Heliyon

Received:
30 May 2018
Revised:
6 November 2018
Accepted:
8 April 2019

Cite as: Anjali Ramachandran, Rabee Rustum, Adebayo J. Adeloye. Anaerobic digestion process modeling using Kohonen self-organising maps. Heliyon 5 (2019) e01511. doi: 10.1016/j.heliyon.2019.e01511

Anaerobic digestion process modeling using Kohonen self-organising maps

Anjali Ramachandran [a], **Rabee Rustum** [a,*], **Adebayo J. Adeloye** [b]

[a] *School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Dubai Campus, Dubai International Academic City, PO Box 294345, Dubai, United Arab Emirates*

[b] *School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh, EH14 4AS, UK*

[*] Corresponding author.

E-mail address: r.rustum@hw.ac.uk (R. Rustum).

Abstract

Anaerobic digestion is a versatile method for wastewater treatment as it not only reduces the waste but also leads to production of renewable energy. Modeling of the anaerobic process requires knowledge of biological and physico-chemical conditions, bacterial growth kinetics, substrate utilization, and product synthesis. However, the complexity of the process calls for highly sophisticated models requiring very high level of expertise and knowledge in the subject. This paper presents an approach for modeling of anaerobic digestion process through which the correlation between various process parameters can be studied, knowledge can be extracted, and system behaviour can be predicted. The datasets have been generated using a synthetic Matlab-Simulink-Excel model and process modelling is done using Kohonen Self organizing maps (KSOM). The resulting KSOM provided a visual interpretation of the inter-relationships between parameters (OLR, Sac, pH, Shco3, Q, Sglu_in, Qgas_out, Sglu_out, and Sch4_gas_out) which would help semi-skilled operators for operation and control of such plants. The model accurately predicts the variations in methane and total gas output with respect to changes in input parameters as the correlation is more than 90% for most of the parameters. This methodology offers a platform for scientists and researchers in comprehending the system behaviour under various operating conditions, even with missing data.

https://doi.org/10.1016/j.heliyon.2019.e01511

Keywords: Computer science, Biotechnology, Environmental science

1. Introduction

Anaerobic Digestion is a process wherein organic substrates, in the form of carbohydrates, proteins, lipids and complex compounds are converted into biogas, which is renewable energy, and anaerobic biomass, which can be used as natural fertilizer or soil conditioner. Biogas generally consists of 55−70% methane, 30−40% carbon dioxide, 1−2 % hydrogen sulphide, hydrogen, ammonia and traces of carbon monoxide, nitrogen and oxygen (Jørgensen, 2009). Biomass is rich in macro and micronutrients as it is the decomposed substrate. It is a flexible process, which can be used for large-scale digesters and can be used for individual homes or family owned, which is done in China, India, Nepal and Vietnam (Seadi et al., 2008). Another major reason why anaerobic digestion process is preferred is due to the global efforts to displace fossil fuels as energy resource and the need to find environmentally sustainable solution for waste management.

Although anaerobic digestion is a mature technology, failure of anaerobic wastewater treatment plants due to inadequate operational management and process control is commonplace. Modeling the anaerobic process helps not only for the purpose of design of wastewater plants and biogas power plants, but also aids in the study of the effect of operational parameters on the process, feasibility of new substrates, assessing of varied operational conditions etc. Moreover, it is advantageous to the construction of anaerobic treatment plants commercially and technically, if the process can be simulated during design stage.

For an understanding of the digestion process, mainly two approaches have been used: the experimental approach in which the parameters that influence the process are measured and the theoretical approach in which mathematical modeling of the process is performed. Due to improvement in research techniques and computational capacity, several models have been developed for anaerobic digestion and extensive description on the digestion process can be found in the literature (Appels et al., 2008; Bjornsson, 2000; Boe, 2006; Jha et al., 2013; Li et al., 2011; Metcalf and Eddy, 2003; Parawira et al., 2005; Yang et al., 2010).

1.1. From ADM1 to KSOM

With the aim of providing a very generic and usable model, Anaerobic Digestion Model No: 1 (ADM1) was developed in 2002 by the International Water Association Anaerobic Modeling Task Group. This has given a platform to apply the model in various anaerobic digestion systems. The model was structured with degeneration, hydrolysis, acidogenesis, acetogenesis and methanogenesis processes (Batstone

https://doi.org/10.1016/j.heliyon.2019.e01511

et al., 2002). The model is based on chemical oxygen demand as a base unit for wastewater classification in a continuous-flow stirred tank reactor. The model uses state variables to illustrate the behaviour of soluble (S) and particulate (X) compounds. Soluble species easily pass through the microbial cell wall and includes sugars, amino acids, long chain fatty acids, volatile organic compounds, hydrogen, and methane. Particulates include active biomass and other particulate substances such as organics from microbial decay or from influent stream. Molar concentration terms are used for nitrogenous species and inorganic carbon (Parker, 2005).

The conversion processes in anaerobic digestion are interlinked biochemical and physico-chemical reactions, which proceed in sequential and parallel steps, both spatially and temporally. The biochemical processes considered in the model are hydrolysis of complex compounds leading to sugars; amino acids and long chain fatty acids production; acidogenesis forming volatile fatty acids including acetic, butyric, propionic and valeric acids; acetogenesis leading to the formation of acetic acid and hydrogen; and lastly methanization. Out of this, the extracellular steps are disintegration and hydrolysis and the intracellular steps are the subsequent steps leading to methanogenesis.

The important difference between ADM1 and other models is the implementation of disintegration step that is different from hydrolysis. Disintegration is mostly a non-biological step wherein the composite particulate substrates are degraded into inerts, particulate carbohydrates, proteins and lipids (Blumensaat and Keller, 2005). Physico-chemical processes involved in the model are chemical equilibria and pH measured by ion association/dissociation (liquid-liquid reactions) and gas-liquid transfer, playing a strong role in biodegradation but is not biologically mediated. Gaseous stripping of compounds such as hydrogen, methane, and carbon dioxide is included to characterize biogas production. The calculation of pH is using six additional physico-chemical processes that explain the acid/base equilibria of valeric acid/valerate, butyric acid/butyrate, propionic acid/propionate, acetic acid/acetate, $NH4^{+}/NH_3$ and CO_2/HCO_3^{-}. The effect of positively and negatively charged ions on pH is also included through addition of dynamic states of cations and anions (Jha et al., 2013).

Uptake of substrate is modelled by Monod-type kinetics. The uptake of inorganic nitrogen is expressed by secondary Monod kinetics and that of butyrate and valerate by single group of organisms. Biomass decay also follows first order kinetics and the dead biomass is considered as composite particulate matter in the system (Batstone et al., 2002). In ADM1, extremes of pH inhibit all microbially mediated substrate conversion, accumulation of molecular hydrogen inhibits anaerobic oxidation processes, and increase in free ammonia inhibits aceticlastic methanogenesis. Mass transfer relationships are used to describe liquid-gas transfer of methane, carbon dioxide and molecular hydrogen (Parker, 2005). The following processes have been

https://doi.org/10.1016/j.heliyon.2019.e01511

excluded from ADM1 such as alternative glucose products like lactate and ethanol; sulphate reduction and sulphide inhibition, nitrate, long chain fatty acid inhibition, acetate oxidation, homoacetogenesis and solids precipitation (Batstone et al., 2002).

Another major domain used for anaerobic digestion modeling is Artificial Neural Networks (ANN) (Holubar et al., 2002; Lauwers et al., 2013; Ozkaya et al., 2007) and fuzzy inference system (Kusiak and Wei, 2014; Pai et al., 2009). ANN models have a reputation of high accuracy but subject to limitations when there are missing or variable data. It thus requires extensive pre-processing of data to achieve completeness before it can be applied. Fuzzy inference systems show lack of learning capabilities, as it requires either expert knowledge while modeling or large data sets. Its performance is also highly dependent on the quality of the training data.

