Mine Water Environ (2016) 35:3–9
DOI 10.1007/s10230-014-0322-4

TECHNICAL ARTICLE

On the Issue of Incomplete and Missing Water-Quality Data in Mine Site Databases: Comparing Three Imputation Methods

Getnet D. Betrie · Rehan Sadiq · Solomon Tesfamariam · Kevin A. Morin

Received: 19 April 2014 / Accepted: 2 December 2014 / Published online: 14 December 2014

Abstract Large water-quality databases are valuable for predicting mine drainage chemistry, identifying optimal measures for mitigation and remediation, and refuting/refining models and theories. However, such databases often have missing values due to periodic lack of sampling and analysis or input errors. These missing values lead to problems in machine learning and statistical analysis of water-quality data from mine sites. Using water-quality data collected from 1971 to 1994 from many locations at a copper-molybdenum-gold-silver-rhenium mine site, we compared three imputation methods to estimate missing water-quality data: iterative robust model-based imputation (IRMI), multiple imputations of incomplete multivariate data (AMELIA), and sequential imputation for missing values (IMPSEQ). These methods were evaluated based on mean absolute error, relative absolute error, and percent bias techniques. The results showed that IMPSEQ and IRMI are suitable to impute missing values in water-quality databases at mine sites, whereas AMELIA is not.

Keywords Missing values · Machine learning · Data-driven · IMPSEQ · IRMI · AMELIA

Introduction

Large water-quality databases are valuable for predicting mine drainage chemistry, identifying optimal measures for mitigation and remediation, and refuting/refining models and theories (Morin et al. 2012). These databases often contain analyses of dozens of chemical elements at many locations over extended periods of time. Some data are often missing within such databases due to periodic lack of sampling and analysis or input errors. Missing data present challenges in data-driven modelling that includes machine learning, soft-computing, and data mining. The challenges associated with missing data include the use and interpretation of partially collected data (Betrie et al. 2013), accuracy of learning algorithms (Bello 1995; Acuna and Rodriguez 2004), and the use of software as statistical software are developed under the assumption that input is provided as a complete data matrix (Schafer and Olsen 1998). In the latter case, the software would delete listwise/pairwise or substitute missing values with mean values if the data matrix has one or more missing values. In listwise deletion, any records containing a missing value in one of the variables would be deleted. On the other hand, a variable containing missing values would be deleted and the other variables used for analysis in pairwise deletion. The downside of such methods is that it leads to a severe loss of information or introduction of bias (Little and Rubin 2002). Therefore, better methods are required to treat missing values before modeling.

Methods for treating missing values can be grouped into list/pairwise deletion, parameter estimation, and imputation (Allison 2001; Little and Rubin 2002). It is worth noting that censored data are not considered as missing values. In list/pairwise deletion, records or variables that contain a missing value are deleted in order to create a complete data matrix. In the parameter estimation method, a joint probability function is defined for the data matrix and parameters are estimated using expectation–maximization algorithm or direct maximum likelihood. In the imputation method, missing values are estimated using single or multivariate

G. D. Betrie (✉) · R. Sadiq · S. Tesfamariam
School of Engineering, The University of British Columbia, Kelowna, BC, Canada
e-mail: getnet.betrie@ubc.ca

K. A. Morin
Mine Drainage Assessment Group, Surrey, BC, Canada

imputation methods. Single imputation substitutes missing values with mean estimation or a value estimated from a model. On the other hand, the multiple imputation method substitutes missing values with estimated values from two or more single imputation results.

We examined methods to impute missing values because the literature (e.g. Allison 2001; Little and Rubin 2002) shows that imputation is superior to other methods such as list/pairwise and parameter estimation. Prior to applying an imputation method, however, the mechanism of handling incompleteness should be determined because different assumptions were considered with the developed methods (Little and Rubin 2002). The literature (e.g. Allison 2001; Little and Rubin 2002; Graham 2012) classifies the mechanism of accounting for missing data into: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). In MCAR, the probability of a missing value for a variable does not depend on the probability of missing values of this variable and observed values of other variables. In MAR, the probability of a missing value does not depend on the missing values of this variable, but it does depend on observed values of other variables. In NMAR, the probability of a missing value depends on the other missing values. It is worth noting that the literature (e.g. Güler et al. 2002) indicates that water-quality data from mine sites tends to follow the MAR mechanism.

