing data: Concepts and new development (Version 9.0)," SAS Inst., Rockville, MD, USA, Tech. Rep., 2010, pp. 1–13.

[33] S. van Buuren, *Flexible Imputation of Missing Data*. New York, NY, USA: Taylor & Francis, 2012, p. 342.

[34] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How many imputations are really needed? Some practical clarifications of multiple imputation theory," *Prevention Sci.*, vol. 8, no. 3, pp. 206–213, 2007.

[35] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *J. Roy. Soc. Interface*, vol. 6, no. 31, pp. 187–202, 2009.

[36] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. B, Stat. Methodol.*, vol. 64, no. 4, pp. 583–639, 2002.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[38] R. Kohavi and F. Provost, "Glossary of terms," *Mach. Learn.*, vol. 30, pp. 271–274, Jun. 1998.
[39] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *IAENG Transactions on Engineering Technologies* (Lecture Notes in Electrical Engineering), vol. 229, G. C. Yang, S. Ao, and L. Gelman, Eds. Dordrecht, The Netherlands: Springer, 2013.
[40] M. Ebden. (2015). "Gaussian processes for regression: A quick introduction." [Online]. Available: https://arxiv.org/abs/1505.02965
[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. London, U.K.: MIT Press, 2006, p. 266.
[42] D. Goldsmith, A. Preis, M. Allen, and A. J. Whittle, "Virtual sensors to improve on-line hydraulic model calibration," in *Proc. 12th Annu. Conf. Water Distrib. Syst. Anal.*, 2010, pp. 1349–1361.
[43] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. Washington, DC, USA: U.S. Government Printing Office 1964, p. 470.
[44] A. Ilin and T. Raiko, "Practical approaches to principal component analysis in the presence of missing values," *J. Mach. Learn. Res.*, vol. 11, no. 7, pp. 1957–2000, 2010.
[45] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Aberdeen, U.K.: Springer, 2002, p. 266.
[46] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[47] P. J. A. Shaw, *Introductory Multivariate Statistics for the Environmental Science*. London, U.K.: Oxford Univ. Press, 2003, p. 233.
[48] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1460–1480, 2012.
[49] G. E. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *Front. Artif. Intell. Appl.*, vol. 87, no. 10, pp. 251–260, 2002.
[50] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[51] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 3, pp. 291–312, May 2012.
[52] I. Yoo *et al.*, "Data mining in healthcare and biomedicine: A survey of the literature," *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, 2012.
[53] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers—A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 4, pp. 476–487, 2005.
[54] S. Zaremotlagh and A. Hezarkhani, "The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran," *J. Afr. Earth Sci.*, vol. 128, no. 4, pp. 37–46, 2017.
[55] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon Press, 1995, p. 518.
[56] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics Intell. Lab. Syst.*, vol. 39, no. 1, pp. 43–62, 1997.
[57] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 109–118, 1990.
[58] M. A. Kramer, "Autoassociative neural networks," *Comput. Chem. Eng.*, vol. 16, no. 4, pp. 313–328, 1992.
[59] D. S. Rizzuto and M. J. Kahana, "An autoassociative neural network model of paired-associate learning," *Neural Comput.*, vol. 13, no. 9, pp. 2075–2092, 2001.
[60] J. Dauwels, K. Srinivasan, R. M. Ramasubba, and A. Cichocki, "Multichannel EEG compression based on matrix and tensor decompositions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2011, pp. 629–632.
[61] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.
[62] L. Garg, J. Dauwels, A. Earnest, and K. P. Leong, "Tensor-based methods for handling missing data in quality-of-life questionnaires," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 5, pp. 1571–1580, Sep. 2014.
[63] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," in *Proc. UCLA Work. Papers Phonetics*, vol. 22, Mar. 1972, pp. 30–44.
[64] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
[65] Y. Wang *et al.*, "Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China," *Environ. Sci. Pollut. Res.*, vol. 20, no. 12, pp. 8909–8923, 2013.
[66] M. J. Mudumbe and A. M. Abu-Mahfouz, "Smart water meter system for user-centric consumption measurement," in *Proc. IEEE Int. Conf. Ind. Inform. (INDIN)*, Jul. 2015, pp. 993–998.
[67] C. P. Kruger, A. M. Abu-Mahfouz, and S. J. Isaac, "Modulo: A modular sensor network node optimised for research and product development," in *Proc. IST-Afr. Conf. Exhib.*, May 2013, pp. 1–9.
[68] S. R. Mounce, R. B. Mounce, T. Jackson, J. Austin, and J. B. Boxall, "Pattern matching and associative artificial neural networks for water distribution system time series data analysis," *J. Hydroinformat.*, vol. 16, no. 3, pp. 617–632, 2014.
[69] K. Adedeji, Y. Hamam, B. Abe, and A. M. Abu-Mahfouz, "Wireless sensor network-based improved NPW leakage detection algorithm for real-time application in pipelines," in *Proc. Southern Afr. Telecommun. Netw. Appl. Conf.*, 2016, pp. 82–83.
[70] P. R. Page, A. M. Abu-Mahfouz, and S. Yoyo, "Parameter-less remote real-time control for the adjustment of pressure in water distribution systems," *J. Water Resour. Planning Manag.*, vol. 143, no. 9, pp. 1–12, 2017.
[71] P. R. Page, A. M. Abu-Mahfouz, and M. L. Mothetha, "Pressure management of water distribution systems via the remote real-time control of variable speed pumps," *J. Water Resour. Planning Manage.*, vol. 143, no. 8, pp. 391–397, 2017.
[72] P. R. Page, A. M. Abu-Mahfouz, and S. Yoyo, "Real-time adjustment of pressure to demand in water distribution systems: Parameter-less P-controller algorithm," *Procedia Eng.*, vol. 154, no. 7, pp. 391–397, 2016.
[73] H. I. Kobo, A. M. Abu-Mahfouz, and G. P. Hancke, "A survey on software-defined wireless sensor networks: Challenges and design requirements," *IEEE Access*, vol. 5, pp. 1872–1899, 2017.
[74] K. M. Modieginyane, B. B. Letswamotse, R. Malekian, and A. M. Abu-Mahfouz, "Software defined wireless sensor networks application opportunities for efficient network management: A survey," *Comput. Elect. Eng.*, vol. 66, no. 2, pp. 274–287, 2018.
[75] J. Quevedo *et al.*, "Validation and reconstruction of flow meter data in the Barcelona water distribution network," *Control Eng. Pract.*, vol. 18, no. 6, pp. 640–651, 2010.
[76] R. Barrela, C. Amado, D. Loureiro, and A. Mamade, "Data reconstruction of flow time series in water distribution systems–a new method that accommodates multiple seasonality," *J. Hydroinformat.*, vol. 19, no. 2, pp. 238–250, 2017.
[77] S. Bennis, F. Berrada, and N. Kang, "Improving single-variable and multivariable techniques for estimating missing hydrological data," *J. Hydrol.*, vol. 191, nos. 1–4, pp. 87–105, 1997.
[78] M. S. Osman, S. Yoyo, P. R. Page, and A. M. Abu-Mahfouz, "Real-time dynamic hydraulic model for water distribution networks: Steady state modelling," in *Proc. 6th IASTED Int. Conf. Environ. Water Resour. Manage.*, 2016, pp. 142–147.
[79] A. M. Abu-Mahfouz, Y. Hamam, P. R. Page, K. Djouani, and A. Kurien, "Real-time dynamic hydraulic model for potable water loss reduction," *Procedia Eng.*, vol. 154, no. 7, pp. 99–106, 2016.
[80] P. R. Page, "The sensitivity of a water distribution system to regional state parameter variations," *Math. Problems Eng.*, vol. 2018, no. 5, Apr. 2018, Art. no. 6938483.
[81] Y. Chen *et al.*, "A global learning with local preservation method for microarray data imputation," *Comput. Biol. Med.*, vol. 77, pp. 76–89, Oct. 2016.
[82] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *Int. Stat. Rev.*, vol. 78, no. 1, pp. 40–64, 2010.

