Suitability of ANN and GP for Predicting Soak Pit Tank Efficiency under Limited Data Conditions

Naresh Sharma^{1,*}, Sivapragasam Chandrasekar², and Krishnan Sundar¹

¹Department of Biotechnology, Kalasalingam Academy of Research and Education, Tamilnadu 626126, India

²Centre for Water Technology, Civil Engineering Department, Kalasalingam Academy of Research and Education, Tamilnadu 626126, India

Abstract. Under Industry 4.0 scenario, smart sensors can be suitably integrated with control systems to monitor the treatment of wastewater systems. Such control systems need a sound modelling or forecasting tool for the real time monitoring. This work reports a pilot study to model the treatment efficiency of soak pit tanks for the treatment of grey water using Genetic Programming (GP) and Artificial Neural Network (ANN). Only the inlet total suspended solid is considered for modelling. The Root Mean Square of Errors for GP and ANN run models were found to be 1.8 and 12.5 respectively for Tank 1. The results indicate that GP is a more promising tool than ANN particularly when modelling under limited data conditions. The difference in performance of both the methods seem to depend on the type learning mode adopted in each case.

1 Introduction

Grey water can be defined as any domestic wastewater (except sewage or black water) which is an inevitable major pollution point source from residential and commercial areas of a community. The grey water is characteristically different from black water in terms of less organic loading, less pathogenic microbes and faster decomposition due to reduced nutrient/carbon content [1]. Although grey and black water vary significantly in their characteristics; the current practice of mixing grey and black water increases the cost and space required for treatment. Grey water treatment is important because it decreases the dependence on fresh water usage by 50-60% and it reduces the burden on STP (sewage treatment plant) by reducing the space and energy requirement for treatment [2].

One of the best ways to manage grey water is either to re-use or recharge the groundwater using treated grey water. The problems with re-use of grey water are that it can pose serious health risks – (i) if watered to fruits or vegetable plants or (2) and if the

* Corresponding author: naresh@klu.ac.in

hard bleach or chemicals used are not treated to standard levels. Of the several techniques available for grey water treatment, soak pit has attracted considerable attention as it is quick and cheap to build, requires very less space (an important consideration for an urban region) and quickly recharges groundwater [3]. Although eventual clogging is inevitable in soak pits; optimization, proper design and effective pre-treatment of grey water can prolong the age of soak pit making it an effective alternative for grey water treatment. The present study focuses on assessing the performance of two soak pits (with difference in the quantity of brick ballasts and gravels used) in the removal of total suspended solids (TSS).

The solids removal in soak pit is a function of many interrelated complex parameters such as concentration of particulates (suspended and dissolved), organic and inorganic chemicals, pH, amount of surfactants etc., and it is relatively difficult to study and interrelate all the parameters using experiments. Modeling of the treatment systems is one way researchers and engineers resort to develop a better and deeper understanding of the process. Of late, data driven modeling tools are more being used to develop models and to mine knowledge of which Artificial Neural Networks (ANN) and Genetic Programming (GP) are more commonly used especially to model complex non-linear processes without considering all details of the process. Under the Industry 4.0 scenario where smart sensors can be integrated to control systems for effective monitoring of treatment efficiency, it is necessary to use a sound forecasting model to support the control system. While ANN has seen application in many aspects of wastewater treatment, GP is a relatively new technique as far as its application in wastewater treatment is concerned. This study is aimed to investigate the suitability of ANN and GP in the modeling of treatment efficiency of soak pit for TSS removal under limited data conditions knowing only the initial solids concentration in grey water as a function of time. From the performance of the modeling tools under limited data conditions, it is intended to gain a confidence in choosing the most appropriate data driven modeling technique for large scale application of wastewater treatment with more complex inter-related input parameters.