We compared three methods to impute missing water-quality data: iterative robust model-based imputation (IRMI), multiple imputations of incomplete multivariate data (AMELIA), and sequential imputation for missing values (IMPSEQ). Note that these methods were selected because their assumptions are based on MAR.

Methodology

Description of Methods

We compared two single imputation (IRMI and IMPSEQ) methods and a multiple imputation (AMELIA) method to impute missing values of multivariate water-quality data from an example minesite. The algorithms of the three methods used to estimate missing values were implemented in RStudio (RStudio 2012).

The IRMI algorithm is a model-based imputation method where missing values are estimated using sequence of regression models (Templ et al. 2011). A brief summary is presented below, but a detailed description of this algorithm, including its development and verification, can be found in Templ et al. (2011). The seven-step procedure for the IRMI algorithm is:

Step 1: Sort the variables based on the number of missing values, as shown in Eq. 1,

$$M(v_1) \leq M(v_2) \leq \ldots M(v_p) \tag{1}$$

where $M(v_i)$ denotes the number of missing values in variable v_i. Denote $I = , \ldots, p$

Step 2: Set $l = 1$

Step 3: Denote $m_l \subset \{1, \ldots, n\}$ the indices of the observations that were originally missing in the variable v_l, and $o_l = \{1, \ldots, n\} \backslash m_l$ the indices corresponding to the observed cells of v_l. Separate the observed and the missing values into two matrices, V_N^{ol} and V_N^{ml}, respectively. The first column of the matrices consists of ones and is used to find an intercept term in the regression problem

Step 4: A regression equation is developed using the matrices with unknown regression coefficients, β, and an error term, ε:

$$v_l^{ol} = V_{I \backslash l}^{ol} \beta + \varepsilon \tag{2}$$

Step 5: Estimate the regression coefficients $\widehat{\beta}$ using Eq. 2, and use $\widehat{\beta}$ to estimate the missing values:

$$\widehat{v}_l^{ml} = V_k^{ml} \widehat{\beta} \tag{3}$$

Step 6: Repeat Steps 2 to 5 for $l = 2$ to $l = p$

Step 7: Repeat Steps 3 to 6 until the estimated values converge

The IMPSEQ algorithm is a covariance-based imputation method where missing values are estimated sequentially by minimizing the determinant of the covariance matrix of the data (Verboven et al. 2007). A brief summary of the IMPSEQ algorithm is presented below; a detailed description of this algorithm can be found in Verboven et al. (2007). The five step procedure for the IMPSEQ algorithm is:

Step 1: Partition the dataset X into complete and missing matrices, which are X_c and X_m, respectively

Step 2: Sort the variables based on the increasing number of missing values, X_m

Step 3: For a variable with the least number of missing values, estimate the missing values by minimizing the determinant of the covariance matrix with respect to X_m, as shown in Eq. 4

$$\begin{aligned} \operatorname{cov}(X_m) = & \frac{c-1}{c} \operatorname{cov}(\mathrm{X_c}) \\ & + \frac{1}{c} (X_m - \overline{X_c})(X_m - \overline{X_c})' \end{aligned} \tag{4}$$

where c is the total number observed values and $\overline{X_c}$ is the mean of observed values

Step 4: Update the missing values in X_c

Step 5: Repeat Steps 2 to 4 for the other variables

The AMELIA algorithm estimates missing values using the expectation–maximization with a bootstrapping algorithm (EMB) (Honaker et al. 2006). This assumes that the complete dataset has multivariate normal distribution and the data are missing at random. A brief summary of the AMELIA algorithm is presented below, but a detailed description of this algorithm can be found in the literature (e.g. Honaker 2006; Honaker and King 2010). The six step procedure for the AMELIA algorithm is:

