Imputation of outliers and missing values for activated sludge dissolved oxygen database using multivariate imputation by chained equations (mice)

IMPUTATION OF OUTLIERS AND MISSING VALUES FOR ACTIVATED SLUDGE DISSOLVED OXYGEN DATABASE USING MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE)

Hatem Nijim[1]and Rabee Rustum[2]
[1,2]Department of Civil Engineering, Heriot-Watt University, United Arab Emirates

ABSTRACT

Activated sludge process (ASP) is the most widely used process in wastewater treatment plants. The concentration of dissolved oxygen (DO) is a crucial controlling parameter in this process as the DO affects the efficiency of the treatment, operational cost and system sustainability. While Field electrodes measure DO, maintaining regular and accurate records remains a challenge, even in the absence of aggressive climate conditions. Most of the DO probes are sensitive and require careful maintenance and calibration that may lead to non-accurate data recording. This paper discussed the validity of statistical models to impute missing DO values. Multivariate Imputation by Chained Equations (MICE) is used in R programming language to predict missing values and replace outliers in the DO measurements. The performance of this technique was very good as the correlation ranges between 0.999 and 0.821 when the missing values range from 1 to 15 % (424 to 6,360 of total data of 42,420). The method is highly effective and ready to be used for many other applications. Kohonen Self Organising Map (KSOM) is applied as an alternative model to verify the accuracy of the MICE.

Keywords: Activated sludge process; Dissolved Oxygen (DO); Multiple Imputations; Chained Equations

INTRODUCTION

The concentration of dissolved oxygen (DO) is a key variable in water resource recovery facilities [1]. DO is monitored to ensure oxygen is sufficient for aquatic species to survive [2]. In wastewater treatment, dissolved oxygen monitoring in aeration tanks in the activated sludge process is essential as it significantly impacts the efficiency of the treatment process.

There are two methods to measure the DO concentrations: Winkler titration method [3].or by oxygen electrode sensors [1]. DO sensors are the most accurate method; however, instrumental errors are inevitable, and a certain procedure should be followed to obtain reliable data [4].

When the DO concentration levels are reduced, the number of filamentous microorganisms increases, which subsequently impacts the ability of the activated sludge to settle. If DO levels continue to decrease, effluent turbidity increases and treatment will be reduced rapidly [5].

This study was based on Seafield wastewater treatment plant data in Edinburgh, UK). The Seafield wastewater treatment plant (WWTP) is the largest WWTP in Scotland and treats more than 300 million litres of wastewater per day. The main areas of Seafield WWTP Comprise of Eight sedimentation tanks with a volume of 9260 m^3, Eight final settlements tanks with a volume of 3974 m^3, and five rectangular aeration lanes. Data of DO vales in these five lanes are used in this paper.

METHODOLGY

Outlier Detection

Data values with magnitudes that differ significantly from the majority of other data values are known as outliers. There are many possible reasons for outliers. Commonly, human errors such as errors in transferring the data or a fault in handling the measuring

sensors can lead to unintentional outliers [6]. Identifying the outliers from the data sets and replacing them so the model will create more uniform sets of data, and accordingly, the results for the model performance will be more accurate. A Z-score test was used to identify the outliers. A Z-score test in Eq. 1 and 2 are used used to obtain a Z-score [6].

$$z_i = \frac{(x_i - \bar{x})}{s} \quad (1)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (2)$$

Where

x_i Observed DO value,

$\bar{x}$ Mean for all observed DO,

S Standard deviation,

N sample size (total number of the data).

As per the Z-score equation the z-score is 3.5 < Z < -3.5. The values beyond this range should be marked as an outlier and be concidered as missing values to be imputed by the models later.

Missing Data

Missing data are two types, ignorable and non-ignorable [7].

1- Ignorable Missing Data are two types:
   a. Missing Data Completely at Random (MCAR): they occur randomly, meaning the value of the missing and observed data is independent.
   b. Missing at Random (MAR): The value of the missing data and observed data might be dependent.
2- Non-Ignorable Missing Data: If the missing data are not at random (MNAR), then they are considered non-ignorable. The MNAR are when the value of the unobserved variable itself predicts missingness [8].

Multiple Imputations

The concept of Multiple Imputation (MI) uses the distribution of the observed data to estimate and predict a set of the possible values for the missing data. With MI, the missing values are replaced by many different values, and accordingly, many different full data sets are performed. Complete data sets will be created by multiple imputations, and these sets should slightly differ. The final step of the MI is to pool out from these imputed steps one complete data set as a final result [9].

Multivariate Imputation By Chained Equations (MICE)

Multivariate imputation by chained equations (MICE), also known as Sequential Regression Imputation [10]. **It** is a practical approach for imputing missing datasets based on a set of imputation models. MICE is a chain of equations used to impute missing values and is fully conditional. The missing values are imputed one by one. For example, in a data set Y_1 ,Y_2 ,Y_3,…….,Y_{10} , if Y_1 is a missing value, MICE will use the information from all other variables Y_2 ,Y_3,…….,Y_{10}, and the same procedure will be repeated for all missing values. The imputed values are modified by adding a residual error which can be added directly to the imputed values or to the parameters estimates of the MICE model. The residual error will improve the sampling variability to the imputations [11]. MICE assumes that missing data are missing at random[12]. MICE procedure is provided in many software, such as SPSS, STATS and S-Plus [13]. However, the most common software is R [11].

