H2Open Journal

IWA PUBLISHING

H2Open Journal Vol 6 No 2, 280 doi: 10.2166/h2oj.2023.013

Infilling missing data and outliers for a conventional sewage treatment plant using a self-organizing map: a case study of Kauma Sewage Treatment Plant in Lilongwe, Malawi

Madalitso H. Mng'ombe[a,b], Brighton Austin Chunga [a,b,*], Eddie W. Mtonga[a], Russel C. G. Chidya[a] and Mphatso Malota[a]

[a] Department of Water and Sanitation, Faculty of Environmental Sciences, Mzuzu University, Private Bag 201, Mzuzu, Malawi
[b] Hydro-Informatics Engineering Centre, Private Bag 40176, Kanengo, Lilongwe, Malawi
*Corresponding author. E-mail: bachunga@gmail.com

BAC, 0000-0002-5681-1537

ABSTRACT

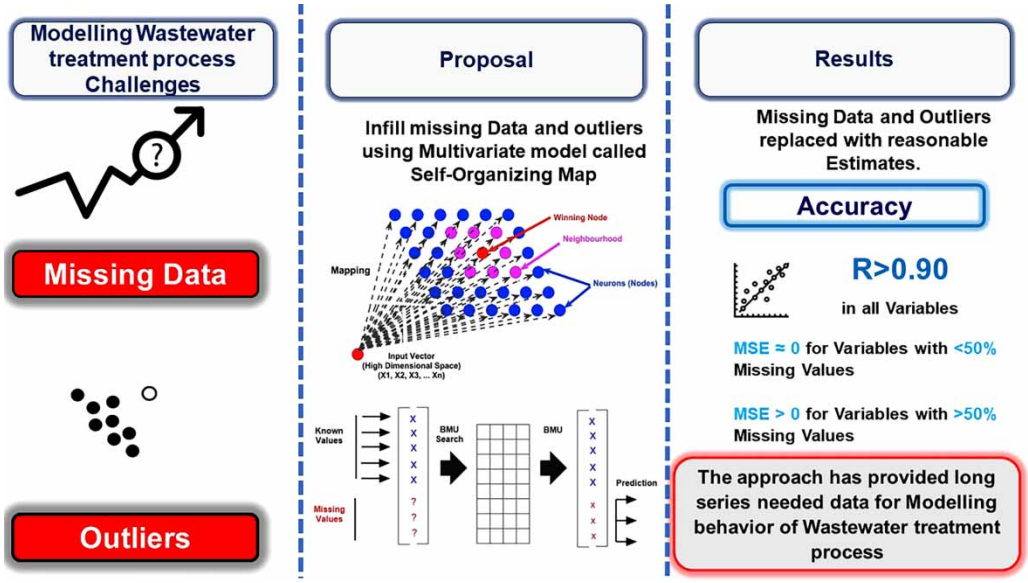
Data availability is key for modeling of wastewater treatment processes. However, process data are characterized by missing values and outliers. This study applied a self-organizing map (SOM) to fill in missing values and replace outliers in wastewater treatment data from Kauma Sewage Treatment Plant in Lilongwe, Malawi. We used primary and secondary wastewater data and executed the SOM algorithm to fill missing values and replace outliers in effluent pH, biochemical oxygen demand, and dissolved oxygen. The results suggest that the SOM algorithm is reliable in filling gaps in wastewater time series data with less than 50% missing values with correlation coefficient (R) values of >0.90. The SOM algorithm failed to reliably fill gaps and replace outliers in time series data with $>50\%$ missing values. For instance, high mean square error (MSE) values of 3,655.57, 10.62, and 2,153.34 for pH, DO, and BOD, respectively, were registered in datasets with more than 50% missing values, while very small MSE values ($MSE \approx 0$) were associated with effluent pH, BOD, and DO data with missing values of $>50\%$. Practitioners can use this approach to improve the planning and management of wastewater treatment facilities where available data records are riddled with missing observations.

Key words: artificial neural network, data management, sanitation, self-organizing map, wastewater treatment

HIGHLIGHTS

- Missing data impinge on wastewater treatment plant processes efficiency.
- The advancement of information technology and artificial intelligence enables the infilling of missing data.
- We proposed to infill missing data and outliers using a Multivariate model called the Self-Organizing Map.
- Missing data and outliers are replaced with reasonable estimates.
- The approach has provided long series data for modelling the behavior of the wastewater treatment process.

GRAPHICAL ABSTRACT



INTRODUCTION

The wastewater treatment (WWT) process aims to achieve effluent and sludge quality that is environmentally safe for disposal/or reuse (Larsen *et al.* 2013). Optimal wastewater treatment plant (WWTP) operation and control can be achieved by developing a robust mathematical tool that enables the prediction of the quality of treated effluent based on past observations of certain key parameters (Hassen & Asmare 2018). However, numerous challenges arise when modeling wastewater treatment processes including lack of reliable wastewater quality data. The lack of good quality data is attributed to the unavailability of equipment, and the high cost of procedures that facilitate the measurement of wastewater quality characteristics (Hassen & Asmare 2018). Equipment malfunction and human errors also lead to gaps and outliers in data records. Gaps and outliers in data records pose a challenge in model identification, calibration, and verification (Rustum 2009). One of the common solutions is simply to remove records containing missing values and outliers, rendering them unreliable for modeling purposes. Considering the time limitations, cost of data collection, and scarcity of available data, removing such records cannot be a viable option (Mwale *et al.* 2012). As a result, significant data pre-processing is required to fill the gaps or identify outliers in the data record (Rustum 2009). This is particularly important, especially in developing countries like Malawi where the availability of data without missing values is a challenge.

The existing methods for dealing with missing values include replacing missing values with the mean, median, or regression models for the data (Rustum & Adeloye 2011). Outliers can be resolved by using trimmed means, and other scale estimators besides standard deviation such as Median of Absolute Deviation (MAD) and Winsorization (Rustum 2009). However, using these models to predict missing values or outliers in a long time series is difficult and frequently unreliable (Zhu *et al.* 2018). This is particularly challenging when the number of values to be filled is relatively high in comparison to the total record length.

These challenges can be resolved by using computing techniques such as artificial intelligence (AI). The most promising approaches in this class of techniques include Artificial Neural Networks (ANNs), Fuzzy Logic (FL), and Genetic Algorithms (GAs). The application of AI in data cleaning and management is well-established in both water resources and hydrology (Rustum 2009; Nkiaka *et al.* 2016). The ANNs are the most popular algorithms among AI classes as they use the same available data to learn about the behaviour of a time series. Furthermore, ANNs have the capability of modeling complex nonlinear systems as they do not require prior knowledge of the system process(s) under study. In addition to that, the ANNs have proven robust even in the presence of missing observations in time series (Mwale *et al.* 2012).

Multilayer Perceptron (MLP) is one of the most widely used algorithms within the ANN family. Despite their robustness in filling missing gaps in time series, the MLP demands a long time series for training (Rustum 2009; Mwale *et al.* 2012). Accordingly, additional pre-processing of the time series is required in order to provide estimates in the input space before the training can begin (Rustum 2009). This is very important especially when

significant portions of the time series to be used for training have incomplete data, or fall short of time series to facilitate training (Nkiaka *et al.* 2016). Dealing with time series data with many missing values is also computationally intensive as it requires additional storage memory (Kalteh & Berndtsson 2011).

Kohonen Self-Organizing Maps (KSOMs) simply called Self-Organizing Maps (SOMs) are another member of the ANN class. The SOMs are competitive and unsupervised ANN (Rustum 2009). SOMs are becoming popular in filling missing values in time series and have proven more effective than ANN-MLP (Mwale *et al.* 2012; Nkiaka *et al.* 2016). Many studies have successfully applied SOMs to fill gaps in time series with satisfactory results (e.g. Mwale *et al.* 2012; Nkiaka *et al.* 2016; Kumar *et al.* 2021a). Despite this widespread use in many studies around the world, the use of SOMs in Malawi, particularly in wastewater treatment processes where data quality is a concern, has been limited. This study tested the reliability of the SOM to predict missing values and replace outliers in time series data for the Kauma Sewage Treatment Plant situated in Lilongwe, Malawi. This work formed part of an ongoing research project that intended to apply AI algorithms in modeling the performance of the Kauma Sewage Treatment Plant.