Step 1: Partition the dataset D into observed D^{obs} and missing D^{mis} elements, which is $D = \{D^{\text{obs}}, D^{\text{mis}}\}$. The missing values are estimated based on the Bayesian concept, as shown below:

$$p\left(D^{obs}|\theta\right) = \int p(D|\theta)dD^{mis} \quad (5)$$

where θ is the distribution parameter, which is equal to μ and $\sum$

Step 2: Define a matrix M that has zero and one elements. The zero and one represent the missing and observed values, respectively

Step 3: Create m samples from M using the bootstrapping algorithm

Step 4: Estimate the posterior θ for the m samples using the EMB algorithm

Step 5: Estimate the missing values in each m sample using the posterior parameter and the conditional distribution of D^{obs}

Step 6: Take the averages of m estimates to obtain each missing value

Data Preparation

The water-quality database was obtained from a copper-molybdenum-gold-silver-rhenium mine site located in British Columbia, Canada. In order to explore missing values, visualization techniques developed by Templ et al. (2012) were used. This dataset consists of 5720 rows and 12 variables, as shown in Fig. 1 (left); such datasets often contain numerous missing values (Morin et al. 2012). The red and grey colors represent, respectively, the missing and observed values in the water-quality dataset. In order to compare the methods (IRMI, IMPSEQ, and AMELIA) previously described, a complete test dataset was extracted from the raw dataset. The test dataset consisted of 8 variables (i.e. pH, electrical conductivity (EC), Cu, Zn, Cd, Ca. Mg, and Al) and 825 rows that were extracted from indices 4700 to 5520 of the raw dataset (Fig. 1, left). It is worth noting that these indices contain no missing values.

Following the missing values pattern in the raw dataset, some observed values in the test dataset were deleted to reproduce the missing values pattern in the raw dataset (Fig. 1, right). This test dataset with its missing values was used to compare the three methods. Before analysis, the dataset was transformed into a normal distribution using a logarithmic transformation since the original dataset was log-normally distributed (Betrie et al. 2013). A statistical summary of this data set is presented in Table 1. The percentage of missing values was highest in Al, followed by EC, Cu, Mg, Cd, Ca, pH, and Zn.

Performance Evaluation

The accuracy between observed and estimated values was evaluated for each imputation method. The evaluation techniques include mean absolute error (MAE), relative absolute error (RAE), and percent bias (PBIAS), as shown in Eqs. 6–8. MAE and RAE compare the match between estimated and observed values. MAE and RAE values equal to zero indicate optimal values; in general, a method that has the smallest values is preferred over the other methods (Betrie et al. 2013). PBIAS measures whether the average tendency of estimated values is larger or smaller than their observed values (Gupta et al. 1999; Betrie et al. 2011). Values of PBIAS equal to zero, positive, or negative indicate, respectively, ideal estimation, overestimation, and underestimation (Gupta et al. 1999). In addition to these three evaluation techniques, visual inspection was used.

$$MAE = \frac{\sum_{i=1}^{n} |o_i - e_i|}{n} \quad (6)$$

$$RAE = \frac{\sum_{i=1}^{n} |o_i - e_i|}{\sum_{i=1}^{n} |o_i - \overline{o}|} \quad (7)$$

$$PBIAS = \frac{\sum_{i=1}^{n} (o_i - e_i)}{\sum_{i=1}^{n} (o_i)} \times 100 \quad (8)$$

where o, e, $\overline{o}$ and n are the observed, estimated, mean of the observed, and size of the dataset, respectively.

Results and Discussion

The performance of the imputation methods is presented in Table 2, and Figs. 2 and 3. For the Zn variable, which had 4 % values missing, the MAE and RAE results show that the IRMI method performed best, followed by IMPSEQ and AMELIA. The PBIAS results showed that the IRMI and IMPSEQ methods underestimated the observed values, whereas AMELIA highly overestimated the observed values. It is interesting to note that the magnitude of underestimation for IRMI is larger than for IMPSEQ. Figure 2 shows that most of the estimated values of Zn by IRMI and