Kohonen Self-Organizing Map (KSOM)

KOSM is a nonlinear classification technique able to recognise a specific group from a complex data set [6]. The KSOM has been used in various research fields, as discussed in [6] [14]. KSOM converts the input signal pattern of complex data into a two-dimensional map. The tuned input pattern are shown as clusters where similar patterns are shown in the same output neurons or with its adjacent one; these

clusters reduce the amount of huge data, and accordingly, the relationship between higher-dimensional data will be in lower dimension [6]. The KOSM consists of two layers, the input layer and output layer, the layers are composed of neurons, and both layers are connected with a weight vector. The input layer consists of multi-dimensional layers, and the output layer consists of two-dimensional layers made of neurons. A specific dimensional weight vector represents each neuron in the two-dimensional map. Similar neurons are located together, and neurons which are not similar will be further apart. Map size varies depending on the number of neurons; the number of neurons affects the accuracy of KOSM. More details on the theory and application of the KSOM can be found in [6][15].

RESULTS AND DISCUSSION

Results for the predicted values of the dissolved oxygen by MICE and KOSM are compared with the observed one. The accuracy of the predicted values was evaluated by three statistical parameters namely the Correlation Coefficient R, the mean square error (MSE), and the Average Absolute Error AAE [6].

Figure 1 illustrates the component planes for the Kohonen self-organising map (KSOM) for all lanes when missing values are 15%; these planes visually show the relationship between DO values for each lane. As shown, lanes 1,2 and 3 are slightly correlated, and lanes 4 and 5 are less correlated. The more correlated vectors between each other lanes indicate better results.

Table 1 shows the results for MICE, and KOSM when 1 % up to 75 % of data are missing. The table summarises the results as an average for the five lanes. The values for R, MSE, and AAE almost have the same values for MICE and KOSM. The performance of the models is very good. The R values for MICE and KOSM are 0.884 and 0.919, the MSE is 0.308 and 0.211, the AAE is 0.158 and 0.124.

CONCLUSION

The performance of MICE is excellent when the missing data are 10% or less, very good performance when the missing data is between 10% and 15%, good performance when the missing data is between 15% and 20% and poor performance when the missing data more than 20%. KSOM was used as an alternative model to validate the MICE results. KSOM is applied in previous case studies, and therefore, KSOM results which are similar to MICE results, emphasise MICE application to predict the missing values.

LIMITATIONS AND RECOMMENDATIONS FOR MICE

Data quality is the major factor which will affect the accuracy of the predicted values. Replacing the outlier will improve the accuracy however if the outliers are more than 10% of the overall data, the accuracy of the predicted values after replacing the outliers will not be improved.

MICE can be applied regardless of the type of missing data, e.g. MAR, MCR, and MNAR. However, better performance for MICE when the missing data is MAR. The number of imputations affects the accuracy of the predicted values; 5 to 10 imputations are the default number of imputations; however, for the data with a big variance, more imputations should be applied. Increasing the number of imputations may have an impact on the software speed; delay is expected when the number of imputations is more than10.

