Fall 2011

# Wastewater's total influent estimation and performance modeling: a data driven approach

Rahilsadat Hosseini
*University of Iowa*

# WASTEWATER'S TOTAL INFLUENT ESTIMATION AND PERFORMANCE MODELING: A DATA DRIVEN APPROACH

by

Rahilsadat Hosseini

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Industrial Engineering
in the Graduate College of
The University of Iowa

December 2011

Thesis Supervisor: Professor Andrew Kusiak

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

MASTER'S THESIS

---

This is to certify that the Master's thesis of

Rahilsadat Hosseini

has been approved by the Examining Committee
for the thesis requirement for the Master of Science degree
in Industrial Engineering at the December 2011 graduation.

Thesis Committee: _____

Andrew Kusiak, Thesis Supervisor

_____

Yong Chen

_____

Albart Ratner

To My Parents and
Moji and Shelley

ACKNOWLEDGMENTS

I would like to thank my graduate advisor, Professor Andrew Kusiak, for providing me the opportunity to work in his laboratory. The attention to detail that he instilled onto his understudies will surely benefit us in our future work. I would also like to acknowledge Professors Marian Muste in Cyber-enabled Discovery and Innovation (CDI) team. The CDI team meetings and discussions were instrumental for inspiring new research ideas. I would like to thank Professor to give me chance of applying data mining techniques in power plant.

I would like to thank Professors Yong Chen and Professor Albert Ratner for serving on my Thesis Committee. I am also grateful for the financial support from the National Science Foundation, who provided the financial support which made my research opportunity possible.

I thank all the members of the Intelligent Systems Laboratory (ISL) who have worked with me and provided me with advice.

Without the absolute support of my family to continue further studies, my work here at the University of Iowa would not be possible. The unconditional support I received from my parents, from Tehran, Iran helped to keep my spirit strong, and provided me with moral support throughout my time in Iowa.

## ABSTRACT

Wastewater treatment plants (WWTP) involve several complex physical, biological and chemical processes. Often these processes exhibit non-linear behavior that is difficult to describe by classical mathematical models. Safer operation and control of a WWTP can be achieved by developing a modeling tool for predicting the plant performance.

In the last decade, many studies were realized in wastewater treatment based on intelligent methods which are related to modeling WWTP. These studies are about predictions of WWTP parameters, process control of WWTP, estimating WWTP output parameters characteristics. In many studies, neural network models were used to model chemical and physical attributes in the flow rate.

In this Thesis, a data-driven approach for analyzing water quality is introduced. Improvements in the data collection of information system allow collection of large volumes of data. Although improvements in data collection systems have given researchers sufficient information about various systems, they must be used in conjunction with novel data-mining algorithms to build models and recognize patterns in large data sets. Since the mid 1990's, data mining has been successfully used for model extraction and describing various phenomena of interest.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1. INTRODUCTION

The availability of quality water is a concern, and human-environment interactions still leave much to be understood. Knowledge about water transport, quality, and quantity awaits further discovery. Water quality has high variance from location to location and time to time, due to its sensitivity to both chemistry (i.e. nutrient loading), and transport (i.e. stream flow). Both human activity such as the application of fertilizers and land management practices, and meteorology play a strong role in water quality.

Accurate water quality prediction would provide us with a better understanding of the human influence on aquatic life and provide knowledge for intelligent decision making in regards to ecological conservation.

The main advantage of applying data driven techniques is that they can eliminate some of sources of errors (human sources' errors or machines' errors) because they do not require a strong physical understanding of the system to be modeled. Data is used *directly* for model building, not for validation of a theoretical physical concept.

Chapter 2 presents the application of data mining techniques for predicting rainfall around the Wastewater Reclamation Authority (WRA). Both radar and rain gauge data are used in constructing prediction models. Model accuracy is estimated using the data from the rain gauges. The models are generated by five data-mining algorithms with the decision tree algorithm produced the highest accuracy predictions.

Chapter 3 presents the application of data-mining techniques for the prediction of total influent of a wastewater treatment plant. Early prediction of total influent will ensure planned and smooth operation of the plant. The author explores radar data (reflectivity at four different heights, rain gauge data and offers influent prediction at three different time stamps in the future. Maximum prediction length being 180 minutes. Six data mining algorithms namely Naïve Bayes, k-nearest neighbor, support vector

machine, logistic regression, neural networks, and decision tree algorithms are employed to build prediction models. Models built using decision tree algorithms yielded the better prediction results. In addition, sensitivity analysis of the proposed model is done by varying the bin size of total influent. Then regression analysis was done with a NN model and maximum prediction length as 3 hours.

In chapter 4 a neural network model for performance optimization of a wastewater treatment plant is presented. The model allows for minimization of operation costs and assessment of the environmental balance (i.e. balanced chemicals' removal in flow rate of a wastewater plant). Neural networks provide effective predictive models for complex processes that are poorly described by the first principle models. The wastewater biological phenomena in wastewater treatment plants fall in such category. The neural network model is developed using the data from the Wastewater Reclamation Authority (WRA) located in Des Moines, Iowa. The model predicts the carbonaceous biological oxygen demand (CBOD) and the total suspended solids (TSS) in the effluent stream.

### 1.1 Radar-based modeling approaches in water quality

With the recent deployment of in situ instrumentation in rivers, streams, and creeks nationwide, as well as real-time data reporting via satellite communication technology, a wealth of data is available that had never before in the past. Data mining can utilize this vast base of data for pattern recognition and machine learning, so as to make accurate predictions.

### 1.2 Data-driven modeling approaches in water quality

Data mining makes models from the "ground up" rather than using the traditional top-down approach of its physics-based counterpart. As data-driven models are derived directly from the data, their accuracy is unparalleled by physics-based models.

In order to achieve high accuracy water quantity estimation, high spatiotemporal resolution precipitation data is highly desirable. There have been a few efforts to utilize

data-driven modeling for precipitation estimation via NEXRAD radar data. There have been fewer attempts to make this link between radar data and tipping bucket data with data-driven techniques. Feed forward neural network (FFNN) have applied for rainfall estimation using radar reflectivity and rain gauge data [1,2]. Trafalis *et al*. considered some different parameters, such as wind speed and bandwidth to complement reflectivity, but with unimproved results. The best performing models in the study all had MSE's less than 0.1mm/hr [3]. In this study, chapter 2, different parameters such as velocity and spectrum width are considered besides reflectivity to make a rainfall predictive classification model but with unimproved results again.

## 1.3 The multilayer perceptron (MLP)

As the algorithm used throughout this Thesis is the multilayer perceptron (MLP), otherwise known as neural network (NN) or artificial neural network (ANN), an in depth algorithm description is justified. It has found widespread success in many areas other than hydrology due to its ability to model noisy data and usefulness for both classification and regression. This section should provide insight to one of the machine learning algorithms that has been so widely labeled a "black box" model.

### 1.3.1 MLP overview

The MLPs applied in this research are feed forward backwardly propagating neural networks. The MLP's structure consists of nodes in an input layer, a hidden layer(s), and an output layer. The concept was biologically inspired to represent the human brain's ability to process in parallel, to learn from experience, and to be highly connective and modifiable. The brain also operates via supervised learning, or the ability to train itself and learn from past experiences. The brain has the ability both to feed connections forward, near sensory input, and feed connections backwards near sensory input. These connections are mimicked by the NN with the use of loops. Feed forward NNs do not have loops, while in a looping, or recurrent NN, information is fed back from

an output node to an input node [4-5]. In both categories of NNs, each input/output parameter is assigned a node in its respective input/output layer.

### 1.3.2 The MLP structure and algorithm

Figure 1.1 is a diagram of a single perceptron with two inputs**,** and a simple binary output. The inputs are multiplied by their respective weights and the products are summed at the junction. If the sum at the junction is greater than the threshold (Θ), the perceptron "fires." In the binary example, firing means outputting a "1." Equation (1.1) describes the summation that occurs at the node.

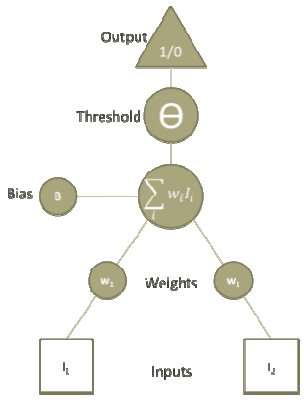



**Figure 1.1 Perceptron**

(1.1)

Where $y_j$ is the output of the $j^{th}$ node, $m$ is the number of inputs to the $j^{th}$ node, $x$ is the input value, $w$ is the input weight, and $b$ is a bias factor.

After each element in the data set, the weights for the inputs are updated, based on error. If the target value was achieved, the weights remain unchanged. Equation (1.2) describes how the neural network updates the $j^{th}$ weight in the $i^{th}$ layer.

(1.2)

Where is the learning rate, is the error attributed to the node and *f* is the activation function.

It is this recalculating of the weights that allows the neural network to "learn" a dataset. Stopping criteria is user defined, usually by limiting the number of epochs, or cycles through the data set, the model continues. The original perceptron was developed by Rosenblatt (1958) in at the Cornell Aeronautical Laboratory**,** but the observation was made that the single layer perceptron was only capable of learning when the data set was linearly separable, such as modeling the XOR gate [6]. However, after further development, multiple perceptrons were placed in layers (see figure 1.2), and the simple stepwise activation function was replaced with a continuous and differentiable sigmoidal one, so that its outputs could be continuous. The resulting structure of the perceptron when put into layers, can be seen below in the Figure (1.2) which is an MLP schematic with two hidden layers and 15 nodes. An example of the new sigmoidal activation function for continuous MLPs, in this case the logistic function, is show in equation 1.3.

—— (1.3)

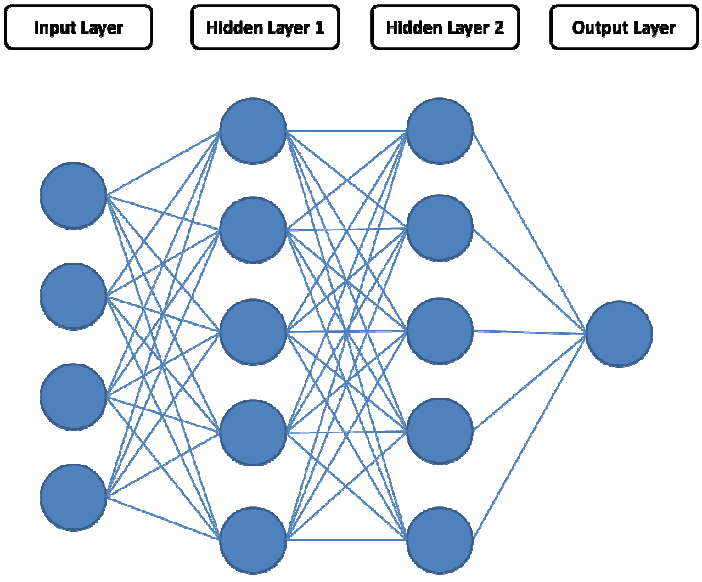



**Figure 1.2 Multilayer perceptron**

The optimal structure of a NN still remains a trial-and-error process, but there are several rules of thumb that previous researchers have found useful. For example, Tarassenko (1998) states that the number of samples in the training set should be greater than the number of synaptic weights in the network, and according to Hecht-Nielsen (1987) the number of hidden nodes, *M*, in a single hidden layer model NN is between *I* and 2*I* +1, where *I* is the number of input nodes [7-8]. Data-mining software, such as Statistica or WEKA can be a useful tool for testing multiple NN structures to find optimal results [9].

### 1.4 Decision tree

Another algorithm used throughout this Thesis is Decision trees. Decision tree is about classification. A decision tree partitions its input space to branches, that will be partitioned repetitively based on the other attributes in the model.

Any node t is splitting based on a criteria called Entropy (t) shown in equation (1.4), where $P_i$ is the probability of class i within node t. Attribute and split selection is to minimize entropy. After node splitting, two or more descendants are produced. Entropy is measured for each child and the sum of it is weighted by its percentage of the parent's cases in computing the final weighted entropy used to decide the best split [10].

$$Entropy(t) = \sum_{i=1}^{n} -\mathrm{P}i \times \log Pi \quad (1.4)$$

Given a node t, the splitting criterion used is the Gain Ratio in equation (1.5).

$$GainRatio = gain(t) \,/\, splitinformation(t) \quad (1.5)$$

This ratio expresses the proportion of information generated by a split that is helpful for developing the classification, and may be thought of as a normalized information gain or entropy measure for the test. A test is selected that maximizes this ratio, as long as the numerator (the information gain) is larger than the average gain

across all tests. The numerator in this ratio is the standard information entropy difference achieved at node t, expressed as in equation (1.6) and the element in equation (1.7) and (1.8)

$$gain(t) = info(T) - info_t(T) \quad (1.6)$$

, where

$$info(T) = -\sum_{i=1}^{k} {C_i}/{C_T} \quad (1.7)$$

$$info_t(T) = \sum_{I=1}^{s} {T_i}/{T} \times info(T_i) \quad (1.8)$$

# CHAPTER 2. PRECIPITATION ESTIMATION WITH DATA DRIVEN MODELING

## 2.1 Introduction

The connection between radar data and tipping bucket precipitation has been a topic of interest in the hydrological and meteorological community for a decade and is motivated by the necessity for higher resolution precipitation for hydrological model input. In this Thesis, a series of algorithms are trained with next generation radar (NEXRAD) and rain gauge data for precipitation estimation at West Des Moines, IA. The resulting DTs have overall accuracy 95.9 %. The vision of the author is to develop this model, which links rain gauge and radar data, to find the radius of accuracy of the model at various locations of rain gauges that benefit from the accuracy of physical tipping bucket rain gauges, and the spatiotemporal resolution of NEXRAD system technology. The system of rainfall predictors at various tipping buckets has been developed to serve as input to the Wastewater Reclamation Authority (WRA).

The high spatiotemporal resolution of next generation radar (NEXRAD) makes it a useful instrument for precipitation estimation. NEXRAD-II data are the three meteorological base data quantities: reflectivity, mean radial velocity and spectrum width. NEXRAD-III data are derived from various algorithms for processing NEXRAD-II data to produce numerous meteorological analysis products, such as storm velocity, one hour precipitation total, storm total precipitation, digital mesocyclone detection, digital precipitation array, wind profiles, and vertical integrated liquid content [11].

Radar data has sources of error which could be mitigated by the aid of a secondary system, such as a rain gauge. Blockage by mountains and hilly terrain, confusion with flocks of birds and swarms of insects, and signal attenuation are all problematic to radar observations. Rain gauges measure rather than estimate precipitation and are thus deemed as the most truthful account of rainfall available. However, rain gauges provide mere point measurements, and their values may be different from those at

another gauge only a few kilometers away. It is common, especially during the convective season when the atmosphere is often unstable, for very high precipitation rates to be measured at one location, and none at another. Should the two technologies be melded together, that is NEXRAD and tipping bucket rain gauge, the strengths of both systems could be utilized.

The aim of this chapter is to use NEXRAD-II reflectivity, velocity, spectrum width data from a weather station in Des Moines, IA and network of seven rain gauges at the Wastewater Reclamation Authority (WRA) plant located in Des Moines, Iowa by a flow monitoring program to train 5 algorithms namely decision Trees (DT), K Nearest Neighbors (K-NN), Naïve Bayes (NB), Multilayer Perceptrons (MLP) and Linear Regression (LR) for precipitation estimation at a rain gauge in Des Moines, IA. The resulting model is verification that rainfall in those locations follows reflectivity so these inputs can be used then for the flow rate prediction model in the next chapter. The National Oceanic and Atmospheric Association's (NOAA) uses an algorithm for converting reflectivity data to hourly precipitation, a NEXRAD-III product. This model could then be used to provide the WRA plant with rainfall data of a 5 minutely observation frequency and a spatial resolution of 1 $km^2$. Currently, the WRA uses seven tipping buckets within the ~250 $km^2$ basin that report rain rates at 15 minute intervals.

There are different approaches to forecast prediction, for example, the algorithms for rainfall estimation were classified into physics-based and statistical/engineering approaches by Chandrasekar [12-13]; radar-derived rainfall products like measurements in reflectivity factor in real-time are used to predict the precipitation; it is transformed into rainfall accumulations and incorporates rain gauge data to improve the radar estimates.

In the early 1980s at the Hydrologic Research Laboratory by executing series of procedures precipitation algorithm was built to estimate rainfall and over time it was developed, and tested [14]. The lowest four elevation angles, 0.58, 1.58, 2.48, and 3.48

for reflectivity are used by the algorithm. Radar estimation has some sources of error which are often hard to quantify [14], while some studies focused on reduction of these errors.

Satellite precipitation algorithm is another approach to generate high spatial and temporal resolutions rainfall estimates by combining the data from Tropical Rainfall Measuring Mission (TRMM) Precipitation Radar (PR) and multispectral Geostationary Operational Environmental Stellite (GOES) imagery. They could predict 30 minutes rainfall estimate by matching PR measurements with four-band GOES image data to to make a data set and train it in neural network [15]. Other data mining techniques like Hybrid or joint PCA are applied in precipitation estimate approaches [3].

In most of these approaches, reflectivity is used to make an NN model for the prediction, while the WSR-88D records digital database contains three native variables: velocity, reflectivity, and spectrum width. These additional radar variables at multiple elevation angles and multiple bins in the horizontal can be used for precipitation prediction. Linear regression models besides feed forward NNs are used for precipitation prediction. New models which contained other radar products were not significantly more accurate than reflectivity alone.

The most commonly used technique of radar-based rainfall estimation is a function between reflectivity (Z) and rain intensity (R) which shows the capability of weather radars to measure rainfall rate using that relationship between radar echo power and rain intensity. A volume is sampled and drop size distribution is identified then Z and R are different moments of that distribution [16]. The most common form of Z and R can be related as follows in equation (2.1) where a and b are empirically estimated [17].

Following the data mining approach in weather data forecasting, different algorithms are built and compared the accuracy and chose the outperforming one to be able to forecast length of short term prediction for precipitation [18]. To do the classification, output has been discretized then encountered class imbalance so synthetic

over sampling techniques has been used. Synthetic Minority Over-sampling Technique (SMOTE) is an over sampling method; it interpolates between some minority-class examples to form new minority class examples that lie together. Thus it is avoided to over-fit and the decision boundaries spread farther for the minority class in to the space of majority class [19].

## 2.2 Radar precipitation estimation (Z-R conversion)

The most common conversion (Z-R) of reflectivity to precipitation rate takes the following relationship:

$$Z = a \cdot R^b \qquad (2.1)$$

Where $Z$ is the reflectivity, $R$ is the precipitation rate, and $a$ and $b$ are constants from empirical studies (calibration). Typically, the values used for $a$ and $b$ are 200 and 1.6 respectively.

## 2.3 Data Acquisition

Two types of data were collected for the building of algorithms in this study, (1) radar reflectivity data and (2) tipping bucket precipitation data. Although other work has considered using reflectivity bandwidth and horizontal wind velocity [20-21] in their models, their experimental results conclude that reflectivity is the only useful input while in this study NEXRAD-II reflectivity, velocity, spectrum width data are used.

### 2.3.1 Doppler WSR-88D Radar

The National Weather Service's (NWS) Next Generation Radar (NEXRAD) system is comprised of 137 radar sites in the contiguous United States, each of with is equipped with Doppler WSR-88D radar capable of reporting high resolution data and making a full 360 degree scan every 5 minutes, with has a range of ~230km and a spatial resolution of about 1km by 1km (Baer, 1991). The weather station used in this study is located in Des Moines, IA (KDMX), which is approximately 32 km far from the tipping bucket locations. Reflectivity and base velocity and base spectrum width were collected

from four elevation angles (tilts) of the antenna the lowest "tilt" angle 0.5° then 1.45° and 2.40° finally 3.35°. As both the intensity and angle of the reflectivity values are required to describe the shape of the approaching storm, it is necessary to provide data from multiple angles [22]. This is also consistent with the literature [23-24].

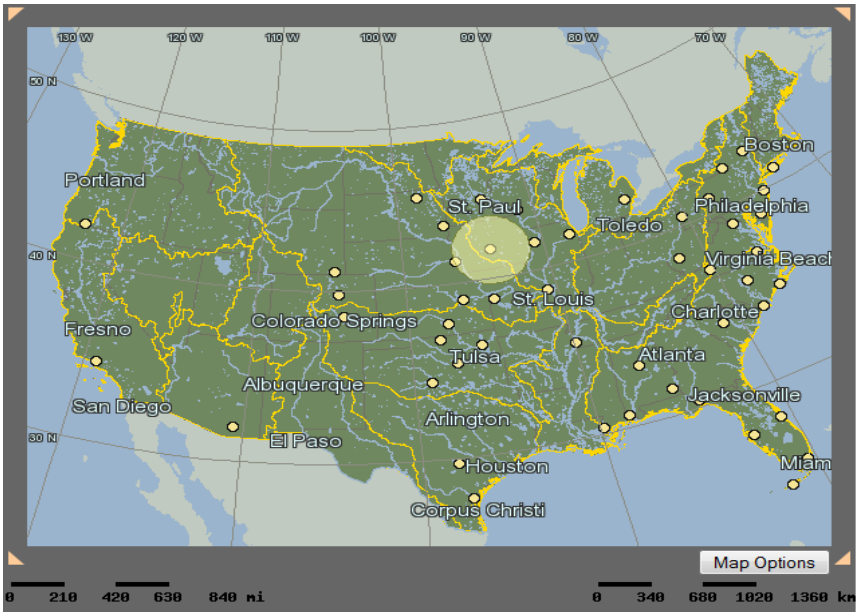



**Figure 2.1 Hydro NEXRAD image of KDMX radar coverage**

Precipitation is detected by the radar earlier than the rain gauge. So, the rain gauge data has been shifted on the time axis to synchronize the attributes in the model.

The rain gauge sites make part of the WRA. Each instrument is equipped with dual buckets for quality checking purposes and redundancy. It records precipitation rate in 0.0001 inches/day, every 15 minutes.

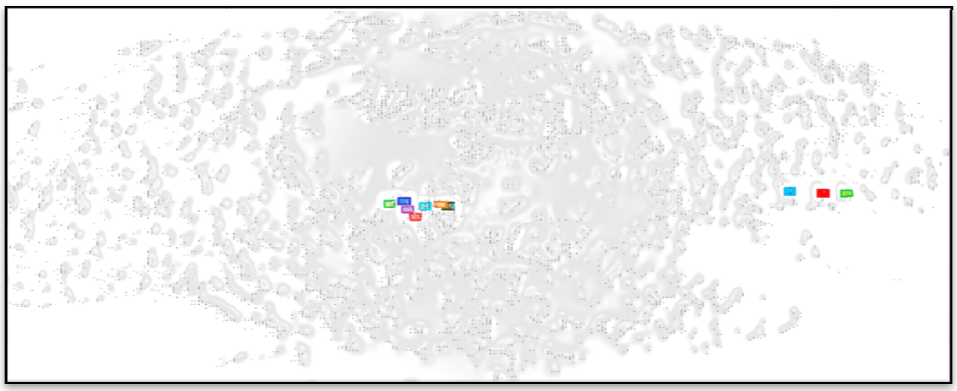


**Figure 2.2 NEXRAD reflectivity raster with WRA (Des Moines) and Iowa City and Amana superimposed**

Figure (2.3) shows the location of the seven tipping buckets out of the 86 stations and the location of the plant itself in the WRA basin and the radar grid superimposed, each cell is equivalent to 1 KM. Figure (2.3) is derived from the reflectivity map.

