ABSTRACT

Water is one of the most important natural resources for all living organisms on earth. The monitoring of treated wastewater discharge quality is vitally important for the stability and protection of the ecosystem. Collecting and analyzing water samples in the laboratory consumes much time and resources. In the last decade, many machine learning techniques, like multivariate linear regression (MLR) and artificial neural network (ANN) model, have been proposed to address the problem. However, simple linear regression analysis cannot accurately forecast water quality because of complicated linear and nonlinear relationships in the water quality dataset. The ANN model also has shortcomings though it can accurately predict water quality in some scenarios. For example, ANN models are unable to formulate the non-linear relationship hidden in the dataset when the input parameters are ambiguous, which is common in water quality dataset.

The adaptive neuro-fuzzy inference system (ANFIS) has been proven to be an effective tool in formulating the complicated linear and non-linear relationship hidden in datasets. Although the ANFIS model can achieve good performance in the water quality prediction, it has some limitations. Firstly, the size of the training dataset should not be less than the number of training parameters required in the model. Secondly, when the data distribution in the testing dataset is not reflected in the training dataset, the ANFIS model may generate out-of-range errors. Lastly, a strong correlation is required between input and target parameters. If the correlation is weak, the ANFIS model cannot accurately formulate the hidden relationship.

In this dissertation, several methods have been proposed to improve the performance of ANFIS-based water quality prediction models. Stratified sampling is employed to cover different kinds of data distribution in the training and testing datasets. The wavelet denoising technique is

used to remove the noise hidden in the dataset. A deep prediction performance comparison between MLR, ANN, and ANFIS model is presented after stratified sampling and wavelet denoising techniques are applied. Because water quality data can be thought as a time series dataset, a time series analysis method is integrated with the ANFIS model to improve prediction performance. Lastly, intelligence algorithms are used to optimize the parameters of membership functions in the ANFIS model to promote the prediction accuracy. Experiments based on water quality datasets collected from Las Vegas Wash since 2007 and Boulder Basin of Lake Mead, Nevada, between 2011 and 2016 are used to evaluate the proposed models.

ACKNOWLEDGEMENT

There are many people have helped me a lot during my Ph. D. study. Firstly, I want to thank my advisor, Dr. Mei Yang, she has given me much advice and offered me as many opportunities as she could in the past five more years. She has also encouraged me when the research did not go well and guided me through the paper writing. I also would like to express my gratitude to my dissertation committee including Dr. Jacimaria Batista, Dr. Yingtao Jiang, and Dr. Shahram Latifi. They helped me a lot when I started the research topic and offered a lot of advice and assistance. Last but not least, I would like to thank my beloved family and friends for their continuous support.

TABLE OF CONTENTS ABSTRACT iii

ACKNOWLEDGEMENT v

TABLE OF CONTENTS vi

LIST OF TABLES ix

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1

* 1. [Background 1](#_TOC_250022)
  2. [Motivation 2](#_TOC_250021)
  3. [Contribution 4](#_TOC_250020)
  4. [Outline of the Dissertation 6](#_TOC_250019)

CHAPTER 2 LITERATURE REVIEW 7

* 1. [MLR 7](#_TOC_250018)
  2. [ANN 8](#_TOC_250017)
  3. [ANFIS 9](#_TOC_250016)
  4. [FTS 11](#_TOC_250015)

CHAPTER 3 STUDY AREA AND METHODS 13

* 1. [Study Area 13](#_TOC_250014)
  2. [Stratified Sampling 14](#_TOC_250013)
  3. [Adaptive Neuro-Fuzzy Inference System 15](#_TOC_250012)
  4. [Wavelet Transform 17](#_TOC_250011)
  5. [Evaluation Metrics 18](#_TOC_250010)

CHAPTER 4 WASTEWATER DISCHARGE QUALITY PREDICTION USING STRATIFIED SAMPLING AND WAVELETDE-NOISING ANFIS MODEL 20

* 1. [Water Quality Parameters 20](#_TOC_250009)
  2. [Wastewater quality predication System 22](#_TOC_250008)
     1. System Overview 22
     2. Network Structure Selection 23
     3. Input Parameter Selection 24
     4. Water Quality Prediction Model 26
  3. [Experimental Results and Discussion 28](#_TOC_250007)
     1. Wastewater Quality Dataset and Experiment Configuration 28
     2. Network Structure Selection and Input Parameter Selection 30
     3. Prediction Performance 31

CHAPTER 5 USING FUZZY MODELS AND TIME SERIES ANALYSIS TO PREDICT WATER QUALITY 36

* 1. [Materials and Methods 36](#_TOC_250006)
     1. Water Quality Parameters 36
     2. Input Parameter Selection 37
     3. Fuzzy Time Series 40
  2. [Results and Discussion 42](#_TOC_250005)
     1. Water Quality Dataset 42
     2. Experimental Configuration 43
     3. Performance Evaluation 43

CHAPTER 6 WATER QUALITY PREDICTION BASED ON THE INTEGRATION OFANFIS MODEL WITH INTELLIGENCE ALGORITHMS 49

* 1. [Materials and Methods 49](#_TOC_250004)
     1. Study Area 49
     2. Water Quality Parameters 49
     3. Genetic Algorithm 51
     4. Particle Swarm Optimization 52
  2. [System Overview 54](#_TOC_250003)
  3. [Experimental Results and Discussion 57](#_TOC_250002)
     1. Water Quality Dataset 57
     2. Experiment Configuration 58
     3. Performance Evaluation 59

CHAPTER 7 CONCLUSIONS AND FUTURE WORK 63

* 1. [Conclusions 63](#_TOC_250001)
  2. [Future Work 64](#_TOC_250000)

APPENDIX 66

REFERENCES 74

CURRICULUM VITAE 81

LIST OF TABLES

Table 1. Statistical measures of water quality parameters at LW3.4 and LW3.7 in the Las Vegas Wash 21

Table 2. Relationship between the number of membership functions and the number of parameters inside the ANFIS model 24

Table 3. Correlation matrix of all wastewater quality parameters listed in Table 1 30

Table 4. The training and testing performance of different models for parameter chloride in the LVW 32

Table 5. The training and testing performance of different models for parameter sulfate in the LVW 33

Table 6. The training and testing performance of different models for parameter EC in the LVW

....................................................................................................................................................... 34

Table 7. The training and testing performance of different models for parameter TDS in the LVW 34

Table 8. A statistical measure of water quality parameters at LW3.4 37

Table 9. A statistical measure of water quality parameters at BB 37

Table 10. Pearson Correlation between water quality parameters in LVW 38

Table 11. Pearson Correlation between water quality parameters in BB 38

Table 12. Pearson Correlation of water quality parameters in LVW with time series analysis 39

Table 13. Pearson Correlation of water quality parameters in BB with time series analysis 40

Table 14. The training and testing performance of different models for parameter EC in LVW 45 Table 15. The training and testing performance of different models for parameter DO in LVW 45

Table 16. The training and testing performance of different models for parameter TDS in LVW

....................................................................................................................................................... 45

Table 17. The training and testing performance of different models for parameter EC in BB 45

Table 18. The training and testing performance of different models for parameter DO in BB 46

Table 19. A statistical measure of water quality parameters at LW3.4 and LW3.7 in the Las Vegas Wash 50

Table 20. Correlation matrix of all water quality parameters listed in Table 19 51

Table 21. The training and testing performance of different models for parameter DO in LVW 59 Table 22. The training and testing performance of different models for parameter TDS in LVW

....................................................................................................................................................... 60

LIST OF FIGURES

Figure 1. The geographical distribution of the six water quality monitoring stations in the Las Vegas Wash 14

Figure 2. Location of Boulder Basin water quality monitoring station 14

Figure 3. A two-input ANFIS for first-order Sugeno fuzzy model with two rules 17

Figure 4. Wastewater quality prediction system based on wavelet de-noised ANFIS model using stratified sampling 23

Figure 5. Organization of water quality prediction model for parameter EC 27

Figure 6. The observed and predicted value of each prediction parameter: (a) Chloride; (b) Sulfate; (c) EC; (d) TDS 35

Figure 7. The observed value of five parameters in LVW between 2005 and 2010 42

Figure 8. The observed value of four parameters in BB between 2011 and 2016 43

Figure 9. The coefficient of determination value of ANFIS-TS and FTS model in the testing dataset of parameter EC, DO and TDS from LVW 47

Figure 10. The coefficient of determination value of ANFIS-TS and FTS model in testing dataset of parameter EC and DO from BB 48

Figure 11. Steps of hybrid GA and PSO with ANFIS model 56

Figure 12. The value of parameter TDS from the six water quality monitoring stations in LVW since 2007 58

Figure 13. The value of parameter DO from the six water quality monitoring stations in LVW since 2007 58

Figure 14. The testing result of parameter DO based on ANFIS-PSO model 61

Figure 15. The testing result of parameter TDS based on ANFIS-PSO model 62

CHAPTER 1 INTRODUCTION

# Background

To protect the environment and human health, treated wastewater discharge must be sampled and monitored in most developed countries to assure discharge permits are met [1]. In the past, scientists had to collect and analyze a large number of wastewater samples to understand how wastewater discharges components impacted the environment. The collection and analysis of treated wastewater effluents is time-consuming and costly. Machine learning methods are proposed to address the problem. The usage of machine learning methods would result in a reduction in sampling frequency and minimization of costs associated with analysis. At first, deterministic models and multivariate linear regression (MLR) analysis were used to speed up the process of evaluating the quality of wastewater effluent discharges [2] [3]. As a water quality dataset can be considered as a time series dataset, which is likely to have a complicated nonlinear relationship, the performance of deterministic and MLR models is expected to be poor.

In the past decade, many machine learning techniques have been proposed to address the problem. Artificial neural networks (ANN) are adopted to explore the non-linear relationships residing in water quality datasets [3][4]. Various ANN models have been designed to predict water and wastewater discharge quality based on previous existing datasets. A comprehensive comparison between ANN and MLR models for oxygen demand prediction has been performed [3]. The experimental results show that a three-layer neural network model outperforms an MLR model. In [4], neural network models are used to predict four parameters in the Qiantang River and the proposed model has higher accuracy and better stability in the experiment.

Although ANN models can effectively improve the prediction accuracy of water quality parameters, shortcomings still exist. Especially in some scenarios where the input parameters are ambiguous, neural networks struggle to formulate a non-linear relationship. Many studies have proven that an adaptive neuro-fuzzy inference system (ANFIS), which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario [5]. The experimental results in [6] show that an ANFIS model worked much better than an ANN model in predicting dissolved oxygen, even though there were only 45 data samples available. The experimental results confirm that the proposed method works. The ANFIS model has also been applied in effluent quality prediction, and an experiment with a dataset of around 150 data samples has proven that the ANFIS model is better than the ANN model [7].

Although the ANFIS model can achieve good performance in the water quality prediction, it has some limitations. Firstly, the size of training dataset should not be less than the number of training parameters required in the model. Secondly, when the data distribution in the testing dataset is not reflected in the training dataset, the ANFIS model may generate out-of-range errors. Lastly, a strong correlation is required between input and target parameters. If the correlation is weak, the ANFIS model cannot accurately formulate the hidden relationship.

On the other hand, the fuzzy time series (FTS) model is an accurate and reliable model to predict time series data. It has been widely used to solve time series dataset prediction problem [8][9]. As a water quality data is a kind of time series dataset, the FTS model is applicable in this scenario.

# Motivation

Among the three types of prediction models introduced in Section 1.1, both the MLR and ANN models have no requirement on the size of data samples and correlation level between

parameters. When training the ANFIS model, researchers need to follow the rules proposed by the model creator, such as the number of data samples should not be less than the number of parameters in the model. In real water quality monitoring systems, the size of the water quality dataset spans between a few hundred to a few thousand [3][4][6][7][8][10][11]. This motivates us to study given different dataset settings, how to choose the right prediction model which can provides good prediction performance for the target parameters.

In this dissertation, the water quality prediction problem is classified into five categories based on the size of a water quality dataset. The following assumptions are used in our classification. First, it assumes that each water quality monitoring station samples the water every half month. Thus, one monitoring station collects 24 data samples in each year. Second, it assumes that a water quality monitoring system has one or more water quality monitoring stations. For a system with five or more water quality monitoring stations and has been running for 20 years, the water quality dataset collected in the system should have more than 2400 water quality samples. In this dissertation, the dataset with 2400 or more data samples is called a large size dataset. A medium- size dataset would be one from a monitoring system with more than two monitoring stations which has been running for 10 more years and has 480 but less than 2400 data samples. The dataset will be treated as a small dataset if the size of the dataset is less than 480.

Five scenarios are envisioned:

**Scenario 1:** Large size dataset  and strong correlation between parameters.

In this scenario, as there are sufficient data samples to cover different water quality conditions, MLR, ANN, and ANFIS models can be directly employed to predict water quality. The most reliable and accurate model can be selected out based on the prediction performance of each model in the testing stage.

**Scenario 2:** Large size dataset and weak correlation between parameters.

As a water quality data is a kind of time series dataset, timing impact should be taken into consideration. The ANFIS model is not applicable for predicting parameters with a weak correlation. Is it possible to expand the input parameters in the time domain to obtain a strong correlation among them? Another option is to apply ANFIS to predict parameters in the timely expanded dataset.

**Scenario 3:** Medium size dataset (480~2400 sample) and strong correlation between parameters For this scenario, one important issue is how to select the membership functions and network

structure to obtain optimum configuration for the ANFIS model. Another issue is that the ANFIS model may have the risk of falling into out-of-range error trap when the size of the dataset is limited.

