7th International Quality Conference
May 24th 2013
Center for Quality, Faculty of Engineering, University of Kragujevac

Aleksandar Novakovic[1)]
Vesna Rankovic[1)]
Dejan Divac[2)]
Nenad Grujovic[1)]
Nikola Milivojevic[2)]

1) Department for Applied Mechanics and Automatic Control, Faculty of Engineering, University of Kragujevac, Serbia

2) Institute for Development of Water Resources "Jaroslav Černi", Belgrade, Serbia

MISSING DATA ESTIMATION IN DAM STRUCTURES USING MULTIPLE IMPUTATION METHOD

Abstract: *The effective dam safety monitoring programs is essential for dam and is widely accepted. Instrumentation as part of dam safety programs is installed to measure a particular parameter of interest. These parameters might include water levels, seepage flows, deformations or displacements, pressures, loading conditions, temperature variations, seepage water clarity, piezometric levels, etc. The aim of the timely detection of abnormal behaviour of the dam does not necessarily imply frequent monitoring or the collection of a great deal of data. It is important that this information is representative and adequately interpreted. Interpretation of the available data is very substantial for dam health monitoring. The data interpretation can be difficult when data are missing or incomplete. In this paper multiple imputation method was used to estimate replacement values for the missing data. The results of simulation show that the multiple linear regression model for prediction of the water level in piezometers with estimated missing values provide better results.*

Keywords: *dam, missing data, multiple imputation method, piezometric water level*

1. INTRODUCTION

Most learning algorithms generally assume that training and test datasets are complete. However, real data sets are often incomplete and they contain a proportion of missing values due to various reasons such as equipment errors, manual data entry procedures, and incorrect measurements.

For example, multiple linear regression techniques are unable to directly handle missing data. Many software implementations of multiple linear regression ommits all instances with missing data before the model is constructed. Such an approach may lead to significant loss of informations, especially as the amount of missing data increases, and if the missing data is not distributed completely randomly, can result in severely biased models [1-2].

Prediction performances are not affected if there is less than 1% missing instances, although 1%-5% is manageable. However, sophisticated handling method is required if there is 5%-15% missing instances, while greater than 15% missing data can severely degrade the prediction performance of learning algorithms [3].

In response to these issues various solutions have been developed in statistics [4-5] and data mining [6-7].

The treatment of missing values is determined by the type of missing data. There are three types of missing data, as follows [8]: a) **Missing completely at random (MCAR)** – There is no dependency between missing value for an atribute and any other observed data or missing attribute; b) **Missing at random (MAR)** – The missing value for an atribute depends on other known data; c) **Missing not at random (MNAR)** – The missing value for an attribute depends on other missing values, and thus missing data cannot be estimated from observed data.

The objective of this study is to develop a multiple linear regression model to predict the piezometric water level in dam. Two models were compared, one with and one without estimated missing values in their dataset. In this paper it is assumed that missing values appear to be MAR, which implies that the missing values are deductible in some complex manner from the remaining data. In order to estimate replacement values for the missing data, multiple imputation method was used.

2. REGRESSION ANALYSIS

Regression analysis is a statistical technique for investigating and modeling the relationship between variables [9]. The multiple linear regression model is widely used for data analysis or prediction in dam engineering [10].

MLR is used for modelling the linear relationship between a dependent variable and one or more independent variables. Consider a training data set $\left\{\left(\boldsymbol{u}_1, z_1\right), \left(\boldsymbol{u}_2, z_2\right), \ldots, \left(\boldsymbol{u}_p, z_p\right)\right\} \in \mathbb{R}^N \times \mathbb{R}$ where $\boldsymbol{u}_i = \left\{u_{1i}\, u_{2i} \ldots u_{Ni}\right\}^T$ is a vector of input variables and z_i is the corresponding output value, p is the number of training data points. The multiple linear regression model is given by:

$$z_m = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \ldots + \beta_N u_N \qquad (1)$$

where β_i represents unknown parameters, β_i can be estimated by which the sum of the squares of the errors:

$$\varepsilon = \left(z_1 - z_{m1}\right)^2 + \left(z_2 - z_{m2}\right)^2 + \ldots + \left(z_p - z_{mp}\right)^2 \; (2)$$

in which z_{mi} denotes the MLR output value from the i-th input element:

$$z_{mi} = \beta_0 + \beta_1 u_{1i} + \beta_2 u_{2i} + \ldots + \beta_N u_{Ni} \qquad (3)$$

The matrix form of Eq. (2) is:

$$\varepsilon = \left(\boldsymbol{z} - \boldsymbol{U\beta}\right)^T \left(\boldsymbol{z} - \boldsymbol{U\beta}\right) \qquad (4)$$

where:

$$\boldsymbol{U} = \begin{bmatrix} 1 & u_{11} & u_{21} & \cdots & u_{N1} \\ 1 & u_{12} & u_{22} & \cdots & u_{N2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_{1p} & u_{2p} & \cdots & u_{Np} \end{bmatrix},$$

$$\boldsymbol{\beta} = \left\{\beta_0\, \beta_1 \; \ldots \; \beta_N\right\}^T, \quad \boldsymbol{z} = \left\{z_1 \; z_2 \; \ldots \; z_p\right\}^T,$$

and the least squares estimator of $\boldsymbol{\beta}$ is given by:

$$\boldsymbol{\beta} = \left(\boldsymbol{U}^T \boldsymbol{U}\right)^{-1} \boldsymbol{U}^T \boldsymbol{z} \qquad (5)$$

3. MULTIPLE IMPUTATION METHOD

Multiple imputation is statistical approach to the analysis of incomplete data, and in this section its main features are summarized. Detailed description is given in the literature [11].