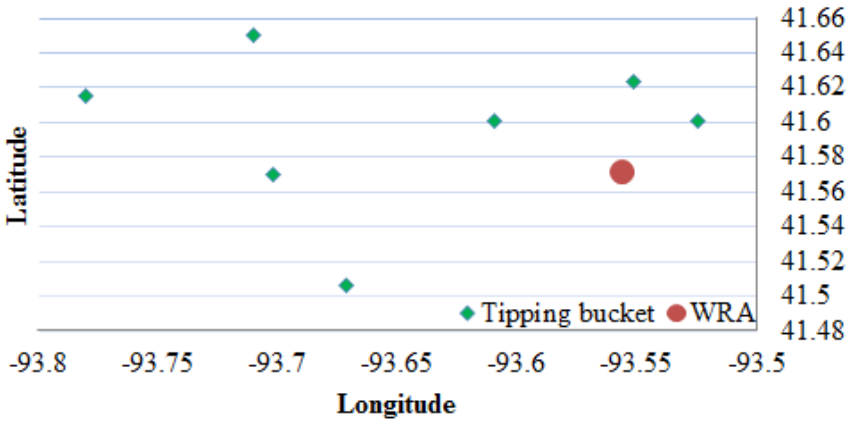



**Figure 2.3 Location of the rain gauges surrounding the plant**

In Table (2.1) the latitude and longitude of each tipping bucket is shown. It gives a better understanding of how distant they are from one another then how correlated they can be.

**Table 2.1 Lat-long of rain gauges**

| Site Name | Lat | Long |
|---|---|---|
| TB1 | 41.62 | -93.55 |
| TB2 | 41.60 | -93.60 |
| TB3 | 41.56 | -93.70 |
| TB4 | 41.61 | -93.78 |
| TB5 | 41.50 | -93.67 |
| TB6 | 41.60 | -93.52 |
| TB7 | 41.6 | -93.70 |
| Plant | 41.57 | -93.55 |

In the Figure (2.4), it can be seen that the amount of rainfall varies greatly in each tipping bucket. That is why one model is made based on the TB2 station data and is examined to see how accurate it will perform on other locations while getting far from the TB2 station i.e. testing it on the other tipping buckets to find the radius of accepted accuracy for the prediction. In Figure (2.4), the X axis stands for time in which each unit is 5 minutes, and the Y axis shows the inches of precipitation.

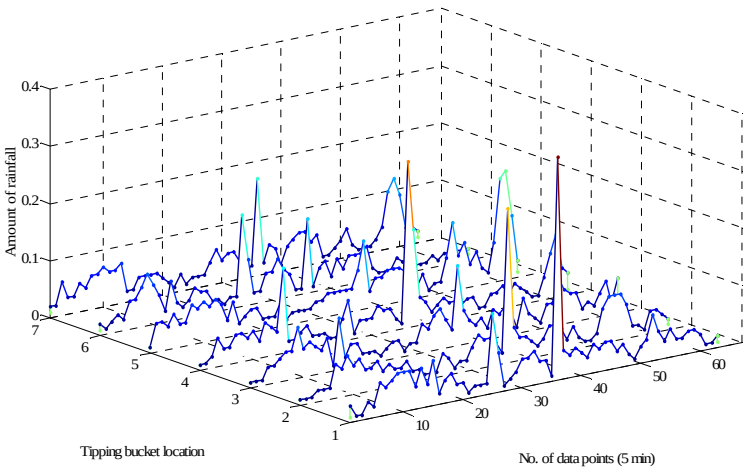



**Figure 2.4 Rain gauge precipitation data in all 7 stations**

## 2.4 Preprocessing

Preprocessing data is a crucial step of the data mining process. Outliers, missing data and unreliable or low quality data all need to be considered before analysis. The NEXRAD data was ordered from the hydro NEXRAD site and downloaded via an FTP connection. A script was written in Matlab to select the closest grid points that corresponded with the WRA tipping bucket location. Nine grid points were selected about the tipping bucket location, in agreement with Liu, Chandrasekar, and Xu (2001) [25]. This is to provide some margin for error in the GPS mapping of the tipping buckets and gridding of the KDMX radar raster map. Also, rain does not fall straight down but may be advected horizontally. Finally, The NEXRAD data was collected at 5-min intervals, which is inconsistent with the temporal resolution of the tipping bucket, reported every 15-min. In order to make the output data suitable for time-series classification, the input data recorded by the different systems, e.g., radar and influent data are time stamped at 5 min intervals. In general, the following processing schemes are used.

- Rain gauge data (15 min) is converted into high frequency data (5 min) by taking the average of the corresponding neighborhood time stamp data, i.e.,

$$X_{new} = {(X_{old-1} + X_{old+1})}/{2} \quad (2.2)$$

In equation (2.2), $X_{new}$ is the new data point to be inserted, whereas, $X_{old-1}$ and $X_{old+1}$ are high frequency neighborhood data points, so this issue was simply dealt with linear interpolating missing tipping bucket data observations.

- The continuous tipping bucket data is discretized into different output class of varying bin size. A description of influent range for different bin size is shown in Table (2.2).

The time series considered was from 7/1/09, 6:30 AM to 10/24/09, 11.55 PM and was formatted to 5-min resolution, for a total of 32,734 data points. The rest of the description is shown in Table (2.4).

**Table 2.2 Discretization of the output data**

| No. of bins | Total influent thresholds |
|---|---|
| 2 | 1.0<br>2.(0 - inf] |
| 3 | 1.(-inf-0.0025]<br>2.( 0.0025-0.0175]<br>3. (0.0175- inf] |
| 4 | 1. (-inf-0.0025]<br>2. (0.0025-0.0125]<br>3. (0.0125-0.0325]<br>4. (0.0325-inf) |

Table (2.3) shows the parameters of the model. The first column is the full name and the second column is the label we used in model construction, then a short description and finally the unit. Overall there are 50 attributes which contains memory parameters for 5, 30, 60 and 90 minutes before real time for radar data and target tipping bucket.

### 2.4.1 Class imbalance

There may be two kinds of imbalances in a data set. One is between-class imbalance; the other is within-class imbalance. By convention, in imbalanced data sets, we call the classes having more examples the majority classes and the ones having fewer examples the minority classes. Simply said, a dataset is imbalanced if the classification categories are not approximately equally represented. There have been attempts to deal with imbalanced datasets in domains such as fraudulent telephone calls, telecommunications management, text classification and detection of oil spills in satellite images. There are different approaches to eliminate this issue. Studies show

**Table 2.3 Parameters of the model**

| Parameter | Label | Description | Unit |
|---|---|---|---|
| **Base reflectivity** | Ref1 | Display of echo intensity transmitted power returned to the radar receiver the lowest angle (0.5°) | dBZ |
| **Base reflectivity** | Ref2 | Angle of 1.45° | dBZ |
| **Base reflectivity** | Ref3 | Angle of 2.40° | dBZ |
| **Base reflectivity** | Ref4 | Angle of 3.35° | dBZ |
| **Base velocity** | Vel1 | The velocity of the precipitation either toward or away from the radar for radar "tilt" angle 0.5° | |
| **Base velocity** | Vel2 | For radar "tilt" angles, 1.45° | |
| **Base velocity** | Vel3 | For radar "tilt" angles, 2.40° | |
| **Base velocity** | Vel4 | For radar "tilt" angles 3.35° | |
| **Base spectrum width** | SW | Spectrum width a measure of velocity dispersion. It is recorded at same tilt angle as reflectivity | |
| **Tipping bucket** | TB | The average of 15 minutes obtained from tipping bucket (TB2) | Inches |

that when over-sampling the minority (abnormal) class and under-sampling the majority (normal) class are combined, better classifier performance is achieved than only under-sampling the majority class; like the study of Ling and Li and Gustavo [34, 26]. The machine learning has dealt with class imbalance in two ways. One way is assigning distinct costs to training examples the other is to re-sample the original dataset, either by oversampling the minority class and/or under-sampling the majority class [27]. Japkowicz discussed the three strategies as under-sampling, resampling and a

recognition-based induction scheme to evaluate the imbalance effect in data set [28]. The resampling methods were consisted of random resampling of the the smaller class till it is as many samples as the majority class and focused resampling which resampled only those minority examples that occurred on the boundary between the minority and majority classes.

**Table 2.4 Dataset sampling description**

| | |
|---|---|
| Time span | 7/1/09, 6:30 AM to 10/24/09, 11.55 PM |
| Frequency | 5 minutes |
| Instances | 32734 |
| Discretized class instances | Tipping bucket data<br>(2 – 3 and 4 bins are modeled) |
| Training dataset | 7/1/09 to 9/15/09 |
| Testing dataset | 9/15/09 to 10/24/09 |

Some studies discussed over-sampling with replacement and have noted that it does not significantly improve minority class recognition [26-28]. A heuristic under-sampling method balanced the data set through eliminating the noise and redundant examples of the majority class [30]. SMOTE (Synthetic Minority Over-sampling Technique) method in Nitesh's study generated new synthetic examples along the line distinguishing minority and majority by producing new minority in their nearest neighbors it makes the decision regions larger [31]. He also improved minority classification by integrating SMOTE into a standard boosting procedure class while the whole accuracy of test set was not sacrificed [32]. Estabrooks proposed a multiple resampling method which selected the most appropriate re-sampling rate adaptively [33]. Technique SMOTE proposes an

over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement and generates synthetic examples in a less application-specific manner, by operating in "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Our implementation currently uses five nearest neighbors. For instance, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each. The effect is that decision trees generalize better [29].

The ROC curve is a helpful metric to evaluate the learners for imbalanced data sets. FP rate denotes the percentage of misclassified negative examples, and TP rate is the percentage of correctly classified positive examples. The ROC curve depicts relative trade-offs between benefits (TP rate) and costs (FP rate). The point (0, 1) is the ideal point of the learners. AUC (Area under ROC) can also be applied to evaluate the imbalanced datasets.

SMOTE is applied in this research a filtering supervised instanced based preprocessing step to resample the minority classes. Results are shown in Table (2.5) for 3 and 4 bins. In each resampling step, the percentage varies. The first bin which contains zero value is the majority so there will be zero percent of resampling for that and for the rest of the bins percentage of resampling will be decided to get closer to the frequency of first bin (majority). Accuracy before and after application of SMOTE is shown below, overall accuracy may decrease while G-mean certainly increases, G-mean is an evaluation index which shows that accuracy of all classes have been satisfactory.

**Table 2.5 Comparing results after application of SMOTE**

| Number of bins | Thresholds | Accuracy (%) Before | Accuracy (%) After |
|---|---|---|---|
| 3 bins | (-inf-0.0025] | 99.50 | 98.1 |
| | ( 0.0025- 0.0175] | 77.10 | 93.5 |
| | (0.0175- inf] | 88.00 | 97.1 |
| | Total | 98.70 | 96.2 |
| 4 bins | (-inf-0.0025] | 99.50 | 97.40 |
| | (0.0025- 0.0125] | 76.20 | 94.10 |
| | (0.0125- 0.0325] | 73.90 | 96.10 |
| | (0.0325- inf) | 86.20 | 95.50 |
| | Total | 98.50 | 95.90 |

The process of resampling the minority classes is shown in Figure (2.5). Resampling is applied for each minority class in a loop until we get to the closest number to the account of the majority class which is 31229 for both 3 and 4 bins.

## 2.1 Parameter selection

While correlation measures the strength of the linear relationship, nonlinear relationships may exist in the data set. Heuristic feature selection algorithms are often used in the field of computational intelligence to find optimal subsets for modeling nonlinear phenomenon. The feature selection algorithms selected is boosted tree algorithm, as in the previous chapters. These algorithms are “wrapped” within the DT algorithm to find the parameters in the data that result in the best model. In other words this algorithm employs a heuristic approach to training and testing data subsets in search of a local optimum.

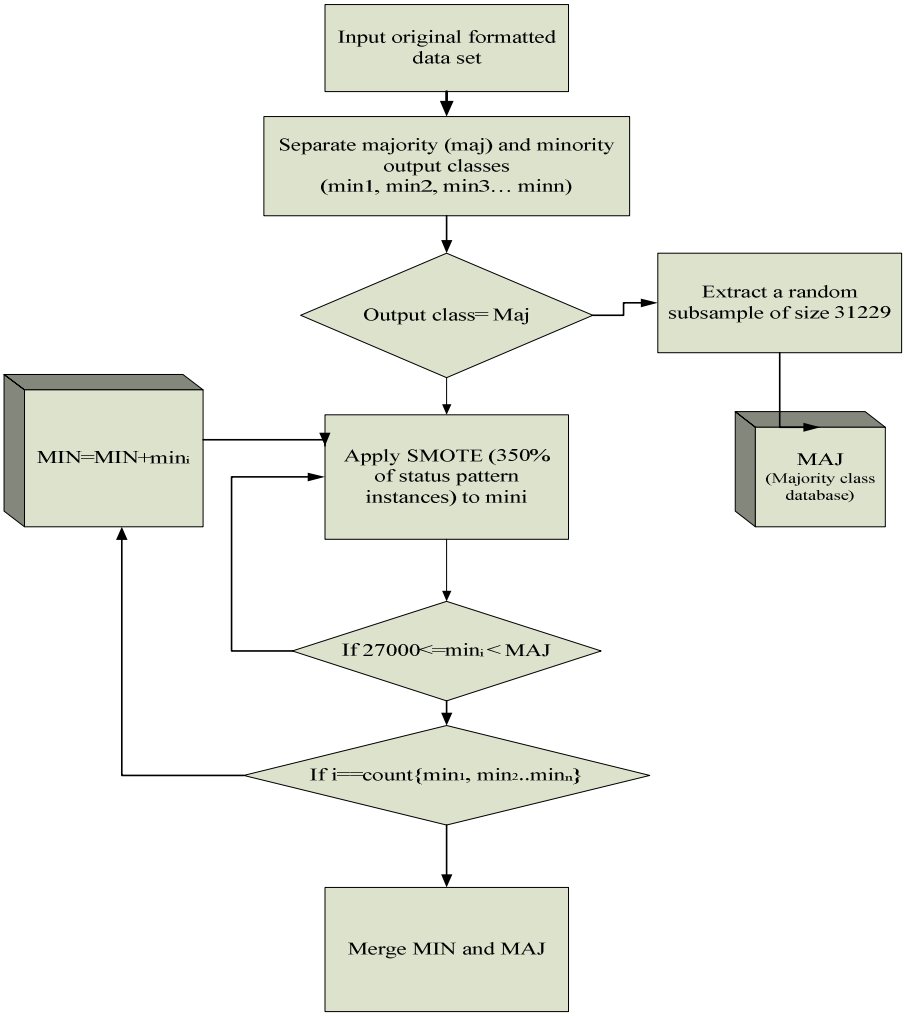



**Figure 2.5 SMOTE process in re-sampling the minority classes**

Tables (2.6) and (2.7) show the results of the feature selection with boosted tree which will be discussed later in this Thesis. To get a better feature selection, we remove the memory parameters of the tipping bucket data to see the ranking of other attributes without the effect of the memory parameter of the tipping bucket since it is so dominant. Elected ones have higher importance than 0.5.

In fact there is no necessity to select some features while with these 50 features the selected algorithm does the classification in a few minutes. But to get a better understanding about the model and the attributes we do the feature selection for 3 bins

and 4 bins as are shown in Tables (2.6) and (2.7) respectively. Finally the attributes with importance rank above 50 % in addition to the memory parameters of the tipping buckets are considered to be more important.

**Table 2.6 Feature selection for 3 bins**

| WMPTB | | | | WAMPTB | | | |
|---|---|---|---|---|---|---|---|
| tb(t-5) | ref1(t-30) | vel1(t-60) | sw(t-90) | ref1(t-5) | vel2(t) | sw(t) | vel1(t-60) |
| tb(t-30) | ref1(t-5) | vel2(t-30) | ref2(t-60) | ref1(t-30) | vel2(t-5) | ref2(t-5) | ref2(t-90) |
| tb(t-60) | ref1(t) | vel1(t-90) | ref2(t) | ref1(t) | vel1(t) | ref2(t) | vel2(t-60) |
| tb(t-90) | vel2(t-90) | vel2(t-60) | ref2(t-90) | ref1(t-60) | vel1(t-5) | sw(t-60) | ref3(t-90) |
| ref1(t-60) | vel2(t) | vel1(t) | vel1(t-30) | ref1(t-90) | sw(t-90) | vel1(t-30) | sw(t-30) |
| ref1(t-90) | vel2(t-5) | ref2(t-5) | | vel1(t-90) | vel2(t-30) | vel2(t-90) | |

### 2.1.1 Boosted tree

It is crucial to have a feature selection mechanism that can find a subset of features that both meets latency requirements and achieves high relevance. Boosted trees (and boosting algorithms in general) have been used widely as a learning algorithm for ranking search results; there are many advantages of using boosted trees as a learning algorithm for ranking. For example, no normalization is needed when using. Different types of data (e.g., categorical and count data); trading off runtime efficiency and accuracy (i.e., relevance for search) can be easily achieved by truncating the number of trees used in the boosted trees model.

**Table 2.7 Feature selection for 4 bins.**

| WMPTB | | WAMPTB | |
|---|---|---|---|
| ref1(t) | ref3(t-90) | tb(t-5) | vel1(t) |
| vel1(t-5) | ref2(t-90) | tb(t-30) | vel1(t-5) |
| ref1(t-5) | sw(t-5) | tb(t-60) | vel1(t-30) |
| ref1(t-30) | vel3(t-60) | tb(t-90) | ref2(t) |
| vel2(t-5) | vel2(t-60) | ref1(t-90) | ref2(t-60) |
| ref1(t-90) | ref3(t-60) | vel2(t-90) | ref2(t-5) |
| vel1(t) | ref2(t-5) | vel2(t-30) | sw(t-90) |
| vel2(t-90) | ref2(t-60) | ref1(t-60) | sw(t-60) |
| vel2(t) | sw(t-30) | ref2(t-90) | ref2(t-30) |
| ref1(t-60) | ref3(t-5) | ref1(t-5) | sw(t-5) |
| vel2(t-30) | ref3(t) | vel2(t-60) | sw(t) |
| ref2(t) | vel4(t) | vel2(t-5) | sw(t-30) |

WMPTB: without memory parameter of tipping bucket, IR: importance rank, WAMPTB: With all memory parameters of tipping bucket

From the perspective of feature selection, a more interesting property of the boosted trees is that (a greedy) feature selection already happens in the algorithm when selecting splitting features (e.g., for regression trees, splitting features and splitting points are found to minimize the squared-error loss for any given partition of the data). Moreover, as a byproduct, a sorted list of relative importance of features (i.e., a feature importance list) is automatically generated for each boosted trees model. The relative

influence of a feature $x_j$ for a single decision tree to boosted trees as an average over all the trees can be generalized as

$$\hat{J}_j^{\,2} = 1/M \sum_{m=1}^{M} \hat{J}_j^{\,2}(Tm) \quad (2.3)$$

where *M* is the number of trees. For each tree the relative importance is calculated as

$$\hat{J}_j^{\,2}(T) = \sum_{t=1}^{L-1} \hat{I}_t^{\,2}\, 1(v_t = j) \quad (2.4)$$

where the summation is over the internal nodes t of a L-terminal node tree T, $v_t$ is the splitting feature associated with node t, and $\hat{I}_t^{\,2}$ is the corresponding empirical improvement in squared-error as a result of the split.

### 2.2 Model training/testing

In training and testing of a data-driven model, there is always a balance between accuracy and overfitting, or lack of generalizability, of the model. Especially for the purpose of this research, which is to establish a model that can be used at other tipping bucket locations, generalizability is of great importance. Following Tan *et al.* (2006), 2/3 of the dataset was used for training, and 1/3 for testing, which is a common split to balance generalizability with accuracy [36]. The networks were tested for predicting the rainfall rate (mm/hr) at the Des Moines tipping buckets. Using Weka's "(J48)" option Decision Trees were generated.

### 2.3 Metrics for Algorithm Evaluation

Three evaluation indexes namely accuracy, sensitivity, and specificity are used to assess the agreement between ground rain gauges and total influent model values. The evaluation is based upon well-known confusion matrix. Table (2.8) displays the confusion matrix of 3 output class.

**Table 2.8 Description of confusion matrix**

| | | Actual | | |
|---|---|---|---|---|
| | | C1 | C2 | C3 |
| Predicted | C1 | TP1 | FP21 | FP31 |
| | C2 | FP12 | TP2 | FP32 |
| | C3 | FP13 | FP23 | TP3 |

$$Accuracy = {\sum_{i=1}^{n} TP_i}\Big/{\sum_{i\&j=1}^{n}(TP_i + FP_{ij})} \quad \text{n is the number of bins}$$

(2.5)

$$Sensitivity\ (recall) = {TP_i}\Big/{\left(TP_i + \sum_{i=1}^{n} FP_{ij}\right)}\ (i \neq j)$$

(2.6)

$$Precision\ (positive\ predicted\ value) = {TP_i}\Big/{\left(TP_i + \sum_{j=1}^{n} FP_{ji}\right)}\ (j = 1,2,3\ and\ j \neq i)$$

(2.7)

$$G - mean = \sqrt[n]{\prod_{i=1}^{n} accuracy_i}$$

(2.8)

In equations (2.5) to (2.8) $TP_i$ is the number of correctly classified instances in class i (i=1, 2, 3), whereas, $FP_{ij}$ is the number of incorrectly classified instances from class i in class j.

### *2.4* Results

The results for this research are categorized to two main parts. First is the prediction ahead for the rain gauge named TB2; second, is finding the radius under which the accuracy of the model stays above 85 %, by testing the algorithm in the other rain gauges' locations.

### 2.4.1 Prediction

The forecasting model by Decision Tree (J48) is shown in the Table (2.9), respectively for 5, 30, 60 and 120 minutes ahead. The result for time t is the total accuracy of the model after feature selection and application of SMOTE. As shown in Table (2.9), the accuracy up to 2 hours is very high.

**Table 2.9 Prediction results by total accuracy**

| No. of Bins | t | t+5 | t+30 | t+60 | t+120 |
|---|---|---|---|---|---|
| 3 | 96.1 | 95.44 | 94.89 | 94.60 | 94.21 |
| 4 | 95.9 | 95.2 | 94.58 | 94.58 | 94.21 |

### 2.4.2 Radius of accuracy

Among all these 7 tipping buckets in the plant, TB7 has the highest correlation with the TB2 station, shown in Table (2.10), because it is the closest one to it. Even though the rain gauges are not that far from each other, the amount of rainfall recorded by them is so variant. Moreover level of correlation cannot be decided only by closeness, height, being located in upstream or downstream level are all effective on tipping bucket records.