2 Experimental setup and tools used

2.1 Soak pit tank

A soak pit is essentially a hole designed with the purpose of allowing wastewater to infiltrate into the ground. These are used for the discharge of domestic and industrial wastewater. Certain design mandates are to be ensured for design of soak pits such as the depth of the soak pit should be between 1.5m to 4m but never less than 2m above the GWL and should be located at a safe distance from drinking water source (minimally 30m). The size of soak pits is dependent on two things, the infiltration rate of the soil and the quantity of waste water being put into it. In this study, two lab scale soak pits of 1 m depth were constructed (Tanks – T_1 and T_2) handling 20L of wastewater (grey water only) each day. The difference between two tanks is the quantity of brick ballast and gravels used. Performance of both the tanks was quantified in terms of removal of solids (TS, TDS, and TSS) and pH. The methods set by APHA, 2005 [4] were followed for measuring TSS. The filter media compositions for the two tanks are shown in Table 1.

Table 1. Filter media composition for soak pit tanks.

Filter Media	Tank 1 (depth in mm)	Tank 2 (depth in mm)	Purpose
20 mm gravel	0.1	0.2	Filtration of suspended solids
40 mm gravel	0.25	0.2	Filtration and structural support to tank/soak pit.
Brick ballast	0.4	0.3	Filtration of dissolved solids.

2.2 Genetic Programming

Genetic Programming (GP) is very similar to Genetic Algorithm (GA), an evolutionary algorithm based on Darwinian theories of natural selection and survival of the fittest with a characteristic difference that while GA operates based on bit strings, GP operates on parse trees to approximate the equation (in symbolic form) that best describes how the output relates to the input variables. By the random combination of input variables and functions, first an initial population of randomly generated equations (programs) is evolved. The functions can include arithmetic operators (plus, minus, multiply, divide), mathematical functions (sine, cosine, exponential, logarithmic), logical/comparison functions (OR/AND) etc. The choice of appropriate functions mandates at least some understanding of the process being modeled. The initial population is processed through an evolutionary process and the quality of the evolved programs are checked by means of a defined ‘fitness’ measure. The best program based on the fitness measure are then selected from the initial population and better programs are evolved by exchange of part of information between them by employing ‘crossover’ and ‘mutation’ as used in GA which is very similar to the natural reproduction process [5]. The crossover and mutation rates need to be carefully decided. A high mutation will introduce high variability in the programs developed and might hinder in the convergence of the evolutionary process. The user must also decide a number of GP parameters before applying the algorithm to model the data, such as population size, number of generations etc. The programs that fitted the data less well as per the fitness measure are discarded. This evolution process is repeated over successive generations and is driven towards finding symbolic expressions describing the data, which can be scientifically interpreted to derive knowledge about the process being modeled [6]. GP is implemented in this study using Discipulus software.

2.3 Artificial neural network

The architecture of ANN is motivated by the structure of the human brain and nerve cells. This technique is used in identifying the statistical pattern present in the time series and applies it to unknown data to predict. A network of countless simple elements called neurons with a small amount of local memory is considered. The neurons are connected through connections which carry numeric data encoded by various means. Each neuron operates only when it receives data through the connections. The architecture is formed by the learning algorithm which is responsible for the extraction of the regularities present in

the data through the finding of a suitable synapses set during the process of observation of the examples. Accordingly, ANNs solve problems by self-learning.

The feed-forward architecture is used in this study. One input layer, one output layer and one or more hidden layers are available in the architecture. The information passes from the input layer to the output layer through hidden layer. Each layer is fabricated by several neurons, and the layers are interconnected by a set of weights. Neurons operate the input and transform it to produce an analog output. More details about ANN have been discussed in [7 - 10].

2.4 Performance Measure

The Root Mean Square of Errors (RMSE) value is taken as the performance measure to check the performance of GP and ANN as shown in Eq (1).

$$RMSE = \sqrt{\frac{1}{n} [(X_m)_i - (X_s)_i]^2}$$

(1)

3 Results and Discussion

The chemical characteristics monitored in grey water were pH, Total solids, Total Suspended solids and Total dissolved solids according to methods set by APHA, 2005 [4]. The readings were taken for 3 weeks and are shown in Table 2.