MATERIALS AND METHODS

Description of SOMs

Kohonen pioneered the use of SOMs (Kangas & Kohonen 1996; Kohonen *et al.* 1996). The success of SOMs in many research disciplines has led to their widespread use in water resources processes and systems research, particularly for data mining, infilling of missing data, estimation and flow forecasting, and clustering (Kalteh & Berndtsson 2011). This is because SOMs have the ability to convert nonlinear statistical relationships from high-dimensional data onto a low-dimensional display (Ismail *et al.* 2011). In the output space, data points with similar characteristics are clustered together or placed close to each other. This mapping method preserves the most important topological and metric relationships that are present in the original data (Rustum & Adeloye 2007). The ability of SOMs to cluster data together makes them robust for data mining and infilling datasets with gaps and outliers as the gaps/outliers are replaced by values that possess similar features to the rest of the values displayed in the map (Adeloye *et al.* 2012). The SOM algorithm executes assigned tasks using an unsupervised and competitive learning approach to discover patterns in data (Kalteh *et al.* 2008). Thus, the entire process is data-driven. A SOM is composed of two layers: an input layer with multiple dimensions and an output layer. Both layers are fully connected by weights that can be adjusted. The output layer is made up of neurons that are arranged in a two-dimensional grid of nodes (Figure 1). Each neuron in the SOM's output layer contains the same set of variables as the input vectors. Further details of SOMs are provided by Rustum & Adeloye (2007).

Description of the study area

This study was conducted at the Kauma Sewage Treatment Plant located in Lilongwe; the capital of Malawi (Figure 2). The plant receives wastewater from the sewer network which covers the following areas of the city (Areas 3, 6, 12, 13, 16, 18, 19, 20, 30, 47, and 48). Septage lagoons within the treatment facility (Figure 3) receive

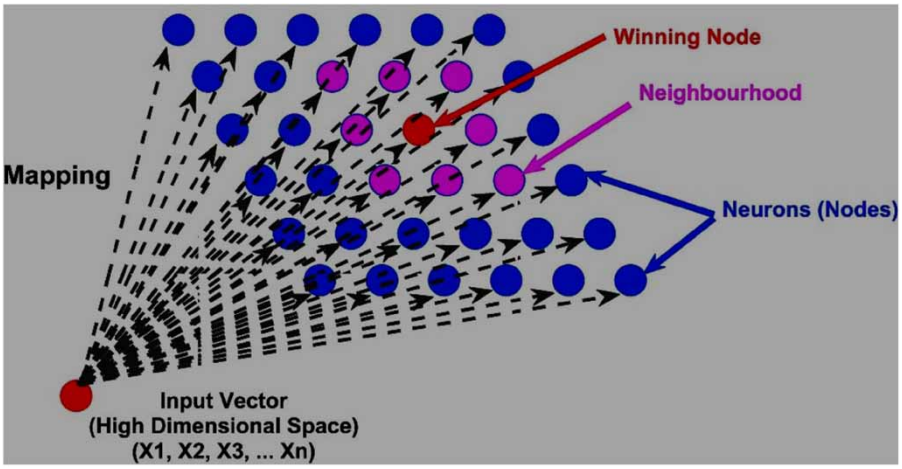


Figure 1 | Architecture of SOM (*source:* Rustum 2009).

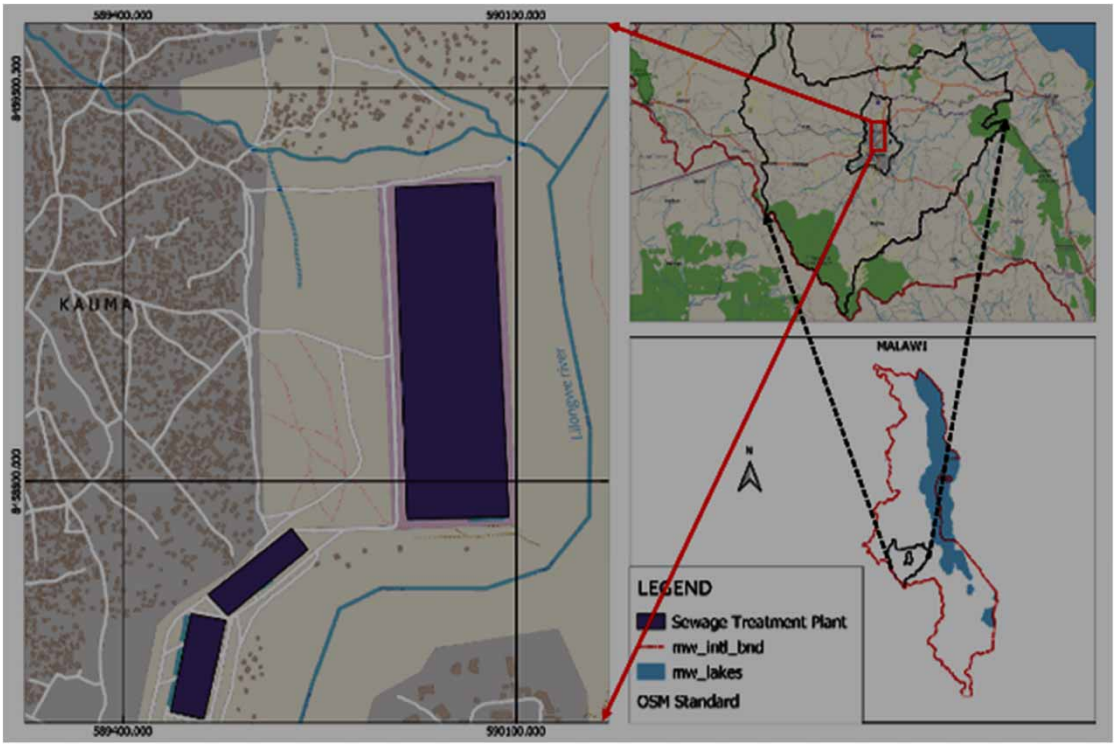


Figure 2 | Map of Malawi showing the study area.

fecal sludge which comes from different non-sewered areas of the city using vacuum trucks that belong to different private operators.

Sampling and data collection procedures

The study used both secondary and primary data. In this study, only domestic sewage was of interest. Secondary data was obtained from document reviews of Kauma sewage treatment plant data. Primary data were collected

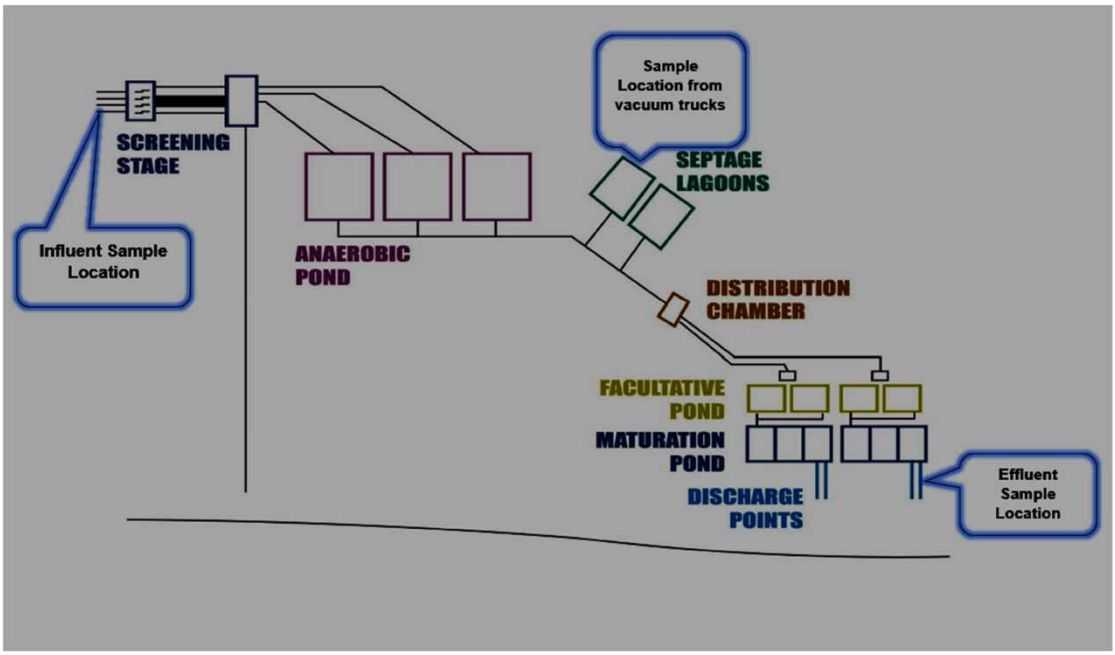


Figure 3 | A schematic of Kauma Sewage Treatment plant. Not drawn to scale. (Adapted from Mtethiwa *et al.* 2008; Ravina *et al.* 2021).