Here the results of the comparison among the other six rain gauges have been discussed. We have 7 tipping buckets in the wastewater plant, so we made the model based on the radar data and rain gauge data of the TB2 station to find out how distant this model works accurately. We tested the other six ones which are distant from the TB2 station as shown in the Table (2.11) and they were depicted in the Figure (2.6).

**Table 2.10 Correlation coefficients among rain gauges.**

| | TB2 | TB1 | TB6 | TB5 | TB3 | TB7 | TB4 |
|---|---|---|---|---|---|---|---|
| TB2 | 1.00 | 0.49 | 0.33 | 0.47 | 0.36 | 0.72 | 0.52 |
| TB1 | 0.49 | 1.00 | 0.63 | 0.56 | 0.40 | 0.30 | 0.29 |
| TB6 | 0.33 | 0.63 | 1.00 | 0.53 | 0.37 | 0.22 | 0.23 |
| TB5 | 0.47 | 0.56 | 0.53 | 1.00 | 0.53 | 0.32 | 0.34 |
| TB3 | 0.36 | 0.40 | 0.37 | 0.53 | 1.00 | 0.27 | 0.35 |
| TB7 | 0.72 | 0.30 | 0.22 | 0.32 | 0.27 | 1.00 | 0.65 |
| TB4 | 0.52 | 0.29 | 0.23 | 0.34 | 0.35 | 0.65 | 1.00 |

In Table (2.11), the results of the total accuracy of the Decision Tree, algorithm (J48) on the other rain gauges around TB2 is shown. The best accuracy belongs to TB7 and the worst to TB3 as we define the accepted threshold for the accuracy as 95% so the maximum distance that this model can be used to result the desired accuracy is the maximum available distance, 33.13 km.

**Table 2.11 Total accuracy of the model tested on the other gauges**

| **No.** | Tipping bucket | Distance from TB2(KM) | Total accuracy (%) |
|---|---|---|---|
| **1** | TB6 | 8.48 | 96.80 |
| **2** | TB1 | 11.04 | 96.12 |
| **3** | TB3 | 18.97 | 95.19 |
| **4** | TB7 | 21.47 | 96.85 |
| **5** | TB5 | 21.63 | 96.55 |
| **6** | TB4 | 33.13 | 96.01 |

## 2.5 Conclusion

This chapter describes the development of a decision tree trained with NEXRAD-II reflectivity, velocity and spectrum width data from a weather station in Des Moines, IA and tipping bucket rain gauge data from WRA in West Des Moines, IA. The model was synced with real time radar and tipping bucket data to provide rainfall estimation. The motivation for a system of rainfall estimation is to provide higher resolution precipitation input for hydrological models. The model compared with previous regression algorithms for converting reflectivity data to precipitation, outperformed in longer prediction.

This paper had an overview on the resampling the majority class "SMOTE" and description of mathematical section for decision trees since it was the most accurate in the classification models for precipitation. Boosted tree was selected as tool for feature selection. Prediction of 120 minutes ahead had the accuracy of 94 %, and the model was accurate for the maximum available radius of 33.13 km, hence the accuracy of the model did not deteriorate by getting far from the tipping bucket which the model was built on, it just varied.

# CHAPTER 3. TOTAL INFLUENT MODELING

## 3.1 Introduction

To maintain stable effluent characteristics in a wastewater treatment plant (WWTP), it is desirable to know in advance the influent flow rate to the wastewater treatment plant. Wastewater characteristics such as biochemical oxygen demand (BOD), total suspended solids (TSS), and pH [56-57] are strongly correlated to the influent flow rate. Prediction of the influent flow rate is helpful in optimally scheduling wastewater pumps.

In practice, the influent flow rates are usually estimated by the operators based on experience and local weather forecasts [58]. However, such estimations are not accurate enough to manage WWTPs, especially for plants that treat both municipal wastewater and storm rainfalls [59]. The precipitation may cause large variability of the influent flow rate, and thus reducing efficiency of WWTPs. Heavy rainfalls overwhelm the wastewater treatment system, causing spills and overflows.

Several studies have been performed to model and predict the influent flow rate to wastewater treatment plants [60-61]. Tan et al. [62] used a direct k-step predictor to forecast the wastewater flow rate and obtained reliable predictions up to 2 h ahead for wet weather sewer flow. Using recursive ARX (autoregressive with exogenous input) filters, a model based on the flow pattern estimation could handle rainy conditions for prediction horizons of a few hours [63].

Data-mining is a promising approach to build prediction models. It is the process of finding patterns from data by algorithms versed on the crossroads of statistics and computational intelligence [64].

Considering the association between the amount of rainfall and influent quality, radar data measuring the rainfall are analyzed in the present study. Using historical radar and rainfall data an accurate prediction model can be built. Data mining techniques such

as neural networks (NN), support vector machine (SVM) etc. are capable of learning complex relationship between input parameters and therefore widely used in literature pertaining to qualitative prediction estimation (QPE) and quantitative precipitation forecast (QPF) of rainfall [37].

Weather radars generally apply a Z-R relationship as mentioned before to relate the measured variable. Reflectivity is mostly chosen among radar products since reflectivity at the lowermost elevation is related to the rain rate [3].

Weather radar has been studied to estimate rainfall quantity in spite of being often considered qualitative. The advantage of weather radars is about high temporal resolution, a full volume scan within to reveal a three-dimensional structure of precipitation. Radars work by sending out an electromagnetic beam and measuring how much of the energy of that beam is reflected back. For precipitation forecasting, researchers use radar-derived rainfall products in real-time. Another approach is a satellite derived precipitation algorithm [15].

The flow monitoring program which includes installation of over 80 flow meters throughout the metro will allow monitoring raw flow in every section of the metro as well as a control for routing raw flow around the metro to storage basins. Knowing the amount of total influent a few hours ahead makes it possible to decrease the impact of diurnal flow.

Data mining has been promising in climatic measurement and the models trained and built by data mining algorithm can be easily updated [38], it is used to build a model for total influent prediction over a short time horizon 30 – 60–120 and 180 minutes ahead. The models are built using the historical data collected by flow monitoring program installed in the WRA plant for 6 locations surrounding the plant and radar data extracted from NCDC website and total influent to the plant.

## 3.2 Problem background

### 3.2.1 Wastewater reclamation authority (WRA)

The treatment plant is located in Des Moines, Iowa and has been operated by the City of Des Moines since 1987. In 2005, WRA was designed to serve a population of 317,930 with the average raw wastewater flow load of 50 million gallons per day. The ultimate goal of WRA for 2020 is to serve a population of 389,200 with the processing capacity of 74 million gallons per day.

WRA is 77 acre over a mile long. It includes preliminary treatment, 6 primary clarifiers, 12 roughing filters, 6 aeration tanks, 12 final clarifiers, disinfection, 2 chlorine contact tanks, 3 rotating drum filters (RDT), 6 anaerobic digesters, solids handling and treatment, bio-solids disposal and 8 bio filters for odor control.

### 3.2.2 Data description

In the research reported in this paper, three years and three months long rain gauge and total influent data from the Wastewater Reclamation Authority (WRA) spanning from 1/1/2005 until 3/31/2008 was used.

The weather radar in the Des Moines area is located about 32 km from the WRA plant. The KDMX station at Des Moines is located 41.7311N, 93.7228W Elevation while plant is 41.5712 N, 93.5862 W as shown in Figure (3.1).

The location of 7 rain gauges out of the 86 stations and the location of the plant are the same as described before in chapter 2, Figure (2.3). In this research only six of these rain gauges are used in making the model since one of them named TB7 was recently installed and could not provide the information which was required.

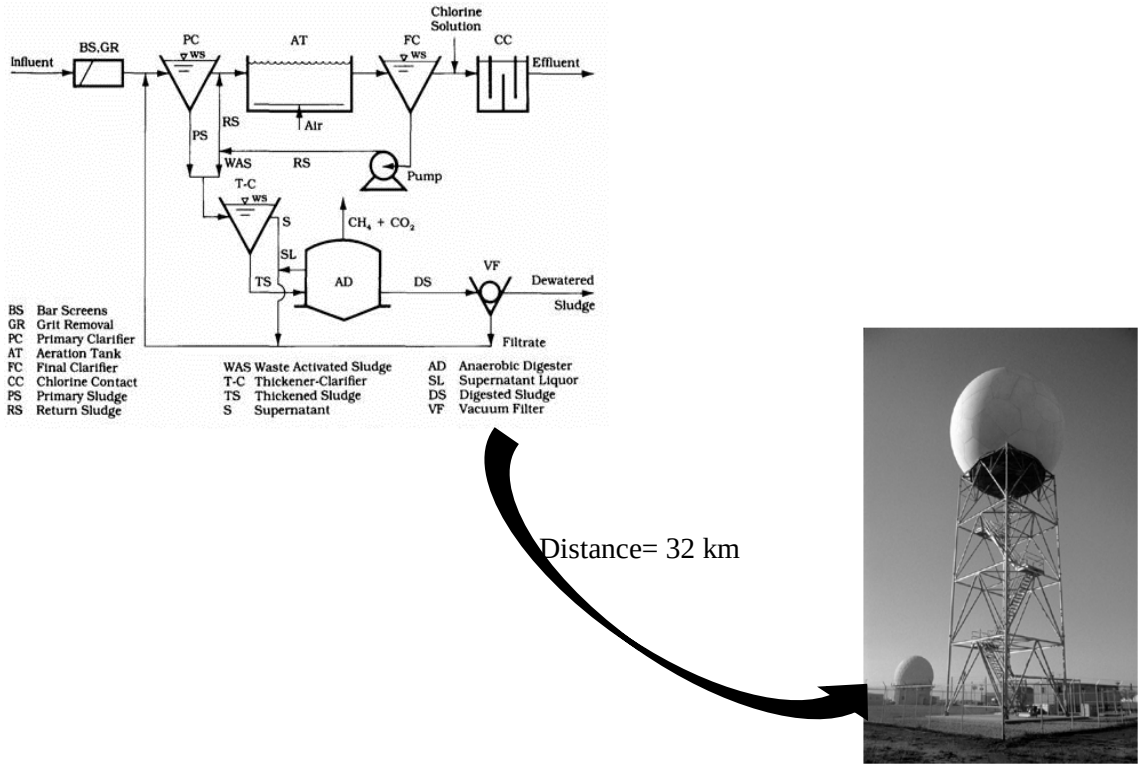



**Figure 3.1 The location of KDMX radar and the WRA plant and the distance between**

Table (3.1) describes the input parameters used in the current research. Each rain gauge has four parameters including reflectivity at 4 different heights (1km, 2 km, 3 km and 4 km height). In addition, to improve the prediction accuracy of the data mining algorithm, memory parameters have been used [39]. The memory parameters for all inputs and influent output recorded 30 minutes ago, 60 minutes ago, 90 minutes ago and 120 minutes before time e.g., t-30, t-60, t-90 and t-120 respectively are included in model construction. Overall, the data consist of 55 input parameters.

**Table 3.1 Parameters of the model**

| Parameter | Label | Description | Unit |
|---|---|---|---|
| Reflectivity at CAPPI height 1 km | c1 | Radar display which gives a horizontal cross-section of data at constant altitude, here it is 1 km | dBZ |
| Reflectivity at CAPPI height 2 km | c2 | Altitude of 2 km | dBZ |
| Reflectivity at CAPPI height 3 km | c3 | Altitude of 3 km | dBZ |
| Reflectivity at CAPPI height 1 4 km | c4 | Altitude of 4 km | dBZ |
| Tipping bucket data, named TB2 | TB2 | The average of 15 minutes obtained from tipping bucket (TB2) | Inches |
| Tipping bucket data, named TB3 | TB3 | The average of 15 minutes obtained from tipping bucket (TB3) | Inches |
| Tipping bucket data, named TB1 | TB1 | The average of 15 minutes obtained from tipping bucket (TB1) | Inches |
| Tipping bucket data, named TB5 | TB5 | The average of 15 minutes obtained from tipping bucket (TB5) | Inches |
| Tipping bucket data, named TB7 | TB7 | The average of 15 minutes obtained from tipping bucket (TB7) | Inches |
| Tipping bucket data, named TB6 | TB6 | The average of 15 minutes obtained from tipping bucket (TB6) | Inches |
| Tipping bucket data, named TB4 | TB4 | The average of 15 minutes obtained from tipping bucket (TB4) | Inches |
| Total influent | Inf | Raw Wastewater Flow (In building05) low:0 and high: 260 | MGD |

The CAPPI is composed of data from each angle that is at the height requested for the cross-section. In the early days, the scan data collected where shown directly on the cathodic screen and a photo sensitive device captured each ring as it was completed. Then all those photographed rings were assembled. Weather radars collect in real time data on a large number of angles.

Information about various memory parameters is shown in Table (3.2). The first two year and three months of the data constitute the training set 1/1/05 to 4/30/07, whereas, remaining about 1 year is used to construct the prediction models at various

time stamps 3/31/07 to 4/30/08. For frequency of 15 minutes there are 114048 numbers of samples, also discretized classes of output for three and four bins.

**Table 3.2 Attribute labels**

| | | | | |
|---|---|---|---|---|
| TB1c1 | TB7c2 | TB6c4 | TB1(t-30) | TB3(t-120) |
| TB2c1 | TB1c3 | TB7c4 | TB3(t-60) | TB6(t-120) |
| TB3c1 | TB2c3 | TB3(t) | TB6(t-60) | TB7(t-120) |
| TB4c1 | TB3c3 | TB6(t) | TB7(t-60) | TB5(t-120) |
| TB5c1 | TB4c3 | TB7(t) | TB5(t-60) | TB1(t-120) |
| TB6c1 | TB5c3 | TB5(t) | TB2(t-60) | TB1(t-120) |
| TB7c1 | TB6c3 | TB2(t) | TB1(t-60) | Infl(t-30) |
| TB1c2 | TB7c3 | TB1(t) | TB3(t-90) | Infl(t-60) |
| TB2c2 | TB1c4 | TB3(t-30) | TB6(t-90) | Infl(t-90) |
| TB3c2 | TB2c4 | TB6(t-30) | TB7(t-90) | Infl(t-120) |
| TB4c2 | TB3c4 | TB7(t-30) | TB5(t-90) | Infl(t) |
| TB5c2 | TB4c4 | TB5(t-30) | TB2(t-90) | - |
| TB6c2 | TB5c4 | TB2(t-30) | TB1(t-90) | Time date |

## 3.3 Research methodology

In this section, description of the proposed approach is provided. In coming sub-section, different data-processing techniques are discussed.

### 3.3.1 Data preprocessing

In order to make the output data suitable for time-series classification, the input data recorded by the different systems e.g., radar and influent data are time stamped at 15 min intervals. In general, following pre-processing schemes are used.

Since the radar image covers the location of all tipping buckets, the reflectivity data at nine surrounding cells (dimension of each cell is 1 km by 1 km) around the center of the tipping bucket on the radar map as well as the reflectivity at the center are extracted and averaged for each tipping bucket. In the original dataset, some null values (denoted as -99) were present, implying that the radar signal has not been detected. These null values are treated as the missing values. When the reflectivity at the center and nine surrounding cells were all nulls, the average value of the preceding and succeeding neighbor values are used as the reflectivity for this particular tipping bucket. The radar data are also averaged over 15 min intervals.

The influent flow rate data is measured at 15 s intervals. It is converted into 15 min average data to bring it to the same frequency as the rainfall rate data. The upper and lower limit on the influent flow rate is 0 and 260 million gallons per day, respectively. The values beyond the limits are considered as outliers and are removed in preprocessing the data. Then UTS offset time is applied for radar data, this shift is done to convert the UTC to local time. The continuous influent data is discretized into different output class of varying bin size. A description of influent range for different bin size is shown in Table (3.3). Discretization is done by WEKA unsupervised attribute filtering with same frequency in each bin.

### 3.3.2 Evaluation metric

Three evaluation indexes namely accuracy, sensitivity, and specificity are used to assess the agreement between ground rain gauges and total influent model values. The

evaluation is based upon well-known confusion matrix that has been described in previous sections.

**Table 3.3 Discretization of output data**

| No of bins | Total influent thresholds |
|---|---|
| 3 | 1. (-inf-38.221052]<br>2. (38.221052-49.149128]<br>3. (49.149128-inf) |
| 4 | 1. (-inf-35.374013]<br>2. (35.374013-43.207119]<br>3. (43.207119-52.403524]<br>4. (52.403524-inf) |

### 3.4 Feature selection and model construction

To obtain a fast and accurate model, the original high dimension data should be reduced to low dimension. In the dataset, there are 63 parameters, including memory parameters recorded 30 minutes earlier (t-30), 60 minutes earlier (t-60), 90 minutes earlier (t-90) and 120 minutes earlier (t-120) for tipping bucket values and total influent data. In the research, different methods are applied to select features, like filtered attribute eval - ranker, filtered subset eval-greedy stepwise, CSFsubset eval-greedy stepwise, info gain attribute evaluator, wrapper subset eval-greedy stepwise and Feature selection in Statistica. These algorithms can greatly reduce the dimension of the input parameters by ranking the parameters based upon their importance to the target output. For example, boosting tree algorithm uses gradient boosting approach to predict the importance of input parameters [39]. A list of top 20 parameters is displayed in Table (3.4) and Table (3.5) for 3 and 4 bins discretization respectively. As anticipated, memory parameters of output class e.g., influent are found to be closely associated with the target output. However, in order to select other potential input parameters, a threshold value of 0.02 is set. Also radar data at 1 km for different tipping bucket locations are more important than

the tipping bucket data. In each table first row indicates the name of the attribute evaluator and second row shows the searching method.

As can be observed in the tables above, memory parameters are selected mostly as the most important ones then radar data for the closest altitudes; 1, 2, 3 and 4 km respectively.

Five promising data mining algorithms namely decision tree (J48), *k*-nearest neighbor (*k*-NN), Support vector machine (SVM), Naïve Bayes (NB), Logistic regression (LR) and Radial Basis Function (RBF) are initially selected to build prediction model at time stamp t.

**Table 3.4 Attribute selection for 3 bins**

| Filtered attribute eval | CFS subset eval | Wrapper subset eval | Feature selection - statistica |
|---|---|---|---|
| Ranker | Greedy stepwise | Greedy stepwise | Chisquare and Pvalue |
| Infl(t-30) | Infl(t-30) | TB1c1 | TB2c1 |
| Infl(t-60) | Infl(t-60) | TB2c1 | TB3c1 |
| Infl(t-90) | Infl(t-90) | TB3c1 | TB4c1 |
| Infl(t-120) | Infl(t-120) | TB4c1 | TB5c1 |
| TB2c1 | TB2c2 | TB5c1 | TB6c1 |
| TB2c2 | TB3(t-90) | TB6c1 | TB7c1 |
| TB2c3 | TB2(t) | TB7c1 | TB1c2 |
| TB2c4 | TB5c4 | TB1c2 | TB2c2 |
| TB7c4 | TB1c2 | TB2c2 | TB3c2 |
| TB7c1 | TB1(t-120) | TB3c2 | TB4c2 |
| TB7c3 | TB4c2 | TB4c2 | TB5c2 |
| TB7c2 | TB5(t) | TB5c2 | TB6c2 |
| TB3c1 | TB2(t-120) | TB6c2 | TB7c2 |
| TB1c4 | TB6c1 | TB7c2 | TB1c3 |
| TB3c2 | TB7c4 | TB1c3 | TB2c3 |
| TB4c4 | TB7(t-30) | TB2c3 | TB3c3 |
| TB4c3 | TB3c1 | TB3c3 | TB4c3 |
| TB6c1 | TB2(t-60) | TB4c3 | TB5c3 |
| TB4c3 | TB2c4 | TB5c3 | TB6c3 |
| TB1c2 | TB7(t-120) | TB6c3 | TB7c3 |

Among the selected set of algorithms, decision tree and logistic regression algorithms outperformed other algorithms Table (3.6). Since, the model constructed by decision tree algorithms are easy to comprehend, therefore, it has been finally selected to build prediction model at all of the time stamps. Various trees were applied for modeling, like random forest, random tree, REP tree and J48 which the most promising result was obtained by J48.

**Table 3.5 Attribute selection for 4 bins**

| Filtered attribute eval | Gain ratio attribute eval | Feature selection - statistica | OneRAttribute eval |
|---|---|---|---|
| Ranker | Ranker | Chisquare and Pvalue | Ranker |
| Inf(t-30) | Inf(t-30) | TB2c1 | Inf(t-30) |
| Inf(t-60) | Inf(t-60) | TB3c1 | Inf(t-60) |
| Inf(t-90) | Inf(t-90) | TB4c1 | Inf(t-90) |
| Inf(t-120) | Inf(t-120) | TB5c1 | Inf(t-120) |
| TB2c3 | TB2c3 | TB6c1 | TB2c3 |
| TB2c4 | TB3(t-120) | TB7c1 | TB2c4 |
| TB2c2 | TB3(t-90) | TB1c2 | TB2c1 |
| TB7c4 | TB2c4 | TB2c2 | TB7c4 |
| TB7c3 | TB3(t-60) | TB3c2 | TB2c2 |
| TB7c2 | TB7(t-120) | TB4c2 | TB6c2 |
| TB4c4 | TB2c2 | TB5c2 | TB7c3 |
| TB7c1 | TB7(t-90) | TB6c2 | TB1(t-30) |
| TB4c3 | TB7(t-60) | TB7c2 | TB1(t) |
| TB2c1 | TB3(t-30) | TB1c3 | TB1(t-120) |
| TB4c2 | TB7c4 | TB2c3 | TB1(t-90) |
| TB4c1 | TB7c3 | TB3c3 | TB1(t-60) |
| TB3c1 | TB6(t-60) | TB4c3 | TB1c4 |
| TB1c4 | TB6(t-120) | TB5c3 | TB6c1 |
| TB1c3 | TB6(t-30) | TB6c3 | TB3c4 |
| TB3c2 | TB6(t-90) | TB7c3 | TB6c4 |

**Table 3.6 Algorithm selection for 3 bins of output class**

| | Prediction Accuracy (%) | | | |
|---|---|---|---|---|
| | Output class | | | Overall |
| Algorithms | (-inf-38.221052] | (38.221052-49.149128] | (49.149128- inf] | |
| **J48** | **94.4** | **90.9** | **93.8** | **93.0** |
| *k*-NN (*k*=5) | 92.0 | 86.1 | 90.4 | 89.5 |
| NB | 89.6 | 83.4 | 91.1 | 87.7 |
| LR | 94.4 | 90.8 | 94.0 | 93.0 |
| RBF | 89.9 | 83.4 | 90.9 | 87.8 |

**Table 3.7 Algorithm selection for 4 bins of output class**

| | Prediction Accuracy (%) | | | | |
|---|---|---|---|---|---|
| | Output class | | | | Overall |
| Algorithms | (-inf-35.374013] | (35.374013-43.207119] | (43.207119-52.403524] | (52.403524-inf) | |
| **J48** | **92.7** | **86.1** | **88.8** | **95.1** | **90.7** |
| *k*-NN (*k*=5) | 89.8 | 78.7 | 82.8 | 92.31 | 85.8 |
| NB | 87.5 | 75.6 | 79.8 | 92.1 | 83.8 |
| LR | 92.8 | 87 | 88.6 | 95.3 | 90.6 |
| RBF | 87.7 | 74.5 | 80.6 | 91.8 | 83 |

Summary of the results obtained through decision tress algorithm over different output bins is shown in Table (3.8). The overall accuracy of the model is always found in the range 90%-93% for different bin size, whereas, g-mean of the output class was also found to be high, indicating algorithm is able to correctly predict all output classes.