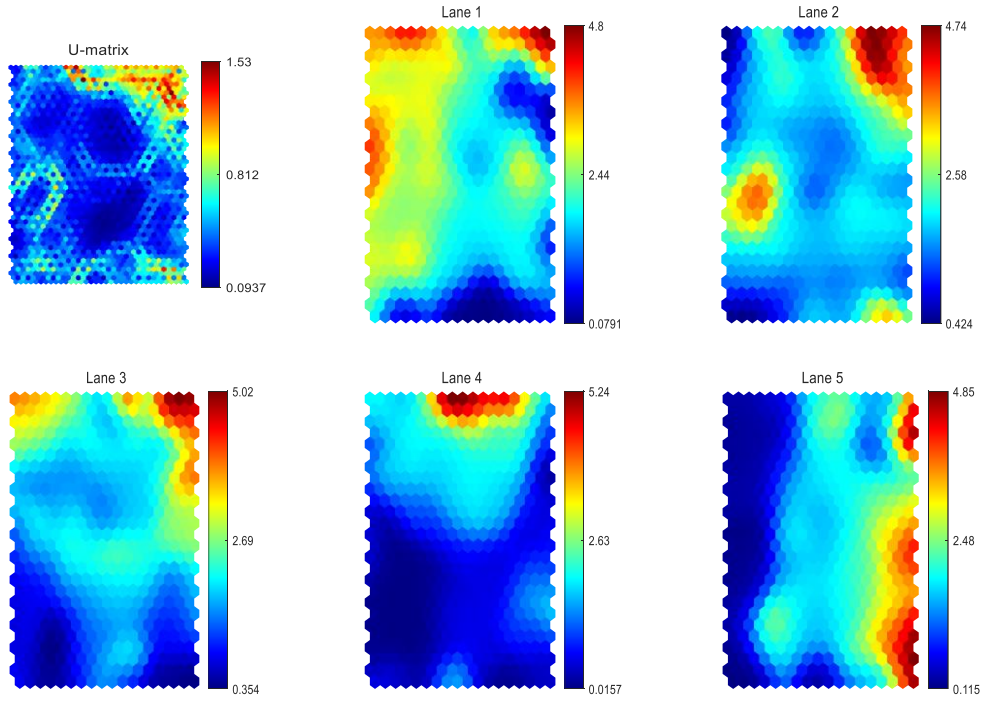

Figure 1 The component planes of the KSOM for all lanes when missing values are 15%

Table 1 Summary of Statistical values for MICE and KOSM performance

	Results when Dates are Missing			
	Mice		KOSM	
Percentage Of Missing Date	Statistical Parameters	Value	Statistical Parameters	Value
Results when 1% of Dates are Missing	R	0.994	R	0.997
	MSE	0.015	MSE	0.009
	AAE	0.009	AAE	0.007
Results when 5 % of Dates are Missing	R	0.96	R	0.975
	MSE	0.107	MSE	0.066
	AAE	0.054	AAE	0.039
Results when 10% of Dates are Missing	R	0.929	R	0.948
	MSE	0.21	MSE	0.136
	AAE	0.108	AAE	0.081
Results when 15% of Dates are Missing	R	0.884	R	0.919
	MSE	0.308	MSE	0.211
	AAE	0.158	AAE	0.124
Results when 20% of Dates are Missing	R	0.837	R	0.884
	MSE	0.435	MSE	0.302
	AAE	0.221	AAE	0.173
Results when 30% of Dates are Missing	R	0.837	R	0.794
	MSE	0.435	MSE	0.506
	AAE	0.221	AAE	0.305
Results when 40% of Dates are Missing	R	0.633	R	0.686
	MSE	0.945	MSE	0.729
	AAE	0.469	AAE	0.422
Results when 50% of Dates are Missing	R	0.63	R	0.791
	MSE	1.16	MSE	0.512
	AAE	0.599	AAE	0.389
Results when 60% of Dates are Missing	R	0.519	R	0.689
	MSE	1.538	MSE	0.739
	AAE	0.752	AAE	0.506
Results when 75% of Dates are Missing	R	0.235	R	0.291
	MSE	1.687	MSE	1.339
	AAE	0.864	AAE	0.782

REFERENCES

1- Samuelsson, O., 2021. *Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities* (Doctoral dissertation, Acta Universitatis Upsaliensis).

2- Holenda, B., Domokos, E., Redey, A. and Fazakas, J., 2008. Dissolved oxygen control of the activated sludge wastewater treatment process using model predictive control. *Computers & Chemical*

3- Winkler, L.W., 1888. The determination of dissolved oxygen in water. *Berlin DeutChem Gas*, *21*, pp.2843-2855.

4- Wei, Y., Jiao, Y., An, D., Li, D., Li, W. and Wei, Q., 2019. Review of dissolved oxygen detection technology: From laboratory analysis to online intelligent detection. *Sensors*, *19*(18), p.3995.

5- Rustum, R., 2009. *Modelling activated sludge wastewater treatment plants using artificial intelligence techniques (fuzzy logic and neural networks)* (Doctoral dissertation, Heriot-Watt University).

6- Rustum, R., Adeloye, A. and Simala, A., 2007. Kohonen self-organising map (KSOM) extracted features for enhancing MLP-ANN prediction models of BOD5. In International Symposium: Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management-24th General Assembly of the International Union of Geodesy and Geophysics (IUGG) (pp. 181-187).

7- *Rubin, D.B., 1987. Statistical analysis with missing data. Wiley.*

8- Bouhlila, D.S. and Sellaouti, F., 2013. Multiple imputation using chained equations for missing data in TIMSS: a case study. *Large-scale Assessments in Education*, *1*(1), pp.1-33.

9- Harel, O. and Zhou, X.H., 2007. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, *26*(16), pp.3057-3077.

10- Van Buuren, S. and Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, pp.1-67.

11- Zhang, Z., 2016. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine*, *4*(2).

12- Ratolojanahary, R., Ngouna, R.H., Medjaher, K., Junca-Bourié, J., Dauriac, F. and Sebilo, M., 2019. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications*, *131*, pp.299-307.

13- Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), pp.40-49.

14- Rustum, R. and Adeloye, A.J., 2007. Replacing outliers and missing values from activated sludge data using Kohonen self-organising map. *Journal of Environmental Engineering*, *133*(9), pp.909-916.

15- Kumar, N., Rustum, R., Shankar, V. and Adeloye, A.J., 2021. Self-organising map estimator for the crop water stress index. *Computers and Electronics in Agriculture*, *187*, p.106232.