Table 2. Inlet and outlet Characteristics of grey water from the soak pit tanks 1 and 2.

Days	Inlet concentration (in mg/L except pH) (average of three weeks)				Outlet concentration (in mg/L) (average of three weeks)	
	pH	TS	TDS	TSS	TSS (Tank1)	TSS (Tank 2)
Day 1	12.4	963.7	523.0	440.7	358.0	308.7
Day 2	12.3	1070.0	602.3	467.7	358.7	430.7
Day 3	12.4	1638.7	723.3	915.3	605.0	577.7
Day 4	12.4	1506.7	542.0	963.7	614.7	500.7
Day 5	12.3	1332.0	512.3	819.7	462.3	446.0

As seen from the table, the inlet characteristics vary substantially as the week progresses and starts decreasing towards weekends. Table 3 shows the training and validation data set taken for GP and ANN modelling. Total of 15 data sets were chosen for this study, of which 10 sets were used to train, three for validation and 2 sets for application. Table 4 and Table 5 show the comparison of actual removal efficiencies with the predicted output values from GP and ANN from soak pit tank1 and tank 2 respectively. Figure 1 and Figure 2 compares the removal efficiency graphically as obtained from soak pit tank 1 and tank 2 respectively on the performance of ANN and GP.

As seen from the Tables (4 and 5) and Figures (1 and 2), GP performs better than ANN for soak pit tank 1 and as good as ANN for soak pit tank 2. An attempt is also made to

mathematically model the inlet and outlet TSS concentration of the functional form as shown in Equation (2) using GP:

$$TSS_{\%out} = f(TSS_{in}, t) \tag{2}$$

Where $TSS_{\%out}$ is the TSS removal in percentage, TSS_{in} is the input TSS and t is the time.

The GP evolved models are given by equation (3) and (4) for soak pit tank 1 and tank 2 respectively.

$$TSS_{\%out} = \sqrt{TSS + t} \tag{3}$$

$$TSS_{\%out} = 0.136 * \sqrt{TSS * t} \tag{4}$$

It is to be noted that these models are only approximate models to provide an overall behaviour of the system and is not the exact equation.

Table 3. Training and validation data sets used for modelling.

Data	Tank 1			Tank 2		
	Time (h)	Inlet TSS (mg/L)	TSS% _{out}	Time (h)	Inlet TSS (mg/L)	TSS% _{out}
Training	24	410	15.9	48	420	5.0
	48	495	21.8	120	898	54.3
	72	942	26.8	24	410	24.9
	96	1002	28.3	48	495	6.5
	120	937	47.5	72	942	33.4
	24	460	26.1	96	1002	47.6
	48	488	20.1	120	937	49.2
	72	972	57.5	96	968	52.8
	96	968	49.9	48	488	11.9
	120	898	49.3	24	460	37.0
Testi ng	24	452	13.9	72	972	50.9
	48	420	28.8	24	452	27.4
	72	832	14.4	120	624	27.6

The prediction of models with actual removal efficiencies from both soak pit tanks are shown in Fig. 1 and 2. It can be seen from the Fig.1, for time 96h, with an inlet TSS of 921 mg/L for tank 1, GP and ANN predicts a removal efficiency of 31.8 and 38.8 while the experimental (actual) efficiency was 30.4; while for tank 2, with an actual removal efficiency of 43.5, GP and ANN showed an removal efficiency of 45.1 and 46.5, making it clear that GP is able to model the soak pit tanks with limited data conditions more efficiently than ANN predictions.

Table 4. Comparison of actual removal efficiency with ANN and GP predicted removal efficiency for validation data from soak pit tank 1.