In this part results which were derived from training data set are shown in tables and figures. In this section models built using decision tree algorithms at various time stamps are discussed. Evaluation criteria namely accuracy, sensitivity, specificity and g-mean are analyzed for models built for various bin size.

Here classification matrix can be seen in Figures (3.2) and (3.3) respectively for three and four bins. These figures are displayed to graphically visualize the accuracy of individual output class.

**Table 3.8 Results obtained using decision tree algorithm on training set**

| **No. of bins** | Total influent threshold | TP rate | Precision (PPV) | Recall (sensitivity) | G-mean |
|---|---|---|---|---|---|
| **3** | (-inf-38.22] | 94.4 | 94.4 | 94.4 | 93.1 |
| | ( 38.22-49.14] | 90.9 | 90.9 | 90.7 | |
| | (49.14- inf] | 93.8 | 93.8 | 94.2 | |
| 4 | (-inf-35.37] | 92.7 | 92.8 | 93.1 | 90.6 |
| | (35.37- 43.20] | 86.1 | 87.1 | 87 | |
| | (43.20- 52.40] | 88.8 | 88.7 | 88.5 | |
| | (52.40-inf) | 95.1 | 95.3 | 95.1 | |

The relative dense accumulation of data points along the actual-predicted axis of output bins indicate the output classes are correctly predicted most of the time. In the figures below, x and y axis represents the predicted and observed class respectively.

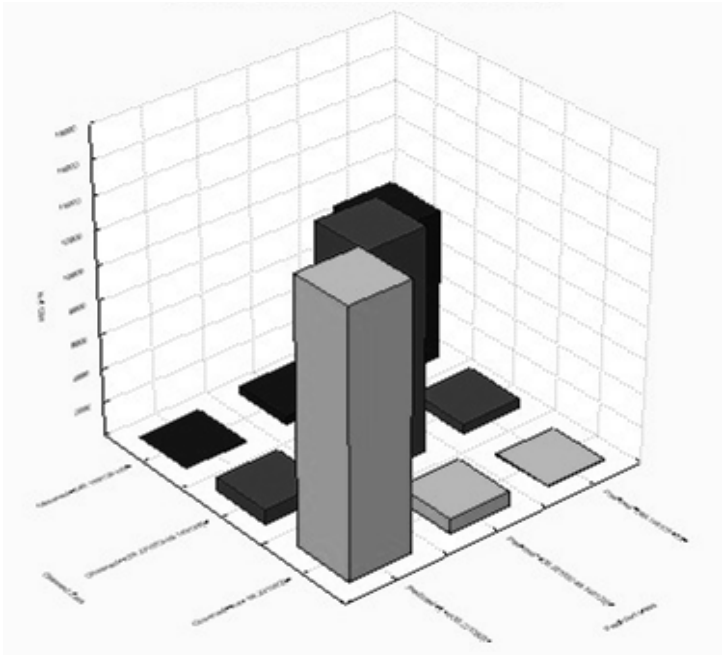


**Figure 3.2 Classification matrix for three bins**

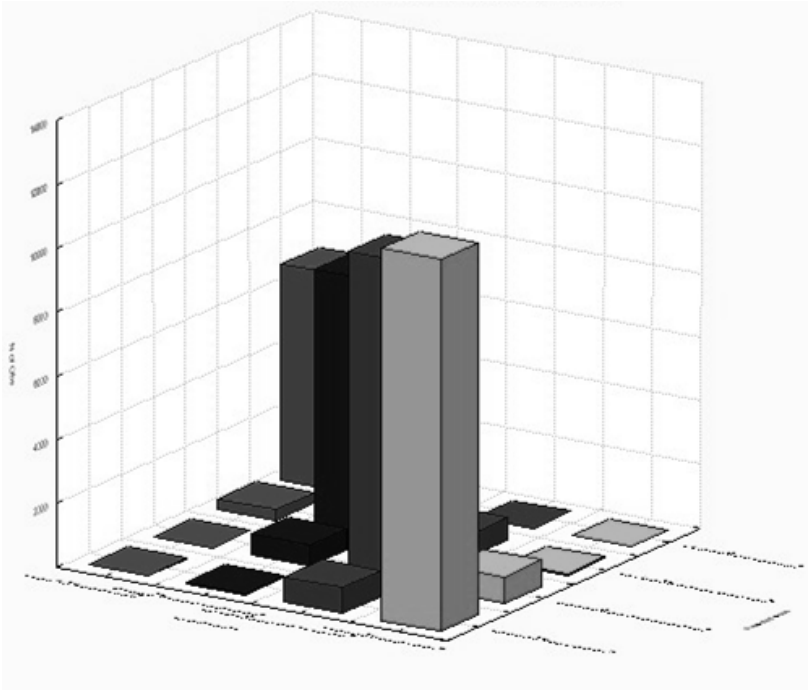


**Figure 3.3 Classification matrix for four bins**

## 3.5 Computational results

### 3.5.1 Test results

The results of applying the algorithm on testing interval are shown in Table (3.9). Testing interval is from 3/31/07 to 4/30/08 the result is as promising as training set.

**Table 3.9 Results obtained using decision tree algorithm on training set**

| No. of bins | Total influent threshold | TP rate | Precision (PPV) | Recall (sensitivity) | G-mean | Total Accuracy |
|---|---|---|---|---|---|---|
| 3 | (-inf-38.22] | 91.3 | 91.3 | 91.5 | 91.4 | 92.8 |
| | ( 38.22-49.14] | 87.7 | 87.7 | 87.2 | | |
| | (49.14- inf] | 95.6 | 95.6 | 95.8 | | |
| 4 | (-inf-35.37] | 90.0 | 90.0 | 90.0 | 89.5 | 90.8 |
| | (35.37- 43.20] | 85.3 | 85.3 | 85.3 | | |
| | (43.20- 52.40] | 88.2 | 88.2 | 88.2 | | |
| | (52.40-inf) | 94.8 | 94.8 | 94.8 | | |

### 3.5.2 Prediction results

Table (3.10) and (3.11) display the overall accuracy obtained using decision tree algorithms by WEKA. Algorithm is accurate enough to predict influent up to 60 minutes in the future; however, the accuracy drops in further time-stamps.

**Table 3.10 Three bins prediction**

| Class label | Total accuracy Y(t) | t+30min | t+ 1 hr | t+ 2 hr | t+ 3 hr |
|---|---|---|---|---|---|
| (-inf-38.221052] | 91.3 | 84.7 | 78.4 | 66.4 | 55.9 |
| (38.221052-49.149128] | 87.7 | 80.2 | 72.8 | 58.7 | 47.0 |
| (49.149128- inf] | 95.6 | 93.6 | 91.6 | 87.5 | 83.5 |
| total | 92.8 | 88.6 | 84.5 | 76.5 | 69.4 |

**Table 3.11 Four bins prediction**

| Class label | Total accuracy t | t+30min | t+ 1 hr | t+ 2 hr | t+ 3 hr |
|---|---|---|---|---|---|
| (-inf-35.374013] | 90.0 | 83.7 | 77.4 | 66.3 | 57.0 |
| (35.374013-43.207119] | 85.3 | 74.4 | 63.8 | 46.6 | 35.8 |
| (43.207119-52.403524] | 88.2 | 80.5 | 72.9 | 58.8 | 48.1 |
| (52.403524- inf) | 94.8 | 91.8 | 88.8 | 83.0 | 77.9 |
| Total | 90.8 | 84.7 | 78.7 | 68.0 | 59.8 |

### 3.6 Regression (time series regression)

A data-mining approach to predict influent flow rate in a wastewater treatment plant for a short-term period (up to 180 min ahead) is presented. The prediction model is constructed by data-mining algorithms using radar reflectivity data, rainfall rate data, and the historical influent flow rate data. Radar reflectivity data can be used to forecast weather several hours or even several days ahead. In the regression model, inputs and output for this model are the same as classification model in previous sections. Sampling interval is 1/1/2007 2:00:00 AM to 3/31/2008 11:45:00 PM, with the frequency of 15 minutes, there are 43768 instances. Training interval is 1/1/2007 2:00:00 AM to 11/1/2007 12:15:00 AM and it is tested over 11/1/2007 12:30:00 AM to 3/31/2008 11:45:00 PM. A multilayer perceptron neural network (MLP) is used to build the prediction model and compare its accuracy with models constructed by three other data-mining algorithms. The best performing algorithm is selected to build the prediction model. The prediction results are evaluated by prediction metrics and discussed in detail.

In Figure (3.4) it can be observed that the amount of rainfall varies from one location to other location and based on correlation coefficient matrix there is a nonlinear relation among them.

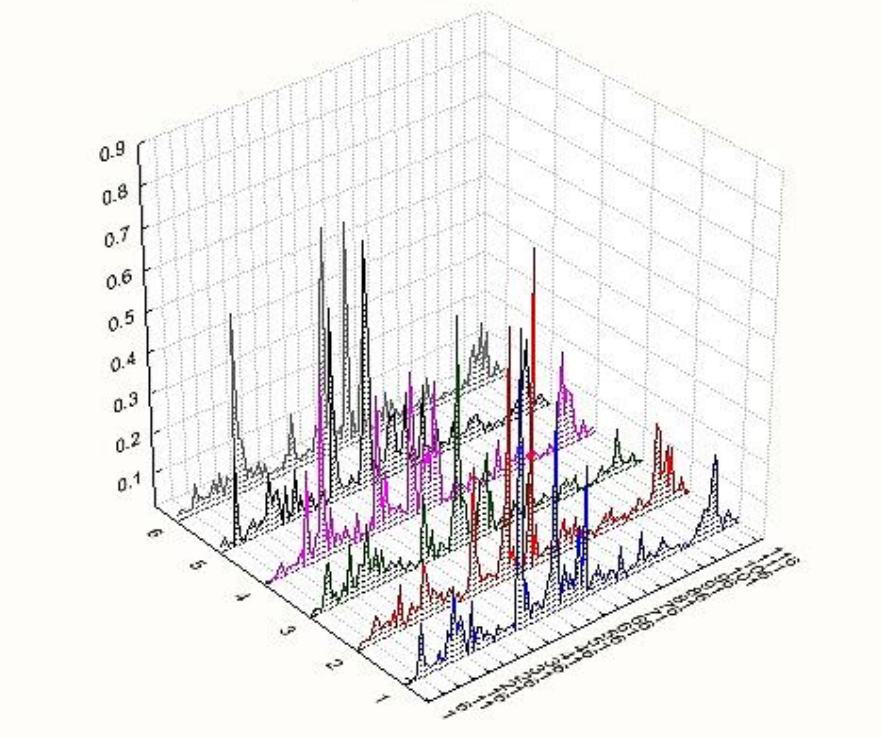



**Figure 3.4 Rainfall comparisons among six tipping buckets in WRA area**

### 3.6.1 Data preprocessing

The historical values of influent flow rate, rainfall rate, and radar reflectivity are used to construct the prediction model. The influent flow rate data was collected at the Des Moines Wastewater Reclamation Facility (WRA), Iowa. WRA processed wastewater from 16 metro area municipalities, counties and sewer districts in the Des Moines area. Also, the tipping bucket values were recorded to the 0.0001 mm/hr, which seemed excessively precise. These values were rounded to the nearest 0.01 mm/hr for modeling purposes. And the rest of the preprocessing steps are the same as the classification except discretization which is not required in this section. Training and testing intervals are the same as previous section.

### 3.6.2 Feature selection and algorithm selection

Three methods are used for feature selection, boosted tree, random forest and feature selection in Statistica software. The results are listed below in the Table (3.11).

**Table 3.12 Feature selection results by different methods**

| Boosted tree | Random forest | Statistica |
|---|---|---|
| infl(t-15) | infl(t-60) | TB5(t-30) |
| infl(t-30) | infl(t-90) | TB7(t-30) |
| infl(t-45) | infl(t-15) | TB2(t-30) |
| infl(t-60) | infl(t-45) | TB7(t-60) |
| infl(t-90) | infl(t-30) | TB7(t-90) |
| infl(t-120) | infl(t-120) | TB1(t-30) |
| TB7(t) | TB2(t-30) | infl(t-60) |
| TB2(t-60) | TB7(t-30) | TB5(t-60) |
| TB5(t) | TB2(t-90) | TB5(t-90) |
| TB1(t) | TB2(t-120) | TB7(t-120) |
| TB7(t-60) | TB5(t) | TB6(t-60) |

As shown in the tables above tipping bucket memory parameters are not effective on the output significantly while memories of total influent radar data of 1, 2, 3 and 4 km of some locations are recognized as important ones. Different data-mining algorithms are used to build the prediction model for prediction of the influent flow rate. Two metrics, the mean absolute error (MAE) and mean squared error (MSE) are used to measure prediction accuracy. MAE is a common used quantity in time series analysis to measure

how close the predictions are to the observations. MSE is a way to quantify the difference between the values implied by the prediction method and the true values. It is a risk function corresponding to the expected value of the squared error loss. The expressions to calculate MAE and MSE are shown in (3.1) and (3.2) and the trained algorithms are shown in Table (3.12). The most promising results are obtained by NN.

**Table 3.13 Regression model accuracy**

| No. | Algorithm | MAE | Correlation coefficient | MSE |
|---|---|---|---|---|
| 1 | NN | 1.095 | 0.988 | 4.215 |
| 2 | Random forest | 3.041 | 0.945 | 20.699 |
| 3 | Boosted tree | 1.776 | 0.970 | 11.162 |
| 4 | SVM | 1.476 | 0.985 | 5.461 |

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | f_i - y_i | \quad (3.1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} | f_i - y_i |^2 \quad (3.2)$$

### 3.6.3 Computational results

MLP is chosen as the best algorithm, here are the results in Table (3.13) for 5 best MLPs. Using Statistica's "Automatic Network Search" option 200 MLP's were generated with random attributes. Some of these characteristics were learning rate, momentum, number of hidden layers, and number of nodes. The activation functions tried in the neurons were the identity, logistic, tanh, and exponential functions. The top 5 performing MLPs were retrained (tuned).

**Table 3.14 Five best MLPs**

| Net. name | Training perf. | Test perf. | Validation perf. | Training error | Test error | Validation error |
|---|---|---|---|---|---|---|
| MLP 60-58-1 | 0.9936 | 0.9922 | 0.9940 | 4.3895 | 4.9278 | 4.0127 |
| MLP 60-19-1 | 0.9929 | 0.9918 | 0.9941 | 4.8712 | 5.1533 | 3.9841 |
| MLP 60-8-1 | 0.9933 | 0.9921 | 0.9940 | 4.6041 | 5.0144 | 4.0164 |
| MLP 60-19-1 | 0.9946 | 0.9925 | 0.9941 | 3.6727 | 4.7152 | 3.9253 |
| MLP 60-11-1 | 0.9932 | 0.9921 | 0.9940 | 4.6870 | 5.0090 | 4.0207 |

Prediction results are shown in the Table (3.14), these results are taken from the MLP network 60-58-1 (Tanh-logistic). Correlation coefficient in training data set was 0.994 and when it was tested over testing period it decreased to 0.988 it kept on decreasing while testing over longer times ahead. Another evaluation metrics which was applied is standard deviation, for predicted values, standard deviation stayed constant as 13.218 while for observed ones for t to t+180 as observed in table below it had these values successively 13.624, 13.624, 13.625, 13.627, 13.628, 13.630, 13.631 and 13.633. Prediction continued till the correlation is higher than 0.85.

**Table 3.15 Prediction results**

| | T (test) | T+15 | T+30 | T+60 | T+90 | T+120 | T+150 | T+180 |
|---|---|---|---|---|---|---|---|---|
| Corr. Coeff. | .988 | 0.983 | 0.976 | 0.958 | 0.934 | 0.905 | .872 | 0.836 |
| MAE | 1.09 | 1.48 | 1.89 | 2.75 | 3.61 | 4.46 | 5.26 | 6.02 |
| MSE | 4.21 | 5.83 | 8.20 | 14.59 | 22.95 | 33.21 | 44.88 | 57.39 |

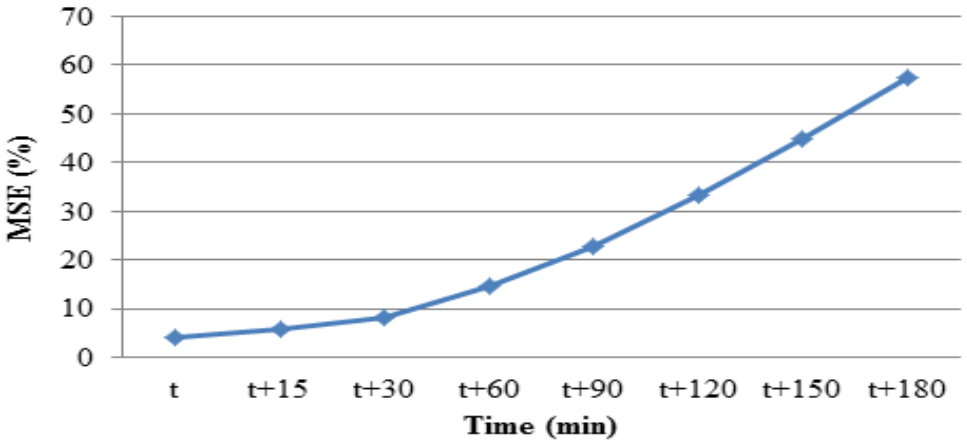



**Figure 3.5 MSE of the model for the prediction of the influent flow rate**

Based on plots below it can be seen that predicted values and observed values are highly correlated and it demonstrates that algorithm is highly accurate. The figures below are test and validation samples plot of predicted versus observed.

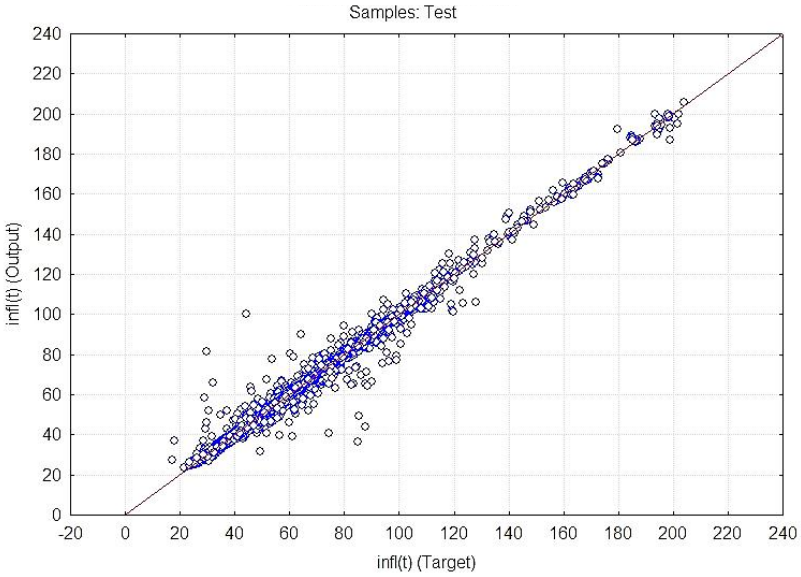



**Figure 3.6 Test samples plot of predicted vs. observed**

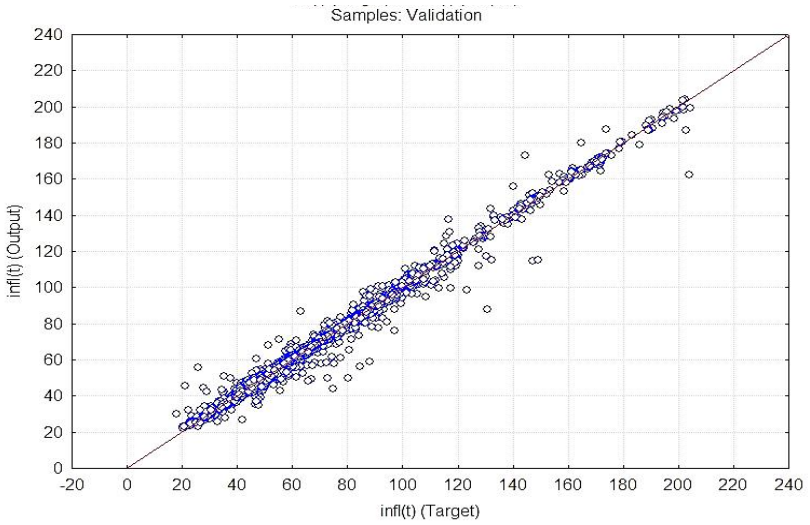



**Figure 3.7 Validation samples plot of predicted vs. observed**

By building 7 MLP prediction models at t + 15 min, t + 30 min, t + 60 min, t + 90 min, t + 120 min, t + 150 min, and t + 180 min respectively, the influent flow rate can be predicted up to 180 minutes ahead. In Figure (3.9) it can be seen that the predicted influent flow rate is close to the observed influent flow rate, and the trend for both predicted and observed values is same. However, there is a slight lag for the predicted values. This lag becomes larger with longer prediction horizon. It can be clearly found in Figure (3.10) which predicts the influent flow rate at time t + 180 min ahead. Even though the trend is successfully predicted, the response for the prediction model is slow.