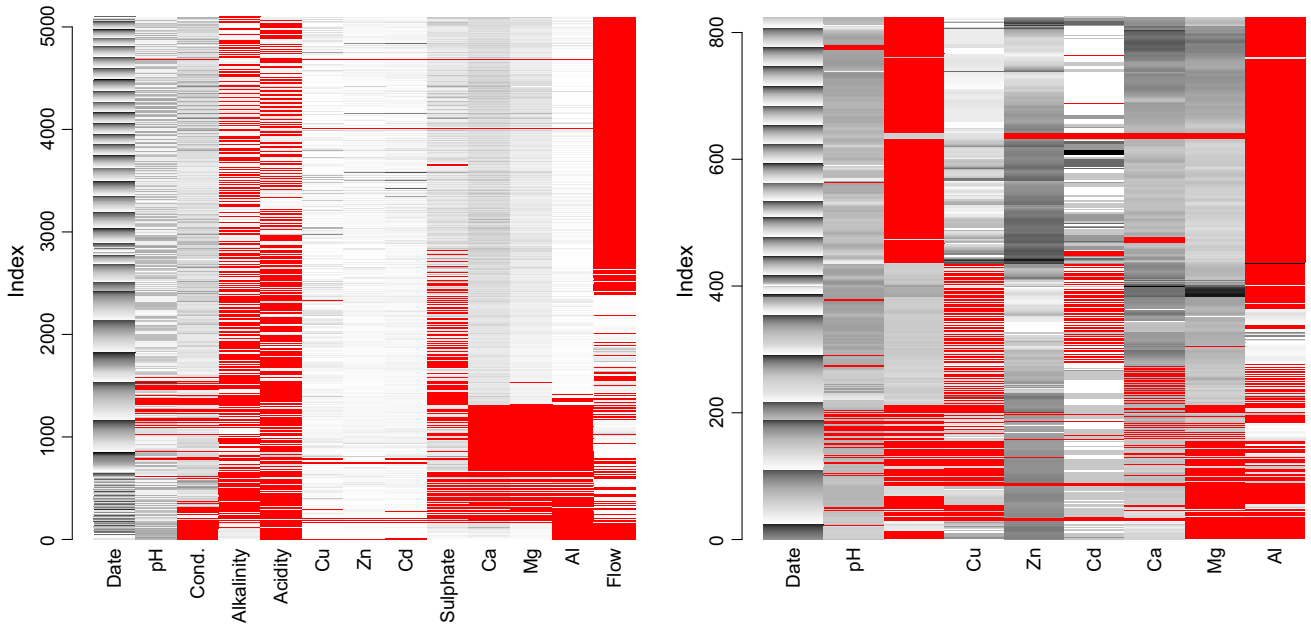

Fig. 1 Raw (*left*) and test (*right*) datasets with their observed (*grey*) and missing (*red*) values

Table 1 Statistical summary of the test data used to evaluate the imputation methods

	Zn	pH	Ca	Cd	Mg	Cu	EC	Al
Minimum	0.1	4.43	140	0	22	0.01	840	0.1
1st quartile	0.7	6.72	240	0.01	35	0.02	1,330	0.1
Median	1.9	7.13	270	0.02	39	0.03	1,440	0.2
Mean	1.91	7.07	274.1	0.02	40.76	0.05	1,435	0.31
3rd quartile	2.7	7.53	310	0.02	45	0.05	1,509	0.2
Maximum	7.3	10.65	430	0.05	72	0.73	3,450	5.2
Missing (%)	4	8	11	15	21	26	62	78

IMPSEQ are above the ideal line (the line of equality along which the estimated and observed values are equal), which indicates that the two methods both underestimated the observed values. On the other hand, the entire estimation of Zn by AMELIA is below the ideal line, which indicates that it highly overestimated the observed Zn values.

For pH, which had 8 % values missing, the MAE, RAE, and PBIAS results show that IMPSEQ performed best, followed by IRMI and AMELIA. The PBIAS results indicate that IRMI, AMELIA, and IMPSEQ slightly overestimated, highly overestimated, and slightly underestimated the observed values, respectively. Figure 2 shows that most of the IMPSEQ and IRMI imputed values are equally distributed below and above the ideal line, which indicates that the imputed values are well estimated. However, the imputed values of AMELIA are far below the ideal line, indicating that they are highly overestimated.