**Scenario 4:** Medium size dataset and weak correlation between parameters

Intelligence algorithms have been proven to be a reliable way to find the near-optimum solution in a search problem. Is it possible to integrate intelligence algorithms with the ANFIS model to find the optimum configuration of parameters in the ANFIS model so that the prediction accuracy can be greatly improved by the hybrid ANFIS model?

**Scenario 5:** Small dataset (<480 samples)

Not applicable, the model built upon small dataset is not reliable or portable.

# Contribution

In this dissertation, to address the water quality prediction problems described in Scenarios 2 to 4, different solutions are proposed and evaluated. Different methods including stratified sampling, wavelet de-noising, time series analysis, and intelligence algorithms are selected to

integrate with the ANFIS model to predict water quality according to the scenario that the collected water quality dataset fits in. The major contributions of our work are listed below:

For issues in Scenario 2:

* + 1. Input parameters are expanded in the time domain to build correlation between parameters so that the ANFIS model can be applied. The FTS model is applied to predict the water quality parameters which have small fluctuation. As the FTS model only uses the historical data of the prediction target, it is more suitable to apply this model to predict the water quality parameter which has a very long record in one single water quality monitoring station.

For issues in Scenarios 3:

1. Stratified sampling strategy is used to mitigate the uneven distribution of training and testing dataset and thus eliminate out-of-range errors of the ANFIS model in the testing stage.
2. A general framework of water quality prediction system based on wavelet de-noised ANFIS model using stratified sampling is proposed.
3. The general input parameter selection method is proposed to identify the most correlated input parameters.
4. The network structure selection method is proposed to increase the robustness and scalability of the ANFIS-based prediction system.

For issues in Scenario 4:

1) Integrate intelligence algorithms combined with the ANFIS model are proposed to improve the water quality prediction performance. Compared with the FTS model,

this model can be used to predict many water quality parameters with no limitation on the number of water quality monitoring stations.

# Outline of the Dissertation

The rest of the dissertation is organized as follows. Chapter 2 reviews different kinds of water quality prediction models. The study area, basic methods and evaluation metrics that are commonly used in the study of both scenarios are introduced in Chapter 3. A wastewater quality prediction system based on stratified sampling and wavelet denoising ANFIS model is presented to solve the Scenario 3 problem in Chapter 4. In Chapter 5, the Scenario 2 problem is considered and fuzzy models and time series analysis are presented to solve it. The integration of intelligence algorithms and the ANFIS model is presented to solve Scenario 4 problem in Chapter 6. Chapter 7 concludes this dissertation.

CHAPTER 2 LITERATURE REVIEW

Modeling the quality of water resources is vitally important for water scheduling and management. In the past, scientists regularly sampled the water in water quality monitoring stations and assessed the components in the water sample in a lab. However, this process takes a long time, and thus, the detected results are not timely. With the emergence of artificial intelligence (AI) techniques since the last decade, researchers have begun to adopt multivariate linear regression (MLR), artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), and Fuzzy time series (FTS) model to predict water quality by exploring the linear and non-linear relationships residing in water quality datasets. In addition, the wavelet denoising method and intelligent algorithms are also proposed to combine with machine learning techniques to enhance the prediction accuracy. In the following, we will review these related work in four categories of machine learning methods.

# MLR

MLR is a kind of statistical analysis method which is used to estimate the target value based on given values collected from a set of independent variables. It is adopted to predict the water quality because of its speed and simplicity. In [3], the MLR model is used to predict biochemical oxygen demand (BOD) and chemical oxygen demand base on four independent variables, temperature, pH, total suspended solid, and total suspended. The system quickly receives relatively good result in BOD prediction with a correlation coefficient value of 0.5. MLR model has also been used to predict the water quality index in [10] and found to be reliable in formulating the relationship excluding the parameter chloride. However, the MLR model can only be used to formulate linear relationship. It is likely to have a large prediction error if the

MLR model is used to predict non-linear relationship.

# ANN

Various ANN models have been designed to predict water and wastewater discharge quality based on previous existing datasets. A two-layer ANN model has been applied to predict the DO concentration in the Mathura River [11], and the experimental result showed that the ANN model worked well. In [12], various neural network types are compared in predicting water temperatures in streams. A radial basis function neural network has also been proposed to describe the water quality parameters in [13]. The summary of the experiment result shows the model outperforms the linear regression model in conductivity, turbidity, and total dissolved solids prediction. A time series prediction model, namely the autoregressive integrated moving average, was integrated with the ANN model to improve the prediction performance. The experimental results showed that the hybrid model provided better accuracy than ARIMA and ANN models [14]. Additionally, a comprehensive comparison between ANN and MLR models in biochemical oxygen demand and chemical oxygen demand prediction has been performed [3]. The experimental results show that a three-layer neural network model outperforms an MLR model. The other comparison between ANN and MLR models in water quality index prediction furtherly proves that the ANN model is a better option [10].

Although ANN models can effectively improve the prediction accuracy of water quality parameters, shortcomings still exist. Especially in some scenarios where the input parameters are ambiguous, neural networks struggle to formulate a non-linear relationship. In [15], wavelet transformation was applied to the ANN model to improve the prediction accuracy of a variety of ocean water quality parameters. An integration of a particle swarm optimization algorithm with ANN models has also been investigated to improve the forecasting performance [16]. In [17],

120 data samples, collected from 2002 to 2012, are used to verify whether the integration of fuzzy logic and ANN models can improve the water quality prediction performance. The experimental results confirm that the proposed method works.

# ANFIS

Many studies have proven that ANFIS, which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario [5]. The experimental results in [6] show that an ANFIS model works much better than an ANN model in predicting dissolved oxygen, even though there are only 45 data samples available. An ANFIS model with eight input parameters is used to predict total phosphorus and total nitrogen, the experiment result based on 120 water samples shows the proposed model is reliable [18]. The ANFIS model has also been applied to estimate the biochemical oxygen demand in the Surma River [19]. The testing results from 36 water samples confirmed that the ANFIS model could accurately formulate the hidden relationship and correlation analysis can improve the prediction accuracy. Two different kinds of ANFIS model, fuzzy c-means and subtractive clustering-based was compared in [20], the experiment result shows the ANFIS model built by fuzzy c-means provides more accurate prediction result. In [21], the ensemble models of wavelet ANNs are found to be superior to the best single model for forecasting chlorophyll and salinity concentrations in coastal water. An ensemble of ANN and ANFIS is proposed in [22] to improve the prediction performance of the ANN and ANFIS model, the test result shows there is a significant improvement in the Ensemble ANN-ANFIS model.

According to the developer of the ANFIS model, the size of the training dataset should be no less than the number of training parameters [23]. In the aforementioned papers, though the ANFIS models have received higher prediction accuracy, the sizes of the training datasets are

me scenarios, especially when the input



data have a large value range and there exist some extreme data value points, an out-of-range error is likely to occur, which happens when the testing dataset cannot find any insight from the training model. A few out-of-range errors can cause a very large prediction error, even though the model can accurately predict most of the data samples. In [24], a dataset collected from 122 wells in Mashhad plain (Iran) is used to investigate the performance of ANFIS, ANN, and geostatistical models in groundwater quality prediction. The experimental result shows that the ANFIS model has poor performance in the testing stage because the limited training dataset cannot build a robust or reliable model.

Recently, a few researchers have tried to integrate a machine learning model with a wavelet de-noising technique to improve prediction accuracy. Wavelet support vector regression and wavelet artificial neural networks have been proposed to model monthly pan evaporation [25]. The experimental results confirm that wavelet artificial neural networks outperform other models. An integrated wavelet de-noising ANFIS model was proposed to predict electrical conductivity (EC) and total dissolved solids (TDS) in [26]. Although the size of the dataset was smaller than the requirement, the model still achieved good prediction performance. In [27], eight different kinds of membership functions, with different wavelet de-noising schemes, were investigated to improve the performance of an ANFIS model. Based on the above two research studies, an optimized wavelet-ANFIS model is proposed in [28] and the experimental results show that a bell-shaped membership function with random sampling has the best prediction performance. In [29], a wavelet-ANFIS model is proposed to predict the groundwater level. Compared to ARIMA and ANFIS, the proposed model provides a more precise prediction result. A comparative study of different wavelet-based ANN models to predict sewage sludge quantity is

given in [30], the experiment result also proves wavelet can improve the accuracy to the ANN models.

On the other hand, many researchers have also tried to integrate intelligence algorithms with the ANFIS model to improve the performance of the proposed model. An application of genetic algorithm (GA), ant colony optimization for continuous domains, and differential evolution is introduced in [31] to improve the performance of the ANFIS model in predicting parameter electrical conductivity, sodium absorption ratio, and total hardness. The experiment result confirms that the proposed model can improve the performance of the ANFIS model for predicting EC and pH and the root mean square error (RMSE) value of the proposed model in the testing stage is 73.03 and 49.55, respectively. In [32], the genetic algorithm and particle swarm optimization (PSO) algorithm are integrated with the ANFIS model to optimize the threshold bank profile prediction. This method is also used in precipitation modeling. The experimental result indicates that the integrated ANFIS models with hybrid GA/PSO achieve better accuracy than the simple ANFIS model [33].

# FTS

A water quality data is a kind of time series dataset which is likely to have complicated linear and nonlinear relationships. The Fuzzy time series (FTS) model was first proposed by Song and Chissom in 1993 to address an enrollment prediction problem [34]. Chen improved this model by replacing complicated max-min composition operations with simplified arithmetic operations [35]. In [8], a Heuristic Gaussian cloud transformation was integrated with an FTS model to forecast water quality. The experimental results showed that the proposed model significantly improved the prediction accuracy. However, there were only 520 water quality samples available to build the cloud, and thus, the model was not reliable or robust. Time series analysis is also

proposed to address dissolve oxygen prediction, and the experimental results show that the proposed analysis method can find out valuable knowledge from water quality historical time- series data [36].

In this dissertation, MLR, ANN, ANFIS, and FTS models are integrated with statistical analysis, wavelet denoising, and intelligence algorithm to explore the prediction of water quality.

CHAPTER 3

STUDY AREA AND METHODS

This chapter introduces the study area, basic methods, and evaluation metrics that are used in the study of both scenarios 2 to 4 in Chapters 4 to 6.

# Study Area

The Las Vegas Wash (LVW) 

wastewater discharges and runoff to Lake Mead, a man-made lake on the Colorado River [37]. With an average flow of 200 million gallons per day, the LVW contributes approximately 1.5% of the total water flow to the lake. The flow in the Wash consists of highly treated wastewater, urban runoff, shallow groundwater, and stormwater. Lake Mead is the largest reservoir in the United States, located in the Colorado River, which is the drinking water source for 30 million people in Nevada, Arizona, and California. Salinity control is a huge issue for the Colorado River, and part of an international treaty with Mexico. Agricultural damage due to salinity costs the US over 454 million per year [38].

Along the LVW, several government organizations, including the Southern Nevada Water Authority (SNWA) and United States Geological Survey (USGS), have built multiple water quality monitoring stations to monitor and track the wastewater quality. Six locations have been selected as the places to build monitoring stations, by nearly all organizations, because of their geographical advantages. These six water quality monitoring stations are labeled in Figure 1 [37], and their geographical distribution is given.

There are three basins occupied by the Lake Mead Reservoir, with the Boulder Basin (BB) as the most western one. It lies within the boundaries of Clark County, Nevada and Mohave County, Arizona. The BB provides drinking water resources for the people living there. The

water in the BB finally joins the Lake Mead. The geographical distribution of the BB in Lake Mead is depicted in Figure 2.



**LW11.0**

**LW0.55**

**LW8.85**

**LW3.4**

**LW3.7**

**LW6.05**

Figure 1. The geographical distribution of the six water quality monitoring stations in the Las Vegas Wash



Figure 2. Location of Boulder Basin water quality monitoring station

# Stratified Sampling

Stratified sampling is a probability sampling method that partitions the entire dataset into different subgroups, or strata, based on the value of each data point. Then researchers proportionally select data points from different strata for different purposes [39]. This method is widely used in classification problems to verify proposed models, in which the training and testing datasets of each fold contain roughly the same proportion of each class label. Compared with the traditional sampling method, which generates *n* random partitions from the original dataset, stratified sampling makes sure the training and testing dataset can evenly cover all of the different categories.

A stratified sampling method works better in scenarios where the problem has data sparsity restrictions. When sampling from a large dataset, randomly sampling data from the collected dataset can cover all scenarios. However, when it comes to a medium or even a small order of magnitude dataset, some types of scenarios probably fall out of the sampled dataset. Since most of the water quality monitoring stations have only been built in the past several decades, and a large proportion of the water quality parameters have only been recordable in the past 20 years, the water quality dataset belongs to this case, which has limited number of data samples.

# Adaptive Neuro-Fuzzy Inference System

The ANFIS is a hybrid learning model, which integrates the neural network and fuzzy logic into an integrated system. The system can achieve high performance in formulating nonlinear relationships and forecasting chaotic time series. It can construct a reliable and accurate input- output mapping relationship based on the fuzzy if-then rules. The ANFIS model used in this study is generated based on the fuzzy model proposed in [40]. Given two input parameters, *x* and *y* , and one output function, *f* , the rule set built upon the model can be expressed as follows:

*Rule 1 : if x is A1 and y is B1 then f1 = a1x+b1 y+c1* (1)

*Rule 2 : if x is A2 and y is B2 then f2 = a2 x+b2 y+c2* (2)

where

*A*1 ,

*A*2 ,

*B*1 , and *B*2 represent four input membership functions for input parameters *x* and

*y* . In this example, each input has two membership functions. The value of the consequent

parameters *ai* ,

*bi* , and *ci*

are calculated by the least square error method.