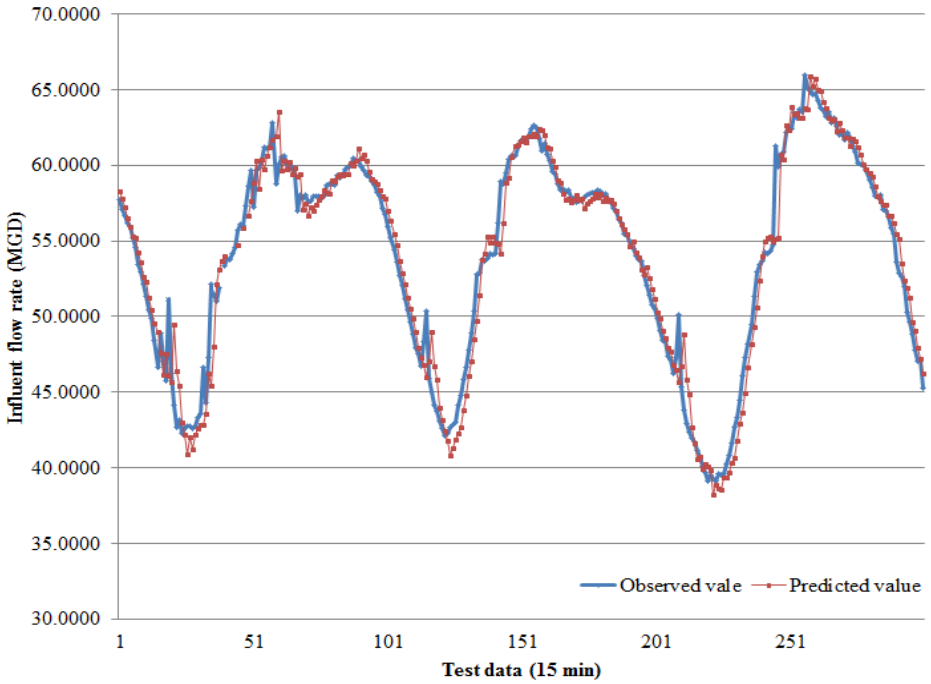



**Figure 3.8 Prediction of the influent flow rate at current time t**

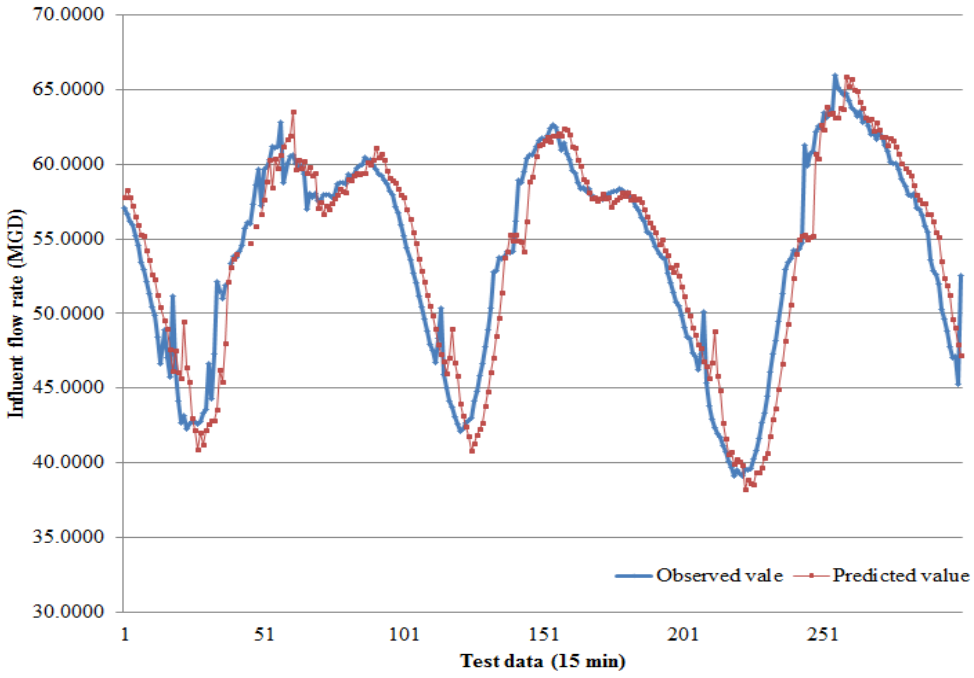



**Figure 3.9 Prediction of the influent flow rate at time t + 30 min**

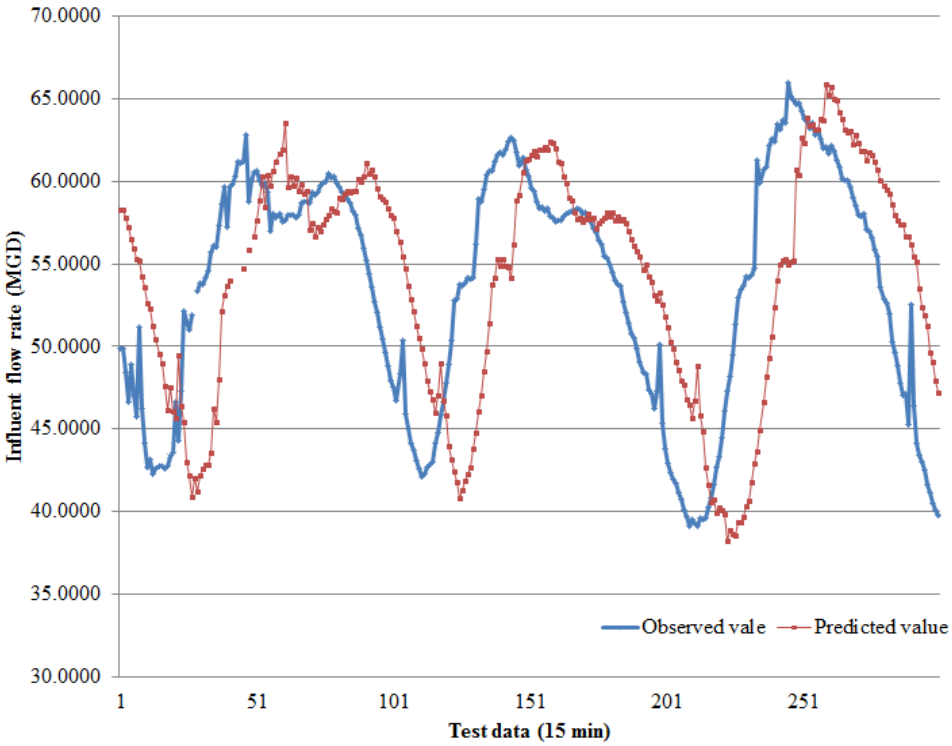



**Figure 3.10 Prediction of the influent flow rate at time t + 180 min**

## 3.7 Conclusion

To maintain stable effluent and optimally arrange wastewater boosting pumps, it is helpful to know in advance the influent flow rate to the wastewater treatment plant. In this chapter, the prediction model of influent flow rate up to 180 min ahead was built using rainfall rate, radar reflectivity, and influent flow rate as predict inputs. The influent flow rate data were collected at Wastewater Reclamation Facility (WRA), the rainfall rate data were recorded by 6 tipping buckets surrounding WRA, and the radar reflectivity data were obtained from the radar map through nearby radar station. The data were converted to have same frequency by taking the average based on different frequencies.

Among four data-mining algorithms used in this paper, Decision trees for classification and the MLP neural network for regression performed better than other algorithms applied to build the prediction model. It was selected to construct the prediction model of influent flow rate for all prediction horizons from t to t + 180 min. The results showed that the prediction model predicted the influent flow rate well till t + 150 min. The predicted influent flow rate was close to the measured influent flow rate, and the trend for both predicted and observed values was same. In addition, there was a lag between predicted and observed influent flow rate after t + 30 min, and the lag became larger with longer time horizons. At t + 180 min, i.e., 3 hours ahead, the prediction accuracy metrics indicated that the prediction model performed not well enough.

Prediction of the influent flow rate 150 min ahead might give enough time for wastewater treatment plant to arrange operators and schedule pumps. However, the prediction accuracy of the prediction model should be improved in future research in order to provide long term prediction with acceptable accuracy. In case of heavy rainfall, long term prediction will give more time to wastewater treatment plant to make operation plans.

# CHAPTER 4. PERFORMANCE PREDICTION OF A WASTEWATER TREATMENT PLANT

## 4.1 Introduction

If a model is developed based on historical observations of main parameters to predict the performance of the plant, there will be a safer operation and easier control of the wastewater plant. Wastewater treatment plants consist of chemical, physical and biological processes. Neural Networks can work as performance predictors for such nonlinear complex processes. To assess the performance, historical data of key parameters are applied in the model. For example, biological oxygen demand (BOD), suspended solid (SS) and chemical oxygen demand (COD) [40]. Quality of influents are deteriorated by wastewaters and plant effluents, hence to increase the potential of water reuse, advanced treatment is needed [48].

Intelligent methods for prediction of WWTP parameters are widely used in the recent decades. Chen, Chang, and Shieh (2003) used a novel approach based on NN model to predict nitrogen contents in treated effluents [42]. Total suspended solid (TSS) is an indication of plant performance. Belanche, Valde´s, Comas, Roda, and Poch (2000) predicted TSS based on Neural Networks [43]. Shetty and Chellam (2003) develop a neural network model to predict long term fouling of nanofiltration membranes that are used to purify contaminated water supplies [44]. Hamed et al (2004).used NN model to predict biological oxygen demand (BOD) and suspended solid (SS) concentrations in plant effluent [39]. Maier, Morgan, and Chow (2004) modeled alum dosing surface waters by NN [46]. Kohonen found that low pH in biological reactor and long solid retention time caused high concentration of BOD and TSS by using the self-organizing feature maps to calssify data [49]. In another NN model, parameter selection for entering the network resulted that porous media porosity, wastewater temperature and hydraulic residence time are the main parameters affecting BOD removal also COD removal was highly correlated to BOD removal [50].

One-line training of the neural network model may improve the prediction accuracy [47]. Oliveira-Esquerrea [51] used multilayer perceptron (MLP) and functional-link neural networks (FLN) and to model and predicts inlet and outlet biochemical oxygen demand (BOD) and developed them using linear multivariate regression techniques.

### 4.2 Plant layout: a case study

The NN model was applied to Wastewater Reclamation Authority (WRA). The models were tested for different configurations of input–output data. Using the results of this modeling process, the plant operator will be able to have an assessment of the expected plant effluent for a given quality of the wastewater stream at input locations. A schematic diagram of the plant is shown in Figure (4.1).

Main outfall, sewer and fourmile are collected in raw wastewater junction chamber then screened for removal of grits. Settled solids are scrapped down in the hoppers and carried to paddle mixers. Aerobic bacteria are activated by aeration and mixing with activated sludge. After roughing filters process, aeration tanks starts working and chlorination system. Rotary Drum Thickeners navigate solid parts to the secondary digesters and gas is generated.

### 4.3 Data collection

It was decided to relate the outputs of the treatment effluent stream to the inputs of the stream (influent). Therefore, measurements of the carbonaceous biochemical oxygen demand (CBOD) and total suspended solid (TSS) in the effluent stream and influent stream were collected over period of 1/1/2008 to 12/31/2010, 3 years of data. This period was satisfactory as it covers all probable seasonal variations. Parameters in the model are introduced in Table (4.1).

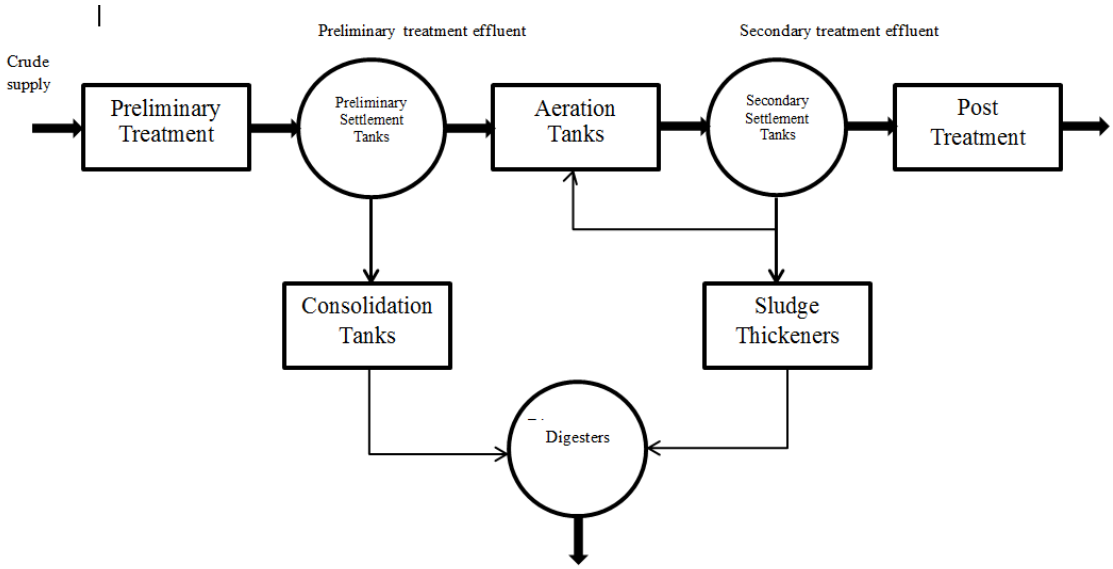



**Figure 4.1 Schematic diagram wastewater processes**

**Table 4.1 Parameters of the model**

| Parameter | Label | Description | Unit |
|---|---|---|---|
| CBOD in influent plant | infl-CBOD | Amount of CBOD in influent stream | mg/l |
| TSS in influent plant | infl-TSS | Amount of TSS in influent stream | mg/l |
| CBOD in effluent plant | efl-CBOD | Amount of CBOD in effluent stream | mg/l |
| TSS in effluent plant | efl-TSS | Amount of TSS in effluent stream | mg/l |
| Total influent to the plant | Infl | Amount of total influent to the wastewater plant (Building 05) | GPD |

The measurements were performed in the plant almost for 3 days a week. CBOD and TSS are measured and sent to the laboratory. Based on the data set they usually do the sampling on Wednesdays, Thursdays and Fridays but it may change in some weeks and sometimes they measure only twice a week even not on successive days, because of inconsistency in the method of sampling, frequency of 7 days is supposed for the model

and the average is taken for all available values in a week and considered as data set1, different data sets are defined for other approaches. Sampling is done over the period 1/1/2008 to 12/31/2010, hence there are 157 instances with 7 days frequency, first two years are used as training data set and the rest which is 1 year is for testing.

In dataset1, it is supposed that measurements are done weekly so the average of three values in each week is calculated and there were 157 data points. In data set 2, it is supposed that the measurements are done in a daily manner, so there are 440 data points and for the whole sampling interval which is 3 years it is required to have 1095 data points, so interpolation is applied to fill 655 missing values and the interpolation is not only based on historical data but also based on the amount of the total influent to the plant.

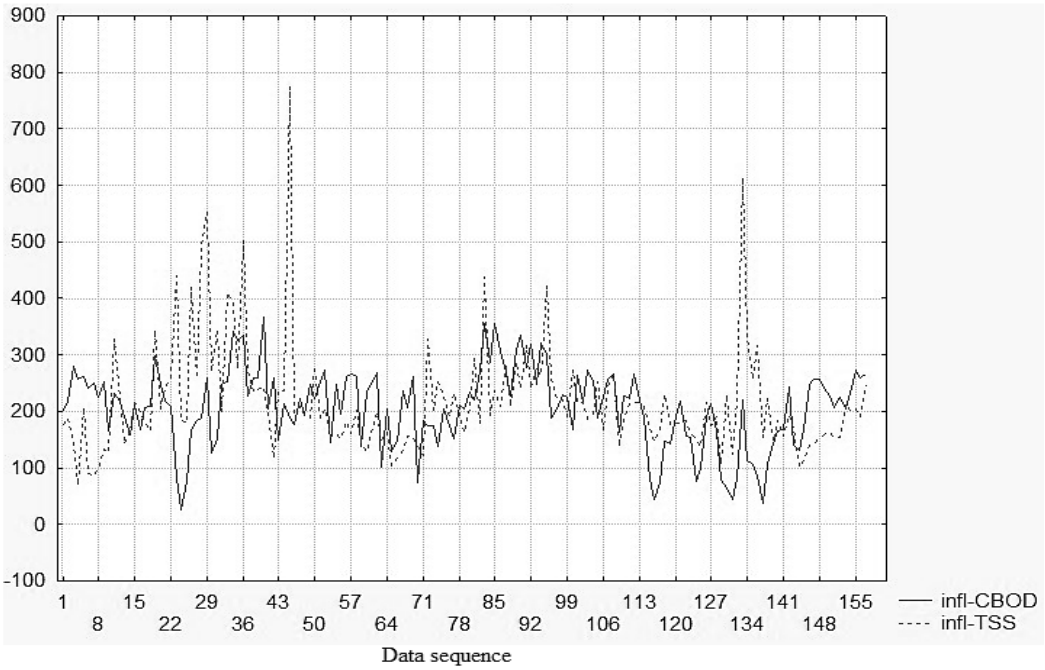



**Figure 4.2 Data sequence of CBOD and TSS in influent, time unit is week and chemicals' unit is mg/l**

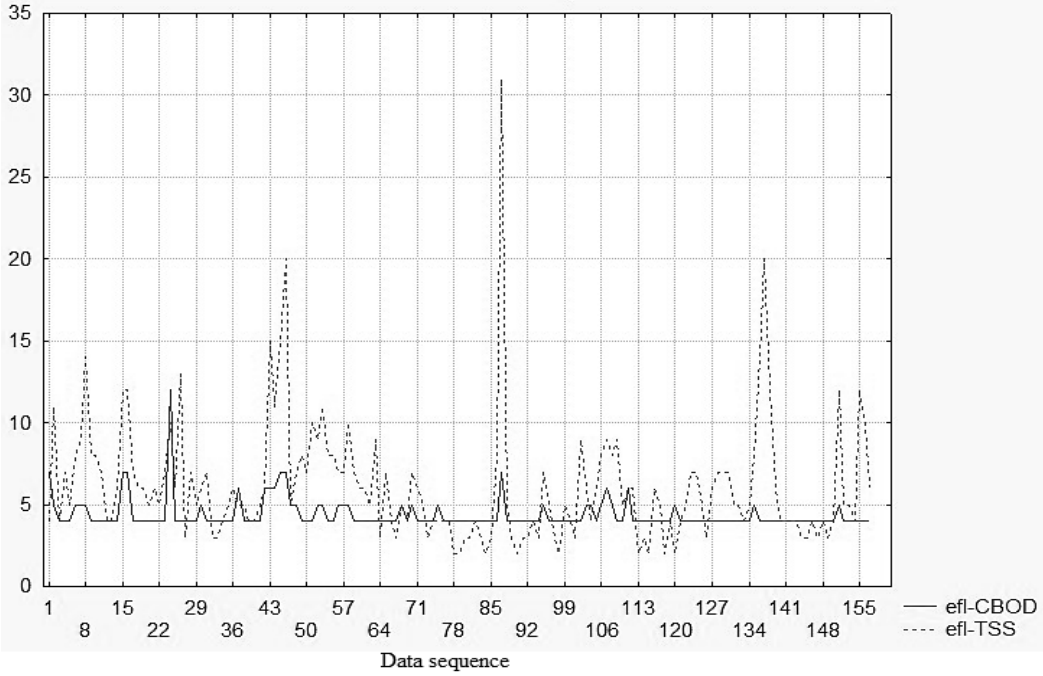



**Figure 4.3 Data sequence of CBOD and TSS in effluent, time unit is week and chemical's unit is mg/l**

The conventions efl-CBOD, efl-TSS, infl-CBOD and infl- TSS hold for CBOD in effluent flow, TSS in effluent flow, CBOD in influent flow and TSS in influent flow, respectively.

### 4.4 Data preparation, preprocessing and statistical analysis

The CBOD and TSS were selected because they can be used as measures for the effectiveness of the wastewater treatment plant. Data refining was done by taking the average of 2 or 3 values in each week then excluding the outliers by considering the acceptable limit in the plant controlling system and excluding the values which were not in the range of $\pm 3\sigma$ around mean.

Various manipulations can be applied to decipher data series, in this research data series are normalized. The main objective here is to ensure that the statistical distribution of the values for each net input and output is roughly uniform. In addition, the values

should be scaled to match the range of the input neurons. The data sets are usually scaled so that they always fall within a specified range or they are normalized so that they have zero mean and unity standard deviation. This is achieved by normalizing the mean and standard deviation of the data set.

The preprocessed data set was analyzed statistically by generating a box and whiskers plot for each variable. These plots summarize each variable by three components; a central line to indicate central tendency or location; a box to indicate variability around this central tendency, median and whiskers around the box to indicate the range of the variable. This is shown in Figure (4.4), which is derived from raw data before any preprocessing. The plots illustrate the extent of outlier density in each variable as indicated by the points extending beyond the whiskers. In addition, it shows the range of each variable and, consequently, the efficiency of the plant treatment.

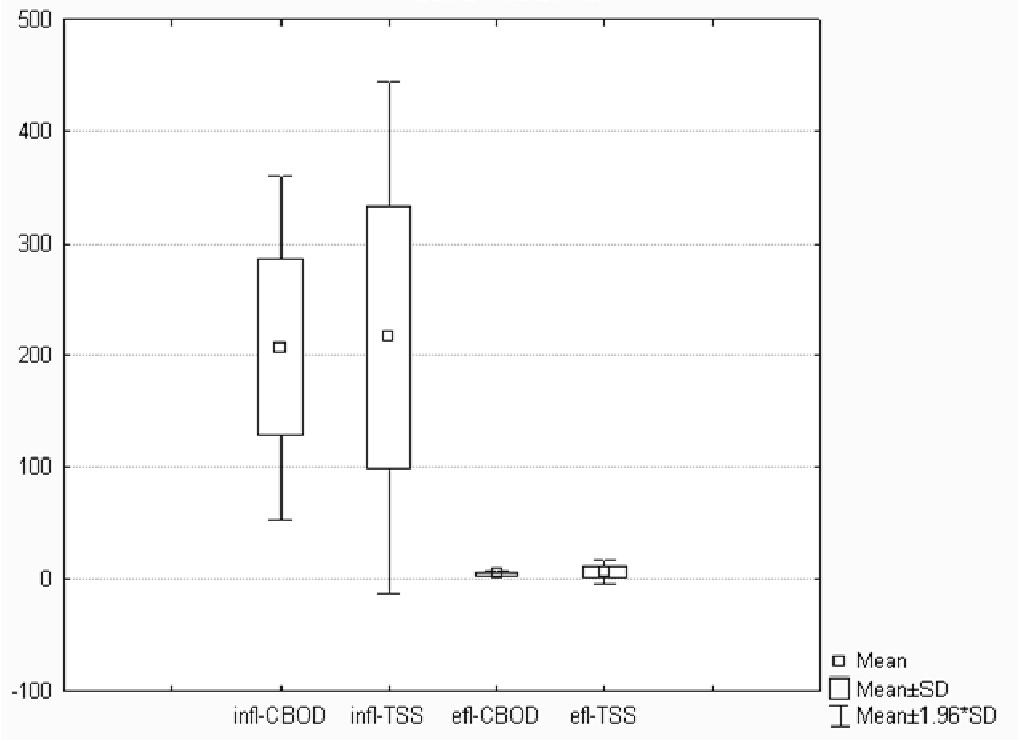



**Figure 4.4 Box diagrams for the plant data for effluent and influent streams**

### 4.5 NN modeling; methodology

The NNs can be categorized in terms of topology such as single and multi-layer feedforward networks (FFNN), feedback networks (FBNN), recurrent networks (RNN), self-organized networks. In addition, they can be further categorized in terms of application, connection type and learning methods. The most commonly used type of networks in the field of modeling and prediction is the FFNN. In this topology, the network is composed of one input layer, one output layer and a minimum of one hidden layer. The term feedforward describes the way in which the output of the FFNN is calculated from its input layer-by-layer throughout the network.

Activation functions for the hidden units are needed to introduce the nonlinearity into the network. The Sigmoidal functions, such as logistic and tanh, and the Gaussian function, are the most common choices for the activation functions. The neural system architecture is defined by the number of neurons and the way in which the neurons are interconnected. In this research Gaussian is applied for data set1 because of its most promising result.

The data are normally divided into three subsets; training, validation and testing subsets. The training subset data are used to accomplish the network learning and fit the network weights by minimizing an appropriate error function. Backpropagation is the training technique usually used for this purpose. It refers to the method for computing the gradient of the case-wise error function with respect to the weights for a feedforward network. The performance of the networks is then compared by evaluating the error function using the validation subset data, independently. The testing subset data are then used to measure the generalization of the network (i.e. how accurately the network predicts targets for inputs that are not in the training set) this is sometimes referred to as holdout validation. Training, test and validation ratio which was applied in this research is 4:2:1.