twice (morning and evening) daily for 30 days from 11 February 2022 to 17 March 2022. Wastewater samples were collected from the influent raw wastewater. This was followed by analysis of the influent raw wastewater composite samples characteristics such as including pH, chemical oxygen demand (COD), total dissolved solids (TDS), total suspended solids (TSS), electrical conductivity (EC), and dissolved oxygen (DO). The analyses were conducted following the standard methods for measuring characteristics of water and wastewater as highlighted by APHA (2017). Only COD and biochemical oxygen demand (BOD) were obtained from samples that were collected from the septage lagoon and effluent-treated wastewater. At the septage, lagoon samples were taken during the emptying of the sludge, and to ensure that the samples were not industrial sewage, a deliberate effort was made to follow the source of the sludge. For each truck, four samples of 2-L each were collected, where one sample was collected at the beginning of emptying and two samples were collected in the middle of emptying, followed by one sample at the end of the emptying exercise. The samples were then mixed, also regarded as double sampling, to obtain the most homogenous characteristics. Double sampling was considered for quality assurance.

Implementation of the SOM algorithm

The implementation strategy of the SOM algorithm used in this study is fully adopted from Rustum (2009). A SOM algorithm is executed in a series of steps. The multidimensional input data is first standardized to ensure that the map is not dominated by variables with extremely high or low values. The standardization process gives equal weight to all input variables because SOMs use Euclidian metrics to measure distances between vectors (Vesanto *et al.* 2000). In this study, normalization was done by deducing the mean, and later dividing it by the standard deviation to calculate a transformed variable with a mean of 0 and variance of 1. Thereafter, input vector was randomly chosen and presented to each neuron for comparison with their respective weight vectors to identify the weight vector with the highest similarity to the presented input vector. The identification of the weight vector employs the Euclidian distance as given in Equation (1).

$$D_i = \sqrt{\sum_{j=1}^{n} m_j (x_j - w_{ij})^2}; \quad i = 1,\ 2,\ 3 \ldots M \tag{1}$$

where D_i refers to the Euclidian distance between the input vector and the weight vector i; $x_j = j$ element of the current vector; $w_{ij} = j$ element of the weight vector i; $n =$ the dimension of the input vector; $m_j =$ 'mask'. Where the input vector contains missing elements, the mask is set to zero for such elements, thereby allowing the SOM algorithm to handle missing elements in the input vector with ease. The winning node or best matching unit (BMU) (Figure 4) is the neuron whose vector closely matches the input vector (i.e. with D_i minimum).

Following the discovery of the BMU, the winner neuron's weight vector is adjusted so that the BMU and its adjacent neurons move closer to the input vectors in the input space, increasing the agreement between the

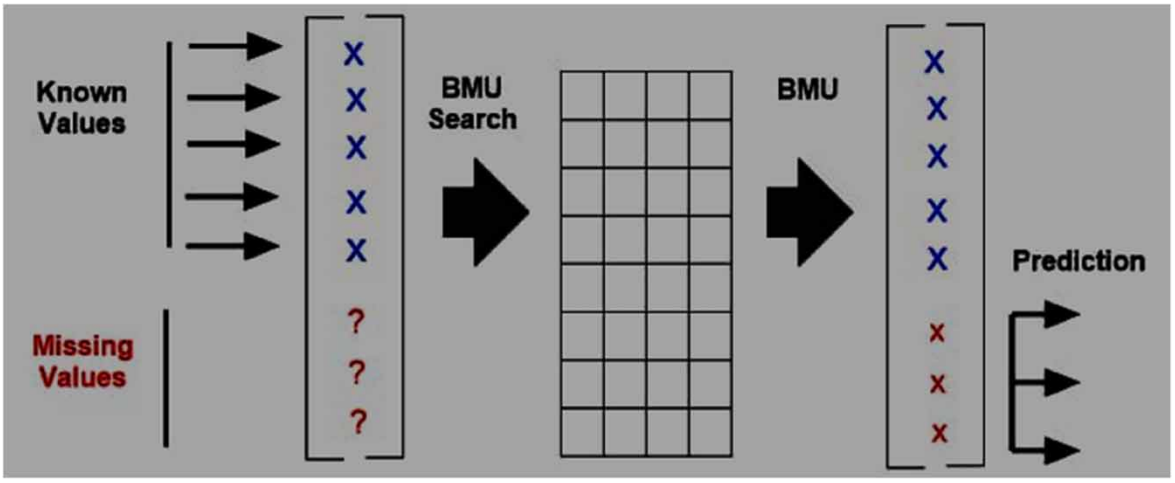


Figure 4 | Prediction of missing components of the input vector using the KSOM (BMU, Best Matching Unit) (*Source:* Rustum & Adeloye 2011).

input vector and the weight vector. The Equation (2) is used to make such adjustment:

$$w_t(t+1) = w_t(t) + \alpha(t)h_c i[x(t) - w_t(t)] \tag{2}$$

where w_t = element of the weight vector; t = time; $\alpha(t)$ = learning rate at time t; $h_{ci}(t)$ = neighborhood function centered in the winner unit c at time t.

Thereafter, each node in the map learns to recognize input vectors that are similar to its own. This ability is known as self-organizing because no external information is required for this process to occur. The learning procedure is repeated until the SOM algorithm achieves convergence. In Equation (3), the learning rate decreases monotonically as the number of iterations increases.

$$\alpha(t) = a_0\left(\frac{0.005}{\alpha_0}\right)^{\frac{t}{T}} \tag{3}$$

where $\alpha(t)$ = learning rate, α_0 = initial learning rate, and T = training length.

In this case, the neighborhood function is Gaussian, centered in the winner unit c, and is calculated using Equation (4).

$$h_{ci}(t) = \exp\left\{-\frac{||r_c - r_i||^2}{[2\sigma^2(t)]}\right\} \tag{4}$$

where $h_{ci}(t)$ = neighborhood function centered on the winner unit c at time t, r_c and r_i = positions of nodes c and i on the SOM grid, and $\alpha(t)$ is the neighborhood radius, which decreases monotonically as the number of iterations increases.

The total average quantization error and total topographic error are used to assess the quality of the trained SOM. The average quantization error indicates how well the map fits the input data (it calculates the average distance between each data vector and its BMU). The smaller the quantization error, the smaller the average distance between the vector data and the prototypes, indicating that the data vectors are closer to their prototypes; it is a positive real number with a value close to zero indicating a good fit between the input and the map. The quantization error is computed as follows:

$$q_e = \frac{1}{N}\sum_{i=1}^{n}||X_i - W_{ic}|| \tag{5}$$

where q_e denotes quantization error, N denotes the number of input vectors used to train the map, X_i denotes the ith data sample or vector, and W_c denotes the prototype vector of the BMU for X; the Euclidian distance is denoted by ||.||.

Topographic error (t_e) (Equation (6)) quantifies how well the map preserves the topology of the data by considering the map structure. The lower the topographic error, the better the SOM preserves the data's topology and has positive real numbers ranging from 0 to 1, with a value close to 0 indicating high quality.

$$t_e = \frac{1}{N}\sum_{i=1}^{N}u(X_i) \tag{6}$$

where N = number of input vectors used to train the map, and u = binary integer equal to 1 if the first and second BMU for X_i is not adjacent units, otherwise zero.