The network structure of an ANFIS model consists of five layers. The data flow in Figure 3 illustrates the process of deriving the output from two inputs. The function of each layer is

*i*

*i*

fuzzification, normalization, merge, and summarization. Assuming function

*A* and

*B* are the

membership functions of fuzzy sets *Ai* and *Bi* , parameters *di* and *i* are the premise parameters

that define the structure of the membership functions in each node. To each layer, *i* , the output of the *jth* node is denoted as *O j* . A brief introduction of the corresponding function in each layer is as follows:

*i*

Layer 1: Generates membership weights for the fuzzy sets based on membership function;

*i*



*O*

1 *Ai*

exp

(*x d* )2

2

*i*

*i*

(3)

Layer 2: Generates rules by executing, AND operator to the two incoming membership weights;

*Oi w*

(*x*) ( *y*)

(4)

2 *i Ai Bi*



Layer 3: Normalizes the weight for the merge process;

*wi*

*O*

*i*

3 *wi* 4

(5)

 *wi*

*i* 1

Layer 4: Computes the output by multiplying the incoming normalized weight with the result of the linear regression model in the current node;

4 *wi zi wi* (*ai x bi y ci* ) (6)

*O*

*i*

Layer 5: Summarizes all of the products from Layer 4.

4

4 *wi yi*



*i* 1



*i* 1

*O*

*i*

5 *wi yi* 4

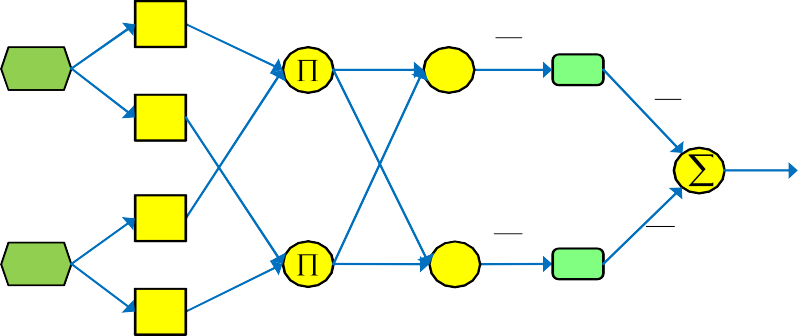
(7)

*wi*



*i* 1

*F*



*layer 1*

*layer 2*

*layer 3*

*layer 4*

*layer 5*

**A1**

*x y*

*xx*

*w1*

*N*

*w1*

*f*

*1*

**A2**

*w1 f1*

**B1**

*y*

*w2*

*N*

*w2*

*f*

*w2 f2*

*2*

**B2**

*x y*

**Fuzzification**

**And Normalization Merge Summarization**

**Backward Pass Forward Pass**

Figure 3. A two-input ANFIS for first-order Sugeno fuzzy model with two rules

# Wavelet Transform

Wavelet Transform (WT) is widely used in the analysis of time series signals due to its effectiveness in smoothing the irregularity of non-stationary data [30]. According to the way the scale parameter is discretized, it is classified into continuous or discrete WT. As continuous WT requires a large number of data samples, discrete WT is selected as the de-noising technique in this study. Discrete WT decomposes the input signal into a mutually orthogonal set of wavelets by using a discrete set of wavelet scales and translations. Compared to continuous WT, it requires much less computation time and is simpler to develop. Given the limited number of the highest coefficients of the discrete WT spectrum, an inverse transform can be performed with the

same wavelet basis to remove the noise hidden in the true signal. The corresponding wavelet

transformation can be defined as:

*WT* (*a*, *b*, )

*f* (*t*)



1

*a*

\* (*t b* )*dt a*

(8)

where the variables *n* and *m* are integers that control the wavelet dilation and translation, *a* is the scale index parameter, and *b* is the time-shifting parameter. All of the points that can be represented as *(am , namb)* are included in the subset of the wavelet scales and translations. (*t*) is a continuous function in both time and frequency domains, called mother wavelet, to get a stably invertible transform; and *f* (*t*) is the input signal, or time series.

# Evaluation Metrics

There are many evaluation metrics available to examine the performance of the proposed model. In this study, the mean average percentage error (MAPE), mean square error (MSE), RMSE, and coefficient of determination (R2) are adopted to compare the performance of different models. MAPE is used to represent the difference between the predicted value and true value, in percentage form. RMSE is the value calculated by rooting the square of the mean of the residuals between the true value and predicted value. R2 is an indicator to show how close the data are to the fitted regression line. The mathematical definitions of the evaluation metrics are

defined as:

| *ytrue ypred* |



1

*n*

*n i* 1



(9)

*MAPE*

*i i*

*ytrue*

*i*

*i i*

2

100%

*MSE*



1

*n*

*n i* 1

( *ypred*

*ytrue* )

(10)

*RMSE* (11)



1

*n*

*n*

( *y*

2

*predi truei*

*y*

)

*i* 1

*n*

( *y y* )2

*predi truei*

2 *i* 1

1

*R*

*n*

2

(12)

 ( *ytrue y*

*i*

*i* 1

*mean* )

where index *i* represents the position of the element in the vector, *ytrue* is a vector holding all of the

observed value,

*ymean*

stands for the average value of vector *ytrue* , *ypred* is a vector storing all the

forecasting value, and *n* is the size of the dataset.

CHAPTER 4

WASTEWATER DISCHARGE QUALITY PREDICTION USING STRATIFIED SAMPLING AND WAVELETDE-NOISING ANFIS MODEL

This chapter is focused on studying the water quality prediction problem fitting in Scenario 3. We aim to predict the parameters related to the salinity levels of wastewater discharge samples from the LVW, which consists of over 90% of treated wastewater discharge. The control of these parameters discharging into the Colorado River is paramount to keep salinity levels low. It is desirable to use some easily measurable parameters to predict a more complicated parameter. The objective of this research is to predict a target parameter, based on other highly correlated parameters, using the ANFIS model. Evaluation of the proposed model is conducted using the treated wastewater discharge quality dataset from LVW since 2007. Compared with other artificial intelligence methods. The ANFIS model has been proven to be reliable in predicting parameters related to salinity levels in experimental results.

The organization of the remainder of the chapter is as follows. The wastewater quality parameters are introduced in Section 1. The wastewater quality prediction system is introduced in Section 2. Section 3 presents the experimental configuration and results.

# Water Quality Parameters

In this research, the water quality dataset managed by the SNWA is adopted because of its high sampling frequency and more effective water quality parameters. It can be seen from Figure 1 that LW3.4 is the key monitoring station, where if the system determines the water quality parameter exceeds the regulation limit, the water still can be treated before it is discharged into Lake Mead. Five hundred and sixty-six wastewater quality samples collected by SNWA at LW3.4 since 2007 are used to evaluate the model.

Table 1. Statistical measures of water quality parameters at LW3.4 and LW3.7 in the Las Vegas Wash.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameters | Unit | Min | Max | Mean | S. D. | MCL |
| **EC** | **uS/cm** | **1118** | **2677** | **2241** | **190** | **2000** |
| DO | mg/L | 6.19 | 10.08 | 8.57 | 0.74 | 5~14 |
| pH | units | 7.51 | 8.89 | 8.22 | 0.19 | 6.5~9.2 |
| TP | mg/L | 0.05 | 0.58 | 0.12 | 0.07 | 0.1 |
| TN | mg/L | 3.97 | 17.84 | 13.72 | 1.75 | 10 |
| TSS | mg/L | 3 | 433 | 25.2 | 50.8 | N/A |
| AT | deg F | 37.3 | 122.1 | 79 | 16.7 | N/A |
| WT | deg C | 12.94 | 30.08 | 22.88 | 3.88 | N/A |
| CT | MPN | 23 | 240000 | 7660 | 24676 | N/A |
| fluoride | mg/L | 0.33 | 1.11 | 0.81 | 0.12 | 2.0 |
| **chloride** | **mg/L** | **104** | **370** | **300.5** | **31.57** | **250** |
| **TDS** | **mg/L** | **754** | **1838** | **1501** | **138** | **500** |
| **sulfate** | **mg/L** | **302** | **678** | **538** | **58.8** | **250** |
| turbidity | NTU | 1.5 | 448 | 15 | 39.4 | 25 |

The SNWA collects and analyzes two separate wastewater samples at LW3.4 every two weeks. In each wastewater sample, there are more than fifty water quality parameters, but only fifteen of them chloride, coliform total (CT), EC, dissolved oxygen (DO), fluoride, total nitrogen (TN), pH, total phosphorus (TP), sulfate, air temperature (AT), water temperature (WT), TDS, total suspended solids (TSS), and turbidity are quantized due to equipment and human limitations. Table 1 lists the statistical measures of the fifteen parameters. The first and second columns list the parameters and corresponding units. Columns three to six show the statistical information of each parameter. The last column lists the maximum contaminant level (MCL) of each parameter allowed by national drinking water regulations [41].

As studied in [42], the salinity level of Colorado River is largely impacted by the urban growth of the Las Vegas Valley. This study is focused on the parameters related to salinity levels. EC is used to measure the ability to carry electric current in wastewater. It is an indication of the total dissolved solids concentration found in that water sample. TDS is the total dissolved

solids contained in the wastewater. The two parameters are the major indicators that quantify the quality of wastewater discharged to control salinity levels in the Colorado River. Other parameters, including fluoride, chloride, and sulfate, are also monitored as part of the TDS. In this study, chloride, and sulfate, together with EC and TDS, are selected as the prediction targets.

# Wastewater quality predication System

* + 1. *System Overview*

The water quality prediction system is built based on the water quality dataset collected by the SNWA. Figure 4 illustrates the workflow of the proposed wastewater quality prediction system. First, the network structure of the ANFIS model needs to be defined based on the size of input dataset. The second step is selecting proper input parameters to predict the target parameter. In this study, four input parameters are selected from each sample to predict the target parameter. Meanwhile, all collected water quality data samples are partitioned into training and testing sets based on the stratified sampling method. To find the optimum ANFIS model to solve this problem, the network structure and membership functions need to be adjusted in each loop. The root mean square error (RMSE) is utilized to evaluate the simulation results. If the RMSE of the current prediction is smaller than the previous simulation error, the new model will be stored; if it is even smaller than the target threshold, the system will automatically terminate the whole process. The function of each part of the whole system is detailed in the following subsections.

**Start**

**Data Collection**

**Stratified Sampling**

**Input Parameters Selection**

**Training**

**Set**

**Testing Set**

**Adjust MF Number and Type**

**Testing Model**

**Input Parameters**

**Output Parameters**

**YES**

**NO**

**Prediction Result**

**Evaluation by RMSE**

**New MF Option Available**

**NO**

**RMSE<=TARGET**

**Observed Value**

**Training Model**

**Split Data**

**Network Structure Selection**

**WT**

**YES**

**Stop**

Figure 4. Wastewater quality prediction system based on wavelet de-noised ANFIS model using stratified sampling

* + 1. *Network Structure Selection*

Table 2 lists the relationships between the number of membership functions to each input, and the required number of datasets to design an ANFIS model. The first column is the number of membership functions for each corresponding input parameter (see Eq. (1) - (2)). The second and third columns list the number of linear and non-linear parameters used inside each model,

respectively. Previous research has found a high performance ANFIS model usually has two or three membership functions for each input parameter [43]. To achieve a good generalization model, the size of the training dataset should be at least as large as the total number of linear and non-linear parameters in the ANFIS model. Otherwise, the constructed ANFIS model will have a high risk of facing out-of-range error; i.e., the testing data is not within the range of input data space. The last column lists the minimum size of the training dataset needed.

Table 2. Relationship between the number of membership functions and the number of parameters inside the ANFIS model

|  |  |  |  |
| --- | --- | --- | --- |
| No. of membership functions No. of linear | | No. of nonlinear | The size of |
|  | parameters | parameters | dataset |
| 2 2 | 12 | 12 | 24 |
| 2 3 or 3 2 | 18 | 15 | 23 |
| 3 3 | 27 | 18 | 45 |
| 2 2 2 | 32 | 18 | 50 |
| 2 2 3 or 2 3 2 or 3 2 2 | 48 | 21 | 69 |
| 2 3 3 or 3 2 3 or 3 3 2 | 72 | 24 | 96 |
| 3 3 3 | 108 | 27 | 135 |
| 2 2 2 2 | 80 | 24 | 104 |
| 2 2 3 3 or 2 3 2 3 or 3 3 2 2 or 3 2 3 2 | 180 | 30 | 210 |
| 2 3 3 3 or 3 2 3 3 or 3 3 2 3 or 3 3 3 2 | 270 | 33 | 303 |
| 3 3 3 3 | 405 | 36 | 441 |
| 2 3 3 3 3 | 972 | 42 | 1014 |
| 3 3 3 3 3 | 1458 | 45 | 1503 |

As most water quality parameters need extra chemical experiments to quantize their net weights, engineers can only collect water quality data samples once or twice in a month. Even in 20 years, at one monitoring station, there are less than 500 data samples available. Therefore, a reasonable network structure selection scheme should be established for researchers to design an accurate model to predict future water quality with limited dataset.

* + 1. *Input Parameter Selection*

Appropriate input parameter selection is the premise used to build an accurate model to predict the output parameters. Table 1 shows that the values of different parameters are in different orders of magnitude, with a very large value range, which will lead to a very large search space. The idea behind machine learning is to find similar previous experiences to predict the current target. If the testing case cannot find a similar pattern from the training model, the model will simply make a random prediction, and the output is likely to be inaccurate. In this study, preliminary experiments in sulfate and chloride prediction have proven this assumption. Therefore, instead of using raw data as the input, a standard score, which can greatly narrow down the search space by transforming the raw value into the number of standard deviations, is

utilized as the input. The standard score vector of a monitoring parameter

*Xss* can be formulated

as:

*Xss*



*X*

*X*

*X*

(13)

where *X* is a vector storing the observed value of a parameter; while *X* and *X* are the mean and standard deviation of the vector, respectively.