There are many reported techniques to avoid underfitting and overfitting such as model selection, jittering, early stopping, weight decay, Bayesian learning, and combining networks, in this research different values for weight decay is applied to avoid over-fitting.

The structure must be optimized to reduce computer processing, achieve good performance and avoid overfitting. The selection of the best number of hidden units depends on many factors. The size of the training set, amount of noise in the targets, complexity of the sought function to be modeled, type of activation functions used and the training algorithm all have interacting effects on the sizes of the hidden layers. There is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each.

### 4.5.1 NN vs. regression

Neural networks extract information from data in the form of predictive input–output models also they provide a very general framework to approximate any type of nonlinearity in the data [52].

Regression equations are very useful, but they simplify a complex system, like wastewater plants, into a few parameters, and may ignore crucial factors [50].

NN are used as nonlinear modeling techniques for CBOD and TSS prediction. Because neural networks are parallel and have better filtering capacity moreover with noisy or incomplete data NN usually perform better than linear models [8]. However, as neural networks function known as black boxes, are difficult to interpret and unknown in physical insight of data [55], in addition, Multilayer perceptron (MLP) have been successfully used in modeling biological wastewater treatment processes [53, 54].

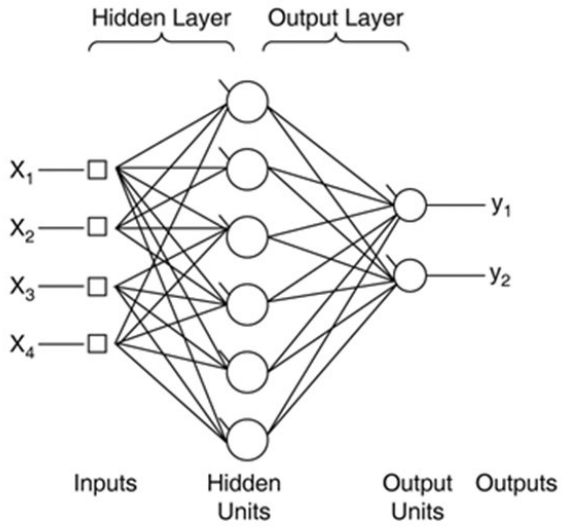



**Figure 4.5 Schematic of the multi-layer NN**

## 4.6 Results and discussion

In this section statistical results for each data set are described, then the modeling and prediction is presented, finally the results of all sections are compared and the most promising one is introduced.

### 4.6.1 Data set 1

#### 4.6.1.1 Statistical analysis

Correlation coefficient table was a preliminary multivariable statistical analysis which was used to explore the degree that a linear model can describe the relation among variables. Correlation matrix is used widely to measure correlation or association. It can give the idea that which variable is better to use to predict the other on based on a linear relation. As shown in Table (4.2), there were some degrees of linear correlation between the variables in influent and in the effluent. The weakness of the values in table proves that conventional regression techniques in modeling such a complex process will give poor results so there is a need to use complex modeling.

**Table 4.2 Correlation matrix for plant variables**

| Correlation Coefficients | | | | |
|---|---|---|---|---|
| | infl-CBOD | infl-TSS | efl-CBOD | efl-TSS |
| infl-CBOD | 1 | **0.193692** | -0.12027 | -0.16956 |
| infl-TSS | **0.193692** | 1 | 0.035442 | 0.069419 |
| efl-CBOD | -0.12027 | 0.035442 | 1 | **0.476558** |
| efl-TSS | -0.16956 | 0.069419 | **0.476558** | 1 |

#### 4.6.1.2 Modeling results

Neural network toolbox in Statistica is utilized for this analysis. The previously described neural networks design procedure is applied to model the WWTP. Two NN input topologies are considered for the plant modeling, different configurations of input–output data. There are six different configuration of input-output as can be seen below in Table (4.3).

In the first approach, each of the influent variables (TSS, CBOD) is used to predict each of the effluent variables. In the second approach multi-input variables are used to predict the corresponding output variables in the effluent stream.

**Table 4.3 Different configurations of input-output**

| Model number | Input | Output |
|---|---|---|
| 1 | Infl-TSS | Efl-CBOD |
| 2 | Infl-CBOD | Efl-CBOD |
| 3 | Infl-TSS | Efl-TSS |
| 4 | Infl-CBOD | Efl-TSS |
| 5 | Infl-TSS, Infl-CBOD | Efl-CBOD |
| 6 | Infl-TSS, Infl-CBOD | Efl-TSS |

To keep the model simple, 1 hidden layer for single input and 1 hidden layer for multiple inputs are used. On the other hand, the number of neurons in the hidden layer is selected after testing the performance of the networks at different combinations. It is

noticed that 40 and sometimes 60 neurons is the least number of neurons, in the hidden layer, which converged to a final solution.- for multi-layer. However, for the multi-input case the hidden layer contains 40 neurons.

The constituents of the network layers, i.e. types of neurons, were taken to be Gaussian after testing different combinations. MLP was applied to all configurations but there was not any promising result. The results below all had RBFT training algorithm and error function was sum of squared error, hidden activation is Gaussian and output activation is Identity.

The computational results are shown in Table (4.4). The table indicates the most promising networks yielded from automated research in Statistica for each configuration.

**Table 4.4 Summary of trained NN results for different input-output variable combinations**

| Input | Output | Net. name | Training perf. | Test perf. | Validation perf. | Training error | Test error | Validation error |
|---|---|---|---|---|---|---|---|---|
| TSS | TSS | RBF 1-40-1 | 61.50 | 16.71 | 24.29 | 5.73 | 6.45 | 3.95 |
| CBOD | TSS | RBF 1-60-1 | 64.68 | 31.32 | 15.24 | 5.40 | 5.86 | 5.29 |
| CBOD | CBOD | RBF 1-60-1 | 51.38 | 17.15 | 25.53 | 0.01 | 0.01 | 0.01 |
| TSS | CBOD | RBF 1-40-1 | 46.64 | 18.35 | 39.72 | 0.01 | 0.01 | 0.09 |
| TSS, CBOD | TSS | RBF 2-40-1 | 58.28 | 19.96 | 46.42 | 6.08 | 5.66 | 2.64 |
| TSS, CBOD | CBOD | RBF 2-40-1 | 21.04 | 25.60 | 10.00 | 0.01 | 0.00 | 0.00 |

### 4.6.2 Data set 2

In preprocessing step outliers are thrown out based on the limits provided by plant; also reinforced by 6sigma principal which was applied before to have an acceptable range for all variables in the model; for influent chemicals minimum and maximum for CBOD are 20 and 443 mg/l respectively and for TSS 60 and 1260 mg/l respectively. On the other hand, in effluent stream minimum and maximum for CBOD are 4 and 20 also for TSS 2 and 88 mg/l respectively, these limits for influent rate to the plant are 0 and 260 GPD. Time span for sampling is 1/1/2008 to 12/30/2009 whereas 1/1/2008 to 4/30/2009 is considered for training and 5/1/2009 to 12/30/2009 for testing, there are 730 points of daily data with many missing values. Table (4.5) shows different configurations of inputs and outputs.

**Table 4.5 Different configurations of input-output**

| Model number | Input | Output |
|---|---|---|
| 1 | Infl-TSS, Infl-CBOD,influent | Efl-CBOD |
| 2 | Infl-TSS, Infl-CBOD, influent | Efl-TSS |

In data set 2, total influent is added to the model, correlation coefficient among total influent and rest of the attributes is shown in Table (4.6). These weak values demonstrate feeble linear relationship among new income and old elements of the model.

**Table 4.6 Correlation coefficient of total influent and other attributes of the model**

| | inflCBOD | infTSS | influent | eflCBOD | eflTSS |
|---|---|---|---|---|---|
| Influent | -0.65 | 0.05 | 1.00 | 0.25 | 0.07 |

Boosted tree algorithm was applied for feature selection, this algorithm is already described in previous sections, and the results are explained for CBOD and TSS output. As can be seen in the table when the output is TSS, influent TSS and total influent are more important than the influent CBOD, and exactly the same for configuration 2.

**Table 4.7 Boosted tree results in feature selection**

| | CBOT out put | | TSS out put | |
|---|---|---|---|---|
| Inputs | Variable Rank | Importance | Variable Rank | Importance |
| inflCBOD | 100 | 1.00 | 100 | 1.00 |
| influent | 67 | 0.67 | 63 | 0.63 |
| infTSS | 35 | 0.34 | 60 | 0.60 |

Based on the most promising results derived from the other data set, MLPs are chosen and trained for this data set. Configuration 1 has the output (Efl CBOD) and best MLP networks obtained by Automated search in Statistica for it is shown in Table (4.8). Configuration 2 has the output (Efl TSS); the best networks are shown in Table (4.9).

**Table 4.8 Best MLP networks for configuration 1**

| Net. name | Test perf. | Validation perf. | Test error | Validation error | Hidden activation | Output activation |
|---|---|---|---|---|---|---|
| MLP 3-163-1 | 0.638295 | 0.829115 | 0.433823 | 4.692967 | Exponential | Exponential |
| RBF 3-50-1 | 0.277232 | 0.501661 | 0.460141 | 5.651888 | Gaussian | Identity |

**Table 4.9 Best MLP networks for configuration 2**

| Net. name | Test perf. | Validation perf. | Test error | Validation error | Hidden activation | Output activation |
|---|---|---|---|---|---|---|
| MLP 3-91-1 | 0.760129 | 0.472630 | 2.673671 | 5.606065 | Tanh | Exponential |
| MLP 3-17-1 | 0.750982 | 0.445657 | 2.746222 | 5.067053 | Tanh | Logistic |
| RBF 3-30-1 | 0.608495 | 0.550173 | 3.955731 | 4.131057 | Gaussian | Identity |

In Figures (4.6) to (4.9) correlation between predicted values and observed values in testing and validation samples is visualized, as can be seen the observed and predicted values do not follow the expected linear relation, that is the reason that data set 3 is defined to see if the error can be decreased by filling the missing values.

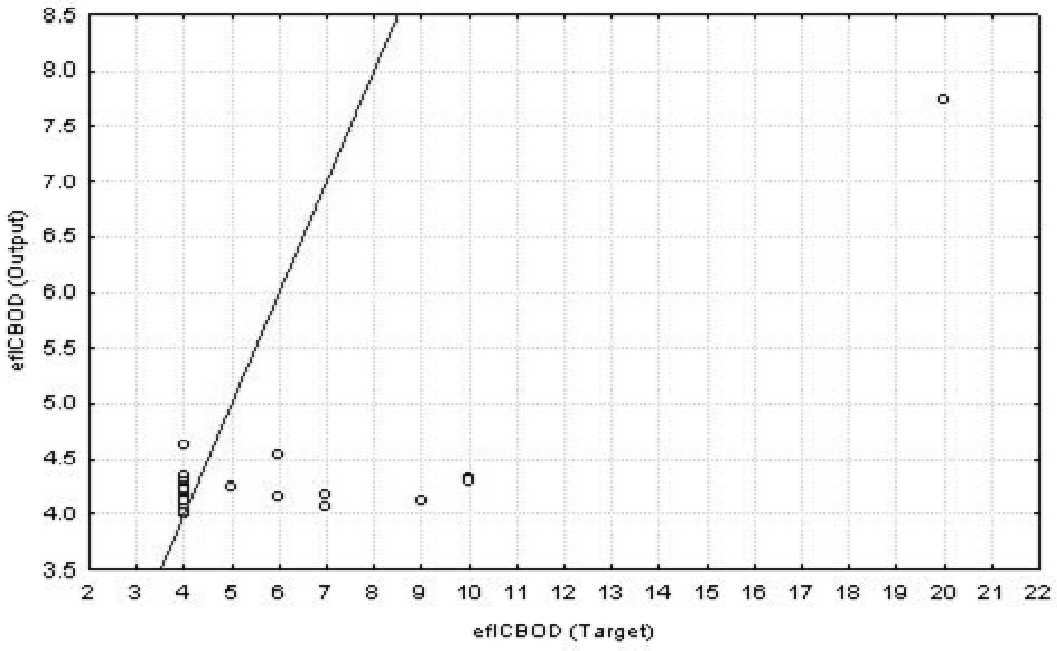



**Figure 4.6 Observed versus predicted values for validation samples - configuration 1**

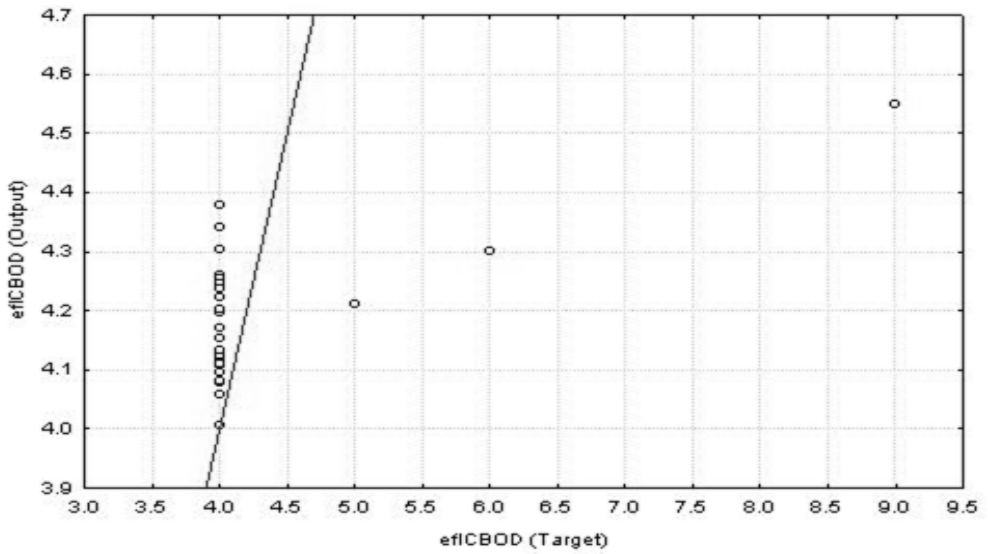



**Figure 4.7 Observed versus predicted values for test samples – configuration 1**

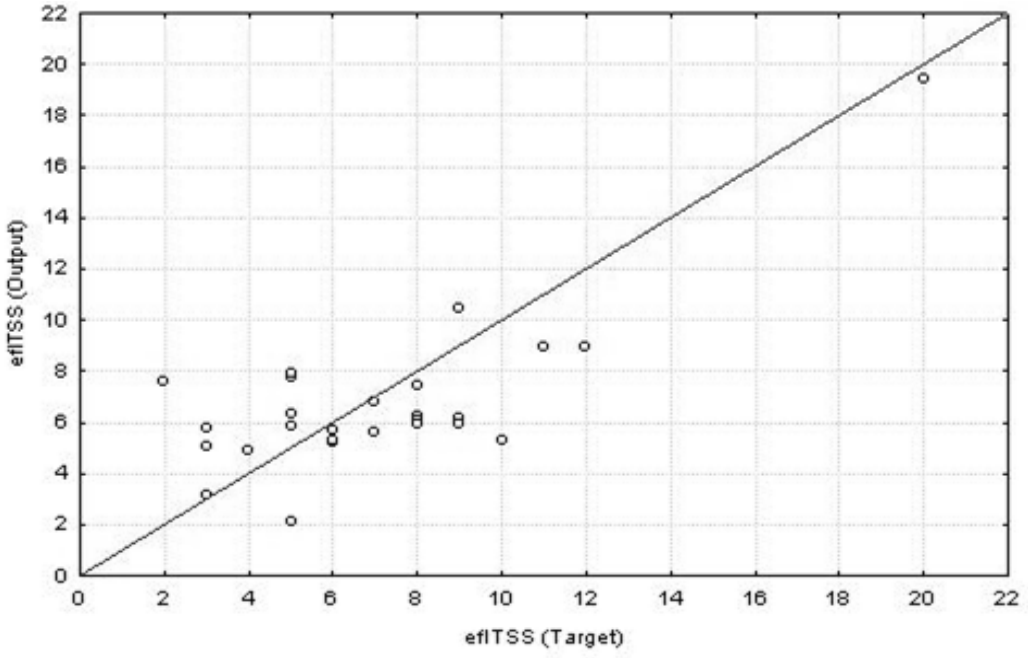



**Figure 4.8 Observed versus predicted values for test samples – configuration 2**

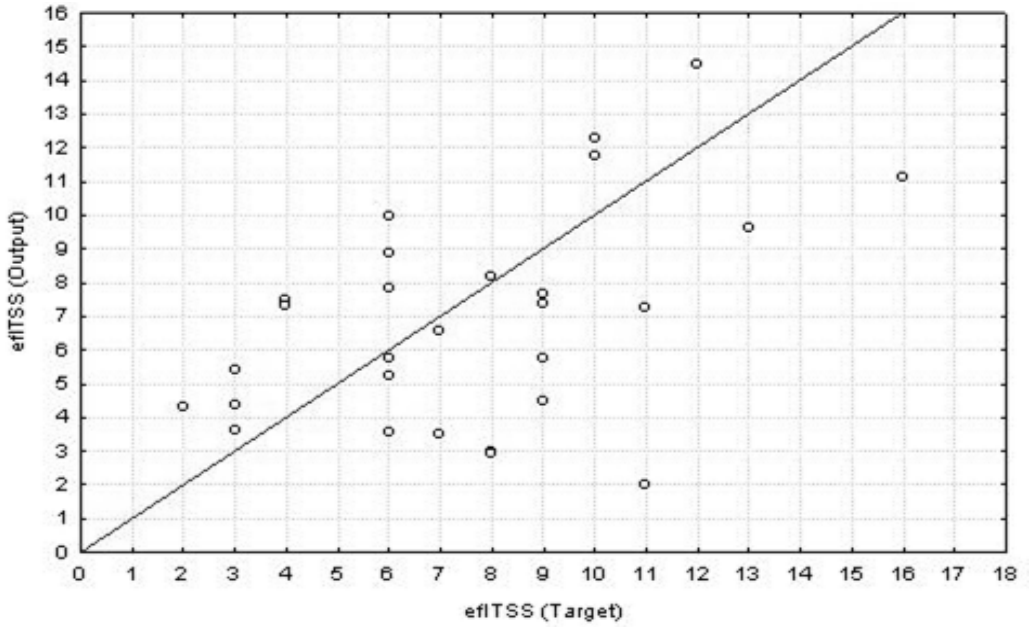



**Figure 4.9 Observed versus predicted values for validation samples – configuration 2**

### 4.6.3 Data set 3

The other approach is to interpolate the missing values in CBOD and TSS based on historical data and total influent to the plant (dataset3), which will change frequency of the data from 7 days to 1 day and more data points for the interval of 2 years (2008 and 2009).

After statistical analysis and modeling of this data set, the results of the weekly data will be compared to the results of daily and interpolated daily data to find out which one is more promising and yields better result.

Interpolation is applied to fill 655 missing values and the interpolation is not only based on historical data but also based on the amount of the total influent to the plant, other sections like feature and algorithm selection, sampling interval is the same as data set2.To fill the missing values, these equations are calculated and applied:

- InflCBOD (Y1) based on total influent (x), Y1=f(x0)
- InflTSS (Y2) based on total influent (x), Y2=f(x0)
- EflCBOD (Y3) based on inflCBOD (x1), Y3=f(x1)
- EflTSS (Y4) based on inflTSS (x2), Y4=f(x2)

To formulate CBOD in influent based on the total influent rate, time series of input and output of the model were applied in software Eureqa and equation for interpolation was produced shown in equation (4.1). Equation (4.2) is the formulation of TSS in influent based on total influent rate to the wastewater plant. Eureqa (pronounced "eureka") is a software tool for detecting equations and hidden mathematical relationships in data. Its primary goal is to identify the simplest mathematical formulas which could describe the underlying mechanisms that produced the data. The chosen equations have the least value for fitness function, most correlation coefficient, least linear residual and least mean and absolute errors.

$$\text{Y1}= f(x0) = 299.15387 + 21.633059 * \sin\left(\frac{376.6543}{x0}\right) - 11.803284 * \sin(0.83268148 * \text{x0}) - x0 \quad (4.1)$$

$$\text{Y2}= f(x0) = 283.78384 - 42.428005 * \sin(0.15061 * \text{x0} - 4.8503766) - \frac{283.78384}{(0.061204407 * x0)} \quad (4.2)$$

Formulation of chemicals in effluent is as below, when TSS in effluent was considered as output, first total influent was applied as variable in function but the results were not that promising, hence TSS amount in effluent in missing points was calculated based on the TSS in inflent, and the equation which was searched in Eureqa software is as below equation ( 4.4), the chosen one has the least fitness, most correlation coefficient, least linear residual and least mean and absolute errors.

$$\text{Y3}= f(x1) = 4.1960101 + \cos\left(2331550.3 + \frac{34.808167}{x1}\right) \quad (4.3)$$

Y4=

$$f(x2) = 4.2334962 + cos\,(0.04909274 * x2 - 2.6754057) + ((344.07095 + 344.07095 * cos\,(0.090452246 * x2 - 3.7199678) - x2 * cos\,(0.090452246 * x2 - 3.7199678)\,)) \,/\, x2 \quad (4.4)$$

Genetic Algorithm approach was applied to interpolate the missing values in TSS-infl, CBOD-infl, TSS-efl and CBOD-efl. Population size for all of them was 512 the rest of the evaluation metrics to generate the final equation are shown in the tables and figures below successively for each output.

**Table 4.10 Evaluation metrics of GA approach for TSS in influent**

| Index | Train | Validation |
|---|---|---|
| Sample size | 306 | 161 |
| Fitness | 0.85 | 0.85 |
| R-squared | 0.07 | 0 |
| Correlation coefficient | 0.35 | 0.38 |
| AIC | 2525.45 | 1351.57 |
| MSE | 3441.08 | 3079.35 |
| MAE | 49.47 | 46.43 |

**Table 4.11 Evaluation metrics of GA approach for CBOD in influent**

| Index | Train | Validation |
|---|---|---|
| Sample size | 288 | 150 |
| Fitness | 0.86 | 0.89 |
| R-squared | 0.04 | 0.01 |
| Correlation coefficient | 0.36 | 0.32 |
| AIC | 2384.71 | 1272.28 |
| MSE | 3433.67 | 3696.73 |
| MAE | 44.61 | 46.7 |

**Table 4.12 Evaluation metrics of GA approach for TSS in effluent**

| Index | Train | Validation |
|---|---|---|
| Sample size | 306 | 161 |
| Fitness | 0.79 | 0.68 |
| R-squared | 0.04 | 0.25 |
| Correlation coefficient | 0.33 | 0.52 |
| AIC | 378.06 | 148.13 |
| MSE | 3.29 | 2.22 |
| MAE | 1.28 | 1.09 |

**Table 4.13 Evaluation metrics of GA approach for CBOD in effluent**

| Index | Train | Validation |
|---|---|---|
| Sample size | 707 | 380 |
| Fitness | 0.91 | 0.92 |
| R-squared | -0.06 | -0.08 |
| Correlation coefficient | 0.21 | 0.18 |
| AIC | -2946.17 | -1565.85 |
| MSE | 0.01 | 0.01 |
| MAE | 0.1 | 0.1 |

The figures below consist of 4 parts (a) to (d), for all the Figures (4.10) to (4.13), part (a) shows “Observed vs. predicted plot”. Part (b) shows “Residual error histogram plot”. Part (c) shows “Accuracy vs. complexity plot of the best solutions”. Part (d) shows “The best solution error over the search time”.