Considering that there is always a trade-off between the quantization error and topographic error, in which one needs to be minimized at the expense of the other, the study focused on reducing topographic error to ensure that the infilled values reflect the seasonal trend of the different time series. To assess the quality of the newly generated time series, the coefficient of determination (R) (Equation (7)) and mean square error (MSE) (Equation (8)) were used. The R values denote the proportion of one variable's variance that is predictable from the other

variable and ranges from 0 to 1.

$$R = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}[(xi = \bar{x})^2]\sum_{i=1}^{n}[(y_1 - \bar{x})^2]} \quad (7)$$

where x_i denotes the ith observed value; y_i denotes the ith trained value, $\bar{x}$ is the the mean of the observed value, $\bar{y}$ is the mean of the trained value, and n denotes the number of observations.

On the other hand, the MSE is the average of the squares of the model predictions' errors.

$$MSE = \frac{\sum (x_i - x_i')^2}{N} \quad (8)$$

where $x_i^{'}$ is the model predicted value; x_i is the actual value; N is the number of samples.

Setting of SOM algorithm parameters

According to Gabrielsson & Gabrielsson (2006), the radius of the SOM should be chosen wide enough at the start of the learning process to allow the map to be ordered globally as the radius decreases monotonically over time. Garcia & Gonzalez (2004) propose that, if M is the total number of input elements, the number of neurons in the output can be calculated as follows:

$$N = 5\sqrt{M} \quad (9)$$

where M denotes the total number of samples and N denotes the number of neurons.

Garcia & Gonzalez (2004) propose that once N is known, the number of rows and columns of N can be calculated by using Equation (10).

$$\frac{l_1}{l_2} = \sqrt{\frac{e_1}{e_2}} \quad (10)$$

where l_1 and l_2 represent the number of rows and columns, e_1 is the largest eigenvalue of the training dataset, and e_2 is the second largest eigenvalue.

Bearing in mind that the learning process involved in the computation of a feature map is stochastic, the accuracy of the map depends on the number of iterations executed by the SOM algorithm during the initialization phase of the algorithm (Gabrielsson & Gabrielsson 2006). These authors recommend that for good statistical accuracy, the number of iterations is at least 500 times the number of network nodes. The default SOM software parameters for map size and lattice (rows and columns) were used, which were the same as using Equations (9) and (10).

The infilling process was completed through the following steps:

Step 1: *Data collection and normalization:* The data to be filled (i.e., wastewater quality data) was gathered and standardized; these were the depleted input vectors.

Step 2*: Training:* To form the SOM, the depleted input vector (data matrix) was introduced into the iterative training procedure. Weight vectors were initialized at the start of training using both a random and a linear initialization method. The comparison and adjustment processes were repeated until the optimal number of iterations was reached or the specified error criteria were met.

Step 3: *Extraction of data from the trained SOM:* All minimum Euclidian distances were examined; this was followed by examining the SOM's BMU for the depleted input vector (i.e. with missing values and outliers). Because the BMU identified in this step was a trained SOM node, it was assumed to have the complete complement of missing values.

Step 4: *Missing value replacement:* At this stage, missing values and outliers of the input depleted vector were replaced with the corresponding values in BMU identified in the above step.

The SOM model was developed and validated using data from 2015 to 2022. Initially, the SOM toolbox was used to train the model with default values of learning rate *($a_0 = 0.5$)* and neighborhood radius ($a_0 = max(l_1,l_2)/4$) where l_1 *and* l_2 are the dimensions of the map computed using Equation (10). The toolbox computes the size (number of units or neurons) of the map using Equation (9), but the final units on the map *(M)* were adjusted to equal the product of l_1 *and* l_2. The map size of the SOM model $M = 126$ units with dimensions of 14×9. The final quantization and topographic errors were 0.955 and 0.247, respectively.

Application of SOM

A SOM toolbox version 2.1 developed at Helsinki University of Technology Finland (www.cis.hut.fi/projects/somtoolbox/) was used in the MATLAB® 2021a environment for the application of the SOM algorithm for infilling missing data and outliers (The Maths Works, Inc. 2021). In this analysis, a batch training algorithm was used because its implementation in MATLAB is considerably more efficient than that of the sequential training algorithm as it requires less time for training and produces less quantization and topographic errors. The information was presented in columns, with each column representing a wastewater quality parameter. To meet MATLAB® data entry requirements, entries without data and outliers were recorded as NaN (Not a Number). To train all of the data in a single simulation, the data entries should overlap so that there is no single day/month with no data entry for all the wastewater quality parameters.

Ethical consideration

The study sought clearance from the Mzuzu University Research Ethics Committee (MZUNIREC) Ref No: MZUNIREC/DOR/21/62. Permission was also obtained from Lilongwe City Council to engage laboratory technicians during data collection processes. Informed consent was also obtained from the laboratory technicians.

RESULTS AND DISCUSSIONS

Descriptive statistics of wastewater quality parameters for Kauma Sewage Treatment Plant

Table 1 presents a summary of the descriptive statistics of Kauma Sewage Treatment Plant Data. A total of 616 data samples obtained covered a period of 6 years (2015–2021). A standard deviation (SD) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out. High standard deviation (SD = 2,798.18) value was observed from influent COD from the septage lagoon while influent pH had the lowest standard deviation (SD = 0.46). As shown in Table 2, there were large numbers of missing values that could not be thrown away. Assuming that the data have a normal distribution, the mean and standard deviation of the entire dataset are used to obtain a *Z*-score of each data point, while in the modified Z-score, the median of absolute deviation about the mean (MAD) is used instead of standard deviation to obtain modified Z-score of each data point (Rustum & Adeloye 2007). In the present study, the modified Z-score method identified more outliers than either the visual inspection or the Z-score method. All identified outliers by the modified Z-score method were deleted and treated as missing values to be estimated. Finally, influent pH, influent DO, effluent BOD, and effluent COD were recorded to have more than 50% proportion of missing data and outliers.

Table 1 | Computed descriptive statistics of Kauma Sewage Treatment Plant data

Parameter	Unit	Mean	SD	SE	Max	Min	UB	LB
pH_{inf}	–	7.01	0.46	0.02	8.00	5.40	7.05	6.97
$Temp_{inf}$	°C	24.73	1.83	0.07	29.00	20.40	24.88	24.58
BOD_{5inf}	mg/l	228.47	41.33	2.08	450.00	74.00	232.57	224.38
COD_{inf}	mg/l	358.34	88.49	4.56	552.70	182.00	367.31	349.37
BOD_{inf} SL	mg/l	821.67	542.71	84.76	2,329.5	109	992.96	650.36
COD_{inf} SL	mg/l	2,615.11	2,798.18	437.00	14,090.88	826.56	3,498.32	1,731.89
TDS_{inf}	mg/l	465.15	86.07	3.71	739.00	230.00	472.44	457.87
TSS_{inf}	mg/l	173.28	11.87	0.73	199.00	146.00	174.72	171.85
EC_{inf}	µS/cm	783.83	118.96	4.86	1,070.00	441.00	793.38	774.28
$TURB_{inf}$	NTU	9.649	0.581	0.023	11	8	9.695	9.603
DO_{inf}	mg/l	1.12	0.99	0.04	3.21	0.07	1.21	1.03
BOD_{5eff}	mg/l	22.06	7.16	0.36	70.00	5.00	22.76	21.36
COD_{eff}	mg/l	40.41	12.46	0.63	58.20	20.00	41.65	39.17

SD, standard deviation; SE, standard error; UB, upper bound of 95%. Confidence interval for the mean; LB, lower bound of 95% confidence interval for the mean; BOD, biochemical oxygen demand; COD, chemical oxygen demand; TDS, total dissolved solids; TSS, total suspended solids; EC, electrical conductivity; DO, dissolved oxygen; Temp, temperature; TURB, turbidity. *Suffixes:* inf, influent; eff, effluent; SL, Septage Lagoons.