With the standard score of each parameter, the covariance, which can show the inner tendency of the linear relationship between any two parameters, can be represented by the following

equation:

cov( *Xssi* , *Xssj* )

*E*[( *Xssi X* )( *Xssj X* )]



*ssi ssj*

(14)

where *X ssi* and

are the vector and mean value of the standard score of the parameter *i* ,

*ssi*

*X*

respectively. To quantize and evaluate the correlation between any two parameters, the Pearson correlation coefficient is calculated by:

( *Xi* , *X j* )

*E*[(*Xssi X* )( *Xssj X* )]

(15)

*Xi X j*



*ssi ssj*

For a given parameter *i* , the most correlated parameters (vector *input* ) can be selected by a selection function:

*input = select( ,i,k)* (16)

where *k* is the number of input parameters to be selected, and is the correlation matrix. If the number of positive correlation parameters is smaller than *k* , then only the parameters with a positive correlation coefficient are selected.

More input parameters can usually more accurately model the relationship hidden in the

dataset, but increased parameter operation also triggers a higher demand for data samples. As described in Section 3.3.1, to avoid the risk of an out-of-range error, the ANFIS model is built under the requirements of the data set. In this study, the size of the wastewater quality dataset enables the ANFIS to select, at most, four most correlated input parameters as input, and each input parameter can be set as two or three membership functions. The MATLAB toolbox is used to configure the ANFIS model.

* + 1. *Water Quality Prediction Model*

The whole system is made up of two parts: the ANFIS model and the wavelet de-noise model. Four input parameters are selected out to predict the target parameter. In this study, for example, TDS, sulfate, chloride, and fluoride are selected as the input parameters because of their higher correlation values with EC in the training dataset. The organization of the WT-ANFIS model to predict EC is depicted in Figure 5. It clearly illustrates the water quality prediction process with the selected dataset, as well as the configuration of the network and membership functions. The WT process de-noises the output from the ANFIS model. If the prediction error is under the

limitation value, the system will terminate the process and save the result as output. Otherwise, a new configuration will be made to repeat the process until no new network configuration or membership functions are available. If no configuration can have an error smaller than the threshold, the system will output the best configuration as the prediction result.

**Run**



Layer 1

Layer 2 Layer 3 Layer 4 Layer 5 Layer 6

**r1**

**... f1**

**A3 B1**

**B2**

**r2**

**... f2**

**B3**

**ri**

**... fi**

**C1**

**Observed TDS/EC**

**C2**

**r80 ... f80**

**C3**

**D1**

**YES**

**D2**

**r**

**81**

**... f**

**81**

**New MF options Available**

**input**

**Backward Pass**

**D3**

**inputmf**

**rule**

**outputmf**

**sum**

**Forward Pass**

**wavelet**

**NO RMSE<Min**

**YES**

**Start**

**Sulfate**

**input**

**Compare**

**WDT**

**Fluoride**

**input**

**EC**

**input**

**Chloride**

**input**

**TDS**

**input**

**A2**

**A1**

**Stop**

Figure 5. Organization of water quality prediction model for parameter EC

Pearson correlation is employed to select the input parameters, and the number of membership functions for each input parameter is specified from empirical studies. All of the selected input parameters will be utilized to train an ANFIS model to predict the target parameter. This process can be defined as:

*X pssi = anfis(model,input,memf)* (17)

where *input* is the return value of *select* function, and *model* is the ANFIS model configured based on the size of the dataset. The argument *memf* is the membership function of this ANFIS model. Usually *gbellmf* has very good performance [6].

The prediction value of the test case can be calculated by a de-normalized process:

*X pred*

*Xpss X X*

(18)

where



*X pss*

is the predicted standard score and

*X pred* is the corresponding prediction value.

The functions of the first five layers are introduced in Section 2.D. The WT method will be used to remove noise in the result in layer six. The configuration of WT involves many different factors, including wavelet types, decomposition levels, thresholding methods, and multiplicative threshold rescaling types [26]. A toolbox has been integrated into MATLAB for users to conveniently configure and execute these functions. There are many wavelet options, and orthogonal wavelets are selected in this study because of their perfect reconstruction ability. Since soft thresholding is more accurate than hard thresholding, it is selected as the default configuration in this system. The value of each input parameter in the ANFIS system is classified into several levels. In this study, the system configuration and the number of input parameters are considered when defining the level of decomposition. The rescaling threshold is implemented with the level of noise analysis.

# Experimental Results and Discussion

* + 1. *Wastewater Quality Dataset and Experiment Configuration*

To evaluate the performance of the proposed wastewater quality prediction model, 566 wastewater quality samples from LW3.1 and LW3.4 are used to investigate the performance of each model. Figure 6 presents the observed values of the four prediction target parameters: chloride, sulfate, EC, and TDS, collected by the SNWA in the past 11 years. All of these

parameters are indicative of salinity levels. From the Figure, it can be seen that the values of these four parameters are correlated, especially in some extreme events. The gradual decreasing trend of each parameter value suggests that the wastewater quality for salinity in the LVW is improving. However, the values of chloride, sulfate, EC, and TDS are still higher than the MCL for drinking water more than 75% of the time. Therefore, timely prediction of the water quality is still very important.

In this study, MLR [3], ANNs [3], and ANFIS, along with an ensemble of ANNs and ANFIS (EANFIS) [22], are implemented and compared with the proposed model. TensorFlow is a widely used machine learning library, and the MLR and a three-layer ANN model developed in this paper are implemented by the library. In the MLR model, four parameters are selected based on correlation values to predict the target parameter. The ANN model built in this study is based on the model proposed in [3]. In the three-layer ANN model, the input layer has four nodes, and the hidden layer has four nodes. The activation function used is linear activation. Gradient descent is used to minimize the root mean square error between the true value and the prediction value in each iteration. The ANFIS model and WT are developed by the MATLAB toolbox. It can be seen in Figure 6 that there are many abnormal values in the whole dataset because of sudden changes in water conditions or human behaviors. A few out-of-range errors occurring in the ANFIS model can cause a very large testing error. Thus, cross-validation is not applicable in this scenario, especially when the size of the dataset is not large enough for each part in the cross-validation set to cover all of the situations. In these experiments, two different kinds of sampling methods, random sampling and statistical stratified sampling, are implemented to evaluate the effectiveness of different sampling strategies. The water quality dataset collected from the SNWA is partitioned into two parts by 75% and 25% for training and testing purposes,

respectively.

* + 1. *Network Structure Selection and Input Parameter Selection*

It can be seen in Table 2 that this dataset can train a reliable ANFIS model with at most four parameters, and each parameter can have at most three membership functions. Something noteworthy is that more input parameters or membership functions cannot guarantee better prediction accuracy. However, in the general case, especially when the input parameters are strongly correlated to the output parameters, more input parameters and membership functions can lead to a more accurate prediction result. For the ANFIS model, the Gaussian, triangular, trapezoidal, sigmoid, spline, and generalized bell-shaped input membership functions are used, and the number of each membership function is dynamically configured under the dataset size limitation. There are two types of output membership functions: constant and linear. In this study, the linear function is chosen because of its higher prediction accuracy [18].

Table 3. Correlation matrix of all wastewater quality parameters listed in Table 1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CH** | CT | **EC** | DO | FL | TN | PH | TP | **SU** | AT | WT | **TDS** | TSS | TU |
| CH | 1 | -0.49 | **0.96** | 0.35 | 0.82 | 0.77 | 0.31 | -0.33 | **0.89** | -0.16 | -0.20 | **0.94** | -0.51 | -0.56 |
| CT | -0.49 | 1 | -0.46 | -0.29 | -0.28 | -0.42 | -0.24 | 0.36 | -0.35 | 0.16 | 0.20 | -0.39 | 0.48 | 0.51 |
| EC | **0.96** | -0.46 | 1 | 0.32 | 0.76 | 0.73 | 0.26 | -0.28 | **0.93** | -0.16 | -0.20 | **0.97** | -0.45 | -0.51 |
| DO | 0.35 | -0.29 | 0.32 | 1 | 0.18 | 0.22 | 0.35 | -0.22 | 0.40 | -0.60 | -0.76 | 0.33 | -0.20 | -0.19 |
| FL | **0.82** | -0.28 | **0.76** | 0.18 | 1 | 0.78 | 0.20 | -0.20 | **0.74** | -0.13 | -0.10 | **0.77** | -0.42 | -0.44 |
| TN | 0.77 | -0.42 | 0.73 | 0.22 | 0.78 | 1 | 0.16 | -0.30 | 0.66 | -0.18 | -0.15 | 0.72 | -0.43 | -0.49 |
| PH | 0.31 | -0.24 | 0.26 | 0.35 | 0.20 | 0.16 | 1 | -0.25 | 0.26 | 0.14 | 0 | 0.25 | -0.25 | -0.24 |
| TP | -0.33 | 0.36 | -0.28 | -0.22 | -0.20 | -0.30 | -0.25 | 1 | -0.19 | 0.02 | 0.05 | -0.26 | 0.73 | 0.71 |
| SU | **0.89** | -0.35 | **0.93** | 0.40 | 0.74 | 0.66 | 0.26 | -0.19 | 1 | -0.23 | -0.31 | **0.95** | -0.32 | -0.37 |
| AT | -0.16 | 0.16 | -0.16 | -0.60 | -0.13 | -0.18 | 0.14 | 0.02 | -0.23 | 1 | 0.93 | -0.18 | 0.05 | 0.06 |
| WT | -0.20 | 0.20 | -0.20 | -0.76 | -0.10 | -0.15 | 0 | 0.05 | -0.31 | 0.93 | 1 | -0.23 | 0.08 | 0.06 |
| TDS | **0.94** | -0.39 | **0.97** | 0.33 | 0.77 | 0.72 | 0.25 | -0.26 | **0.95** | -0.18 | -0.23 | 1 | -0.41 | -0.47 |
| TSS | -0.51 | 0.48 | -0.45 | -0.21 | -0.42 | -0.43 | -0.25 | 0.73 | -0.32 | 0.05 | 0.08 | -0.41 | 1 | 0.93 |
| TU | -0.56 | 0.51 | -0.51 | -0.19 | -0.44 | -0.49 | -0.24 | 0.71 | -0.37 | 0.06 | 0.06 | -0.47 | 0.93 | 1 |

The input parameters for each prediction parameter are selected based on the proposed input parameter selection method. The correlation matrix between the 15 wastewater quality parameters is given in Table 3. For each prediction target parameter (column), the four most correlated parameters (rows) are selected as the input parameters.

* + 1. *Prediction Performance*

The proposed prediction system will iteratively test the model with the training and testing datasets until the ANFIS model satisfying the accuracy requirement is found. Table 4 shows the experimental results of the parameter chloride, based on both random and stratified sampling strategies. For each model, the upper/lower sub-rows represent the random/stratified sampling strategy results, respectively. The values of the stratified sampling method are bold for ease of view. It can be seen in Table 4 that, with a random sampling strategy, the ANFIS, EANFIS, and WT-ANFIS models have smaller training errors than the MLR and ANN models. This proves that ANFIS models can more accurately model the relationships than traditional models. Moreover, WT can effectively remove the noise hidden in the dataset. However, the shortcomings of ANFIS models are exposed when testing the model. Both ANFIS and WT- ANFIS have much larger prediction errors than other models.

As a comparison, the stratified sampling strategy can effectively address the out-of-range error problem in the ANFIS model. The prediction errors of MLR and ANN models are improved a small amount in both training and testing stages through this sampling method. The ANFIS model obtains a slightly larger training error (5.383 vs. 4.602) and a much smaller testing error than using the random sampling strategy (20.67 vs. 6.849). This is because the training dataset, split by the stratified sampling strategy, covers nearly all possible scenarios. The values of some scenarios are extremely large or small, which do not fit the model very well but are also vitally

important for upcoming prediction. Meanwhile, all of the testing samples can gain insight from the training model with the stratified sampling strategy, which can greatly reduce prediction error. Wavelet transformation also effectively removes the noise in the prediction results, which furtherly reduces the RMSE to 3.324 and 4.993 in the training and testing stages, respectively.

Table 4. The training and testing performance of different models for parameter chloride in the LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Chloride** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **MLR** | 8.028 | 2.039 | 0.931 | 8.784 | 2.220 | 0.933 |
| **8.006** | **1.988** | **0.934** | **8.129** | **1.980** | **0.936** |
| **ANNs** | 7.791 | 1.923 | 0.935 | 8.675 | 2.196 | 0.934 |
| **7.927** | **2.010** | **0.935** | **8.336** | **2.036** | **0.932** |
| **ANFIS** | 4.602 | 1.114 | 0.977 | 20.67 | 4.209 | 0.627 |
| **5.383** | **1.311** | **0.970** | **6.849** | **1.678** | **0.954** |
| **EANFIS** | 5.610 | 1.436 | 0.966 | 12.258 | 3.021 | 0.869 |
| **6.120** | **1.558** | **0.961** | **7.011** | **1.720** | **0.952** |
| **WT-ANFIS** | 2.654 | 0.634 | 0.992 | 18.51 | 3.319 | 0.701 |
| **3.324** | **0.795** | **0.989** | **4.993** | **1.173** | **0.976** |

The training and testing results for the parameter sulfate, based on random sampling and stratified sampling, are given in Table 5. In the training stage of random sampling, the ANFIS model has the smallest RMSE, which is only 9.567. However, in the testing stage, the RMSE error increases to 19.24, which is larger than those of the MLR, ANNs, and EANFIS models. As explained before, the reason is that part of the testing data from random sampling could not find a similar pattern from the training model. These testing samples are likely to incur out-of-range prediction errors. When it comes to stratified sampling, the testing RMSE is reduced by 23.9% at the cost of the training error increasing by 9.19%. This is because the training dataset is evenly distributed across all value ranges. In other words, all continuously changing water conditions have been covered in the training model. Some of the training data samples, like extreme

weather conditions, rarely occur; therefore, these training data samples will have larger training errors. However, all of the testing data samples can find similar patterns from the training model, which leads to a much smaller testing error. Wavelet transformation, combined with the ANFIS model, also greatly improves training and testing prediction results. It can be seen from Table 5 that the WT-ANFIS model is the only model in which the RMSE is smaller than 10. This proves that wavelet de-nosing techniques can work very well in a model where the training dataset can cover all of the different scenarios.