For Ca, which had 11 % values missing, IRMI performed best, followed by IMPSEQ and AMELIA. The PBIAS results show that the imputed values of IRMI are slightly underestimated, whereas the imputed values of IMPSEQ and AMELIA slightly and highly overestimated the observed values, respectively. Visual inspection of Fig. 2 shows that most of the imputed values of IMPSEQ and IRMI are distributed around the ideal line, indicating that the imputations are quite good. On the other hand, the imputed values of AMELIA are way below the ideal line, indicating that this method tended to highly overestimate the observed values.

For Cd, which had 15 % values missing, the evaluation techniques showed that IMPSEQ performed the best, followed by IRMI and AMELIA. In addition, the PBIAS results show that imputed values of all three methods overestimated the observed values, but the overestimation of IRMI and IMPSEQ is slight, whereas the overestimation of AMELIA is very high. Visual inspection of Fig. 2 shows that the estimations of IMPSEQ and IRMI are well distributed above and below the ideal line. This indicates that the central tendency of imputed values and the observed values would tend to be similar, but that the estimates could have larger error on a sample-by-sample basis. On the other hand, the imputed values of AMELIA are way

Table 2 The performance of the imputation methods

Variable	Missing (%)	MAE	RAE	PBIAS
IRMI Zn	4	0.84	1.71	−32
IMPSEQ Zn		0.90	1.84	−17
AMELIA Zn		34	71	1,483
IRMI pH	8	0.28	0.82	0.25
IMPSEQ pH		0.22	0.65	-0.13
AMELIA pH		64	184	893
IRMI Ca	11	18	0.47	−0.63
IMPSEQ Ca		20	0.29	3
AMELIA Ca		2,359	56	944
IRMI Cd	15	0.004	1.02	2.60
IMPSEQ Cd		0.003	0.86	2.82
AMELIA Cd		0.139	35	1,115
IRMI Mg	21	5.97	1.43	18
IMPSEQ Mg		3.20	0.77	8
AMELIA Mg		297	71	1,006
IRMI Cu	26	0.02	0.92	−3
IMPSEQ Cu		0.01	0.76	−14
AMELIA Cu		0.41	19	1,048
IRMI EC	62	230	1	5
IMPSEQ EC		158	0.68	7
AMELIA EC		12,771	55	966
IRMI Al	78	0.26	1.09	−1
IMPSEQ Al		0.18	0.77	−16
AMELIA Al		2.40	10	681

below the ideal line, indicating that the method highly overestimated the observed values. The imputed values are clustered into three values since most of the observed values are close to the analytical-method detection limit.

For Mg, which had 21 % values missing, IMPSEQ was the best, followed by IMRI and AMELIA. PBIAS shows that the imputed values are overestimated by all methods, but the overestimation of IMPSEQ and IRMI is slight, whereas the overestimation of AMELIA is very high. Visual inspection of Fig. 3 shows that most of the imputed values of IMPSEQ and IRMI are distributed below the ideal line, indicating that these methods slightly overestimated the observed values. On the other hand, the imputed values of AMELIA are clustered way below the ideal line, indicating that this method is highly overestimated the observed values.

For Cu, which had 26 % values missing, IMPSEQ performed the best, followed by IRMI and AMELIA. The PBIAS values show that the imputed values of IMPSEQ and IRMI are slightly underestimated, whereas the imputed values of AMELIA are highly overestimated.Figure 3 shows that the imputed values of IMPSEQ and IRMI are distributed along the ideal line, indicating that imputation is reasonable. All of the imputed values of AMELIA are below the ideal line, indicating that this method highly overestimated the observed values.