Table 2 | Proportion of missing data and outliers

Parameter	Unit	Number of missing values	Number of outliers			The proportion of missing data and outliers (%)
			Visual inspection	*Z*-score	Modified *Z*-score	
pH_{inf}	–	298	8	22	76	60.71
$Temp_{inf}$	°C	81	45	29	76	25.40
BOD_{5inf}	mg/l	221	4	88	83	49.43
COD_{inf}	mg/l	239	0	0	44	46.02
BOD_{inf} SL	mg/l	44	3	1	81	20.33
COD_{inf} SL	mg/l	114	1	0	30	23.41
TDS_{inf}	mg/l	77	2	2	74	24.55
TSS_{inf}	mg/l	17	0	11	73	14.63
EC_{inf}	µS/cm	209	4	2	82	47.32
$Turb_{inf}$	NTU	28	22	25	49	12.50
DO_{inf}	mg/l	298	15	31	66	59.09
$BOD5_{eff}$	mg/l	351	2	0	28	61.63
COD_{eff}	mg/l	226	0	0	101	53.17

SOM component planes

The development of the component planes, which allows visualization of the correlation between the variables, is a significant feature of the SOM. Figure 5 depicts the component planes for each variable in the SOM. Each plane is a sliced SOM with a single vector variable that represents its value in each map unit (Kalteh *et al.* 2008). The component planes are filled with colored or grey shades to reflect the feature values of each SOM unit in the 2-D lattice, with the darker the color indicating a lower relative value of the corresponding variable component. The component planes visually indicate the regions where a variable is high, low, or average in this manner (Kumar *et al.* 2021a). This facilitates visual interpretation of the correlation between SOM simulated values of selected wastewater parameters.

Visual analysis of the component planes shows that the color (or gray) gradient of the plane of BOD_{5inf} is parallel to the gradient of COD_{inf} with high values of BOD_{5inf} being correlated with high values of COD_{inf} and vice-

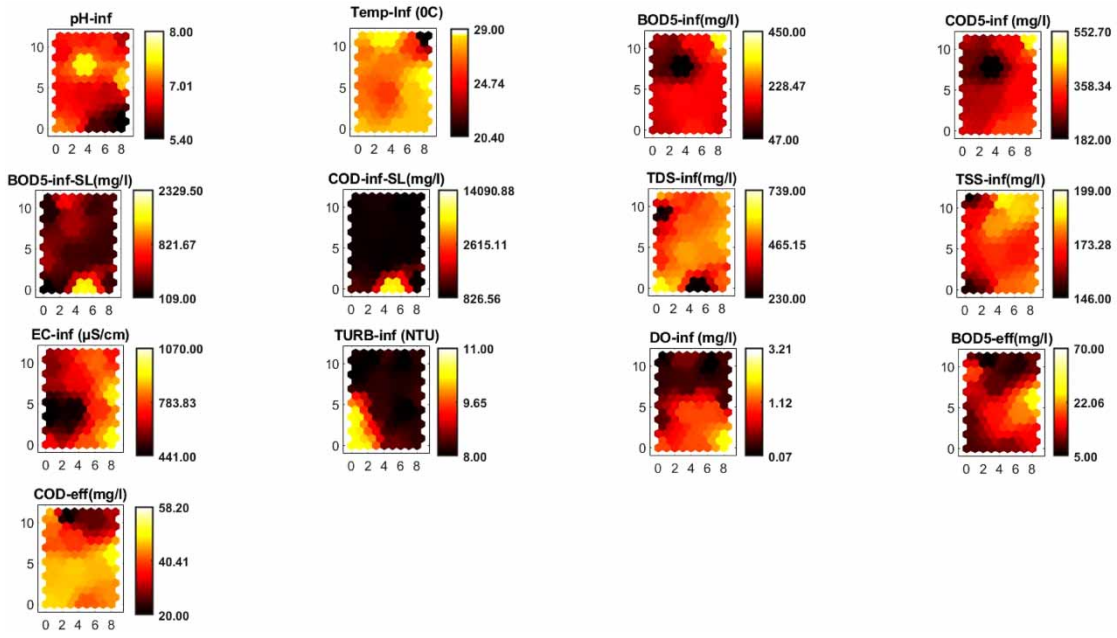


Figure 5 | SOM component planes.

Table 3 | Correlation matrix for variables in code vectors

	pH_{inf}	T_{inf}	BOD_{5inf}	COD_{inf}	BOD_{inf} SL	COD_{inf} SL	TDS_{inf}	TSS_{inf}	EC_{inf}	$TURB_{inf}$	DO_{inf}	BOD_{Seff}	COD_{eff}
pH_{inf}	1												
T_{inf}	0.161	1											
BOD_{5inf}	−0.461	−0.606*	1										
COD_{inf}	−0.615*	−0.510	.922**	1									
BOD_{inf} SL	−0.251	0.167	0.043	0.061	1								
COD_{inf} SL	−0.307	0.153	0.144	0.215	.834**	1							
TDS_{inf}	0.368	−0.125	0.131	0.088	−0.620*	−0.599*	1						
TSS_{inf}	−0.199	−0.319	0.339	0.275	0.249	0.035	−0.196	1					
EC_{inf}	−0.135	0.210	0.193	0.411	0.041	0.130	0.086	0.409	1				
$TURB_{inf}$	0.167	0.067	−0.009	−0.048	−0.048	0.125	0.162	−0.386	−0.320	1			
DO_{inf}	−0.587*	0.204	0.098	0.269	0.110	0.301	0.003	0.024	0.234	0.163	1		
$BOD5_{eff}$	−0.013	0.154	0.033	0.066	−0.309	−0.157	−0.008	0.003	0.252	−0.237	0.196	1	
COD_{eff}	0.014	0.166	−0.131	−0.052	−0.285	0.050	0.114	−0.441	−0.100	0.380	0.344	.625*	1

*Correlation is significant at the 0.05 level (two-tailed).

**Correlation is significant at the 0.01 level (two-tailed).

SD, standard deviations; se, standard error; UB, upper bound of 95% confidence interval for the mean; LB, lower bound of 95% confidence interval for the mean; pH, power of hydrogen; BOD, biochemical oxygen demand; COD, chemical oxygen demand; TDS, total dissolved solids; TSS, total suspended solids; EC, electrical conductivity; DO, dissolved oxygen. *SUFFIXES:* inf, influent; eff, effluent; SL, Septage Lagoon.

versa. Similarly, high values of BOD_{5eff} were correlated with high values of COD_{eff} and vice-versa. The component planes also confirm the negative correlation between pH and BOD_{5inf}, COD_{inf} and DO_{inf} with low values of pH associated with the high values of the BOD_5inf, CODinf, and DOinf. The positive correlation between BOD and COD was expected as COD values are typically higher than BOD values, and the ratio between them varies depending on the characteristics of the wastewater (Rai *et al.* 2019). Table 3 displays the complete correlation matrix for all 11 variables of the prototype vectors. Although this is a simple tool for examining the linear relationship between various variables, its results appear to agree with the indications of cross-correlation provided by the much more complex SOM analysis that resulted in the component planes.

Model evaluations

Figures 6(a) and 6(b), respectively, show a visual comparison of the estimated and measured values for BOD5 and COD effluent concentrations. Figures 6(a) and 6(b), in contrast to Figure 7, help to illustrate how well the SOM outputs have temporally matched the observed data. In general, the SOM outputs accurately reproduced the observed time series data's peaks and troughs. The predicted missing values are also shown (Figures 6(a) and 6(b)), and their trend is consistent with the overall trend of the observed data series.

The performance of the SOM in predicting the various characteristics is depicted in Figure 7 and Table 4. The associated correlation coefficient values are all greater than 0.90, indicating that the overall performance is good. Variables with a high proportion of missing values, on the other hand, had high MSE values. The mean square values of influent pH, DO, and effluent BOD, for example, were 3.655×10^3, 1.062×10^3, and 2.153×10^3, respectively. The MSE is the average of the squares of the model predictions' errors. When there is no error in a model, the MSE is zero. As model error increases, so does its value.

An additional analysis was performed to see if the sample skewness coefficients of the residuals are statistically zero. This is required to ensure that the residuals are distributed normally. Equation (11) was used to estimate the sample skew for a variable x_i at a 95% confidence interval.

$$Skew = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left\{ \frac{x - \bar{x}}{s} \right\}^3 \tag{11}$$

where N is the sample size, $\bar{x}$ is the sample mean, and S is the sample standard deviation.

Based on the null hypothesis that the skew coefficient will be zero, the skew coefficient which was outside this CI was considered to be not normally distributed. Table 5 displays the results of this hypothesis testing. Based on these findings, it is clear that the residuals associated with the majority of characteristics are normally distributed.