As expected, the experimental results of parameters EC and TDS, shown in Tables 6 and 7 have similar patterns as the parameter sulfate. In the case of random sampling, the ANFIS model outperforms the other four models in the training stage. In the testing stage, the ANFIS model has the best performance in parameter EC prediction, while the EANFIS model outperforms the other models in parameter TDS prediction. Wavelet transformation worsens the prediction results of the ANFIS model in both training and testing stages. This further proves that the ANFIS model is inaccurate and unreliable when the model is trained with a random sampling strategy.

Table 5. The training and testing performance of different models for parameter sulfate in the LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Sulfate** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **MLR** | 18.18 | 2.544 | 0.899 | 18.50 | 2.563 | 0.911 |
| **18.62** | **2.567** | **0.898** | **17.90** | **2.499** | **0.911** |
| **ANNs** | 18.20 | 2.553 | 0.899 | 18.42 | 2.566 | 0.912 |
| **18.42** | **2.608** | **0.899** | **17.80** | **2.510** | **0.912** |
| **ANFIS** | 9.567 | 1.309 | 0.972 | 19.24 | 2.431 | 0.904 |
| **10.19** | **1.422** | **0.969** | **14.64** | **1.983** | **0.940** |
| **EANFIS** | 12.36 | 1.771 | 0.953 | 15.42 | 2.136 | 0.938 |
| **12.83** | **1.857** | **0.951** | **13.61** | **1.908** | **0.948** |
| **WT-ANFIS** | 47.49 | 7.084 | 0.311 | 55.54 | 8.057 | 0.199 |
| **4.933** | **0.707** | **0.993** | **9.434** | **1.132** | **0.975** |

Table 6. The training and testing performance of different models for parameter EC in the LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **EC** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **MLR** | 38.96 | 1.075 | 0.957 | 29.28 | 1.026 | 0.977 |
| **38.83** | **1.063** | **0.957** | **29.30** | **1.039** | **0.977** |
| **ANNs** | 38.63 | 1.069 | 0.958 | 30.57 | 1.057 | 0.975 |
| **38.77** | **1.066** | **0.957** | **29.50** | **1.050** | **0.976** |
| **ANFIS** | 21.46 | 0.725 | 0.987 | 26.57 | 0.897 | 0.981 |
| **20.82** | **0.691** | **0.987** | **30.04** | **0.898** | **0.976** |
| **EANFIS** | 26.82 | 0.852 | 0.980 | 27.44 | 0.932 | 0.966 |
| **26.48** | **0.831** | **0.980** | **28.09** | **0.941** | **0.979** |
| **WT-ANFIS** | 150.9 | 5.243 | 0.357 | 160.0 | 5.584 | 0.302 |
| **12.21** | **0.401** | **0.996** | **21.11** | **0.715** | **0.988** |

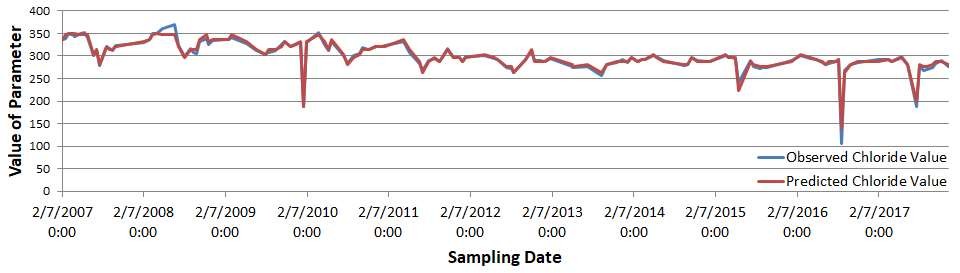
Table 7. The training and testing performance of different models for parameter TDS in the LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **TDS** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **MLR** | 32.74 | 1.599 | 0.948 | 34.52 | 1.730 | 0.911 |
| **28.09** | **1.398** | **0.959** | **34.52** | **1.563** | **0.933** |
| **ANNs** | 26.51 | 1.236 | 0.966 | 29.58 | 1.369 | 0.934 |
| **25.57** | **1.230** | **0.966** | **32.03** | **1.396** | **0.943** |
| **ANFIS** | 18.31 | 0.901 | 0.984 | 31.61 | 1.422 | 0.925 |
| **18.71** | **0.910** | **0.982** | **29.47** | **1.397** | **0.951** |
| **EANFIS** | 20.80 | 1.023 | 0.979 | 27.77 | 1.313 | 0.942 |
| **20.81** | **1.022** | **0.977** | **25.95** | **1.250** | **0.962** |
| **WT-ANFIS** | 117.0 | 6.313 | 0.335 | 102.4 | 5.659 | 0.213 |
| **12.14** | **0.597** | **0.992** | **15.99** | **0.827** | **0.986** |

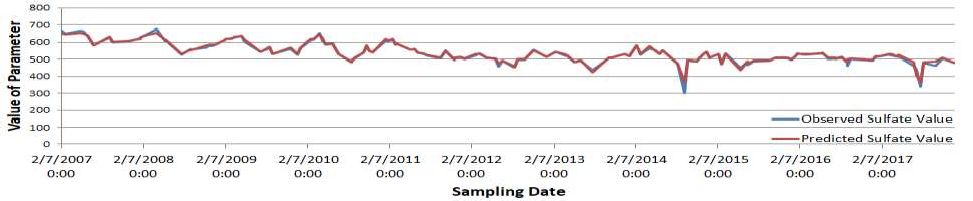
As a comparison, the WT-ANFIS model trained with a stratified sampling strategy achieves the best prediction performance among the five models. The RMSE achieved in training/testing stages is 12.21/21.11 for EC prediction and 12.14/15.99 for TDS prediction. The experimental results from Tables 6 and 7 show that the proposed model, with the stratified sampling strategy, is more reliable and robust.

Figure 6 plots the testing result of the parameters chloride, sulfate, TDS, and EC in the proposed model. The blue line represents the observed value of each parameter every two weeks. The red line indicates the predicted value given by the proposed model every two weeks. It can

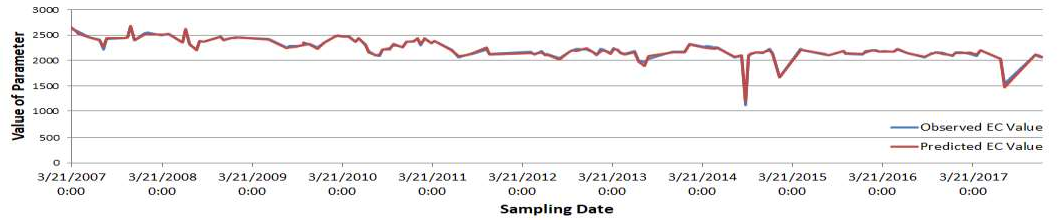
be seen from the Figure that the two lines almost overlap, which demonstrates that the proposed model is reliable.



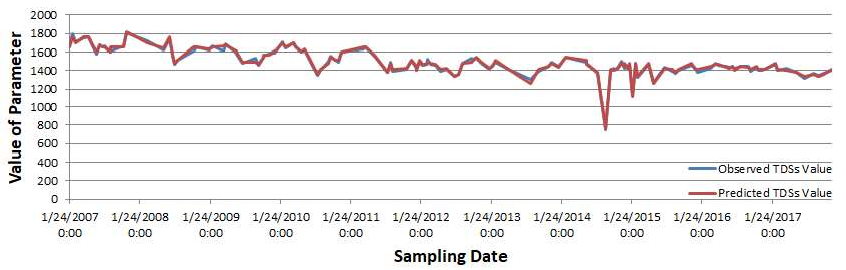
(a)



(b)



(c)



(d)

Figure 6. The observed and predicted value of each prediction parameter: (a) Chloride; (b) Sulfate; (c) EC; (d) TDS.

CHAPTER 5

USING FUZZY MODELS AND TIME SERIES ANALYSIS TO PREDICT WATER QUALITY

In this chapter, fuzzy models and time series analysis are introduced to predict the water quality for problems fitting in Scenario 2 where the dataset has a very long history record in one single water quality monitoring station. Both ANFIS and FTS models are employed to predict water quality when the input and output parameters have weak correlations. The time series analysis method is used to preprocess the water quality dataset to figure out appropriate input parameters. A stratified sampling strategy is employed to evenly partition the whole dataset for training and testing purposes.

# Materials and Methods

* + 1. *Water Quality Parameters*

In the current study, the water quality datasets collected from the LVW and BB are adopted because of high sampling frequency. LW3.4 is the key monitoring station at which if the system determines the water quality parameter exceeds the regulation limit, the water still can be treated before it is discharged into Lake Mead. The water quality datasets monitored at LW3.4 between 2005 and 2010 by LVW Coordination Committee are used to evaluate the model. There are five water quality parameters, temperature (T), pH, EC, DO, total dissolved solids (TDS) in the collected dataset. Table 8 lists the statistical properties of these parameters. The statistical measurement of parameters, depth, pH, T, EC, and DO, in the dataset collected from the BB between 2011 and 2016 are given in Table 9. The first and second columns list the parameter label and corresponding unit, while the third column to the sixth column show the statistical

properties of each parameter. The last column lists the maximum contaminant levels (MCL) permitted by national drinking water regulations [41].

EC is used to lity to carry electricity, and TDS is the combination of items that are dissolved in the water. The two parameters are major indicators that quantify the quality of water. Further, DO is a necessity for all living organisms in the water. In this paper, these three parameters are selected as target parameters.

Table 8. A statistical measure of water quality parameters at LW3.4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Unit** | **Min** | **Mean** | **Max** | **S. D.** | **MCL** |
| **T** | C | 0.9 | 54.49 | 111.2 | 32.38 | N/A |
| **pH** | unit | 5.27 | 8.20 | 8.79 | 0.29 | 6.5~9.2 |
| **EC** | uS/cm | 1569 | 2463.59 | 2921 | 178.90 | 2000 |
| **DO** | mg/L | 2.42 | 8.28 | 17.95 | 1.80 | 5~14 |
| **TDS** | mg/L | 1000 | 1580 | 1870 | 110 | 500 |

Table 9. A statistical measure of water quality parameters at BB

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Unit** | **Min** | **Mean** | **Max** | **S. D.** | **MCL** |
| **Depth** | m | 0.9 | 54.49 | 111.2 | 32.38 | N/A |
| **T** | C | 11.1 | 14.24 | 31.5 | 3.74 | N/A |
| **EC** | uS/cm | 810 | 927.6 | 1160 | 51.81 | 2000 |
| **DO** | mg/L | 2.30 | 7.63 | 11.30 | 1.17 | 5~14 |
| **pH** | unit | 6.90 | 7.91 | 9.40 | 0.25 | 6.5~9.2 |

* + 1. *Input Parameter Selection*

Selecting the appropriate input parameters to build a model is fundamental to receiving accurate prediction results. It can be seen in Table 8 and 9 that the value of each parameter has a

different order of magnitude, and some have a very large range. Instead of using raw data as the input, feature scaling is adopted to normalize the value into the range [0, 1]. The process can be

defined as:

*x*' *xi x*min

*i*

(19)

*x*max *x*min

where *xi* is the observed value,

*x*min and

*x*max represent the minimum and maximum value of this

kind of parameter, and ' is the normalized observed value.

*x*

*i*

The ANFIS model requires that the input parameters have strong correlations with the target parameters. Pearson correlation is used to calculate the correlation between parameters. Tables 10 and 11 list the correlation values between parameters in the dataset collected from the LVW and BB.

Table 10. Pearson Correlation between water quality parameters in LVW

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **T** | **EC** | **PH** | **DO** | **TDS** |
| **T** | 1 | -0.22655 | 0.126015 | -0.4488 | -0.22628 |
| **EC** | -0.22655 | 1 | 0.208367 | 0.15019 | 0.999247 |
| **PH** | 0.126015 | 0.208367 | 1 | -0.09475 | 0.208139 |
| **DO** | -0.4488 | 0.15019 | -0.09475 | 1 | 0.14924 |
| **TDS** | -0.22628 | 0.999247 | 0.208139 | 0.14924 | 1 |

Table 11. Pearson Correlation between water quality parameters in BB

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Depth** | **T** | **EC** | **pH** | **DO** |
| **Depth** | 1 | 0.029396 | -0.02194 | -0.013 | -0.05289 |
| **T** | 0.029396 | 1 | 0.508715 | 0.1725 | -0.63209 |
| **EC** | -0.02194 | 0.508715 | 1 | -0.03386 | -0.29121 |
| **pH** | -0.013 | 0.1725 | -0.03386 | 1 | 0.109696 |
| **DO** | -0.05289 | -0.63209 | -0.29121 | 0.109696 | 1 |

As shown in the two tables, the correlation between parameters is very weak, except for parameters TDS and EC from the LVW. As discussed in Section 1, the ANFIS model is not applicable for cases with parameters of fairly low correlation levels. Water quality data is a type of time series data. Therefore, in this study, the timing effect of the dataset has been taken into

account. When calculating the correlation between the parameters, data collected in *t*

1. , *t*
2. and

*t* 3 are also taken into account. The new correlation between the parameters in the LVW and

BB are given in Table 12 and 13, respectively.