For EC, which had 62 % values missing, MAE and RAE shows that IMPSEQ performed the best, followed by IRMI

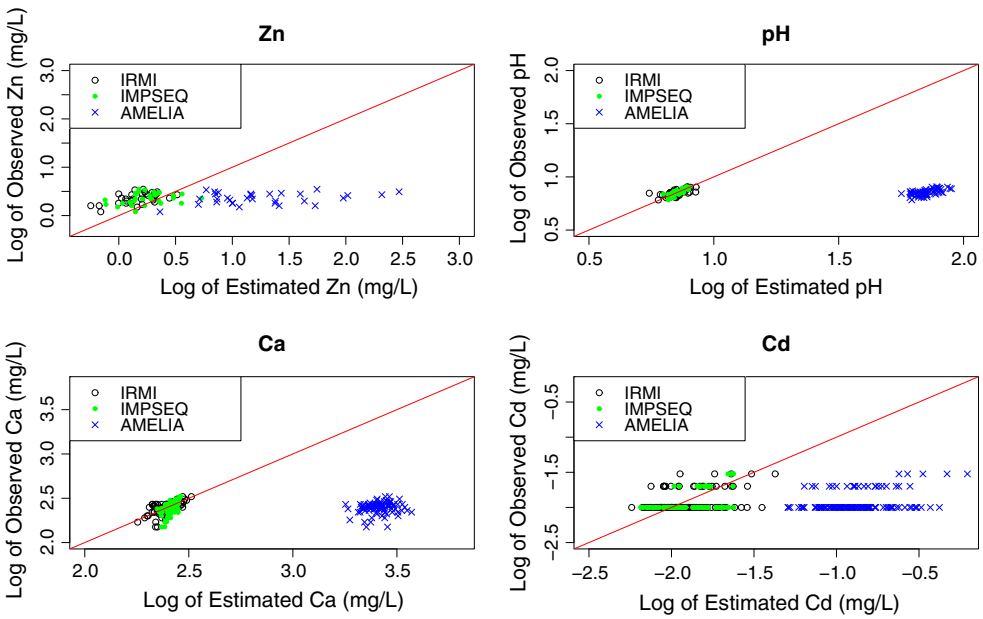

Fig. 2 Comparing the logarithm of observed and estimated values of variables pH, Zn, Ca, and Cd; the *diagonal line* indicates the perfect match between observed and imputed values

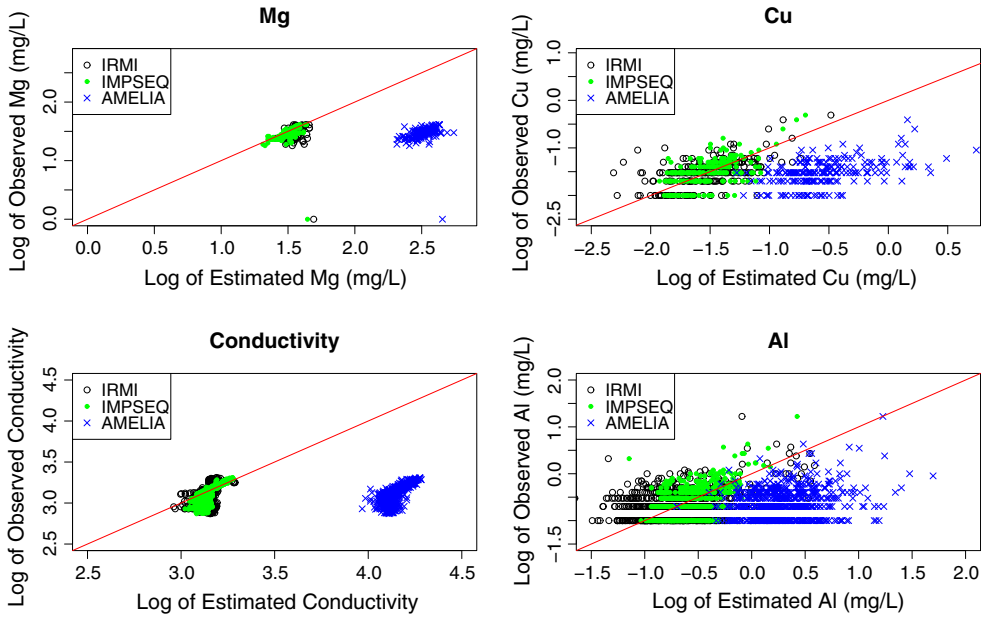

Fig. 3 Comparing the logarithm of observed and estimated values of the variables Mg, Cu, EC, and Al; the *diagonal line* indicates a perfect match between observed and imputed values

and AMELIA. PBIAS shows that the imputed values are all overestimates, but the overestimation of IMPSEQ and IRMI is slight, whereas overestimation by AMELIA is very high compared to the other methods. Figure 3 shows that most of the imputed values of EC by IMPSEQ and IRMI are below the ideal line, indicating that the imputed values slightly overestimated the observed values. The imputed values of AMELIA are way below the ideal line, indicating that the imputed values highly overestimated the observed values.