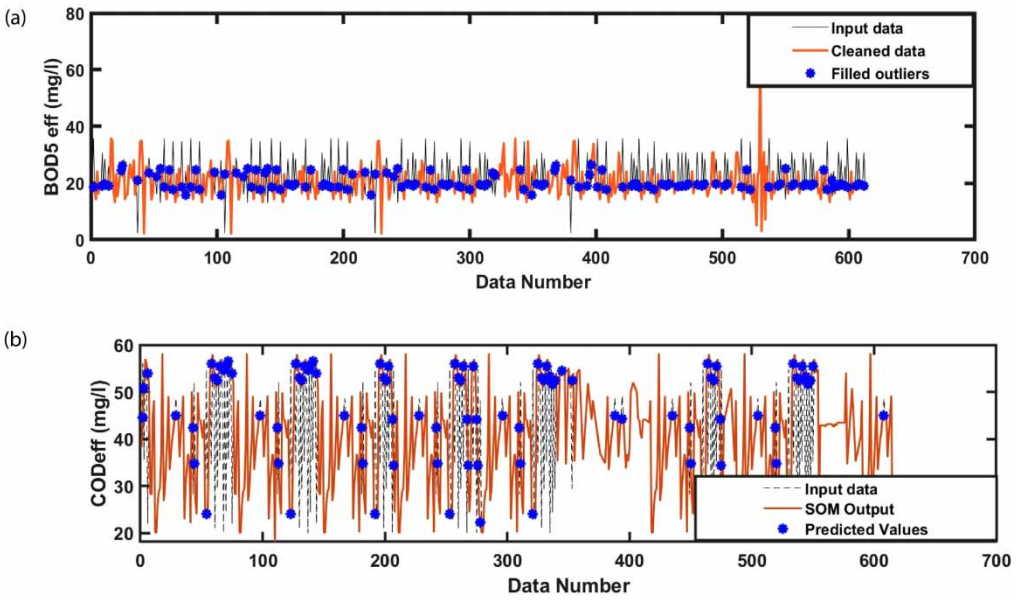


Figure 6 | Time series plots for predicted and observed values (a) for effluent BOD5, (b) for effluent COD.

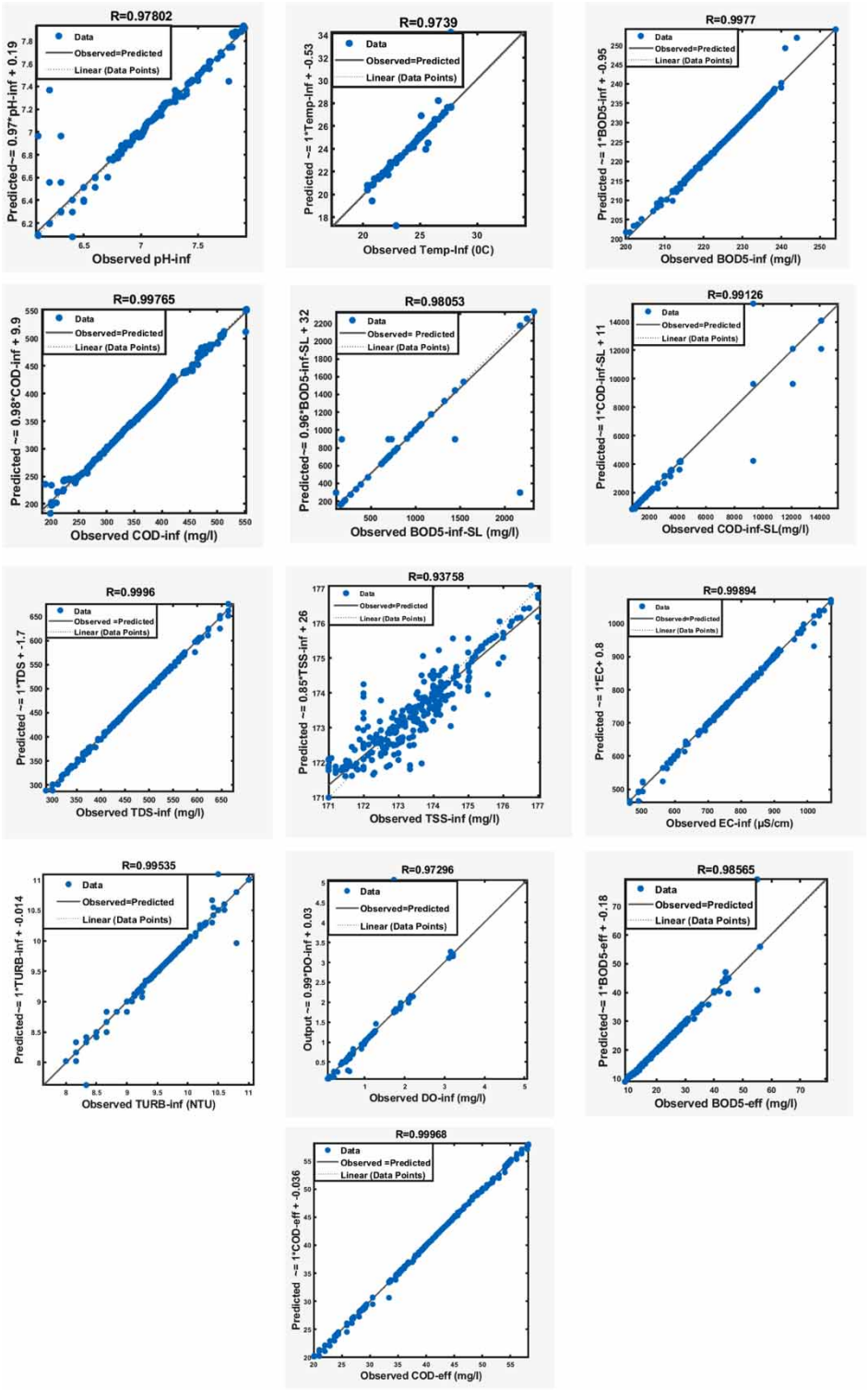


Figure 7 | Performance of SOM in predicting selected wastewater parameters.

The only exceptions are influent pH, influent DO, and effluent BOD, which have test statistics that fall just outside the 95% confidence interval.

These results indicate that, although SOM algorithm is quite robust for infilling gaps in wastewater time series, it cannot be used for infilling gaps in time series with a high proportion of missing data owing to the reduced model performance that was observed in time series that had more than 50% of missing data. This could be explained by the insufficient data from which the model is expected to learn, and thus cannot correctly replicate the pattern in the

Table 4 | Performance indices of SOM

Parameter	Unit	MSE	*R*
pH_{inf}	–	3.655×10^{3}	0.978
$Temp_{inf}$	°C	7.65×10^{-3}	0.974
BOD_{5inf}	mg/l	6.702×10^{-3}	0.997
COD_{inf}	mg/l	6.262×10^{-4}	0.998
BOD_{inf} SL	mg/l	4.950×10^{-5}	0.981
COD_{inf} SL	mg/l	2.301×10^{-3}	0.991
TDS_{inf}	mg/l	1.038×10^{-2}	0.999
TSS_{inf}	mg/l	3.320×10^{-1}	0.938
EC_{inf}	μS/cm	7.192×10^{-3}	0.999
$Turb_{inf}$	NTU	3.61×10^{-4}	0.995
DO_{inf}	mg/l	1.062×10^{2}	0.973
BOD_{5eff}	mg/l	2.153×10^{3}	0.986
COD_{eff}	mg/l	1.159×10^{-3}	0.999

MSE, mean square error; *R*, correlation coefficient.