Therefore, due to the complexity and uncertainty of measuring model parameters, Kohonen Self-Organising Map has been employed in this study for modeling the process. This is due to the aspects of KSOM in dealing with such uncertainty in a similar way of human thinking and its power of dealing with missing values. KSOM has been successfully used to model activated sludge wastewater treatment plant (Asadi et al., 2017; Begum et al., 2016; Liukkonen et al., 2013; Machón-González et al., 2017; Rustum and Adeloye, 2007; Rustum, 2009; Rustum and Adeloye, 2013a, b; Rustum et al., 2008; Rustum and Forrest, 2017; Szelag et al., 2017). Thus, this inspired the authors to try model the anaerobic system using the same approach.

Compared to ADM1, KSOM map itself can be used to predict missing values and as it is not affected by such incomplete data, there is no need of pre-processing the records which is time saving, and most importantly, KSOM has the power to extract valuable information from noisy data. KSOM requires fewer number of process parameters than ADM1 while still yielding accurate results. Measuring the parameters and co-efficients of ADM1 is a difficult task specially while trying to optimize functioning anaerobic digestion plants through simulating process parameters at site. The complexity of models like ADM1 with its numerous input parameters and stochastic/kinetic equations is negated by the simpler KSOM model that provides ease in identification of parameters and subsequent manipulation.

The main aim of this paper is to model anaerobic process using Kohonen self-organising maps that has the power to visualise the correlation between system parameters in a simple way that can be easily interpreted by operators who can adjust the operational parameters to obtain the desired output. Furthermore, KSOM model development and validation which will help to predict the system behaviour and also help in knowledge extraction of the process variables are other key objectives.

https://doi.org/10.1016/j.heliyon.2019.e01511

2. Methodology

2.1. Kohonen self organising maps

Self-organising map is an efficient tool that helps in visualization of data with high dimension. This algorithm relies upon unsupervised learning which is competitive and entirely data driven. Self-organising maps have an exceptional feature of creating internal representation of various aspects of input signals in a spatially organised and effective manner. Hence, the resulting maps strongly resemble or mimic the topographically structured maps (Kohonen et al., 1996). They work in a self-study mode wherein patterns are recognised and clustered into groups. As this network cannot fathom the meaning of the clusters, the users need to interpret the map in a meaningful and useful manner (Rustum, 2009). Self-organising maps take its instigation from neural networks that form the basis of nervous system. The signal progression and network constitution of the nervous system is divided into several categories depending upon various philosophies. In one, the nearby cells in neural network mutually interact and compete with each other and adaptively develop into specific detectors of diverse signal prototypes. In this classification, the learning is unsupervised or self-organising, which forms the base for the development of self-organising maps.

The main working strategy of such maps is the conversion of non-linear and complex correlation among data with high dimension into relatively simple low-dimensional view through geometrical relationships. KSOM consists of neurons arranged on standard one or two dimensional grids wherein every neuron i is characterised by n-dimensional weight/reference/codebook vector given by, $m_i = [\, m_{i1}, \ldots m_{in}]$ in which n is the input vector dimension. These weight vectors form the codebook, which portrays the features of the data or process. From Fig. 1, it can be seen

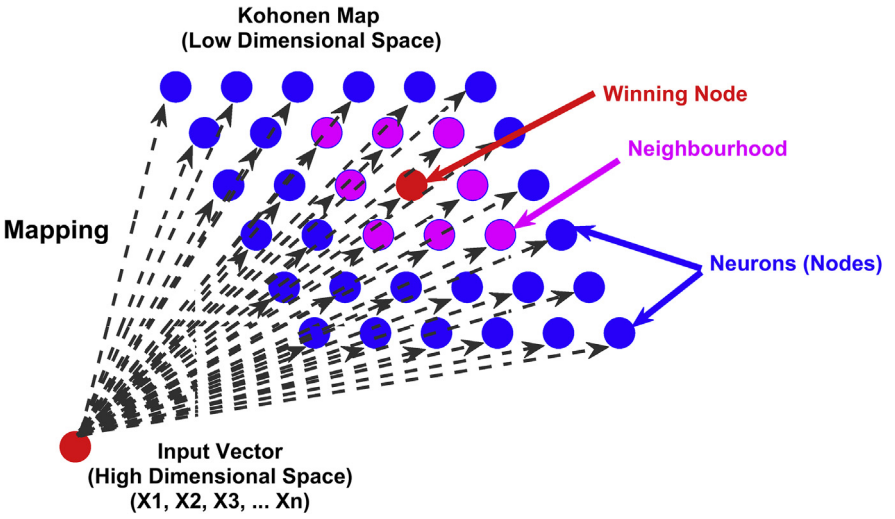

Fig. 1. Demonstration of 2 dimensional input and output layered sheets in KSOM (Rustum, 2009).

https://doi.org/10.1016/j.heliyon.2019.e01511

that every neuron has two locations namely one in the prototype vector, which is the input space, and another in the map grid or output space (Vesanto, 2000; Vesanto et al., 2000). Thus, self-organizing maps can be recognized as a vector projection method, which maps high-dimensional input to low-dimensional output. Neighbourhood relation dictates the connection between adjacent neurons.

The mapping is done from input Euclidean data space $\Re^n$ on to lattice of nodules in two-dimensional space. A characteristic reference vector $m_i \in R^n$, is linked with every node i. When there is an input data $x \in \Re^n$, it is evaluated against all the m_i to arrive at a best match or response. Through this process, input is mapped onto specific locations. Euclidean distances $\|x\text{-}m_i\|$ is used to identify node that matches best, m_c, also known as Best Matching Unit (Kangas and Kohonen, 2003).

$$\|x - m_c\| = min_i\{\|x - m_i\|\} \tag{1}$$

As per the SOM toolbox developed by Helsinki University of Technology for (Vesanto et al., 2000), the basic steps in the development of the map are initialization, training, and validation. Normalisation is a process wherein process variables are prevented from having larger impact than other variables which guarantees that the entire set of variables have same significance in the construction of maps. Initialisation helps the algorithm to converge sooner to a good result in which weight vectors are given values either randomly or linearly. In this process, each neuron is assigned random weight vectors generally between zero and one (Vermasvuori et al., 2002). The central aim of training is to establish the Best Matching Unit (BMU) or winning node from the map units for each input prototype. This unit is largely analogous the input pattern. A distance function is generally used to measure the similarity wherein closer distances define more similarity given by Euclidean distance function. The next step in training is reducing the difference between these units and input pattern by updating the best matching unit and its neighbouring units (Hsu, 2006). The updating is done by two types of algorithm namely sequential training algorithm and batch training algorithm. In sequential training, once the best matching unit is found out, its weight vectors are shifted nearer to the input vector in the input space, a process known as updating. The topological neighbour units of the best matching unit are also treated in the same manner. Distance of these neighbourhood neurons or units from the winner output array determines size of adjustment of the weight vector. More details about training the map are available in Vesanto (2000), (Garća and González, 2004), and Rustum (2009).

The quality of the KSOM is given by mainly two error measurements: the quantization error (q_e) and topographic error (t_e) (Ramos et al., 2013). Quantisation error is calculated by taking the mean Euclidean distance from input vector to its best matching unit. This in turn gives the map resolution and helps to identify outliers. High quantization error indicates a high probability that those input patterns are outliers.

https://doi.org/10.1016/j.heliyon.2019.e01511

Topologic or topographic error is given by the percentage of input vectors for which the best matching unit and the next best are not neighbouring nodes on the grid. This error indicates the degree of preservation of data topology while map is fitted into original dataset.