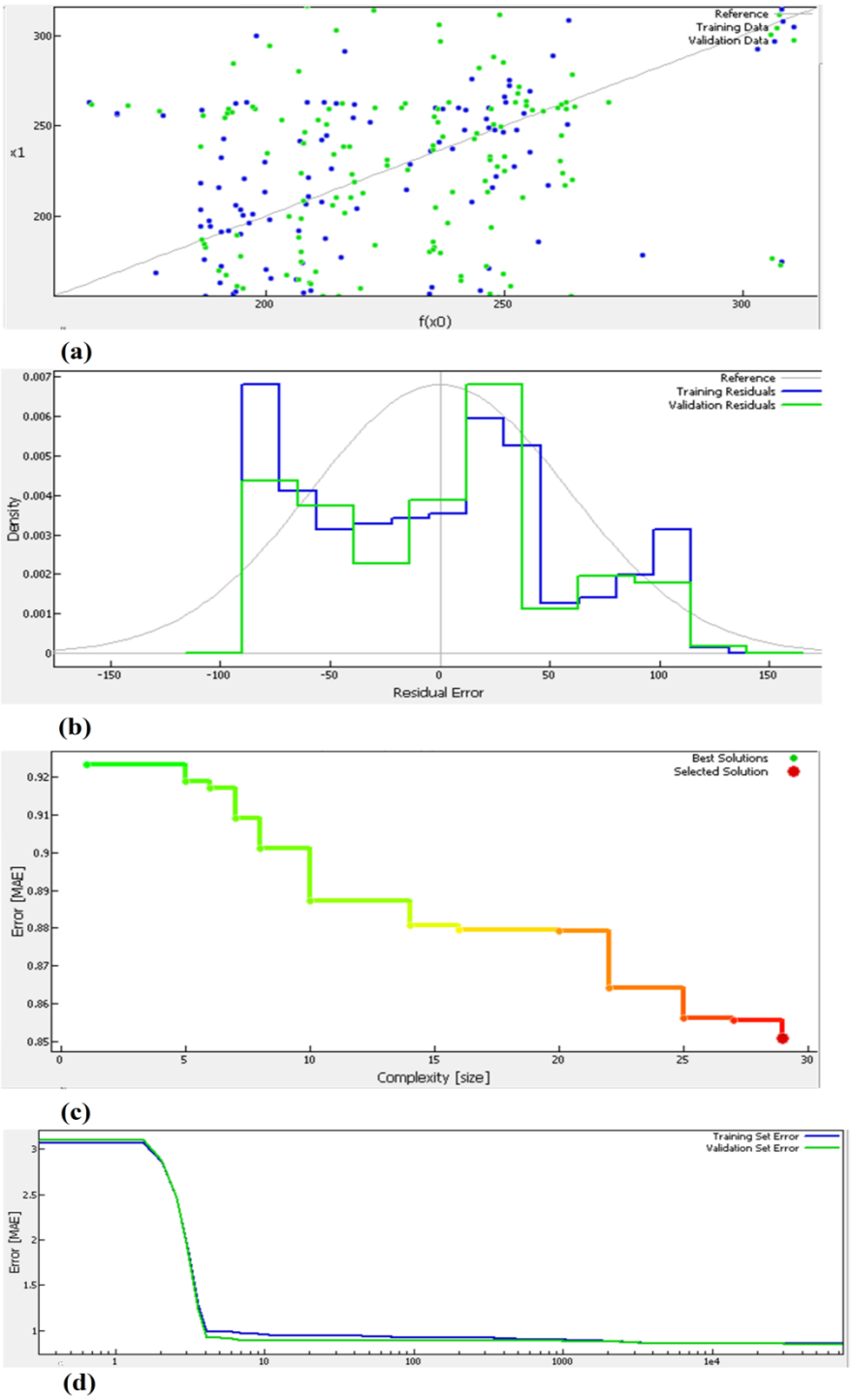



**Figure 4.10 GA approach for interpolation of TSS in influent based on total influent values**

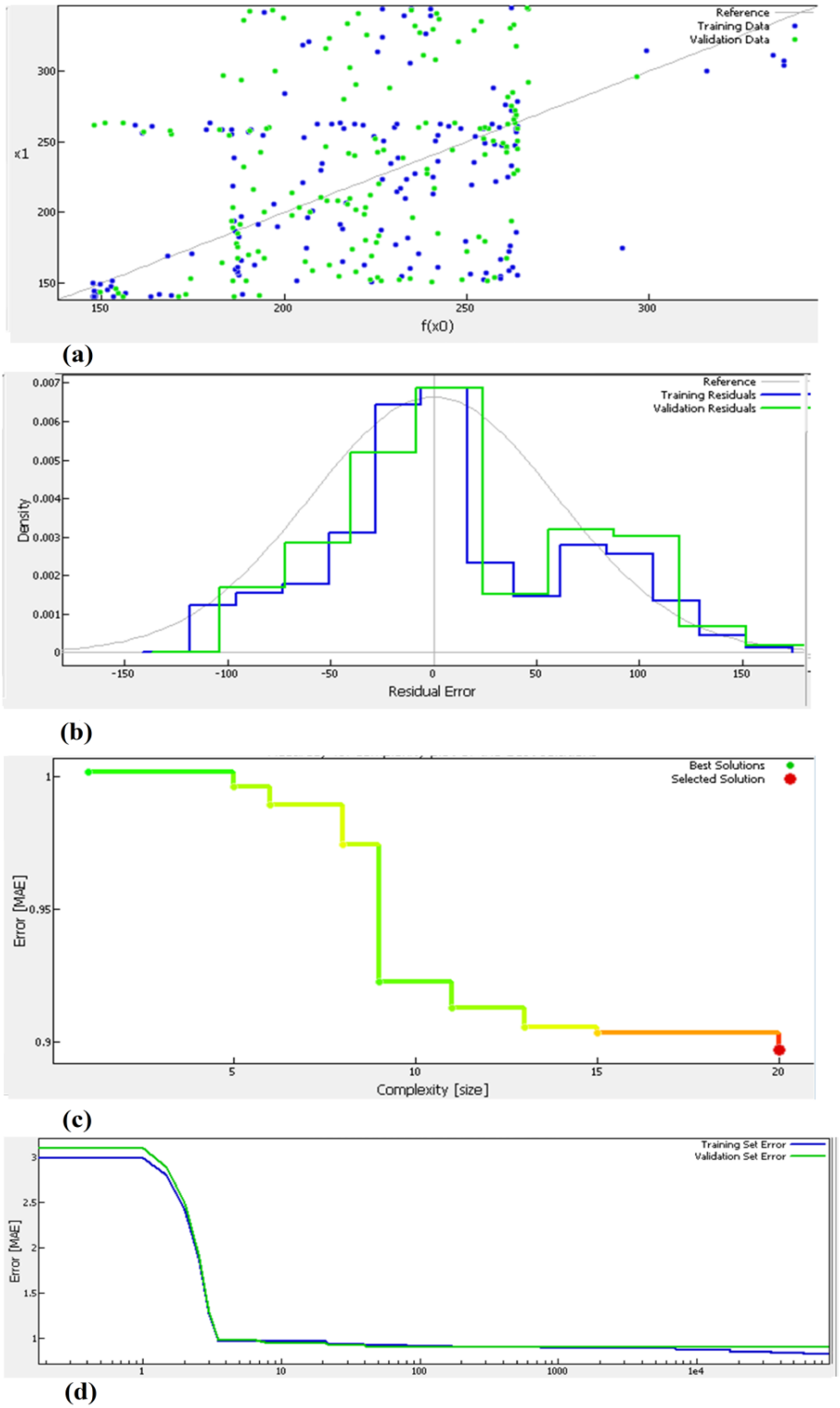



**Figure 4.11 GA approach for interpolation of CBOD in influent based on total influent values**

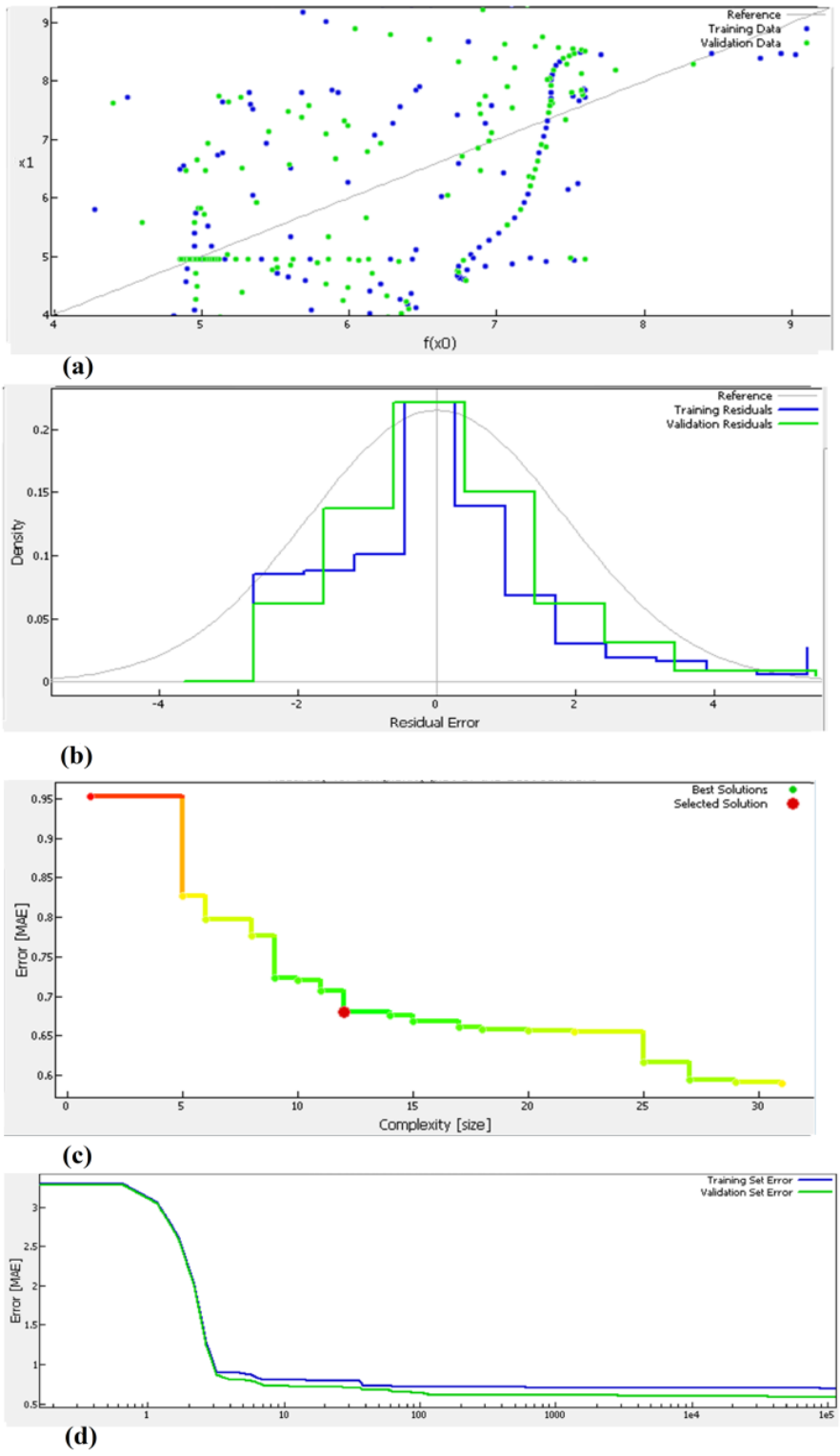
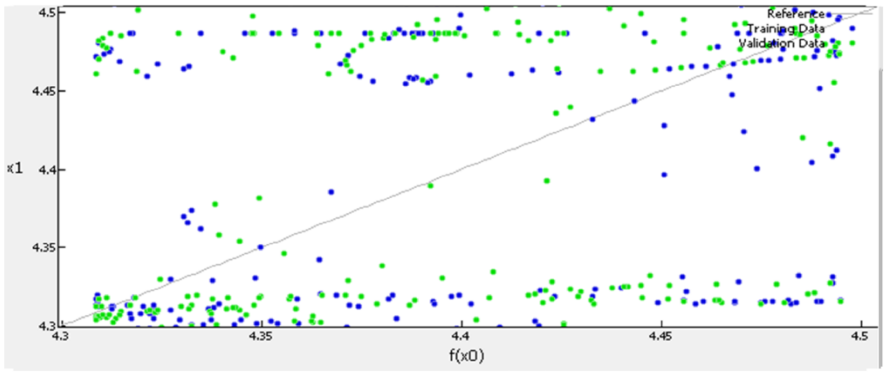
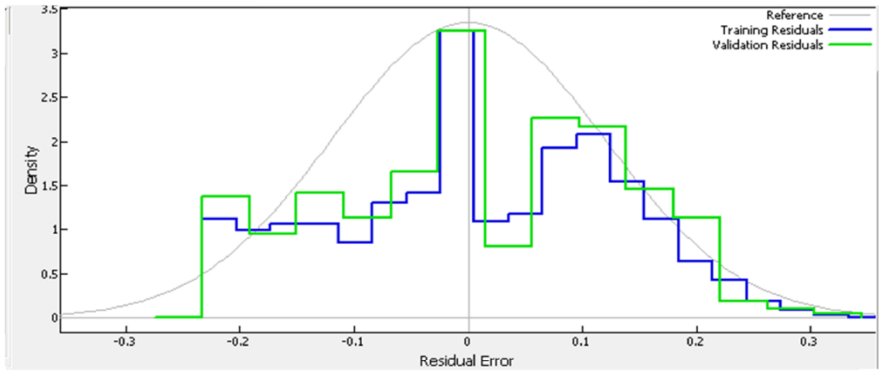


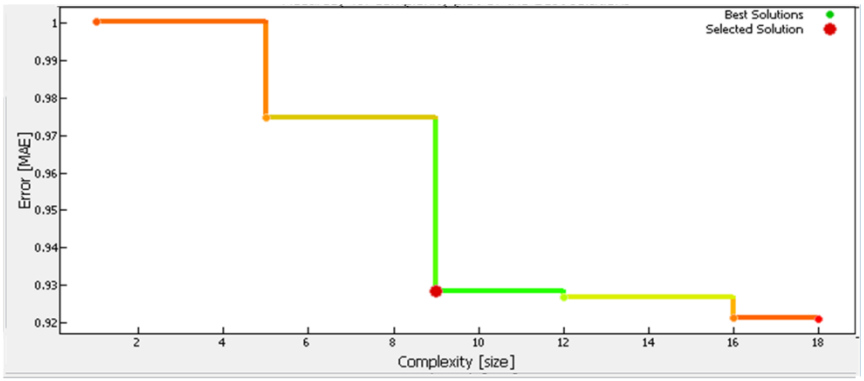

**Figure 4.12 GA approach for interpolation of TSS in effluent based on TSS in influent**



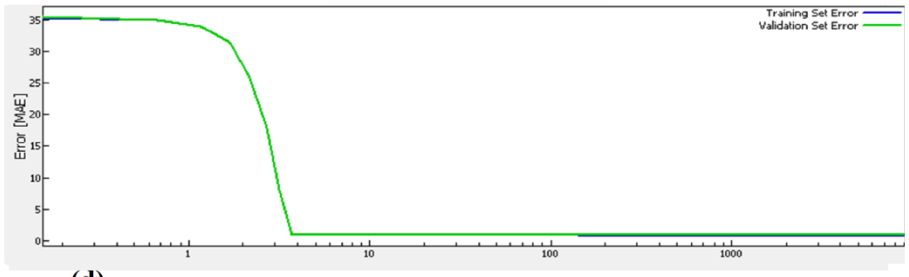

(a)




(b)




(c)




(d)

**Figure 4.13 GA approach for interpolation of CBOD in effluent based on CBOD in influent**

#### *4.6.3.1 Interpolated points in influent*

In this section interpolated points and available values are visualized in plots, red dots show interpolated while blue line is for available data.

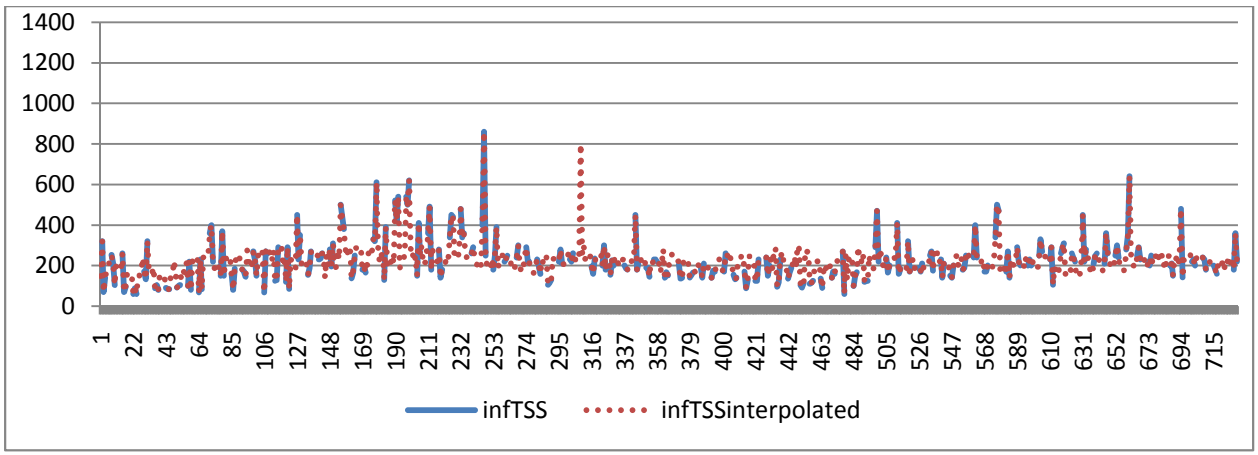



**Figure 4.14 Plot of times series of available values and interpolated data for TSS in influent.**

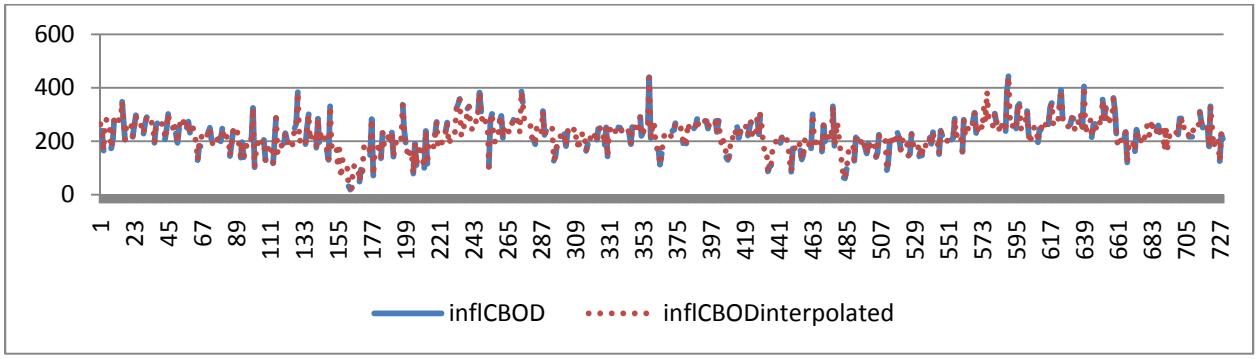



**Figure 4.15 Plot of times series of available values and interpolated data for CBOD in influent.**

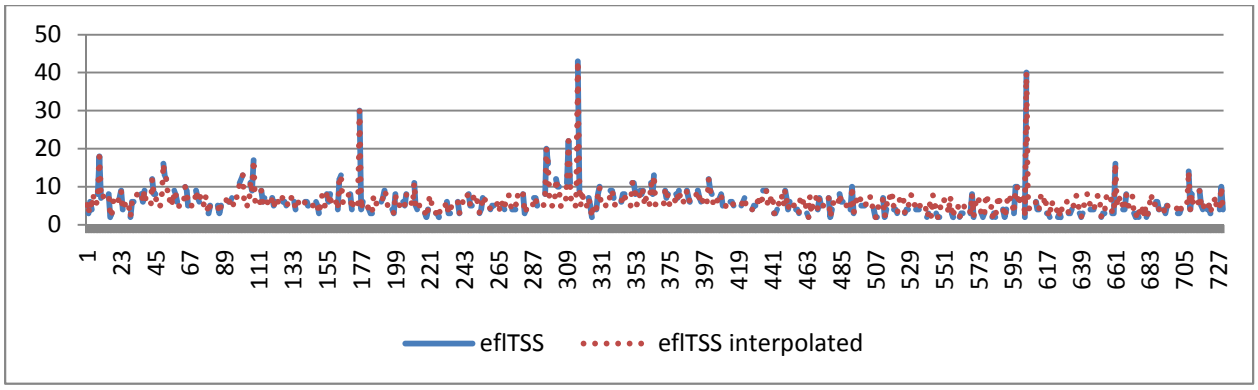



**Figure 4.16 Plot of times series of available values and interpolated data for TSS in effluent.**

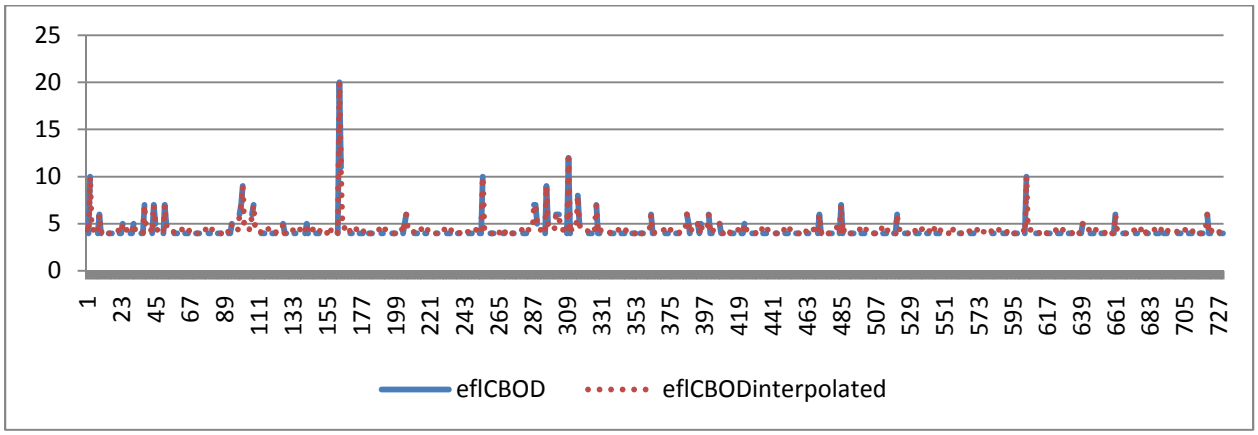



**Figure 4.17 Plot of times series of available values and interpolated data for CBOD in effluent.**

Two best networks are shown in Tables (4.10) and (4.11) and respectively for configuration 1 (output was CBOD in effluent) and configuration 2 (output was TSS in effluent); the results are based on the automated search in *Statistica*.