Data	Tank 1				
	Time	Inlet TSS	TSS% _{out}	GP TSS% _{out}	ANN TSS% _{out}
Valid ation	96	921	30.4	31.8	38.8
	120	624	29.4	27.2	13.9
	RMSE			1.8	12.5

Table 5. Comparison of actual removal efficiency with ANN and GP predicted removal efficiency for validation data from soak pit tank 2.

Data	Tank 2				
	Time	Inlet TSS	TSS% _{out}	GP TSS% _{out}	ANN TSS% _{out}
Valid ation	72	832	24.3	28.3	27.4
	96	921	43.5	45.1	46.5
	RMSE			3.0	2.9

A closer look at the equations (3) and (4) reveal that while for soak pit tank 1, the removal is directly proportional to the sum of TSS and t while for soak pit tank 2; it is a product of the two terms. This indicates that in soak pit tank 2, the removal process proceeds faster than soak pit 1.

While back propagation ANN functions on the principle of redistribution of error to update the weights, GP functions on the evolving functions to model the process. If the data set has no clear regularities but some functional relationship exist between input and output vectors, ANN might not succeed with limited data whereas GP can map the input and output relatively easily if the functions used for evolving the models are correctly chosen. The simple model obtained from GP also indicates that the process of TSS removal through filtration in the tanks is a rather simple process.

Figure. 1. Inlet concentration and soak pit tank 1 actual removal efficiency along with predicted model removal efficiency.

Figure. 2. Inlet concentration and soak pit tank 2 actual removal efficiency along with predicted model removal efficiency.

4 Conclusions

The following conclusions can be arrived at based on this study:

- (a) GP can be an effective modelling and forecasting tool in wastewater treatment when compared to ANN for modelling under limited data conditions.
- (b) The modelling/forecasting tool can be integrated with the control system for real time monitoring of the effective functioning of the treatment system.

The authors feel grateful to the Management of Kalasalingam Academy of Research and Education for funding this current study. The authors also wish to acknowledge Mrs. Mahalakshmi, PG Student and Mr. Subash, UG student, Department of Civil Engineering, KARE, Tamilnadu for their sincere efforts and timely help.

References

1. L. A. Ghunmi, G. Zeeman, M. Fayyad, and J. B. van Lier, Grey Water Treatment Systems: A Review, *Crit. Rev. Environ. Sci. Technol.*, **41** (2011)
2. F. Li, K. Wichmann, and R. Otterpohl, Review for Technological approaches for Grey Water Treatment and Reuses, *Sci. Total Environ.*, **407** (2009)
3. E. Friedler, Quality of Individual Domestic Grey water Streams and its implications for on-site treatment and reuse possibilities. *Environ. Technol.*, **25** (2004)
4. APHA, *Standard methods for the examination of water and wastewater* (American Public Health Association, Washington DC, New York, 2005)
5. J. R. Koza, *Genetic Programming: On the Programming of Computers by Natural Selection* (MIT Press, Cambridge, 1992)
6. C. Sivapragasam, N. Muttill, M. C. Jeselia, and S. Visweshwaran, Infilling of Rainfall Information Using Genetic Programming, *Aquat. Procedia - International Conference on Water Resources, Coastal and Ocean Engineering, (ICWRCOE 2015)*, 1-5 March, Surathkal, India (2015)
7. C. Sivapragasam, S. Vanitha, N. Muttill, K. Suganya, S. Suji, M.T. Selvi, R. Selvi, and S.J. Sudha, Monthly flow forecast for Mississippi River basin using artificial neural networks, *Neural Comput. Appl.*, **24** (2014).
8. S. Govindaraju Rao, Artificial Neural Networks in Hydrology II. Hydrologic Applications, *J. Hydrol. Eng.*, **5** (2000)
9. S. Haykin, *Neural networks: a comprehensive foundation* (Tsinghua University Press, Beijing, 2001).
10. C. Sivapragasam, R. Maheswaran, V Venkatesh, Genetic programming approach for flood routing in natural channels, *Hydro. Proc.* **22**, 5 (2008)