MUHAMMAD S. OSMAN received the B.Eng. degree in chemical engineering and the M.Eng. degree in water utilization engineering from the University of Pretoria in 2006 and 2010, respectively. He is currently a professional licensed Research Engineer with the Hydraulic Infrastructure Engineering Group, Council for Scientific and Industrial Research. He has authored or co-authored several papers that are published in journals and in conference proceedings. His research interests are the treatment of industrial effluents using innovative membrane technologies and statistical numerical modeling of water systems.

ADNAN M. ABU-MAHFOUZ (M'12–SM'17) received the M.Eng. and Ph.D. degrees in computer engineering from the University of Pretoria. He is currently a Principal Researcher with the Council for Scientific and Industrial Research, Research and Innovation Associate, Tshwane University of Technology, and also an extraordinary Faculty Member with the University of Pretoria. He participated in the formulation of many large and multidisciplinary R&D successful proposals (as a Principal Investigator or main author/contributor). His research interests are wireless sensor and actuator network, low power wide area networks, software defined wireless sensor network, cognitive radio, network security, network management, and sensor/actuator node development. He is a member of many IEEE Technical Communities. He is an Associate Editor at the IEEE Access, the IEEE Internet of Things, and the IEEE Transactions on Industrial Informatics. He is the Founder of the Smart Networks collaboration initiative that aims to develop efficient and secure networks for the future smart systems, such as smart cities, smart grid, and smart water grid.

PHILIP R. PAGE received the Certificate of Advanced Study in mathematics from the University of Cambridge, U.K., and the D.Phil. degree from the University of Oxford, U.K., in 1996. He is currently a Senior Researcher with the Council for Scientific and Industrial Research. He held research positions in the U.K. and USA. He has authored or co-authored over 70 papers in published journals and conference proceedings. His current focus is on water infrastructure engineering. Specifically, he performs mathematical and numerical modeling of water distribution systems and smart water infrastructure. He has extensive background in the field of quantitative modeling and computing.