Table 5 | Approximate normality test for residuals

Variable	No.	Lower limit	Upper limit	Skew coefficient	Normal (Y/N) (skewed)
pH_{inf}	572	−0.009	0.009	−1.054	N
$Temp_{inf}$	459	−0.6273	0.00265	−0.312	Y
BOD_{5inf}	395	−0.182	0.182	−0.164	Y
COD_{inf}	377	−3.079	3.079	0.806	Y
BOD_{5inf} SL	242	−10.8931	10.89306	−0.95502	Y
COD_{5inf} SL	252	−19.6372	19.63723	−0.10421	Y
TDS_{inf}	539	−0.791	0.791	0.134	Y
TSS_{inf}	265	−8.83	8.83	−0.64	Y
EC_{inf}	599	−1.839	1.839	−0.077	Y
$Turb_{inf}$	529	−0.775	1.193	0.209	Y
DO_{inf}	502	−0.021	0.021	3.072	N
BOD_{5eff}	407	−0.450	0.450	−1.411	N
COD_{eff}	390	−0.408	0.408	0.325	Y

data. For example, measured influents pH, DO had 60.71 and 59.09% proportions of missing data produced mean square values of 3.655×10^{3} and 1.062×10^{3} in that order. Similarly, effluent BOD had a proportion of 61.63% missing data that produced an MSE value of 2.153×10^{3}. This implies that time series with extended periods of missing observations should not be used as the model may infill the missing observations but still fail to replicate the pattern in the data. In the context of rainfall–runoff modeling, Mwale *et al.* (2012) propose that such inconsistencies can be resolved by training time series data with the data from the same spatial zones. However, this can be a very challenging task in the context of wastewater systems due to the complex processes involved. The correlation coefficient values of more than 0.9 obtained in this study are comparable to those reported by other researchers summarized in Table 6. Similarly, the quantification errors that were very close to zero are also comparable to those reported by the aforementioned authors. This similarity in the results could largely be explained by similarity in the implementation procedure of SOM algorithm in MATLAB.

CONCLUSIONS

This study sought to apply a SOM algorithm in filling missing values and replacing outliers in wastewater data for the Kauma sewage treatment plant. Results showed that the SOM algorithm is reliable for infilling gaps and

Table 6 | Research in engineering-related problem optimization using SOM algorithm

Authors	Location	Optimisation problem	Parameters	Software	Fitness functions	Major findings
Nijim & Rustum (2022)	United Kingdom (Seafield wastewater treatment plant data in Edinburgh)	Apply SOM algorithm as alternative model to verify the accuracy of the Multivariate Imputation by Chained Equations (MICE)	DO	Not specified	R, MSE, AAE	Performance of MICE Model was excellent with less proportion of missing values and poor when proportion was high. Results were similar to those produced by the SOM Model in previous case studies.
Juboori *et al.* (2022)	Desk Reviews	Analyze reinforced concrete structures employing Self-organizing maps	Elements of Reinforced Cement Concrete (RCC)	MATLAB	AAE, RAAE, NRMSE, MSE, R, CE	The Self-Organizing Map (KSOM) is an attractive tool for modeling reinforced concrete structures. Moreover, this technique offers a magnificent tool for high-dimensional data visualization.
Kumar *et al.* (2021b)	India (National Institute of Technology, Hamirpur)	Develop self-organizing map (SOM), feed-forward neural network (FFNN), and multiple linear regression (MLR) models were for estimating the well-watered canopy temperature (Tc-ww) using air temperature and relative humidity as input predictor variables	Relative Humidity, Air Temperature, Well-watered canopy Temperature	MATLAB	MBE, MAE, MSE, PE, R	The findings indicated that the SOM-modeled values presented a better agreement with the measured values in comparison to MLR- and FFNN-based estimates, with R^2 values of 0.978, 0.924, and 0.923 for KSOM, MLR, and FFNN, respectively, during model validation.
Kumar *et al.* (2021a)	India (National Institute of Technology, Hamirpur)	Develop a self-organizing map (SOM) based model to predict the Crop Water Stress Index (CWSI) using microclimatic variables, namely air temperature, canopy temperature and relative humidity	CWSI, air temperature, solar radiation, wind speed, and relative humidity	MATLAB	NSE, BE, AE, R	The SOM predicted CWSI presented a good agreement with the baseline computed CWSI values during model training (R = 0.98, NSE = 0.97, AE = 0.018, BE = 0.0004) and testing (R = 0.98, NSE = 0.98, AE = 0.018, BE = 0.002).
Ramachandran *et al.* (2019)	N/A	Use SOM to predict anaerobic digestion system behavior, study correlation between various process parameters, and extract Knowledge.	Glucose, Biogass flowrate, Methane gas, pH,	MATLAB (Synthetic MATLAB–Simulink–Excel model)	R, AAE, MSE, RMSE	The model accurately predicted the variations in methane and total gas output with respect to changes in input parameters as the correlation was more than 90% for most of the parameters.
Rizvi & Rustum (2018)	California (wastewater		Precipitation, Q, SS, BOD, COD, NH_4,	MATLAB	N/A	The results of the case study showcased SOM as a tool which was able to

(*Continued.*)

Table 6 | Continued

Authors	Location	Optimisation problem	Parameters	Software	Fitness functions	Major findings
	treatment plant in San Diego)	Use SOM to study the effects of precipitation on the performance of wastewater treatment plant	NO_3, PO_4, Temp, pH,			recognize the relationship among different parameters with rain in the wastewater treatment system.
Nkiaka *et al.* (2016)	Cameroon (Logone catchment, Lake Chad basin)	Use SOM to infill missing data in hydro -meteorological time series	Rainfall data, River discharge data	MATLAB	R	SOMs are a robust and efficient method for infilling missing gaps in hydro-meteorological time series as indicated by coefficient of determination values which were all above 0.75 and 0.65 for rainfall and river discharge time series respectively.
Mwale *et al.* (2014)	Malawi (Shire River Basin)	Use SOM to extract features from the raw data, which then formed the basis of infilling the gap-riddled data to provide more complete and much longer records those enhanced predictions	Rainfall data, River discharge data	MATLAB	NSE, R, MSRE	SOM is quite robust to infill missing data and can therefore be used to infill large gaps, something that would be impossible with traditional infilling methods, thus presenting a relatively long series needed data for hydrological modeling.
Adeloye & Rustum (2012)	Nigeria (Osun basin)	Use SOM to model rainfall–runoff relationship	Rainfall data, River discharge data	MATLAB	R	The study demonstrated the successful use of emerging tools to overcome practical problems in sparsely gauged basins.
Rustum & Adeloye (2012)	United Kingdom (Seafield ASP in Edinburgh	Used SOM to enhance the performance of a multi-layered perceptron, feed-forward back propagation artificial neural networks	Flow Rate, COD, SS. Ammonia, Blanket Depth	MATLAB	R, MSE, AAE	The study clearly demonstrated the usefulness of the clustering power of the SOM in helping to reduce noise in observed data to achieve better modeling and prediction of environmental systems behavior.
Rustum & Adeloye (2007)	United Kingdom (Seafield ASP in Edinburgh	Using SOM to replace outliers and missing values from activated sludge plant data	Flow rate, BOD, SS, WAS, MLSS, RAS, SSVI, sludge age, F/M	MATLAB	R, MSE	Results demonstrated that the SOM is an excellent tool for replacing outliers and missing values from a high-dimensional dataset.

R, coefficient of determination; MSE, Mean Square Error; AAE, Average Absolute Error; MBE, Mean Bias Error; MAE, Mean Absolute Error; PE, Percent Error; NSE, Nash–Sutcliffe efficiency; BE, Bias Error; AE, Absolute Error; RAAE, Relative Average Absolute Error; NRMSE, Normalized Root Mean Square Error; CE, Classification Error; MSRE, Mean Squared Relative Error; SS, Suspended Solids; WAS, Waste Activated Sludge; MLSS, Mixed liquor suspended solids; RAS MLSS, Return Activated Sludge Mixed Liquor Suspended Solids; SSVI, Stirred sludge Volume Index; F/M, Food to microorganisms Ratio.

replacing outliers in wastewater time series data with less than 50% proportion of missing data. The SOM performance registered a deteriorating trend with missing values of more than 50% in the time series data. Overall R values of >0.90 obtained in this study are within the range of performance prediction reported in literature. This approach can be used by practitioners to enhance the planning and management of wastewater treatment facilities where available records are infested with missing observations. We recommend further research to ascertain the accuracy of the SOM algorithm in filling and replacing outliers in extended data records measured at different time scales such as hourly or/and daily measured data.