The bold value in Table 12 and 13 are the three strongest correlation values to the parameter named in each column. For example, to the parameter DO in Table 12, the value of DO in *t* has

the top three correlation with the value of DO collected in *t*

1. , *t*
2. and *t*
3. . Compared with

Table 10, in which no qualified correlation pair exists for parameter DO, each parameter in Table 12 can find out the appropriate input for itself. The FTS and ANFIS models were used to model the prediction of water quality with the new dataset.

Table 12. Pearson Correlation of water quality parameters in LVW with time series analysis

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LVW | **pH(t)** | EC(t-2) | EC(t-1) | **EC(t)** | DO(t-2) | DO(t-1) | **DO(t)** | TDS(t-2) | TDS(t-1) | **TDS(t)** |
| T(t-3) | 0.105 | -0.233 | -0.239 | -0.247 | -0.446 | -0.443 | -0.439 | -0.233 | -0.239 | -0.246 |
| T(t-2) | 0.110 | -0.227 | -0.233 | -0.239 | -0.449 | -0.446 | -0.443 | -0.226 | -0.233 | -0.239 |
| T(t-1) | 0.116 | -0.232 | -0.227 | -0.233 | -0.445 | -0.449 | -0.446 | -0.231 | -0.226 | -0.233 |
| T(t) | 0.126 | -0.238 | -0.232 | -0.226 | -0.443 | -0.445 | -0.449 | -0.238 | -0.231 | -0.226 |
| EC(t-3) | 0.183 | 0.850 | 0.732 | 0.631 | 0.147 | 0.146 | 0.145 | 0.849 | 0.732 | 0.631 |
| EC(t-2) | 0.190 | 1 | 0.850 | 0.732 | 0.150 | 0.147 | 0.146 | 0.999 | 0.849 | 0.732 |
| EC(t-1) | 0.199 | 0.850 | 1 | **0.850** | 0.148 | 0.151 | 0.147 | 0.850 | 0.999 | **0.849** |
| EC(t) | 0.208 | 0.732 | 0.850 | 1 | 0.144 | 0.148 | 0.151 | 0.731 | 0.849 | **0.999** |
| pH(t-3) | **0.880** | 0.196 | 0.185 | 0.174 | -0.097 | -0.095 | -0.094 | 0.196 | 0.185 | 0.173 |
| pH(t-2) | **0.915** | 0.208 | 0.196 | 0.185 | -0.095 | -0.097 | -0.095 | 0.208 | 0.196 | 0.185 |
| pH(t-1) | **0.953** | 0.199 | 0.208 | 0.196 | -0.095 | -0.095 | -0.097 | 0.199 | 0.208 | 0.196 |
| pH(t) | 1 | 0.190 | 0.199 | 0.208 | -0.097 | -0.095 | -0.095 | 0.190 | 0.199 | 0.208 |
| DO(t-3) | -0.096 | 0.148 | 0.144 | 0.146 | 0.958 | 0.940 | **0.934** | 0.147 | 0.143 | 0.145 |
| DO(t-2) | -0.097 | 0.150 | 0.148 | 0.144 | 1 | 0.958 | **0.940** | 0.149 | 0.147 | 0.143 |
| DO(t-1) | -0.095 | 0.147 | 0.151 | 0.148 | 0.958 | 1 | **0.958** | 0.146 | 0.150 | 0.147 |
| DO(t) | -0.095 | 0.146 | 0.147 | 0.154 | 0.940 | 0.958 | 1 | 0.145 | 0.146 | 0.150 |
| TDS(t-3) | 0.182 | 0.850 | 0.731 | 0.630 | 0.146 | 0.145 | 0.145 | 0.849 | 0.731 | 0.630 |
| TDS(t-2) | 0.190 | 0.999 | 0.850 | 0.731 | 0.149 | 0.146 | 0.145 | 1 | 0.849 | 0.731 |
| TDS(t-1) | 0.199 | 0.849 | 0.999 | **0.849** | 0.147 | 0.150 | 0.146 | 0.849 | 1 | **0.849** |
| TDS(t) | 0.208 | 0.732 | 0.849 | **0.999** | 0.143 | 0.147 | 0.150 | 0.731 | 0.849 | 1 |

Table 13. Pearson Correlation of water quality parameters in BB with time series analysis

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BB | **T(t)** | EC(t-2) | EC(t-1) | **EC(t)** | pH(t-2) | pH(t-1) | **pH(t)** | DO(t-2) | DO(t-1) | **DO(t)** |
| Depth(t) | 0.034 | -0.021 | -0.020 | -0.021 | -0.008 | -0.010 | -0.006 | -0.053 | -0.051 | -0.050 |
| T(t-3) | **0.992** | 0.508 | 0.507 | 0.507 | 0.173 | 0.173 | 0.173 | -0.633 | -0.636 | -0.637 |
| T(t-2) | **0.993** | 0.509 | 0.508 | 0.507 | 0.173 | 0.173 | 0.173 | -0.632 | -0.633 | -0.636 |
| T(t-1) | **0.996** | 0.507 | 0.509 | 0.508 | 0.173 | 0.173 | 0.173 | -0.629 | -0.632 | -0.633 |
| T(t) | 1 | 0.506 | 0.507 | 0.509 | 0.1727 | 0.173 | 0.173 | -0.626 | -0.629 | -0.632 |
| EC(t-3) | 0.506 | 0.996 | 0.995 | **0.994** | -0.0346 | -0.035 | -0.035 | -0.292 | -0.293 | -0.294 |
| EC(t-2) | 0.506 | 1 | 0.996 | **0.995** | -0.0337 | -0.035 | -0.035 | -0.291 | -0.292 | -0.293 |
| EC(t-1) | 0.507 | 0.996 | 1 | **0.996** | -0.0341 | -0.034 | -0.035 | -0.290 | -0.291 | -0.292 |
| EC(t) | 0.509 | 0.995 | 0.996 | 1 | -0.0344 | -0.034 | -0.034 | -0.289 | -0.290 | -0.291 |
| pH(t-3) | 0.173 | -0.034 | -0.034 | -0.034 | 0.9674 | 0.948 | **0.935** | 0.106 | 0.103 | 0.102 |
| pH(t-2) | 0.173 | -0.034 | -0.034 | -0.034 | 1 | 0.967 | **0.948** | 0.109 | 0.106 | 0.102 |
| pH(t-1) | 0.173 | -0.035 | -0.034 | -0.034 | 0.9673 | 1 | **0.967** | 0.107 | 0.109 | 0.105 |
| pH(t) | 0.173 | -0.035 | -0.035 | -0.034 | 0.9482 | 0.967 | 1 | 0.106 | 0.107 | 0.108 |
| DO(t-3) | -0.621 | -0.290 | -0.289 | -0.288 | 0.1076 | 0.106 | 0.108 | 0.990 | 0.985 | **0.984** |
| DO(t-2) | -0.626 | -0.291 | -0.290 | -0.289 | 0.1091 | 0.107 | 0.106 | 1 | 0.990 | **0.984** |
| DO(t-1) | -0.629 | -0.292 | -0.291 | -0.290 | 0.1057 | 0.109 | 0.107 | 0.990 | 1 | **0.990** |
| DO(t) | -0.632 | -0.293 | -0.291 | -0.291 | 0.1024 | 0.105 | 0.108 | 0.982 | 0.990 | 1 |

* + 1. *Fuzzy Time Series*

FTS models are widely used in business and environmental forecasting. Compared to conventional time series analysis models, in which each intermediate output has only one real value, there is a fuzzy set to represent the intermediate output in the FTS model [34]. A brief definition of an FTS model is given below.

## Definition 1: fuzzy time series

Let

*w*(*t*)(*t*

..., 0,1, 2,...) be the water quality dataset, which is a subset of *R* , the universe of

discourse in which fuzzy sets

*fi* (*t*)(*i*

1, 2, 3,...) are defined. Assume that *F* (*t*)

is a subset of

*fi* (*t*)(*i*

1, 2, 3,...) , then *F* (*t*) is called a fuzzy time series based on *w*(*t*)(*t*

..., 0,1, 2,...) .

In Definition 1, *F* (*t*) can be treated as a linguistic variable, and *fi* (*t*) is one of the possible

linguistic values of *F* (*t*) , where *fi* (*t*) are represented by fuzzy sets. With the changing of the

universe of discourse at different times, the value of *F* (*t*) also changes.

## Definition 2: fuzzy time series relationship

Let *F* (*t*) and

*F*(*t*

1) be two fuzzy sets expressed in time series. Assuming that

*F* (*t*) is only

caused by *F*(*t* 1) , then the fuzzy logical relationship between the current state and the next state

can be represented as *F* (*t*)

*F* (*t*

1)\* *R*(*t*,*t*

1) , where \* stands for an operator.

Let *F* (*t* 1)

*Ai* and *F* (*t*)

*Aj* ; the fuzzy logical relationship between the current state and the

next state can be denoted as *Ai Aj* . The steps to predict water quality with the FTS model are as

follows:

Step 1: Define the universe of discourse and intervals based on the collected water quality

dataset, which can be represented as: *U* [min, max] . The variable min and max are the minimum

and maximum values of the target parameter.

Step 2: Define the fuzzy sets according to *U* and fuzzify the historical data of the target parameter.

Step 3: Fuzzify the observed rules of the target parameter.

Step 4: Establish fuzzy logic relationships, and group them according to the current state of the target parameter. For example, there is a fuzzy time series, *A*1 , which has three fuzzy logic

relationships: *A*1

*A*2 , *A*1

*A*3 , *A*1

*A*4 . The fuzzy logic relationships can be grouped to:

*A*1 *A*2 , *A*3 , *A*4 .

Step 5: Predict the target parameter in the testing dataset. There are two scenarios. Scenario 1:

If there is only one fuzzy logic relationship:

*Ai Aj* , then the prediction value of

*F* (*t*) is *Aj* .

Scenario 2: If there exists more than one fuzzy logic relationship: *Ai Aa* , *Ab* ,..., *An* , the prediction

value of

*F* (*t*) is equal to the mean value of

*Aa* , *Ab* ,..., *An* .

Step 6: De-fuzzify. Apply the the final prediction result [44].

# Results and Discussion

* + 1. *Water Quality Dataset*

The water quality dataset collected from the LVW between 2005 and 2010, and the BB between 2011 and 2016, are used to measure and compare the performance of different models. There are 4869 and 7502 data samples available in the two water quality datasets, respectively. The observed value of parameters of EC, TDS, and DO, obtained from the water quality dataset in LVW, are given in Figure 7. It can be seen from the figure that only parameters EC and TDS have a strong correlation. Figure 8 presents the observed value of the two parameters, EC and DO, collected at the BB water quality monitoring station. The value of EC has been divided by 1000 and 100 in the LVW and BB data, respectively, for visualization convenience. The whole dataset is split into two parts by 75% and 25% for training and testing purposes, which is the general data division percentage in data-driven research experiments.

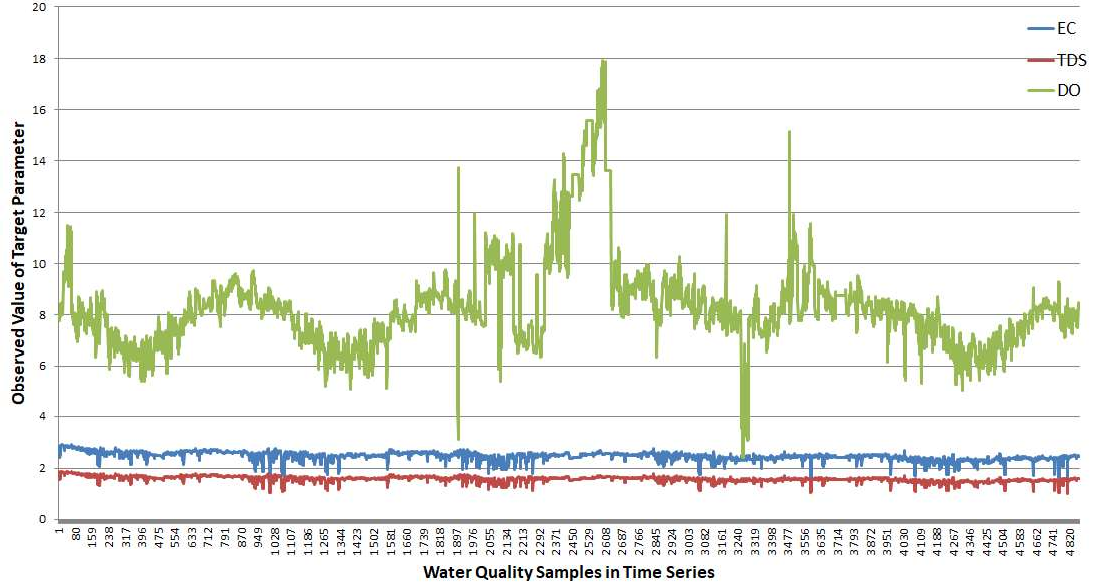


Figure 7. The observed value of five parameters in LVW between 2005 and 2010

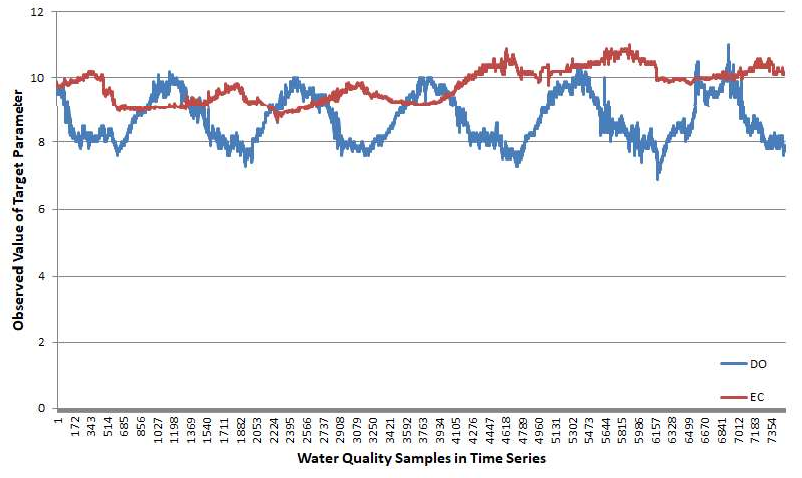


Figure 8. The observed value of four parameters in BB between 2011 and 2016

* + 1. *Experimental Configuration*

Four different kinds of models, ANN with a time-series dataset (ANN-TS), FTS, ANFIS with an original dataset, and ANFIS with time-series analysis dataset (ANFIS-TS) are implemented to investigate the prediction performance of each model. The ANN model built in this study is based on the model proposed in [3]. In the three-layer neural network, the input layer has three nodes and the hidden layer has four nodes. The activation function used is the linear activation function. Gradient descent is used to minimize the root mean square error between the true value and the prediction value in each iteration. The TensorFlow machine learning library is used to implement the ANN models. The FTS model is developed by the weighted FTS model proposed in [44]. The ANFIS model and wavelet transformation are developed by the MATLAB toolbox. A stratified sampling strategy is employed to split the collected dataset into the training and the testing subsets. The experimental results of the ANFIS and ANFIS-TS models are compared to verify the effectiveness of the time series analysis method in a dataset with a weak correlation.