2.2. Modeling evaluation criteria

After fitting the input data into the KSOM or any other model, it is necessary to evaluate how well the model performs. MATLAB offers "goodness of fit" which has a set of parameters that describes the model accuracy. Evaluation can be done graphically using residual plots and prediction bounds and numerically using statistical parameters explained below. Graphical measures help the evaluation of the entire dataset at once and can display a wide range of relationships between the model and the data (Mathworks, 2011). Numerical evaluation measures include correlation co-efficient (R), average absolute error (AAE), mean square error (MSE) and Root mean square error (RMSE).

2.3. Data

The first step towards model generation is data gathering. Due to lack of sufficient data from anaerobic treatment plants in the region, focus has been done to generate data through other models such as ADM1 or from several anaerobic digestion modeling techniques found in literature. However, most of these systems are highly complex in terms of variables and parameters, and require high programming expertise. Even commercial user-friendly software packages are available such as *GPS-X*®, *WEST*® or *AQUASIM*® but offer less flexibility in terms of structural changes to the embedded process models (Henze et al., 2008). Here, data has been created synthetically using the initialized and fully calibrated dynamic simulation model. The model was developed by Rodríguez et al. (2009) based on *MS Excel* and *Matlab-Simulink*® platform. The model is flexible and can be used for implementation of mathematical models with less programming expertise. The key feature of this model is the use of Excel to modify the model structure and high flexibility of Simulink block for interlinkage of various process parameters and even controllers.

The main input variables considered are *Sglu_in* (glucose in) which is described in terms of glucose equivalents indicative of chemical oxygen demand and flow rate (*Q*) which has also been defined as combined input namely organic loading rate (*OLR*). System variables studied due to variations in input parameters are *pH*, *Sac* (acetic acid) and *Shco3* (bicarbonate). Output variables considered are *Q gas_out* (biogas flow rate), *Sglu_out* (glucose out) and *Sch4_gas_out* (methane gas output). Constant volume is assumed for modeling and the concentration of inlet glucose and flow rate are varied to achieve the required *OLR*. *pH*, *Sac* and *Shco3* are system parameters considered for the model which are the stability indicators of the process.

https://doi.org/10.1016/j.heliyon.2019.e01511

Highly alkaline (>11) and acidic pH (<3.5) values are excluded. Even though methanogenesis stops below pH 5.5, those values are considered to study the system behaviour during inhibition circumstances.

2.4. Data generation by simulation

The Matlab-Simulink-Excel model was simulated in Matlab using a range of input parameters to produce datasets required for developing the anaerobic model using Kohonen self-organising Maps. The details of the state variables and parameters are defined in the Excel file that can be simulated in Matlab to produce a suitable model structure and output variables. The model used herein is that of a continuous anaerobic system fed with glucose as substrate over a perfectly mixed volume given by Eq. (2).

$$\frac{dC_i}{dt} = \frac{Q}{V}(C_{i,in} - C_{i,out}) + R_i \left[mol\ or \frac{kg}{L.h}\right] \tag{2}$$

where C_i is the state variables consisting of concentrations of all chemical and biological species given by vector n x 1; Q is the feed flow rate; V is the constant reactor volume; $C_{i,in}$ is the influent concentration of species i; $C_{i,out}$ is the vector of effluent concentration of i; R_i is the vector of terms consisting of a number of reactions and transport processes which contributes to the concentration of each species i.

Rest of the algebraic state variables like reaction and transfer rates are functions of the concentration in the systems and can be computed depending upon definite kinetic and transport equations. The model does not consider biomass decay and assumes constant temperature, pressure, and acid-base equilibrium. The reaction rates are expressed with Monod type Kinetics with an inhibition of pH included as a term I_{ph}. These reactor rates (r_j) are a function of the concentration of all the elements/variables in the system as defined in Eq. (3):

$$r_i = q_{j,max} \times \frac{S_i}{K_{s,i} + S_i} \times I_{ph} \times X_j \quad \left[\frac{molSi}{L.h}\right] \tag{3}$$

where S_i is the liquid phase species i concentration; X_j is the solid phase species i concentration; $q_{j,max}$ is the maximum specific uptake rate; Ks,i is the substrate half saturation constant of process j. The I_{ph} term indicates the reactor pH, which is computed by balancing the charges of the different species in the system by using their property of acid-base equilibrium.

The model framework is structured using specified files such as Simulink file containing model block diagram, Matlab files for codes, stoichiometry, kinetics, feed program and Excel files for inputting feed parameters with description of variables (Rodríguez et al., 2009).

https://doi.org/10.1016/j.heliyon.2019.e01511

The time varying influent feed flow and concentrations together with their reaction rates need to be defined in the Excel spreadsheet before simulation of the continuous reactor using Eqs. (2) and (3). Once the simulation starts, the Simulink function spontaneously uploads and creates structure of the model from the Excel file. Simulink file also manages all the necessary tasks to calculate and integrate mass balances of the model. The simulation can be run up to any time interval required to study the process. After the simulation finishes, the complete data for every time interval consisting of all the state and algebraic variables, input and output conditions, rates and generation terms, are produced in the Matlab workspace. This information is processed and saved into an Excel File, which will form the input database for the KSOM model. This process is repeated with different feed flow and concentrations until the adequate number of training data is achieved. However, the KSOM anaerobic model is based on steady state conditions. Hence, the data output from the dynamic model is further studied and processed to generate steady state data.

For this study, data for simulation has been generated by varying Sglu_in and Q in various combinations, each with 1500 time interval. As it is a dynamic model, for each run it was found that at every 300 time interval, steady state values were achieved. These steady state values have compiled into creating the training data. The range of generated data can be seen in Table 1.

2.5. Modeling

Two sets of data were required for Modeling using KSOM, training and validation. 600 points of training data has been generated as described above. The data for validation has been collected from three different sources.

- Validation data 1 [D1-66 Nos] is assimilated from the research conducted by (Waewsak et al., 2010) which deals with monitoring process response and control in anaerobic hybrid (combination of suspended and attached growth) reactor.
- Validation data 2 [D2-132 Nos] is generated from the Matlab-Simulink model described above. This data have not been used during training.

Table 1. Statistical Analysis of Training data.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
Min.	0.004	0.001	3.77	0.000	0.003	0.015	0.018	0.000	0.001
Max.	0.163	0.377	10.6	0.034	0.110	0.210	0.901	0.019	0.039
Avg.	0.039	0.059	5.893	0.018	0.022	0.097	0.211	0.001	0.016
Stdev.	0.036	0.088	1.235	0.015	0.023	0.055	0.200	0.002	0.010

https://doi.org/10.1016/j.heliyon.2019.e01511

Table 2. Statistical analysis of Validation dataset, D1.

	OLR (Mol/L.day)	pH	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)
Min.	0.002	6.7	0.079	0.013	0.007
Max.	0.030	6.9	0.287	0.068	0.074
Avg.	0.011	6.8	0.118	0.041	0.035
Stdev.	0.005	0.04	0.032	0.017	0.016

- Validation data 3 [D3-2761 Nos] is the dynamic values from the Matlab-Simulink Model which considers time as one of the parameter.

The purpose of using different validation data is to assess the flexibility and efficiency of the KSOM model in predicting the process stability and performance in various scenarios.

Statistical descriptions of the training and validation datasets used are presented in Tables 1, 2, 3, and 4.