ACKNOWLEDGEMENTS

The structure of methodology and results section of this work followed the work of Rustum & Adeloye (2007). The authors are grateful to the following people, Dr Linda Strande (PhD) and the entire team from Eawag (Swiss Federal Institute of Aquatic Science and Technology) for the technical guidance and material support. Lilongwe City Council for the support rendered during data collection particularly to the following people Mr Obvious Nyirenda, Mr Phyllis Mkwezalamba Mr John Thyoka, Mr Chimango Mweso and Mr Orymo Nyirenda

FUNDING

This work received funding from the National Commission for Science and Technology (NCST) under NCST Small Grants Scheme. It also received funding from the Malawi Ministry of Education, Science and Technology under Higher Education Research and Development for Young Researchers (postgraduate) scheme (Ref. No. EDU/HE/21/74).

SOFTWARE AVAILABILITY STATEMENT

The SOM Toolbox (Version 2.2) for MATLAB used in this study is freely available for download from GITHUB (https://github.com/ilarinieminen/SOM-Toolbox)

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Adeloye, A. J. & Rustum, R. 2012 Self-organising map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins. *Hydrology Research* **43** (5), 603–617. https://doi.org/10.2166/NH.2012.017.

Adeloye, A. J., Rustum, R. & Kariyama, I. D. 2012 Neural computing modeling of the reference crop evapotranspiration. *Environmental Modelling & Software* **29** (1), 61–73. https://doi.org/10.1016/J.ENVSOFT.2011.10.012.

APHA 2017 *Standard Methods for the Examination of Water and Wastewater*, 23rd edn. American Public Health Association, Washington, DC.

Gabrielsson, S. & Gabrielsson, S. 2006 The use of Self-Organizing Maps in Recommender Systems: A survey of the Recommender Systems field and a presentation of a State of the Art Highly Interactive Visual Movie Recommender System. Master's Thesis. Uppsala University, Uppsala.

García, H. L. & González, I. M. 2004 Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence* **17** (3), 215–225. https://doi.org/10.1016/J.ENGAPPAI.2004.03.004.

Hassen, E. B. & Asmare, A. M. 2018 Predictive performance modeling of Habesha Brewery's wastewater treatment plant using artificial neural networks. *Journal of Environmental Treatment Techniques* **6** (2), 15–25. Available from: http://www.jett.dormaj.com

Ismail, S., Shabri, A. & Samsudin, R. 2011 A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting. *Expert Systems with Applications* **38** (8), 10574–10578. https://doi.org/10.1016/J.ESWA.2011.02.107.

Juboori, A., Al Juboori, O. & Rustum, R. 2022 *Analysis of Reinforced Concrete Structures Employing Kohonen Self Organizing Map*. Available from: https://researchportal.hw.ac.uk/en/publications/analysis-of-reinforced-concrete-structures-employing-kohonen-self

Kalteh, A. M. & Berndtsson, R. 2011 Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). **52** (2), 305–317. https://doi.org/10.1623/HYSJ.52.2.305.

Kalteh, A. M., Hjorth, P. & Berndtsson, R. 2008 Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environmental Modelling & Software* **23** (7), 835–845. https://doi.org/10.1016/J.ENVSOFT.2007.10.001.

Kangas, J. & Kohonen, T. 1996 Developments and applications of the self-organizing map and related algorithms. *Mathematics and Computers in Simulation* **41** (1–2), 3–12. https://doi.org/10.1016/0378-4754(96)88223-1.

Kohonen, T., Hynninen, J., Kangas, J. & Laaksonen, J. 1996 *SOM_PAK: The Self-Organizing map Program Package*. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.455.8698&rep = rep1&type = pdf

Kumar, N., Rustum, R., Shankar, V. & Adeloye, A. J. 2021a Self-organizing map estimator for the crop water stress index. *Computers and Electronics in Agriculture* **187**, 106232. https://doi.org/10.1016/J.COMPAG.2021.106232.

Kumar, N., Shankar, V., Rustum, R. & Adeloye, A. J. 2021b Evaluating the performance of self-organizing maps to estimate well-watered canopy temperature for calculating crop water stress index in Indian Mustard (Brassica Juncea). *ASCE Journal of Irrigation and Drainage Engineering* **147** (2), 4020040. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001526.

Larsen, T. A., Udert, K. M. & Lienert, J. 2013 *Source Separation and Decentralization for Wastewater Management*. Available from: www.iwappublishing.com

Mtethiwa, A., Munyenyembe, A., Jere, W. & Nyali, E. 2008 Efficiency of oxidation ponds in wastewater treatment. *International Journal of Environmental Research* **2** (2), 149–152. Available from: https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=103029

Mwale, F. D., Adeloye, A. J. & Rustum, R. 2012 Infilling of missing rainfall and streamflow data in the Shire river basin, Malawi – a self-organizing map approach. *Physics and Chemistry of the Earth* **50–52**, 34–43. https://doi.org/10.1016/J.PCE.2012.09.006.

Mwale, F. D., Adeloye, A. J. & Rustum, R. 2014 Application of self-organising maps and multi-layer perceptron-artificial neural networks for streamflow and water level forecasting in data-poor catchments: the case of the Lower Shire floodplain. *Malawi. Hydrology Research* **45** (6), 838–854. https://doi.org/10.2166/NH.2014.168.

Nijim, H. & Rustum, R. 2022 *Imputation of Outliers and Missing Values for Activated Sludge Dissolved Oxygen Database Using Multivariate Imputation by Chained Equations (Mice)*. Available from: https://researchportal.hw.ac.uk/en/publications/imputation-of-outliers-and-missing-values-for-activated-sludge-di

Nkiaka, E., Nawaz, N. R. & Lovett, J. C. (2016). Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. *Environmental Monitoring and Assessment* **188**(7), 1–12. https://doi.org/10.1007/S10661-016-5385-1

Rai, A., Singh, S., Zia, S., Manikpuri, P. & Alexander, K. 2019 *Relation Between COD and BOD in Sangam Water Samples for pre and Post Bath During Kumbh*. Available from: https://www.entomoljournal.com/archives/2019/vol7issue3/PartS/7-3-187-712.pdf

Ramachandran, A., Rustum, R. & Adeloye, A. J. 2019 Anaerobic digestion process modeling using Kohonen self-organising maps. *Heliyon* **5** (4), e01511. https://doi.org/10.1016/J.HELIYON.2019.E01511.

Ravina, M., Galletta, S., Dagbetin, A., Kamaleldin, O. A. H., Mng'ombe, M., Mnyenyembe, L., Shanko, A. & Zanetti, M. 2021 Urban wastewater treatment in African countries: evidence from the hydroaid initiative. *Sustainability* **13** (22), 12828. https://doi.org/10.3390/SU132212828.

Rizvi, S. A. H. & Rustum, R. 2018 Study the effect of precipitation on the performance of wastewater treatment plant using KSOM. *Proceedings of the Annual International Conference on Architecture and Civil Engineering*. https://doi.org/10.5176/2301-394X_ACE18.61.

Rustum, R. 2009 *Modelling Activated Sludge Wastewater Treatment Plants Using Artificial Intelligence Techniques (Fuzzy Logic and Neural Networks)*. Available from: https://www.ros-test.hw.ac.uk/xmlui/handle/10399/2207

Rustum, R. & Adeloye, A. 2007 *Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map Map*. https://doi.org/10.1061/(ASCE)0733-9372(2007)133:9(909).

Rustum, R. & Adeloye, A. 2011 *Artificial Intelligence Modeling of Wastewater Treatment Plants: Theory, Applications and Limitations*. VDM Verlag Dr. Müller, Riga.

Rustum, R. & Adeloye, A. 2012 Improved modelling of wastewater treatment primary clarifier using hybrid ANNS. *International Journal of Computer Science and Artificial Intelligence* **2** (4), 14–22. Available from: https://researchportal.hw.ac.uk/en/publications/improved-modelling-of-wastewater-treatment-primary-clarifier-usin

The Maths Works, Inc. 2021 *MATLAB (Version 2021a)*. https://www.mathworks.com/.

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. 2000 *SOM Toolbox for MATLAB 5*. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.25.7561&rep = rep1&type = pdf

Zhu, J., Ge, Z., Song, Z. & Gao, F. 2018 Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control* **46**, 107–133. https://doi.org/10.1016/J.ARCONTROL.2018.09.003.

First received 24 January 2023; accepted in revised form 16 May 2023. Available online 2 June 2023