**Table 4.14 Best MLPs for configuration 1**

| Net. name | Test perf. | Validation perf. | Test error | Validation error | Hidden activation | Output activation |
|---|---|---|---|---|---|---|
| MLP 3-58-1 | 28.55 | 30.90 | 0.1767 | 0.8422 | Exponential | Identity |
| MLP 3-6-1 | 32.65 | 32.61 | 0.1716 | 0.9922 | Exponential | Identity |

**Table 4.15 Best MLPs for configuration 2**

| Net. name | Test perf. | Validation perf. | Test error | Validation error | Hidden activation | Output activation |
|---|---|---|---|---|---|---|
| MLP 3-68-1 | 36.90 | 19.37 | 2.4710 | 8.653 | Logistic | Identity |
| MLP 3-67-1 | 32.91 | 18.62 | 2.544 | 8.649 | Exponential | Tanh |

## 4.7 Conclusion

Modeling a wastewater plant is difficult to accomplish due to high level of complexity and nonlinearity of the plant and non-uniformity of the available data as well as the nature of the biological treatment. An NN modeling approach was implemented to solve the problem and discover the relation of input-output to be able to predict the behavior of the plant performance. It really involves a great degree of complexity and uncertainty. When CBOD was defined as output the result was so poor because of lab manipulation whereas using TSS as output yielded better results. Error was so low when using CBOD as output, the reason was that the mode and median value for CBOD was 4 mg/l and the algorithm could predict it very well since it was majority.

NN modeling technique has many advantages in modeling complex systems, simplicity, efficiency and generalization which was so useful in modeling a wastewater plant performance prediction model.

The other approach was to interpolate the missing values in CBOD and TSS based on historical data and total influent to the plant, which will increase frequency of the data from 7 days to 1 day and more data points for the interval of 3 years.

In this section the results of the weekly data will be compared to the results of daily interpolated data to find out which one is more promising and yields better result. Comparison of Network results for weekly data (data set1), daily data with missing values besides total influent (data set2) and filled daily data (data set3) are compared for two configurations mentioned before. The computational results are demonstrated in Tables (4.12) and (4.13).

**Table 4.16 Compared networks for three defined datasets for TSS concentration**

| Data set | Input | Output | Net. name | Test perf. | Validation perf. | Test error | Validation error | Frequency | Missing values |
|---|---|---|---|---|---|---|---|---|---|
| 1 | TSS, CBOD | TSS | RBF 2-40-1 | 19.96 | 46.42 | 5.66 | 2.64 | Weekly | No-averaged |
| 2 | Tss, CBOD, tot inflent | TSS | RBF 3-60-1 | 49.87 | 53.29 | 7.10 | 7.42 | Daily | Yes |
| 3 | Tss, CBOD, tot inflent | TSS | MLP 3-68-1 | 36.9 | 19.37 | 2.47 | 8.65 | Daily | No-interpolated |

**Table 4.17 Compared networks for three defined datasets for CBOD concentration**

| Data set | Input | Output | Net. name | Test perf. | Validation perf. | Frequency | Missing values |
|---|---|---|---|---|---|---|---|
| 1 | TSS, CBOD | CBOD | RBF 2-40-1 | 25.6 | 10 | Weekly | No-averaged |
| 2 | Tss, CBOD, tot inflent | CBOD | RBF 3-50-1 | 27.72 | 50.16 | Daily | Yes |
| 3 | Tss, CBOD, tot inflent | CBOD | MLP 3-6-1 | 32.65 | 32.61 | Daily | No-interpolated |

## 4.8 Discussion

Nonlinear interpolation (curve fitting with noisy data) is not that meaningful while we ignore some crucial questions about data set and make some assumptions like data can be assumed to be continuous, smooth, possibly periodic, so it is subject to uncertainty.

It is suggested to model daily regular measurements for CBOD and TSS without laboratory filtering to be able to measure performance more accurately.

# CHAPTER 5. CONCLUSION

This Thesis explores some practical applications of data mining techniques and heuristic search methods by using concepts in hydrology in the field of Wastewater Plant Process. Data sets considered for study included water quality parameters, influent rate, radar reflectivity, and tipping bucket. Statistical analysis, in particular correlation-based analysis, was used for the selection of input parameters for the modeling challenges tackles throughout this work.

Chapter 1 provided background information and a literature review of past applications of data mining in hydrology, as well as an introduction to the multilayer perceptron (MLP), decision tree (DT) which was extensively applied throughout this Thesis.

The Second Chapter proved data mining and the DTs competence at making a prediction at a different spatial location. In this data driven model, rainfall at a downstream location was predicted with reflectivity, velocity and spectrum width data from other tipping bucket locations, as well as the rainfall data from all other tipping buckets in surrounding area. Rainfall is a particularly difficult water quantity parameter to predict due to its erratic and fluctuating behavior and its tendency to zero value which caused class imbalance problem and erroneous recording in rain gauges. The DT model derived in this chapter makes a rainfall prediction at a gauge up to 120 minutes with accuracy of 94.21 %. The model's robustness is analyzed as it is tested outside of its training domain, at six other locations along the wastewater plant. The results show that the model is highly robust. The radius that model could maintain while tested on other rain gauges was 33.13 km with the accuracy above 95%.

Chapter 3 combines tipping bucket data and Next Generation Radar data to build a predictive model of the total influent rate to the wastewater plant, by way of MLP and

also DT. For classification model prediction could be done up to 1 hour successfully and for regression model accuracy decrease to about 86 % up to 3 hours.

Chapter 4 considered multiple water quality parameters, and provided a methodology toward a very practical use of data mining; data gap filling. One method for filling missing data was presented, called non-linear interpolation, considered complimentary water quality parameters, to predict carbonaceous biochemical oxygen demand and total suspended solids in influent rate. The other water quality parameters in effluent were measured concurrently with influent concentrations. This method may be useful at a location that missing values do not outnumber available values. The methodology introduced utilizes time series data mining, or the use of historical data. The method was used to model the current CBOD, TSS concentrations in influent, and also to make a short term forecast. The behavior of the model is analyzed when making longer term forecast, as well. The steps in chapter 1 to 4 are summarized in Figure (5.1); the flowchart demonstrates the main goal of this Thesis and shows the relation of each forecast model to the other one.

Future research will focus primarily on using radar data in other predictive models. Upon research there appeared to be many other areas for analysis and model improvement, like using reflectivity in solar energy predictive model in solar plant. Also, the apparent usefulness of such high spatiotemporal resolution precipitation data to hydrological models, namely, flood forecasting models, makes this an exciting area of research. Some other topics that will be studied in the future are; further robustness testing, such as testing the rainfall prediction models' performance in other regions farther away with different terrain properties, and the rainfall prediction models' dependency on nearby actual tipping buckets for accurate longer predictions. Also more accurate method to fill the missing values of flow rate chemicals.

In Figure (5.1) whole process of this thesis is demonstrates in the flow chart.

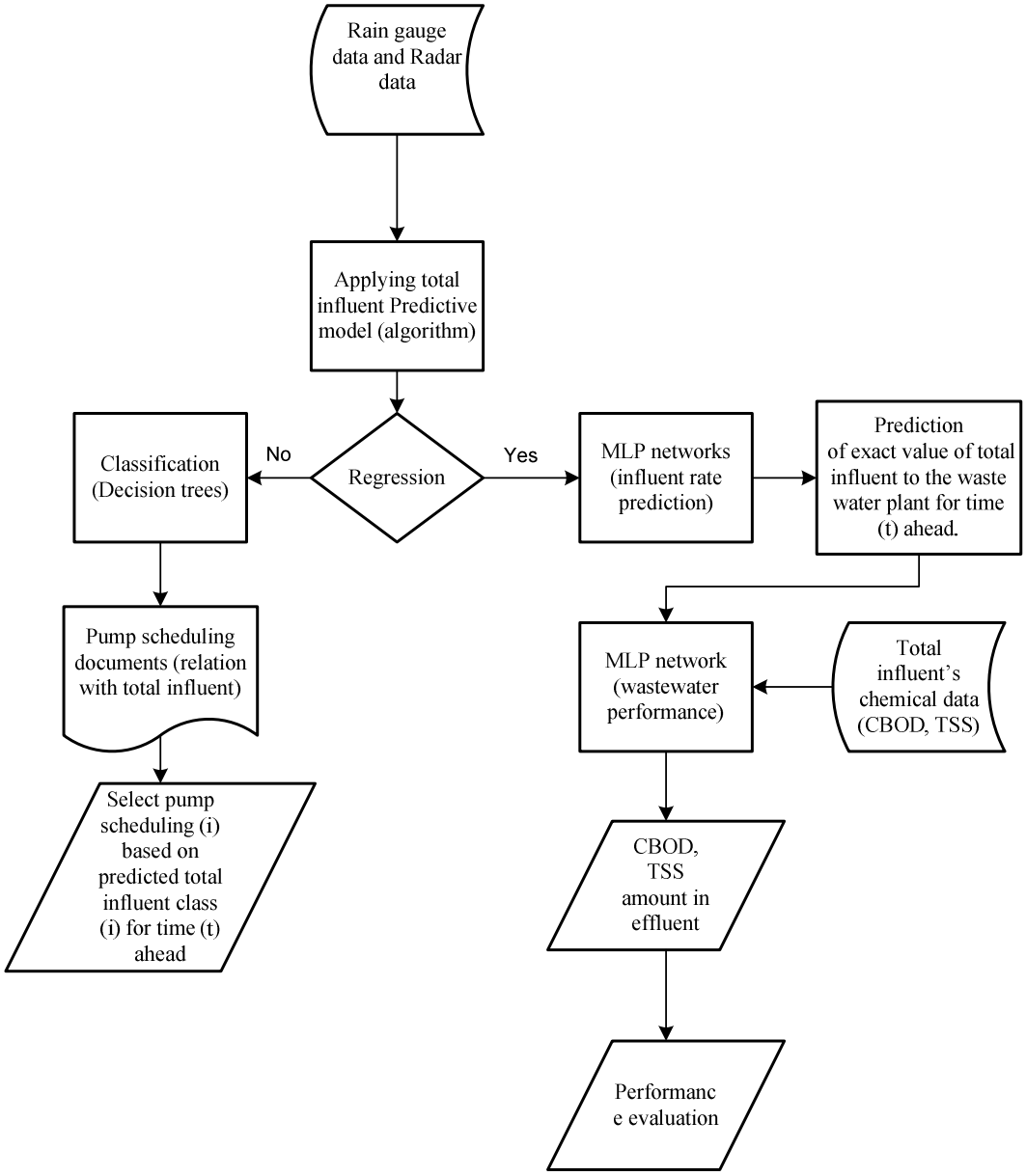



**Figure 5.1 Thesis summary**

# REFERENCES


[1] G. Tiron and S. Gosav, "*The July 2008 rainfall estimation from barnova wsr-98 d radar using artificial neural network*", Romanian Reports in Physics, Vol. 62, pp. 305–313, 2009.

[2] R. Teschl, W. Randeu, and F. Teschl, "*Improving weather radar estimates of rainfall using feed-forward neural networks*", Neural Networks, Vol. 20, pp. 519–527, 2007.

[3] T.B. Trafalis, M.B. Richman, A. White, and B. Santosa, "*Data mining techniques for improved WSR-88D rainfall estimation*" , Computers & Industrial Engineering, Vol. 33, pp. 775–786, 2002.

[4] R. Chattamvelli, "*Data Mining Methods*", Alpha Science International Ltd., Oxford, U.K., 2009.

[5] T. Masters, "*Advanced Algorithms for Neural Networks*", A C++ Sourcebook. John Wiley and Sons, New York, 1993.

[6] F. Rosenblatt, "*The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain*" Psychological Review, Vol. 65, No. 6, pp. 386-308, 1958.

[7] L. Tarassenko, "*A Guide to Neural Computing Applications*", Arnold Publishers, London, UK, 1998.

[8] R. Hecht-Nielsen, "*Kolmogorov's mapping neural network existence theorem*" Proceedings of 1st IEEE International Joint Conference of Neural Networks, Institute of Electrical and Electronics Engineers, New York, NY, pp. 11-13, 1987.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "*The WEKA data mining software: an update*" SIGKDD Explorations, Vol. 11, No. 1, pp. 10-18, 2009.

[10] C Apte, S Weiss, "*Data mining with decision trees and decision rules*", Future Generation Computer Systems, Vol.13, pp 197-210, 1997.

[11] V.E. Baer, "*The Transition from the Present Radar Dissemination System to the NEXRAD Information Dissemination Service (NIDS)*" American Meteorological Society Bulletin, Vol. 72, No. 1, pp. 29-33, 1991.

[12] Chandrasekar b and L. Baldini, "*Rainfall estimation from X-band dual polarization radar using reflectivity and differential reflectivity*", Atmospheric Research, Vol. 82, pp 164-172, 2006.

[13] V. Chandrasekar , G Scarchilli and S Bolen, "*Variation of mean raindrop shape derived from polarimetric radar measurements*", Atmospheric Research, Vol. 59-60, pp. 283-293, 2001.

[14] R.A. Fulton, j p. Breidenbach, DJ Seo and D A. Miller, "*The WSR-88D Rainfall Algorithm*", Weather and Forecasting, Vol. 13, pp. 377-395, 1998.

[15] T Bellerby, M Todd, D Kniveton, C Kidd, "*Rainfall estimation from a combination of TRMM precipitation radar and GOES multispectral satellite imagery through the use of an Artificial Neural Network*", Journal of Applied Meteorology and Climatology, Vol. 39, pp. 2115-2128, 2000.

[16] H. Sauvageot, "*Rainfall measurement by radar: a review*", Atmospheric Research, Vol. 35, pp. 27-54, 1994.

[17] Y.M Chiang, F.J. Chang, B. Jong-Dao Jou, P.F Lin, "*Dynamic ANN for precipitation estimation and forecasting from radar observations*", Journal of hydrology, Vol. 334, pp250-261, 2000.

[18] A. Kusiak, H Zheng, Z Song, "*Wind farm power prediction: A Data-Ming Approach*", Wind Energ., Vol. 12, pp. 275- 293.

[19] S Suman, K Laddhad, U Deshmukh, "*Methods For Handling Highly Skewed Datasets*", 2005.

[20] J.A. Smith, and W.F. Krajewski, "*A modeling study of rainfall rate-reflectivity relationships*," Water Resour. Res., Vol. 29, pp. 2505–2513, 1993.

[21] L.J. Battan, "*Radar observation of the atmosphere, III*." Univ. of Chicago Press, Chicago, 1973.

[22] P.M Austin, "*Relation Between Measured Radar Reflectivity and Surface Rainfall*" Mon. Weather Rev., Vol. 115, pp. 1053–1070, 1987.

[23] J. B. Sérodes, M. J. Rodriguez, and A. Ponton, "*Chlorcast: a methodology for developing decision-making tools for chlorine disinfection control*" Environmental Modelling & Software, Vol. 16, No. 1, pp. 53-62, 2000.

[24] C. Damle and A. Yalcin, "*Flood prediction using time series data mining*" Journal of Hydrology, Vol. 333, No. 2-4, pp. 305-316, 2007.

[25] H. Liu, V. Chandrasekar, and G. Xu, "*An adaptive neural network scheme for radar rainfall estimation from WSR-88D Observations*" Journal of Applied Meteorology, Vol. 30, pp. 2038-2050, 2001.

[26] Ling, C., & Li, C. "*Data Mining for Direct Marketing Problems and Solutions*". In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press, 1998.

[27] Domingos, P." *Metacost: A General Method for Making Classifiers Cost-sensitive*". In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA. ACM Press, pp. 155–164, 1999.

[28] Japkowicz, N. "*The Class Imbalance Problem: Significance and Strategies*". In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning Las Vegas, Nevada, 2000.

[29] N V. Chawla, K W. Bowyer, L O. Hall, W. P Kegelmeyer, "*SMOTE: Synthetic Minority Over-sampling Technique*", Artificial Intelligence Research, Vol. 16, pp. 321–357, 2002.

[30] Kubat, M., and Matwin, S. "*Addressing the Course of Imbalanced Training Sets: One-sided Selection*", ICML, Vol. 97, pp. 179-186, 1997.

[31] Chawla, N.V., Bowyer,K.W., Hall, L.O., Kegelmeyer W.P." *SMOTE: Synthetic Minority Over-Sampling Technique*" Journal of Artificial Intelligence Research Vol. 16, pp. 321-357, 2002.

[32] Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K. "*SMOTEBoost: Improving prediction of the Minority Class in Boosting*", 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat Dubrovnik, Croatia, pp. 107-119, 2003.

[33] Andrew Estabrooks, Taeho Jo and Nathalie Japkowicz, "*A Multiple Resampling Method for Learning from Imbalanced Data Sets*" Computational Intelligence, Vol. 20, pp. 18-36, 2004.

[34] Gustavo, E.A., Batista, P.A., Ronaldo, C., Prati, Maria Carolina Monard "*A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*". SIGKDD Explorations 6, Vol. 1, 2004, pp. 20-29.

[35] H Han, W-Y Wang1, B-H Mao, "*Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*".

[36] P.N Tan, M. Steinback, and V. Kumar, "*Introduction to Data Mining, Pearson Education/Addison-Wesley*", Reading, Mass, 2006.

[37] Y.M Chiang, F.J. Chang, B. Jong-Dao Jou, P.F Lin, "*Dynamic ANN for precipitation estimation and forecasting from radar observations*", Journal of hydrology, Vol. 334, pp. 250-261, 2007.

[38] H. Sauvageot, "*Rainfall measurement by radar: a review*", Atmospheric Research, Vol. 35, pp. 27-54, 1994.

[39] A. Kusiak, H. Zheng, and Z Song, "*Wind farm power prediction: a data mining approach*".

[40] T. Kudo, and Y. Matsumoto. "*A Boosting Algorithm for Classification of Semi-Structured Text*" EMNLP, 2004.

[41] Farouq S. Mjalli_, S. Al-Asheh1, H.E. Alfadala, "*Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance*", Journal of Environmental Management, Vol. 83, pp. 329–338, 2007.

[42] Chen, J. C., Chang, N. B., & Shieh, W. K. "*Assessing wastewater reclamation potential by neural network model*". Engineering Applications of Artificial Intelligence, Vol.16, pp. 149–157, 2003.

[43] Belanche, L., Valde´s, J. J., Comas, J., Roda, I. R., & Poch, M. "*Prediction of the bulking phenomenon in wastewater treatment plants*", Artificial Intelligence in Engineering, Vol. 14, pp. 307–317, 2000.

[44] Shetty, R. G., & Chellam, S. "*Predicting membrane fouling during municipal drinking water nanofiltration using artificial neural networks*". Journal of Membrane Science, Vol. 217, pp. 69–86, 2003.

[45] Hamed, M. M., Khalafallah,M. G., & Hassanien, E. A. "*Prediction of wastewater treatment plant performance using artificial neural networks*", Environmental Modelling and Software, Vol. 19,pp. 919–928, 2004.

[46] Maier, H. R., Morgan, N., & Chow, C. W. K. "*Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters*". Environmental Modeling and Software, Vol. 19, pp. 485–494, 2004.

[47] Zhu, J., Zurcher, J., Rao, M., Meng, M.Q-H., "*An on-line wastewater quality prediction system based on a time-delay neural network*", Engineering Application of Artificial Intelligence, Vol. 11, pp. 747–758, 1998.

[48] A Taebia, R L. Droste, "*Performance of an overland flow system for advanced treatment of wastewater plant effluent*", Journal of Environmental Management, Vol. 88, pp. 688–696, 2008.

[49] O¨zer Cs inar, "*New tool for evaluation of performance of wastewater treatment plant: Artificial neural network*", Process Biochemistry, Vol.40, pp. 2980–2984, 2005.

[50] Christos S. Akratos, John N.E. Papaspyros, Vassilios A. Tsihrintzis, "*An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands*", Chemical Engineering Journal, Vol. 143, pp. 96–110.

[51] Karla Patricia Oliveira-Esquerrea, Dale E. Seborgb, Milton Moria, Roy Edward Bruns, "*Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part II. Nonlinear approaches*", Chemical Engineering Journal, Vol. 105, pp.61–69, 2004.

[52] K. Hornik, M. Stinchcombe, H. White, "*Multilayer feedforward networks are universal approximators*", Neural Networks, Vol. 2, pp. 359–366, 1989.

[53] M. Cote, B.P.A. Grandijean, P. Lessard, J. Yhibault, "*Dynamic modeling of the activated sludge process: improving prediction using neural networks*", Water Res. Vol. 29, pp. 995–1004, 1995.

[54] H. Zhao, O.I. Hao, A.S.C.E. Fellow, T.J. McAvoy, C.H. Chang, "*Modeling nutrient dynamics in sequencing batch reactor*", J. Environ. Eng. Vol. 123, pp. 311–319, 1997.

[55] S. Chen, S.A. Billings, "*Neural networks for nonlinear dynamic system modeling and identification*", Int. J. Contr. Vol. 56, pp. 319– 346, 1992.

[56] G.M. Tillman, *Primary Treatment at Wastewater Treatment Plants*. Chelsea: Lewis Publishers, Inc, 1991.

[57] P.A. Vesilind, "*Wastewater Treatment Plant Design"*. Alexandria: International Water Association Publishing, 2003.

[58] J.R. Kim, J.H. Ko, J.H. Im, S.H. Lee, S.H. Kim, C.W. Kim, and T.J. Park, "*Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model*," Water Science & Technology, vol. 53, no. 4, pp. 185-192, 2006.

[59] G.E. Kurz, B. Ward, and G.A. Ballard, "*Simple method for estimating I/I using treatment plant flow monitoring reports – a self help tool for*

*operators*," Proceedings of the Water Environment Federation, Collection systems, pp. 568-576(9), 2009.

[60] P. Young, and S. Wallis, "*Recursive estimation: a unified approach to the identification estimation, and forecasting of hydrological systems*" Applied Mathematics and Computation, vol. 17, no. 4, pp. 299-334, 1985.

[61] Z. Vojinovic, V. Kecman, and V. Babovic, "*Hybrid approach for modeling wet weather wesponse in wastewater systems*," Journal of Water Resources Planning and Management, vol. 129, no.6, pp. 511-521, 2003.

[62] P.C. Tan, C.S. Berger, K.P. Dabke, and R.G. Mein, "*Recursive identification and adaptive prediction of wastewater flows*," Automatica, vol. 27, no.5, pp. 761-768, 1991.

[63] J. Lindqvist, T. Wik, D. Lumley, and G. Aijala, "*Influent load prediction using low order adaptive modeling*," In 2nd IWA Conference on Instrumentation, Control and Automation, Busan, South Korea, 2005.

[64] I.H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.