Analysis of the box plots (Figs. 2, 3, 4, and 5) show that in the training data OLR, Acetic acid, flow rate, gas production, and glucose output have quite a few outliers, indicating a high variation in data points in these parameters. This is beneficial in terms of modeling as the range of data that can be considered for the model is more. Bicarbonate, glucose input and methane production shows a wide range between the interquartile length indicating a large spread of data, but with no outliers. For the pH values, 50% of the data is below pH 7, with only less outliers in alkaline range. The outliers present a degree of uncertainty in the dataset but still modeling can be proceeded using KSOM (Rustum, 2009). The validation dataset D1 and D2 presents a consistent dataset with very few outliers mainly because the sample size is smaller with very few extreme data for the former and steady data distribution in the latter. The dataset D3 shows a very high variation in the dataset unlike the others. This is mainly due to the dynamic nature of the data. The data is highly uncertain and has high outliers in the parameters considered. It can be seen that the 50% of the data range between the interquartile lengths is very small and a wide span of the space in the plot is filled with outliers. The data is highly diverse and

Table 3. Statistical analysis of Validation dataset, D2.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
Min.	0.008	0.001	3.841	0.000	0.005	0.013	0.030	0.000	0.001
Max.	0.148	0.326	11.93	0.034	0.095	0.207	0.791	0.008	0.036
Avg.	0.046	0.091	5.437	0.012	0.026	0.097	0.257	0.001	0.011
Stdev.	0.037	0.091	1.484	0.015	0.023	0.053	0.205	0.002	0.010

https://doi.org/10.1016/j.heliyon.2019.e01511

Table 4. Statistical analysis of Validation dataset, D3.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
Min.	0.014	0.001	3.988	0.000	0.005	0.115	0.061	0.0001	0.002
Max.	0.162	0.242	6.790	0.034	0.050	0.135	0.901	0.0019	0.026
Avg.	0.038	0.041	6.185	0.027	0.013	0.124	0.207	0.0003	0.018
Stdev.	0.035	0.084	1.007	0.013	0.012	0.007	0.196	0.0004	0.008

inconsistent. This data has been chosen for validation to test the efficacy of the model in dealing with independent dataset.

3. Results and discussion

Comprehension of the behaviour of the system with study of correlation between parameters was done using correlation matrix, U-matrix and component planes of KSOM. Expression and verification of the KSOM model using different type of data sets that have not been used during model development are shown under scatter plots and time-series plots. Prediction of system behaviour through process stability and performance can be deciphered from the component planes of KSOM. This whole process helped in knowledge extraction of the anaerobic digestion process successfully.

Pre-processing of the input data has been done in *MS Excel* and *Matlab-Simulink*® developed by Rodríguez et al. (2009). Missing data have been replaced by NAN (Not a number) to satisfy the Matlab requirements. The data was divided into training (600 data points) and validation-D1 consisting of 66 data points, D2 consisting 132 data points and D3 consisting 2761 data points. Training datasets express the

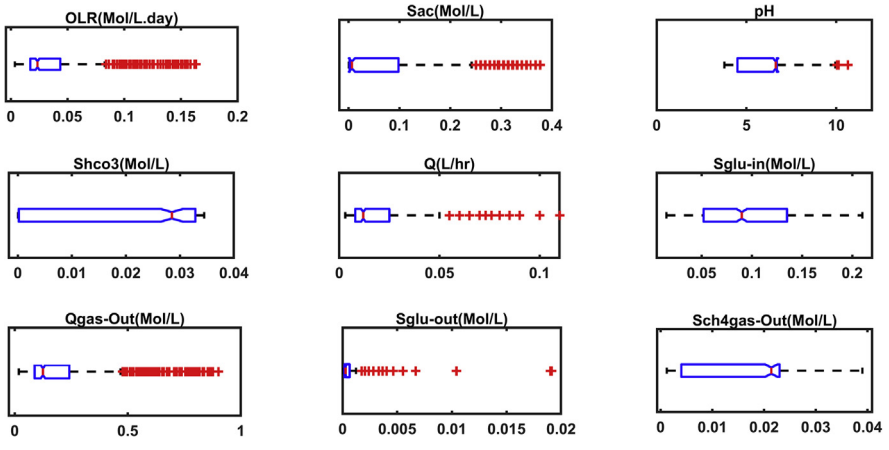

Fig. 2. Box Plots of Training data.

https://doi.org/10.1016/j.heliyon.2019.e01511

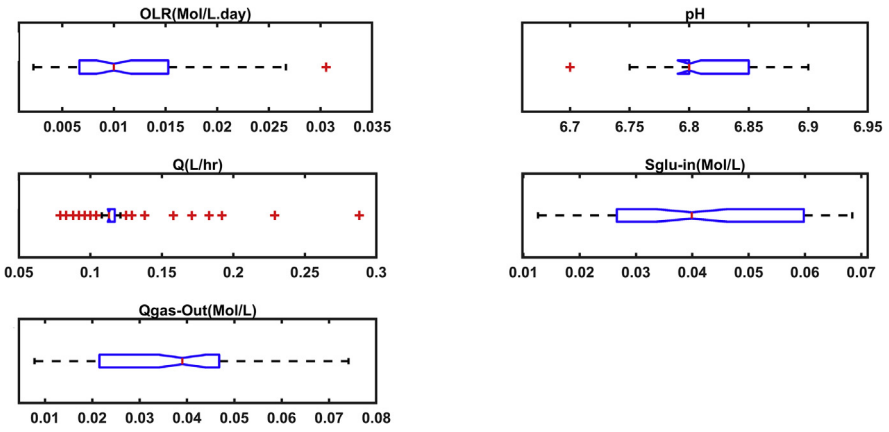

Fig. 3. Box Plots of Validation data D1.

effectiveness of learning and validation dataset is used to assess the efficiency of the model. The training dataset represent the entire operation range values of the process. The best matching unit was found out for the training dataset starting with the default value of 0.5 for learning rate in the SOM Toolbox. The map size calculated using empirical formulae gave M of 122 units. This is slightly different from the map size computed in SOM toolbox, which adjusts itself to the final n_1 and n_2, giving a result of M = 117 map units. The other characteristics of the trained SOM are given in Table 5. The topographic and quantization errors are small indicating that the model is well suited for prediction purposes.

The correlation matrix, Table 6, was generated which gives the relationship between parameters in the dataset. This correlation formed the base of parameter selection for modeling, and gave a crude indication on how variation among parameters influences the process.

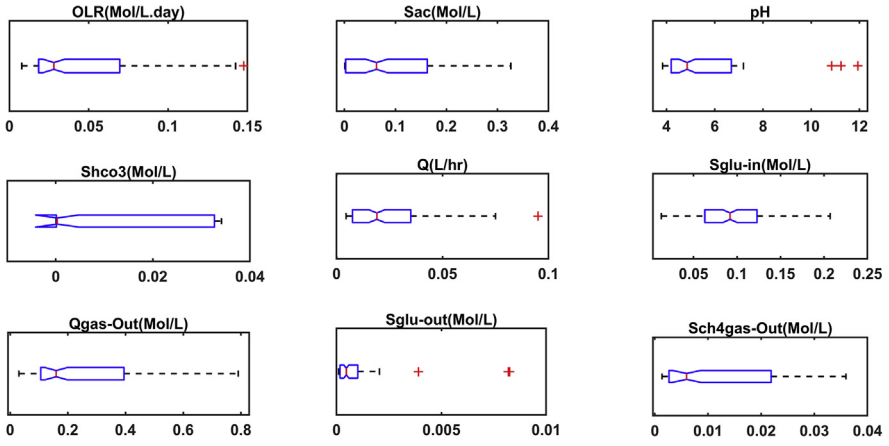

Fig. 4. Box Plots of Validation data D2.

https://doi.org/10.1016/j.heliyon.2019.e01511

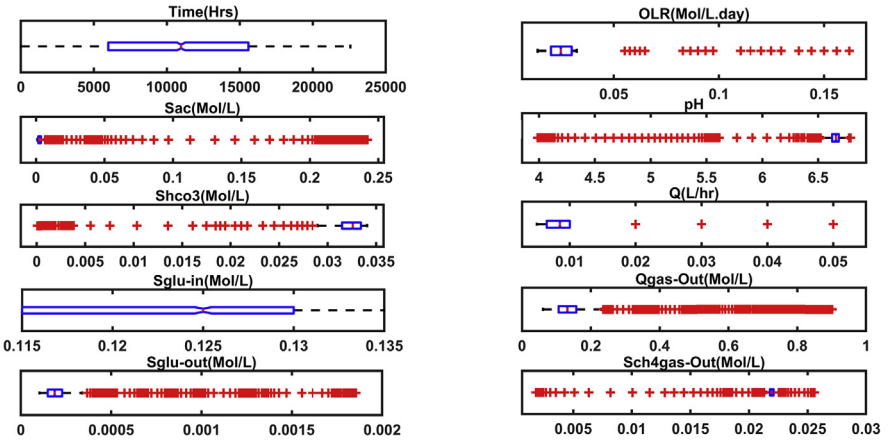

Fig. 5. Box Plots of Validation data D3.