* + 1. *Performance Evaluation*

The experimental results of parameters EC, DO, and TDS prediction from the LVW are given

in Table 14, 15, and 16. The top three correlation coefficients of parameter EC in *t* include TDS in *t,* EC in *t*-1 and TDS in *t*-1, which are used as the input parameters to predict EC. The experimental result from Table 14 shows that the ANFIS-TS model has the best performance, which had the smallest training and testing errors, 6.73 and 4.70 measured by RMSE, respectively. Table 12 shows that the three strongest correlation inputs of parameters DO are the data itself, collected in the past three time units. In this scenario, the FTS model has a much higher prediction accuracy than the other three models, achieving 0.23 and 0.17 in RMSE in the training and testing stages, respectively. Compared to the other three models, even the most accurate one has training and testing errors of 0.37 and 0.73 in RMSE, respectively. The parameter TDS has a similar correlation pattern with parameter EC. It has a stronger correlation with the value of EC in *t*, EC in *t*-1 and TDS in *t*-1. The experimental results show that the ANFIS-TS model also has the best training performance, and the ANFIS has the smallest testing error, which is 0.0063, compared with 0.0064 of ANFIS-TS.

The correlation coefficients between the water quality parameters from BB are presented in Table 13. Each parameter has a stronger correlation with its historical record than the other parameters. This correlation pattern is similar to the correlation of parameter DO from the LVW. The aforementioned four water quality prediction models are implemented to investigate the prediction performance with the selected input and target parameters. The experimental results are listed in Table 17 and 18. The FTS model has the smallest testing error, which could greatly reduce the prediction error. In the prediction of parameter EC from the BB dataset, the FTS model achieves the best testing performance, even though the ANFIS-TS model has a smaller training error. It shows that the FTS model is more reliable than the ANFIS-TS model in the testing stage. For parameter DO, the FTS model has the smallest error in both training and

testing stages. This experimental result proves that the FTS model works better than the ANFIS and ANFIS-TS if the target parameter only has a strong correlation with its historical data.

Table 14. The training and testing performance of different models for parameter EC in LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **EC** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **ANN-TS** | 32.92 | 1.321 | 0.966 | 53.67 | 2.114 | 0.909 |
| **FTS** | 35.00 | 0.977 | 0.959 | 44.91 | 1.310 | 0.911 |
| **ANFIS** | 7.48 | 0.159 | 0.998 | 5.12 | 0.164 | 0.999 |
| **ANFIS-TS** | **6.73** | **0.160** | **0.999** | **4.70** | **0.160** | **0.999** |

Table 15. The training and testing performance of different models for parameter DO in LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **DO** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **ANN-TS** | 0.68 | 7.290 | 0.856 | 0.90 | 8.018 | 0.746 |
| **FTS** | **0.23** | **2.015** | **0.987** | **0.17** | **1.909** | **0.960** |
| **ANFIS** | 1.49 | 11.398 | 0.315 | 1.52 | 11.503 | 0.284 |
| **ANFIS-TS** | 0.37 | 2.721 | 0.954 | 0.73 | 3.275 | 0.834 |

Table 16. The training and testing performance of different models for parameter TDS in LVW

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **TDS** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **ANN-TS** | 0.020 | 1.242 | 0.970 | 0.035 | 2.173 | 0.903 |
| **FTS** | 0.022 | 0.998 | 0.961 | 0.0279 | 1.334 | 0.916 |
| **ANFIS** | 0.004 | 0.158 | 0.999 | 0.0063 | 0.169 | 0.997 |
| **ANFIS-TS** | **0.003** | **0.162** | **0.999** | **0.0064** | **0.175** | **0.997** |

Table 17. The training and testing performance of different models for parameter EC in BB

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **EC** | | | | | |
| **Training** | | | **Testing** | | |
| **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **ANN-TS** | 11.57 | 1.060 | 0.952 | 11.35 | 1.061 | 0.954 |
| **FTS** | **4.03** | **0.358** | **0.993** | **4.70** | **0.391** | **0.975** |
| **ANFIS** | 36.99 | 3.089 | 0.510 | 37.85 | 3.163 | 0.488 |
| **ANFIS-TS** | 3.89 | 0.228 | 0.994 | 11.34 | 0.267 | 0.954 |

Table 18. The training and testing performance of different models for parameter DO in BB

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **DO** | | | | | |
|  | **Training** | | | **Testing** | | |
|  | **RMSE** | **MAPE** | **R2** | **RMSE** | **MAPE** | **R2** |
| **ANN-TS** | 0.418 | 4.428 | 0.687 | 0.629 | 6.102 | 0.293 |
| **FTS** | **0.053** | **0.551** | **0.995** | **0.075** | **0.655** | **0.989** |
| **ANFIS** | 0.406 | 3.623 | 0.705 | 0.408 | 3.655 | 0.702 |
| **ANFIS-TS** | 0.095 | 0.764 | 0.984 | 0.126 | 0.814 | 0.972 |

The predicted values of the parameters EC, DO, and TDS from the LVW using ANFIS-TS and FTS models vs. the observed values are depicted in Figure 9. The left part is the testing result obtained by the ANFIS-TS model, and the right part is the testing result obtained by the FTS model. For parameters EC and TDS, which have a strong correlation with other parameters, the experimental results from Tables 7 and 9 show that the ANFIS-TS model is a better choice in this kind of scenario. The observed value and the predicted value of the ANFIS-TS model are very close to the regression line, except for a few errors. The parameter DO only has a strong correlation with itself. The training and testing data are split in a time manner. The FTS model achieved the best performance as compared to the other three models. The prediction result of the FTS model fluctuates around the observed value, which proves that this model is accurate and reliable in this scenario.

Similar to the scenario of the DO of the LVW, the DO and EC from the BB only have strong correlations with themselves. The training and testing results from Tables 10 and 11 furtherly prove that the FTS model can perform accurate prediction for this type of parameter. Figure 10 shows that the predicted value fluctuates around the observed value except in some extreme scenarios. Compared to the ANFIS-TS model which has a few out-of-range errors, the FTS model is more accurate in this scenario.

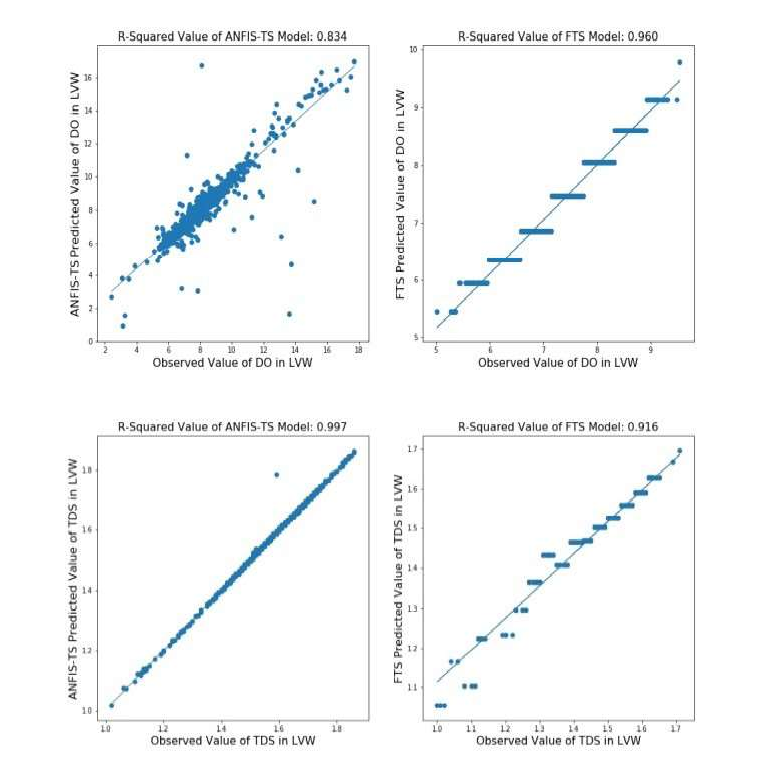
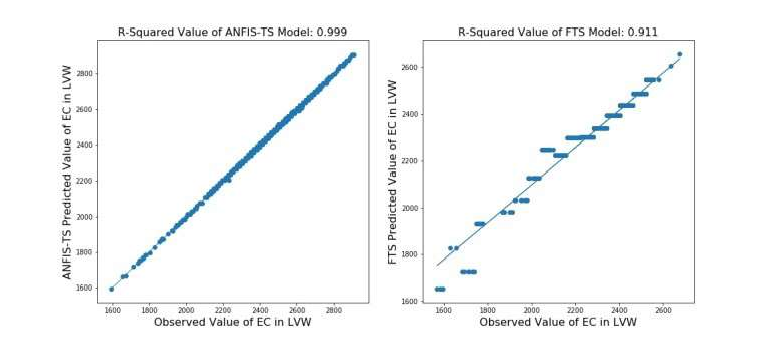


Figure 9. The coefficient of determination value of ANFIS-TS and FTS model in the testing dataset of parameter EC, DO and TDS from LVW

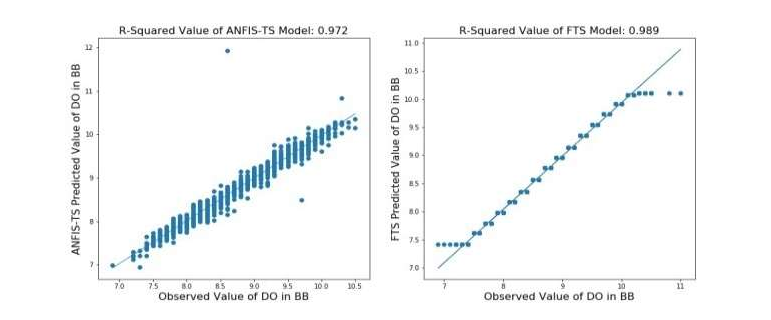
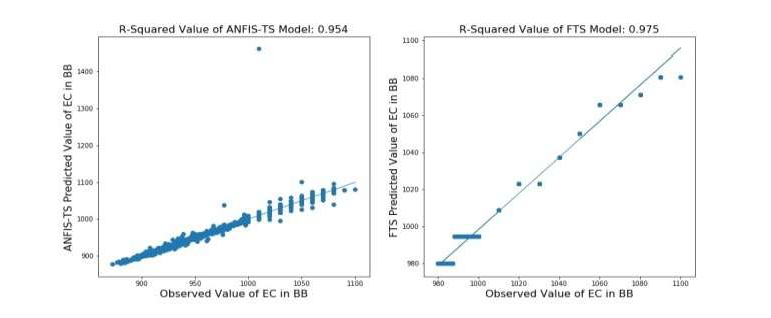


Figure 10. The coefficient of determination value of ANFIS-TS and FTS model in testing dataset of parameter EC and DO from BB

CHAPTER 6

WATER QUALITY PREDICTION BASED ON THE INTEGRATION OFANFIS MODEL WITH INTELLIGENCE ALGORITHMS

Intelligence algorithms have been proved to be an effective way to find the optimum solution in the search problem in many research fields. This chapter is focused on solving problems in Scenario 4 using intelligent algorithms. Genetic algorithm (GA) and particle swarm optimization (PSO) are the two most widely used intelligence algorithms. Both of them are integrated with the ANFIS model to improve the prediction accuracy in predicting water quality parameters when input parameters and the output parameter have a weak correlation and the datasets are collected from different monitoring stations. Instead of finding out a strong correlation input parameter by time series analysis as in Chapter 5, a raw dataset will be used to predict the target parameter.

The organization of the remainder of the chapter is as follows. Materials and methods are given in Section 2. The wastewater quality prediction system is introduced in Section 3. Section 4 presents the experimental configuration and results.