Lets assume that X is the $n \times p$ data matrix, which can be thought of as $X = \left(X_{obs}, X_{mis}\right)$, where X_{obs} and X_{mis} are the observed and the missing parts, respectively. It is considered a model $P\left(X \mid \theta\right)$ for the data X, where θ is a vector parameter.

In the beginning, M complete datasets are created, using an appropriate imputation model to generate a plausible values for the missing observations. In order to obtain the imputed values data augmentation, [12], is used. Practically, it is a MCMC (Markov Chain Monte Carlo) procedure in which, given the values $\theta^{(k)}$

and $X_{mis}^{(k)}$ at the k-th iteration, these values are updated by drawing random values from the conditional distributions as follows:

$$X_{mis}^{(k+1)} \sim P(X_{mis} \mid X_{obs}, \theta^{(k)}) \qquad (6)$$

$$\theta^{(k+1)} \sim P(\theta \mid X_{obs}, X_{mis}^{(k+1)}) \qquad (7)$$

Step (6) is called the Imputation step, while (7) is known as the Parameter step. When $k \to \infty$, the sequence $\left(\theta^{(k)}, X_{mis}^{(k)}\right)$ has a stationary distribution whose marginals are $P\left(\theta \mid X_{obs}\right)$ and $P\left(X_{mis} \mid X_{obs}\right)$, respectively. After convergence, the imputations are acquired from (6).

Finally lets $\widehat{X}^{(i)} = (X_{obs,} X_{mis}^{(i)}),\ i = 1,...M$ denote the imputed-data estimates of X. Under general condition, multiple imputation estimate of X is calculated as follows [11]:

$$\bar{X} = \sum_{i=1}^{M} \widehat{X}^{(i)} \qquad (8)$$

4. CASE STUDY: PRVONEK DAM

The dam and the reservoir Prvonek (Fig. 1) were built in 2005, in order to solve the water supply problem of the towns Vranje, Bujanovac, and the surrounding villages in south-east Serbia. They are located on the Vranjsko-Banjska River, the right tributary of the river Južna Morava, 9 km upstream of Vranjska Banja-Spa, near the village Prvonek. Prvonek dam is rockfill embankment dam, with sloped central clay core within the dam body**.** The height of the dam is 90 m. At maximum water levels, the volume of the reservoir is 20 million m^3.

For the purpose of this paper, one of the piezometers, installed on the section of the dam, was observed. The water level in examined piezometer have been measured every day. The data collected from June 2010 to April 2011 were used for training and testing MLR models.

Figure 1. The view of Prvonek dam

5. SIMULATION RESULTS

For the purpose of constructing the MLR model, a program was written in R by the authors. The program implemented classes provided by the Amelia II package, which offers a comprehensive range of functions, necessary for implementation of multiple imputation method.

Accuracy of the MLR model depends on the appropriate choice of the input variables. The input variables of both MLR models were measurements of the tailwater levels taken on the same day (hl$_1$), 1 day before (hl$_2$) and 2 days before (hl$_3$) the measurements taken by piezometers.

For the purpose of training and testing MLR models, respectively, 70% and 30% of randomly chosen data points collected during the period of June 2010-April 2011 were used. Collected dataset contained 24% of missing data. The MLR models for prediction of water level in the examined piezometer, one without and one with estimated missing values in their dataset, are respectively:

$$hp_{col}^{MLR} = 194.52 + 0.64 \cdot hl_1 + 0.08 \cdot hl_2 - 0.06 \cdot hl_3 \qquad (9)$$

$$hp_{est}^{MLR} = 169.50 + 0.68 \cdot hl_1 + 0.07 \cdot hl_2 - 0.05 \cdot hl_3 \qquad (10)$$

The performance of two MLR models was evaluated by comparing the estimates of the models with experimental data. The performance parameters of the training and test sets are presented in Table 1.

Table1. The performance parameters of MLR models for prediction of water levels in the examined piezometer

Piezometer		r
hp_{col}^{MLR}	Training	0.93
	Test	0.96
hp_{est}^{MLR}	Training	0.97
	Test	0.99

5. CONCLUSION

In this paper, two MLR models, one with and one without estimated missing values in their dataset, were developed to predict water level in one of the installed piezometers installed on the section of the dam. The performance of the two MLR models were tested using correlation coefficients. As it can be easily observed, both models are capable of predicting water levels in piezometers with reasonable accuracy, although model with estimated missing values in its dataset gives a slightly higher coefficient of correlation values for training and test sets.

REFERENCES:

[1] Allison, P. D. (2001). *Missing Data.* Thousand Oaks, Sage, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences.

[2] Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons.

[3] Acuna, E., & Rodriguez, C. (2004). *The treatment of missing values and its effect in the classifier accuracy*. In: Banks, D., House, L., McMorris, F. R., Arabie, P., Gaul, W. (Eds.), Classification, Clustering and Data Mining Applications, Springer, 639-648.

[4] Dempster, A. P., Laird, N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39,* 1-38.

[5] Wang, Q., Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics, 30,* 896-924.

[6] Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence, 11*, 259-275.

[7] Chan, K., Lee, T. W., Sejnowski, T. J. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation, 15*(8), 1991-2011.

[8] Shafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London, UK: Chapman and Hall.

[9] Montgomery, D. C., & Runger, G. C. (1994). *Applied statistics and probability for engineers.* John Wiley & Sons, Inc.

[10] Rocha, M., Serafim, J., & Silveira, A. (1958). *A method of quantitative interpretation of the results obtained in the observation of the dams*. R.84, Q.21, In: VI ICOLD congress, New York, 2927-60.

[11] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

[12] Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of American Statistical Associacion, 82*, 528-550.

Acknowledgment: The part of this research is supported by Ministry of Science in Serbia, Grants III41007 and TR37013.