From the correlation matrix, it can be seen that pH, Shco3 and methane output have more negative correlation with other parameters. The input glucose stands uninfluenced by other factors. Strong positive correlation is seen in OLR.

The KSOM model consists of nine parameters pre-processed from the Matlab-Excel dynamic model. These components were used to create the U-Matrix and clusters, Fig. 6. The entire quantity of nodes in the map is 117, which are demonstrated in hexagonal grids. After the creation of the map, correlations between components is analysed by placing the component planes serially as shown in Fig. 7.

Every hexagon in individual component plane characterises a single map node and the colour coding gives respective values. Hexagons located in identical place but on various component planes signify similar map node and depict component values in weight vector of that particular node. Colour coding adjacent to each component planes gives the connection between the colour and the values. From Fig. 7, it can be seen that high values are marked by brown, mid-range values by yellow-green and low values by blue. This shows the best clustering structure.

Table 5. Characteristics of trained SOM.

Characteristics	Values
Normalization Method	"var": x' = (x - $\bar{x}$) /σ_x
Codebook	117 * 19
Neighbourhood Function	Gaussian
M Size	13 * 9
Lattice	"Hexa"
Shape	Sheet
Final Quantization Error	0.5653
Final Topographic error	0.0420

https://doi.org/10.1016/j.heliyon.2019.e01511

Table 6. Correlation matrix for variables in the features.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
OLR	1.00								
Sac	0.78	1.00							
pH	−0.77	−0.89	1.00						
Shco3	−0.64	−0.74	0.84	1.00					
Q	0.63	0.31	−0.45	−0.71	1.00				
Sglu_in	0.19	0.21	−0.06	0.380	−0.47	1.00			
Qgas_out	0.99	0.82	−0.80	−0.64	0.57	0.24	1.00		
Sglu_out	0.52	0.12	−0.30	−0.51	0.92	−0.41	0.43	1.00	
Sch4Gasout	−0.77	−0.84	0.98	0.82	−0.49	−0.09	−0.79	−0.37	1.00

Prime importance is given to Methane gas output (Sch4Gas_Out). From Fig. 7, it can be seen that the Sch4Gas_Out component plane is distinctly divided. The figure can be divided into 3 sects with the upper left portion covering the low values; bottom & central left covering the mid-range vales and the central right covering the high methane output values. By comparing this with the other component planes, the high methane output has a positive correlation with pH and bicarbonate (Shco3). The pH values between 6 and 7.92 favours methane production of up to 0.0329 Mol/Lg. Bicarbonate concentration of 0.017−0.0339 mol/L in the system favours methane production. Acidic pH of less than four has an adverse effect on methane production as it dips to as low as 0.0016 Mol/Lg. Hence, pH plays a pivotal role in methane production as authenticated by the study conducted by Rizzi et al. (2006) and (Zuo et al., 2013).

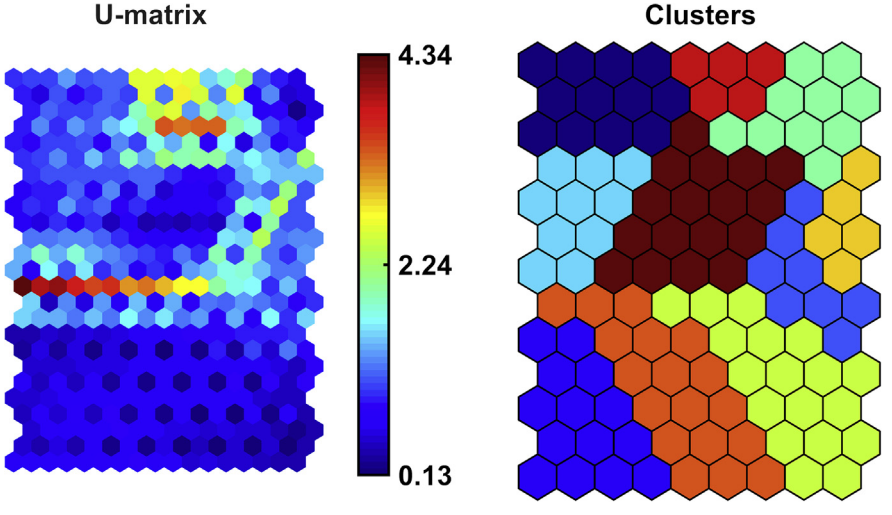

Fig. 6. U-matrix and clusters.

https://doi.org/10.1016/j.heliyon.2019.e01511

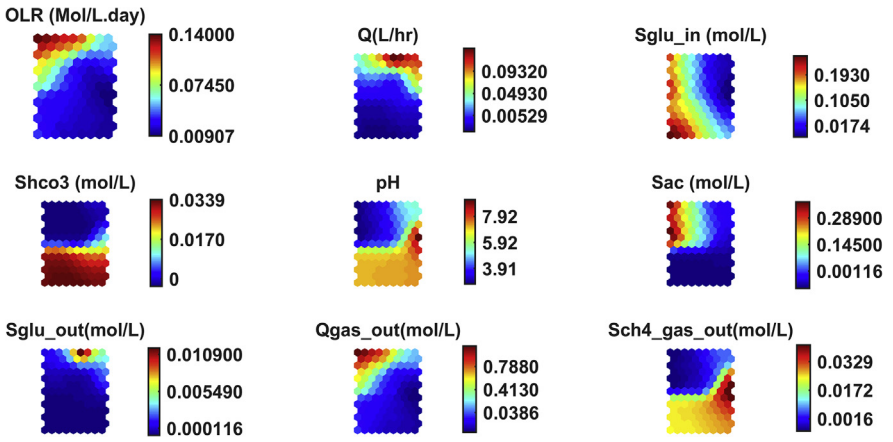

Fig. 7. Component planes of the KSOM Model.

Sch4gas_Out component plane have opposite distribution on the map with OLR and acetic acid (Sac) indicating a negative correlation. Acetic acid (volatile fatty acid) in small amounts of 0.1−0.001 mol/L is beneficial for methane production, but higher amounts are inhibitory as it can subsequently lead to drop in pH that has been studied in the research conducted by Brown and Li (2013). Similarly, high OLR (glucose substrate) of 0.07−0.14 Mol/L. day was found to reduce methane production. An optimal OLR range for efficient methane generation was 0.009−0.06 Mol/L. day from this research model.

The upper middle part of the Sglu_Out component plane has high glucose content in the effluent whereas the rest of the map shows very low glucose concentration indicating that the process is highly efficient in glucose reduction. This has a strong positive correlation with the flowrate. High flow rate of 0.09 L/hr causes an imbalance in the process by reducing the HRT and glucose reduction. Even a high input glucose has not affected the output levels. Most of the other parameters also have very less influence on the output glucose concentration indicating a stable reduction.