For Al, which had 78 % values missing, MAE and RAE show that the imputation of IMPSEQ performed the best, followed by IRMI and AMELIA. PBIAS shows that imputed values of IRMI and IMPSEQ are slightly underestimated, whereas the imputed values of the AMELIA are highly overestimated. Figure 3 shows that the imputed values of IRMI and IMPSEQ are distributed along the ideal line, indicating that imputation of these methods are reasonable. Most of the imputed values of AMELIA are below the ideal line, indicating that AMELIA overestimated the observed values.

Summary and Conclusion

Mine site water-quality databases often have missing values due to periodic lack of sampling and analysis or input errors. These missing values often limit the use and interpretation of data in making appropriate decisions. This study presents advanced methods to treat missing values in water-quality databases to improve decision-making in mine sites. Three methods were compared to impute missing values of multivariate water-quality data from mine sites: IRMI, sequential imputation for missing values (IMPSEQ), and multiple imputations of incomplete multivariate data (AMELIA). The performance of these imputation methods were evaluated using mean absolute error (MAE), relative absolute error (RAE), and percent bias (PBIAS) techniques. In addition, visual inspections were used to evaluate the imputation methods.

This study showed that IMPSEQ and IRMI are suitable to impute missing values of water-quality databases at mine sites, whereas AMELIA is not. The first two methods could also be used to treat the problem of missing values in other databases, such as climate and water resources. Although these two methods are superior to traditional methods of imputation that introduce serious bias and loss of information, the uncertainty associated with imputed values is not addressed in this study. This shortcoming will be addressed in future work.

References

Acuna E, Rodriguez C (2004) The treatment of missing values and its effect on classifier accuracy. In: Banks D, House L, McMorris FR, Arabie P, Gaul W (eds) Classification, clustering and data mining applications. Springer, Berlin, pp 639–647

Bello AL (1995) Imputation techniques in regression analysis: looking closely at their implementation. Comput Stat Data Anal 20(1):45–57

Betrie GD, Mohamed YA, van Griensven A, Srinivasan R (2011) Sediment management modelling in the Blue Nile Basin using SWAT model. Hydrol Earth Syst Sci 15(3):807–818

Betrie GD, Tesfamariam S, Morin KA, Sadiq R (2013) Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. Environ Monit Assess 185(5):4171–4182

Graham JW (2012) Missing data: analysis and design. Springer Science + Business Media, USA

Güler C, Thyne G, McCray J, Turner K (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol J 10(4):455–474

Gupta HV, Sorooshian S, Yapo PO (1999) Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. J Hydrol Eng 4(2):135–143

Honaker J, King G (2010) What to do about missing values in time-series cross-section data. Am J Polit Sci 54(2):561–581

Honaker J, King G, Blackwell M (2006) Amelia II: a program for missing data. J Stat Softw 45(7):1–47

Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley, USA

Morin KA, Hutt NM, Aziz M (2012) Case studies of thousands of water analyses through decades of monitoring: selected observations from three mine sites in British Columbia, Canada. Proceeding, 2012 international conference on acid rock drainage, Ottawa, Canada

RStudio (2012) RStudio Software. http://www.rstudio.org/

Schafer JL, Olsen MK (1998) Multivariate behavioral multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivar Behav Res 33(4):545–571

Templ M, Kowarik A, Filzmoser P (2011) Iterative stepwise regression imputation using standard and robust methods. Comput Stat Data An 55(10):2793–2806

Templ M, Alfons A, Filzmoser P (2012) Exploring incomplete data using visualization techniques. Adv Data Anal Classi 6(1):29–47

Verboven S, Branden KV, Goos P (2007) Sequential imputation for missing values. Comput Biol Chem 31(5–6):320–327