From Fig. 7, pH has an influence on almost all the parameters of the process. A balance between pH and Shco3 can be seen wherein higher pH is complemented by higher Shco3. This could be due to the nature of the substrate or alkalinity addition in the process. Higher OLR caused a dip in pH mainly because of accumulation of volatile fatty acids (Rincón et al., 2008). This has been reflected by the negative correlation between acetic acid and pH. For optimal pH range, acetic acid levels of 0.1−0.001 are advisable. These cross-correlations from visual interpretations of the KSOM seem to agree with the much more simple correlation matrix showing the linear relationship between process variables.

The errors during training and validation are summarised in Tables 7, 8, 9, and 10. It can be seen that for most of the parameters the correlation co-efficient is above 0.9. Exception to this can be seen in validation data D1 suggesting that the data does not

https://doi.org/10.1016/j.heliyon.2019.e01511

Table 7. Model evaluation criteria of Training Data.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
R	0.9857	0.9895	0.9728	0.9931	0.9810	0.9873	0.9877	0.9475	0.9889
AAE	0.0039	0.0064	0.1062	0.0009	0.0027	0.0063	0.0202	0.0002	0.0007
MSE	4.4E-05	0.0002	0.0849	0.00	2.5E-05	0.0001	0.0011	0.00	0.00
RMSE	0.0066	0.0138	0.2914	0.0018	0.0051	0.0095	0.0336	0.0008	0.0016

Table 8. Model evaluation criteria of Validation Data D1.

	OLR (Mol/L.day)	pH	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)
R	0.8692	0.6725	0.8598	0.8840	0.7836
AAE	0.0240	1.1816	0.0992	0.0498	0.1566
MSE	0.0010	1.7762	0.0102	0.0034	0.0381
RMSE	0.0312	1.3327	0.1011	0.0586	0.1951

Table 9. Model evaluation criteria of Validation Data D2.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
R	0.9548	0.9536	0.921	0.9767	0.9359	0.9686	0.9552	0.8693	0.9371
AAE	0.0077	0.0164	0.303	0.0024	0.0051	0.0100	0.0436	0.0003	0.0024
MSE	0.0001	0.0011	0.406	0.0000	0.0001	0.0002	0.0038	6.8E-07	1.6E-05
RMSE	0.0110	0.0335	0.637	0.0039	0.0086	0.0137	0.0620	0.0008	0.0040

fully comply with the developed model. The AAE for methane gas production and glucose output are very minimal for all the datasets which shows that the model performs well in the predicting the parameters. All the RMSE values are also below 0.5 (except D1) which further proves that the efficiency of the model. However, pH value errors seem to be varying with some showing slightly higher values that prove that this parameter is highly non-linear and difficult to control in an anaerobic digestion process.

Table 10. Model evaluation criteria of Validation Data D3.

	OLR (Mol/L.day)	Sac (mol/L)	pH	Shco3 (mol/L)	Q (L/hr)	Sglu_in (mol/L)	Qgas_out (mol/L)	Sglu_out (mol/L)	Sch4_gas_out (mol/L)
R	0.9664	0.8956	0.960	0.9683	0.9713	0.7620	0.9573	0.9464	0.9490
AAE	0.0061	0.0249	0.195	0.0033	0.0054	0.0312	0.0361	0.0004	0.0019
MSE	0.000	0.0016	0.088	0.00	0.00	0.0014	0.0035	0.00	0.00
RMSE	0.0092	0.0406	0.297	0.0044	0.0094	0.0370	0.0588	0.0013	0.0025

https://doi.org/10.1016/j.heliyon.2019.e01511

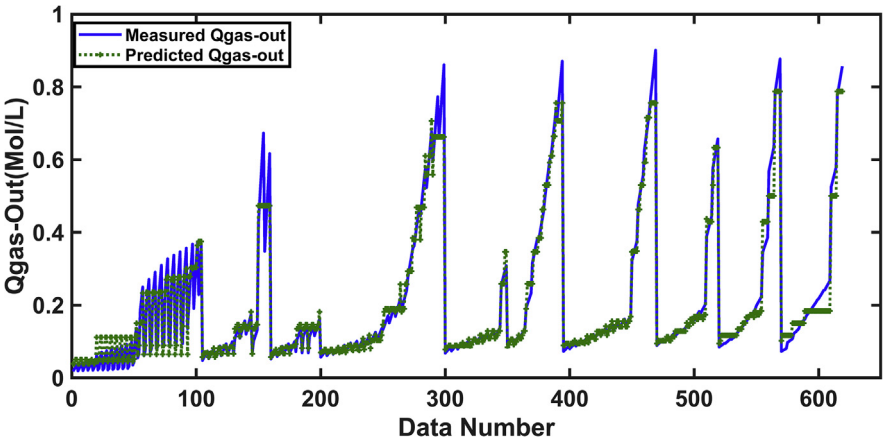

Fig. 8. Training data Time Series Plot of Biogas Output.

Further analysis was conducted due to the satisfactory model performance to evaluate the efficiency of the KSOM in predicting process through time series and scatter plots. It can be deciphered that the performance is good comparing with the model evaluation criteria tables presented above. The methane and glucose output has also been predicted well, which are key parameters in controlling anaerobic digestion process. The values, predicted and observed, of training and validation data for Biogas Output is shown in Figs. 8 and 9. Time series plots of training and validation datasets for rest of the parameters were done and evaluated.

The training and validation D2 plots show an excellent match between measured and predicted data on a temporal scale. This shows that for steady state values within range of the training data, the model performs the best. The predicted values for pH, Sac, Shco3, glucose output, gas and methane production closely follows the observed values even when there are variations in organic loading to the system. However, the model prediction seems to be compromised for validation data D1

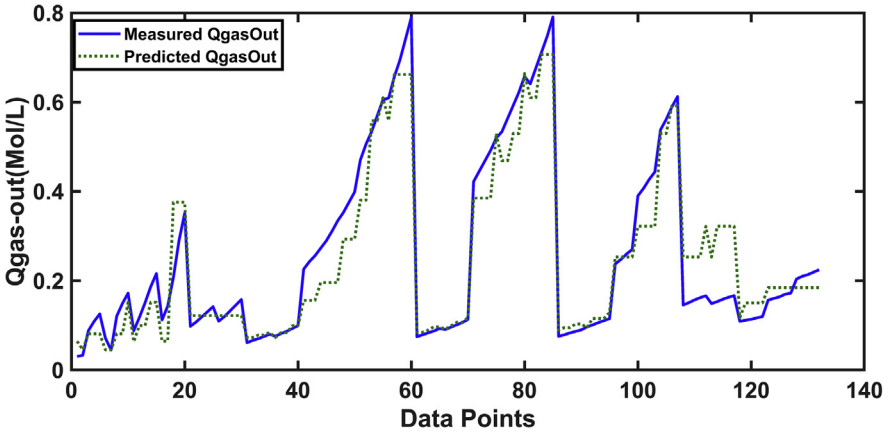

Fig. 9. Validation data D2-Time Series Plot of Biogas Output.

https://doi.org/10.1016/j.heliyon.2019.e01511

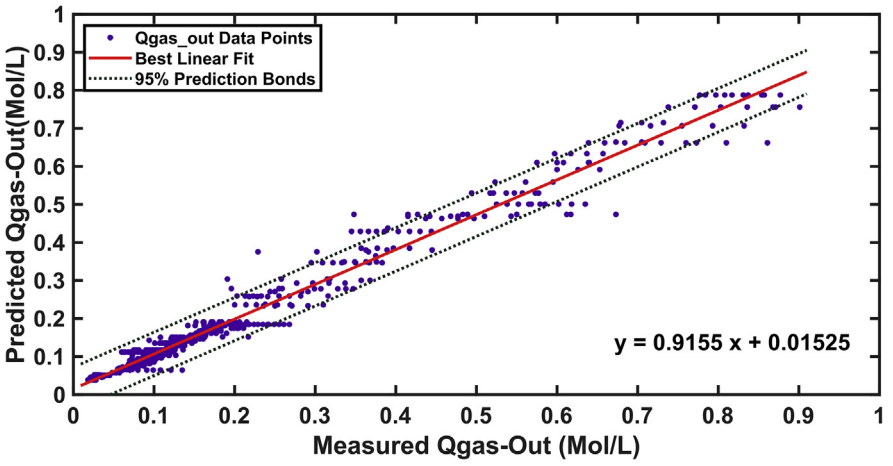

Fig. 10. Training data Scatter Plot of Biogas Output.

that has also been reflected in the model evaluation criteria, Table 8. The predicted values follows the general trend of the observed values for pH, OLR, Q and gas production, however there is a range gap between the two. This could be mainly due to the fact that the data for validation has been generated from a hybrid model of anaerobic digestion and not a complete stirred tank reactor process as is used for model generation. For the final set D3, for an initial period, the model predicts well for pH, gas output, as well as glucose output. With time, there is variation in the performance but still follows the general trend of the process. This proves that the model can be used for dynamic data, but with less certainty than steady state values. These plots demonstrate a good match with the corresponding evidence presented in the model evaluation criteria.

Figs. 10 and 11 show the scatter plots of training and validation data for Biogas Output. The remaining scatter plots of the training and validation datasets were also generated but not shown due to limited space. Concerning the scatter plots,

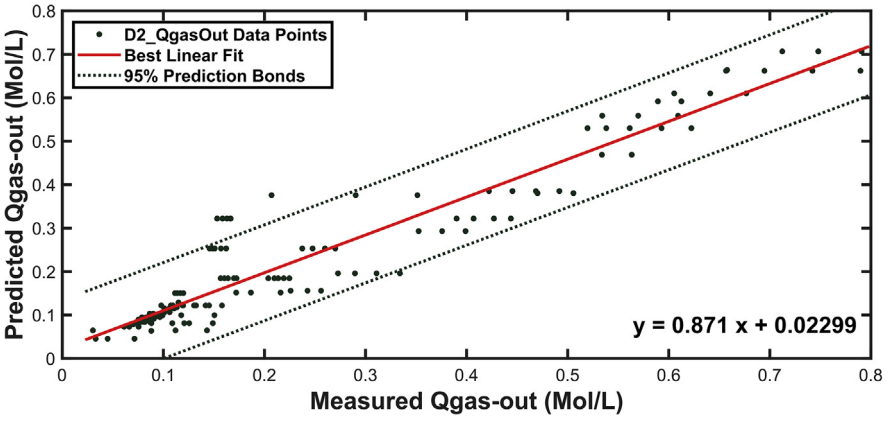

Fig. 11. Validation data D2 scatter plot of biogas output.

https://doi.org/10.1016/j.heliyon.2019.e01511

almost all of the data points lie within the 95% prediction limits. This matches the model evaluation criteria portraying minimal errors for the variables. The training and validation plot D2 show very good prediction capability. Unlike the time series plot, the D1 scatter plots also shows that most of the points are within the 95% prediction bonds. From this, we can draw the conclusion that the model has average prediction capability even for different kinds of anaerobic process, but should be cross-examined from case to case. Validation data D3 scatter points illustrate a slightly different scenario with some points outside the 95% prediction bonds and also clustered. This is because of dynamic nature of the data, which is reflected in the time series plots as well.

In total, the scatter plots of training and validation data indicates the non-linear nature of the datasets and anaerobic process parameters in general.

4. Conclusion

Modeling of anaerobic digestion process is useful for design and efficient functioning of anaerobic wastewater treatment plants. This model developed herein has been, to a varying degree, successful in predicting digestion operation, failure, and possible remedies. The Matlab-Simulink model offered good simulation platform for the anaerobic digestion process that in turn helped in generation of datasets for the KSOM model. Process variables were clustered using KSOM and the variation of effluent characteristics with its relation with the other parameters can be easily evaluated by the map despite process complexity. KSOM was successful in defining multifaceted associations between variables with no previous information about the mechanisms of the anaerobic process. Given below are the specific findings of the research:

- From the results, it can be seen that this model can be used for anaerobic wastewater treatment plants with COD loading from 2 to 12 gCOD/day. The biogas generation is found to increase with higher OLR. It could be seen that even when gas flow rate is high, as methanogenesis is inhibited, methane production is lower.
- The difference in predicted and measured values showed that this model would not be an ideal solution for hybrid reactors employing both suspended and attached growth for anaerobic digestion.
- The validation of the model with less data has proved that KSOM can be used even if there are missing data.
- Although the model is based on steady state continuously stirred tank reactor, it can handle a dynamic system to a good degree.

https://doi.org/10.1016/j.heliyon.2019.e01511

The feasibility of application of this model in the field has a better chance than other models like ADM1 due to consideration of only key parameters of the process. However, a downside of this is that its effectiveness is also limited by these parameters.

The KSOM model developed in this work is an adequate tool for modeling the process of anaerobic digestion, which is ultimately the main aim of the research. Comprehension of the system behaviour through parameter correlation, expression, and verification of model using KSOM, prediction of system behaviour through process stability and performance of future circumstances and knowledge extraction of anaerobic digestion process could be done as illustrated in the Results and Discussion. This model could offer some promise in process control, design, and study of anaerobic wastewater treatment using the KSOM.

This research would add to the knowledge database of modeling anaerobic digestion process using KSOM, which is an area not very, researched upon. However, the knowledge gathering needs to continue to include more parameters and process conditions such as temperature. The efficiency of the model in prediction of higher COD load conditions is low which needs to be inculcated. Dynamic data handling capacity of the model also needs to be improved, as this would be very useful during on-field conditions of process monitoring and control. With the knowledge extracted from the model together with data correlation, fuzzy logic can be used to develop rules (Mamdani system) and further add features. A software or GUI (graphical user interface) can be developed to predict and monitor process parameters for anaerobic wastewater treatment plants. The model can be extended to various kinds of anaerobic treatments and effect of inhibitors on the process.

Declarations

Author contribution statement

Rabee Rustum: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Anjali Ramachandran: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Adebayo Adeloye: Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

https://doi.org/10.1016/j.heliyon.2019.e01511

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Appels, L., Baeyens, J., Degrève, J., Dewil, R., 2008. Principles and potential of the anaerobic digestion of waste-activated sludge. Prog. Energy Combust. Sci.

Asadi, A., Verma, A., Yang, K., Mejabi, B., 2017. Wastewater treatment aeration process optimization: a data mining approach. J. Environ. Manag.

Batstone, D.J., Keller, J., Angelidaki, I., Kalyuzhnyi, S.V., Pavlostathis, S.G., Rozzi, A., et al., 2002. The IWA anaerobic digestion model no 1 (ADM1). Water Sci. Technol. 45 (10), 65−73.

Begum, S.F., Kaliyamurthie, K.P., Rajesh, A., 2016. Comparative study of clustering methods over Ill-structured datasets using validity indices. Indian J. Sci. Technol.

Bjornsson, L., 2000. Intensification of the Biogas Process by Improved Process Monitoring and Biomass Retention. Dept. of Biotechnology.

Blumensaat, F., Keller, J., 2005. Modelling of two-stage anaerobic digestion using the IWA Anaerobic Digestion Model No. 1 (ADM1). Water Res.

Boe, K., 2006. Online Monitoring and Control of the Biogas Process. Institute of Environment & Resources Technical University of Denmark.

Brown, D., Li, Y., 2013. Solid state anaerobic co-digestion of yard waste and food waste for biogas production. Bioresour. Technol.

Garća, H.L., González, I.M., 2004. Self-organizing map and clustering for wastewater treatment monitoring. Eng. Appl. Artif. Intell.

Henze, M., van Loosdrecht, M.C.M., Ekama, G.A., Brdjanovic, D., 2008. Wastewater treatment development. Biol. Wastewater Treat.

Holubar, P., Zani, L., Hager, M., Fröschl, W., Radak, Z., Braun, R., 2002. Advanced controlling of anaerobic digestion by means of hierarchical neural networks. Water Res.

Hsu, C.C., 2006. Generalizing self-organizing map for categorical data. IEEE Trans. Neural Netw.

https://doi.org/10.1016/j.heliyon.2019.e01511

Jha, A., Li, J., Nies, L., Zhang, L., 2013. Research advances in dry anaerobic digestion process of solid organic wastes. Afr. J. Biotechnol.

Jørgensen, P.J., 2009. Biogas-green Energy. Faculty of Agricultural Sciences, Aarhus University, Tjele, Denmark.

Kangas, J., Kohonen, T., 2003. Developments and applications of the self-organizing map and related algorithms. Math. Comput. Simulat.

Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J., 1996. Engineering applications of the self-organizing map. Proc. IEEE.

Kusiak, A., Wei, X., 2014. Prediction of methane production in wastewater treatment facility: a data-mining approach. Ann. Oper. Res.

Lauwers, J., Appels, L., Thompson, I.P., Degrève, J., Van Impe, J.F., Dewil, R., 2013. Mathematical modelling of anaerobic digestion of biomass and waste: power and limitations. Prog. Energy Combust. Sci. 39 (4), 383−402.

Li, Y., Park, S.Y., Zhu, J., 2011. Solid-state anaerobic digestion for methane production from organic waste. Renew. Sustain. Energy Rev.

Liukkonen, M., Laakso, I., Hiltunen, Y., 2013. Advanced monitoring platform for industrial wastewater treatment: multivariable approach using the self-organizing map. Environ. Model. Softw 48, 193−201.

Machón-González, I., Rodríguez-Iglesias, J., López-García, H., Castrillón-Peláez, L., Marañón-Maison, E., 2017. Knowledge extraction from a nitrification denitrification wastewater treatment plant using SOM-NG algorithm. Environ. Technol.

Mathworks, 2011. MATLAB: Getting Started Guide. R2011b.

Metcalf, W., Eddy, C., 2003. Metcalf and Eddy wastewater engineering: treatment and reuse. In: Wastewater Engineering: Treatment and Reuse. McGraw Hill, New York, NY.

Ozkaya, B., Demir, A., Bilgili, M.S., 2007. Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors. Environ. Model. Softw.

Pai, T.Y., Wan, T.J., Hsu, S.T., Chang, T.C., Tsai, Y.P., Lin, C.Y., et al., 2009. Using fuzzy inference system to improve neural network for predicting hospital wastewater treatment plant effluent. Comput. Chem. Eng.

Parawira, W., Murto, M., Read, J.S., Mattiasson, B., 2005. Profile of hydrolases and biogas production during two-stage mesophilic anaerobic digestion of solid potato waste. Process Biochem.

https://doi.org/10.1016/j.heliyon.2019.e01511

Parker, W.J., 2005. Application of the ADM1 model to advanced anaerobic digestion. Bioresour. Technol.

Ramos, M.J.C., Gonzalez, I.M., García, H.L., Rolle, J.L.C., 2013. Visual supervision of a waste water biological reactor using artificial intelligence algorithms. In: Conference and Exhibition - 2013 International Conference on New Concepts in Smart Cities: Fostering Public and Private Alliances, SmartMILE 2013.

Rincón, B., Borja, R., González, J.M., Portillo, M.C., Sáiz-Jiménez, C., 2008. Influence of organic loading rate and hydraulic retention time on the performance, stability and microbial communities of one-stage anaerobic digestion of two-phase olive mill solid residue. Biochem. Eng. J.

Rizzi, A., Zucchi, M., Borin, S., Marzorati, M., Sorlini, C., Daffonchio, D., 2006. Response of methanogen populations to organic load increase during anaerobic digestion of olive mill wastewater. J. Chem. Technol. Biotechnol.

Rodríguez, J., Premier, G.C., Dinsdale, R., Guwy, A.J., 2009. An implementation framework for wastewater treatment models requiring a minimum programming expertise. Water Sci. Technol.

Rustum, R., 2009. Modelling Activated Sludge Wastewater Treatment Plants Using Artificial Intelligence Techniques (Fuzzy Logic and Neural Networks). PQDT – UK & Ireland. Retrieved from. http://easyaccess.lib.cuhk.edu.hk/login?url=http://search.proquest.com/docview/1774184029?accountid=10371%5Cnhttp://findit.lib.cuhk.edu.hk/852cuhk/?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:P.

Rustum, R., Adeloye, A., 2013a. Improved modelling of wastewater treatment primary clarifier using hybrid anns. Int. J. Comput. Sci. Artif. Intell.

Rustum, R., Adeloye, A.J., 2007. Replacing outliers and missing values from activated sludge data using kohonen self-organizing map. J. Environ. Eng. 133 (9).

Rustum, R., Adeloye, A.J., 2013b. Hybrid unsupervised-supervised artificial neural networks for modelling activated sludge wastewater treatment plants. In: Bacciga, A., Naliato, R. (Eds.), Recent Advances in Artificial Intelligence Research, first ed. Nova Science Publishers, pp. 1–26.

Rustum, R., Adeloye, A.J., Scholz, M., 2008. Applying Kohonen self-organizing map as a software sensor to predict biochemical oxygen demand. Water Environ. Res.: A Res. Publ. Water Environ. Fed. 80 (1), 32–40.

Rustum, R., Forrest, S., 2017. Fault detection in the activated sludge process using the kohonen self-organising map. In: 8th International Conference on Urban Planning, Architecture, Civil and Environment Engineering. Dubai, UAE.

https://doi.org/10.1016/j.heliyon.2019.e01511

Seadi, T.A., Rutz, D., Prassl, H., Köttner, M., Finsterwalder, T., Volk, S., Janssen, R., 2008. Biogas Handbook. University of Southern Denmark Esbjerg.

Szelag, B., Gawdzik, A., Gawdzik, A., 2017. Application of selected methods of black box for modelling the settleability process in wastewater treatment plant. Ecol. Chem. Eng. S.

Vermasvuori, M., Endén, P., Haavisto, S., Jämsä-Jounela, S.L., 2002. The use of Kohonen self-organizing maps in process monitoring. In: 2002 1st International IEEE Symposium.

Vesanto, J., 2000. Neural network tool for data mining: SOM toolbox. In: Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET 2000).

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM toolbox for Matlab 5. Tech. Rep. A57 2 (0), 59.

Waewsak, C., Nopharatana, A., Chaiprasert, P., 2010. Neural-fuzzy control system application for monitoring process response and control of anaerobic hybrid reactor in wastewater treatment and biogas production. J. Environ. Sci.

Yang, Q., Luo, K., Li, X. ming, Wang, D. bo, Zheng, W., Zeng, G. ming, Liu, J. jin, 2010. Enhanced efficiency of biological excess sludge hydrolysis under anaerobic digestion by additional enzymes. Bioresour. Technol.

Zuo, Z., Wu, S., Zhang, W., Dong, R., 2013. Effects of organic loading rate and effluent recirculation on the performance of two-stage anaerobic digestion of vegetable waste. Bioresour. Technol.

https://doi.org/10.1016/j.heliyon.2019.e01511