# Materials and Methods

* + 1. *Study Area*

In this study, the dataset collected from the LVW since 2007 is used to verify the effectiveness of the proposed model. As shown in Figure 1, there are six water quality monitoring stations along the LVW. Unlike the previous two chapters, only the dataset from station LW3.4 is collected for training and testing purposes. In this chapter, all the datasets from the following six water quality monitoring stations, LW11.0, LW8.85, LW6.05, LW3.7 LW3.4 LW0.55, are collected to evaluate the performance of the different model.

* + 1. *Water Quality Parameters*

In this chapter, the water quality dataset collected by the SNWA is selected as the testing dataset because of the high sampling frequency and more effective water quality parameters. The SNWA samples and analyzes the water in LVW every two weeks. Totally 2218 wastewater quality samples collected by SNWA at the six water quality monitoring stations since 2007 are used to evaluate the model. In each water sample, there are more than fifty water quality parameters. But only eleven of them, chloride, EC, DO, fluoride, TN, pH, sulfate, AT, WT, TDS, and TSS, are selected because the eleven parameters have continuous and effective record in all six water quality monitoring stations. Table 19 lists the statistical measures of the eleven parameters. The first and second columns list the parameters and corresponding units. Columns three to six show the statistical information of each parameter. The last column lists the MCL allowed by national drinking water regulations [41].

Table 19. A statistical measure of water quality parameters at LW3.4 and LW3.7 in the Las Vegas Wash.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Unit** | **Min** | **Max** | **Mean** | **Median** | **SD** | **MCL** |
| **Chloride** | **mg/L** | 10 | 464 | 281.6794 | 286 | 46.68805 | 250 |
| **EC** | **uS/cm** | 18 | 5968 | 2390.52 | 2206 | 621.0961 | 2000 |
| **DO** | **mg/L** | 5.08 | 23.11 | 8.951776 | 8.74 | 1.818756 | 6.5~12 |
| **Fluoride** | **mg/L** | 0.09 | 1.67 | 0.772101 | 0.77 | 0.153872 | 2.0 |
| **TN** | **mg/L** | 0.65 | 18.98 | 11.04379 | 12.65 | 4.395493 | 10 |
| **pH** | Units | 6.79 | 11.12 | 8.068395 | 8.12 | 0.356384 | 6.5~9.2 |
| **Sulfate** | **mg/L** | 52 | 2597 | 705.9179 | 531 | 408.0432 | 250 |
| **AT** | deg F | 30.1 | 122.1 | 73.84243 | 73.75 | 18.06446 | N/A |
| **WT** | deg C | 1.21 | 32.37 | 21.5918 | 21.99 | 5.317761 | N/A |
| **TDS** | **mg/L** | 215 | 5084 | 1725.757 | 1479 | 654.5504 | 500 |
| **TSS** | **mg/L** | 2 | 2646 | 29.16758 | 9 | 109.7146 | N/A |

This study is focused on the key parameters that will strongly affect the health of human life and aquatic organism. The parameter TDS is used to measure the combined total of organic and

inorganic substances dissolved in the water. Monitoring and predicting the TDS level is very important to measure the quality of drinking water resources. The parameter DO means the amount of oxygen existed in the water. Many aquatic organisms cannot live when the level of DO is too high or too low. The two parameters, TDS and DO, are the major indicators that quantify the quality of wastewater. They are selected as the prediction target in this study.

Table 20. Correlation matrix of all water quality parameters listed in Table 19

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Chloride | EC | **DO** | Fluoride | N | pH | Sulfate | AT | WT | **TDS** | TSS |
| Chloride | 1 | 0.571 | 0.127 | 0.375 | 0.229 | 0.316 | 0.366 | 0.051 | -0.112 | 0.463 | -0.202 |
| EC | 0.571 | 1 | 0.438 | -0.202 | -0.539 | 0.255 | 0.939 | -0.185 | -0.473 | 0.955 | -0.05 |
| DO | 0.127 | 0.438 | 1 | -0.088 | -0.496 | 0.525 | 0.467 | -0.36 | -0.595 | 0.443 | -0.062 |
| Fluoride | 0.375 | -0.202 | -0.088 | 1 | 0.584 | -0.038 | -0.359 | 0.083 | 0.253 | -0.292 | -0.229 |
| TN | 0.229 | -0.539 | -0.496 | 0.584 | 1 | -0.207 | -0.687 | 0.181 | 0.395 | -0.619 | -0.184 |
| pH | 0.316 | 0.255 | **0.525** | -0.038 | -0.207 | 1 | 0.21 | 0.158 | -0.165 | **0.223** | 0.008 |
| Sulfate | 0.366 | 0.939 | **0.467** | -0.359 | -0.687 | 0.21 | 1 | -0.238 | -0.53 | **0.986** | 0.029 |
| AT | 0.051 | -0.185 | -0.36 | 0.083 | 0.181 | 0.158 | -0.238 | 1 | 0.827 | -0.216 | 0.041 |
| WT | -0.112 | -0.473 | **-0.595** | 0.253 | 0.395 | -0.165 | -0.53 | 0.827 | 1 | **-0.507** | 0.004 |
| TDS | 0.463 | 0.955 | 0.443 | -0.292 | -0.619 | 0.223 | 0.986 | -0.216 | -0.507 | 1 | 0.005 |
| TSS | -0.202 | -0.05 | -0.062 | -0.229 | -0.184 | 0.008 | 0.029 | 0.041 | 0.004 | 0.005 | 1 |

The correlation between all the water quality parameters listed in Table 19 is given in Table 20. The selection of input parameters for predicting DO and TDS needs simultaneously consider the correlation value and the convenience of collecting these parameters. Based on the consideration, parameters pH, sulfate, and WT are selected as the input parameters. It is very easy to quantize pH and WT. Out of the three parameters, only parameter sulfate takes some time to quantize the value. The correlation between DO and three selected parameters is very weak. All three correlation value is smaller than 0.75. Regarding parameter TDS, only sulfate has strong correlation with TDS. Both AT and pH are weakly correlated with TDS and WT even has negative impact on the value of parameter TDS. On the other hand, the dataset study here is a

mix of data samples collected from five independent water quality monitoring stations. FTS model is more suitable to the dataset which has a long record in one single station. Therefore, the models proposed in Chapters 4 and 5 are not applicable in this scope. In this chapter, an integration of intelligence algorithms with the ANFIS model is proposed to predict parameters DO and TDS in all monitoring stations of LVW.

* + 1. *Genetic Algorithm*

The GA is a kind of metaheuristic algorithm which is proposed by Holland [45] in 1975. It is

    -



optimum solutions in many search problems. The GA algorithm consists of five independent steps: initial population, fitness function evaluation, selection, crossover, and mutation. In the beginning, a set of individuals is randomly generated or provided by the problem. Then, a set of solutions is generated and the whole set is named as *initial population*. Each solution is made up of a set of parameters and each parameter is named a *gene*. In Step 2, the *fitness* function will be used to evaluate the performance of each solution based on the fitness score generated by the fitness function. Then, a part of the solutions with the higher fitness score are selected out for reproduction. In the *crossover* step, the selected solutions are paired with each other to generate new solutions. A pair of solutions exchanges a part of genes with each other at a random crossover point within the solution. The last step is the *mutation* operation. The system will randomly change the genes of a solution with a low probability which is set up based on the mutation probability in nature evolution [46]. A new solution is generated with a few genes changed if a mutation operation happens. After the mutation process, the system will go back to Step 2 until the termination criteria is met or the maximum loop is reached.

* + 1. *Particle Swarm Optimization*

The PSO is a kind of metaheuristic algorithms which optimizes the solution of a problem by iteratively improving the candidate solutions based on the given evaluation metrics. It is inspired by the movement of fish and flock when they try to find the best path to reach the food. This algorithm has been proved to be robust and reliable in the optimization of nonlinear problems [47]. Many researchers have used this algorithm in many different fields.

Each particle in a PSO represents a candidate solution of a problem. The particle randomly moves around the given search space according to the current location and the velocity calculated by mathematical formulation. The velocity of the particle is influenced by the local best known position and global best position found by other particles. The particle stops moving until the position cannot be optimized any more.

A *particle* represents a candidate solution in a problem. Assuming there are *n variables* in a particle and each variable can be treated as an option in a dimension. A swarm of *m* particles is randomly generated in the *n* dimensional space as the beginning candidate solution. Let

*xi* (*t*

1) ( *xi*, 1 (*t*

1), *xi*, 2 (*t*

1),..., *xi*, *n* (*t*

1)) and *vi* (*t*

1) (*vi* , 1 (*t*

1), *vi* , 2 (*t*

1),..., *vi* , *n* (*t*

1)) be the location and

velocity of the particle *i* ( *i*

[1, *n*] ) in the cycle (*t*

1) , respectively. Then the location and

velocity in the cycle *t* can be represented as:

*xi* (*t*)

*xi* (*t* 1)

*vi* (*t*)

(20)

*vi* (*t*)

*wvi* (*t* 1)

1 (*xPbest*

*xi* (*t*))

2 (*xGbest*

*xi* (*t*))

(21)

where *w* , 1 , and 2 are the inertia weight, personal learning coefficient, and global learning

*i*

*i*

coefficient, respectively. The

*xPbest* and

*xGbest* stand for the best position found by the particle *i* and

the best position found by the particle swarm so far, respectively. In this study, the default value of *w* , 1 , and 2 are set up as 1, 1, and 2, respectively.

*i*

*i*

# System Overview

The flowchart of the water quality prediction system built in this study is given in Figure 11. In the beginning, the dataset collected from SNWA is used to generate an initial ANFIS model based on fuzzy c-means clustering. The number of clusters will determine the number of rules and membership functions in the generated fuzzy inference system. In this study, the number of parameter *cluster* \_ *n* is set up as *auto* when generating a fuzzy inferences system. It means that

the system will find out the optimum number of clusters in the dataset by fuzzy c-means clustering then decide the number of rules and membership functions to be used in the generated ANFIS model. Then the system has two independent options, PSO or GA. If GA is selected to optimize the population set, PSO will be disabled, vice versa. One particle in the PSO has the same meaning as one population in GA in this problem. Both of them correspond to one type of configuration for an ANFIS model. The parameters of the input and output membership functions are used as the initial population. The initial population will be used to generate a set of candidate populations in the system by a set of uniformly distributed random numbers in both the GA and PSO model. The iteration number is checked before the start of the optimization process. Then the candidate populations are evaluated by a scoring function. One population can stay in the population set only when it has a score higher than the minimum requirement.

The GA or PSO will be used to optimize the selected population set. If GA is selected, the crossover and mutation operation will be used to optimize the selected population set. Otherwise, PSO will be selected to improve the selected population set. After the optimization operation by GA and PSO, each population in the population set will be used to generate an independent ANFIS model. The system will evaluate the performance of each solution in the current iteration and only the configuration and experimental result of the best solution will be stored. The system

will repeat this process until the iteration number is larger than the maximum limit. Then the system will be terminated and the best result will be output as the final result.



Data collection

Iter < MAX

Fitness evaluation of the population set

Candidate population selection

Optimize each population which contains the configuration parameters of an ANFIS model based on the selected method

Performance evaluation

Generating initial population set

Optimization method selection: PSO or GA

Generating initial ANFIS model

Generate an ANFIS model based on the parameters in each population

Train the ANFIS model by the training dataset

Evaluating the performance of each model by testing dataset

Store the best result and the population

Terminate the process, output the best result and the population.

Figure 11. Steps of hybrid GA and PSO with ANFIS model.

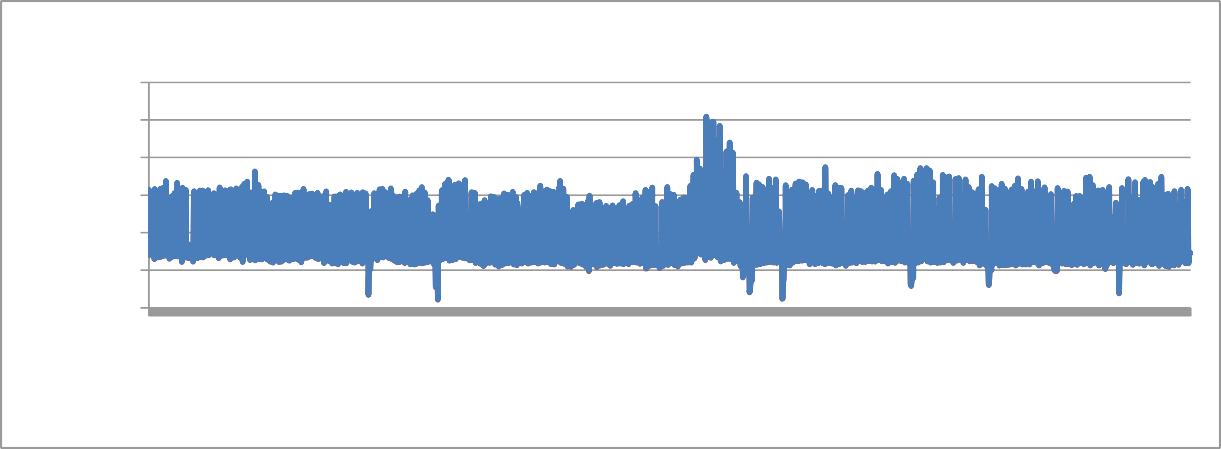
# Experimental Results and Discussion

* + 1. *Water Quality Dataset*

The water quality dataset collected by SNWA in water quality monitoring stations in LW11.0, LW8.85, LW6.05, LW3.7, LW3.4, and LW0.55 is used to evaluate the prediction accuracy of the proposed wastewater quality prediction model. There are 2218 water quality data samples available in the above six water quality monitoring stations. As explained in 6.1.2, parameters DO and TDS are selected as the target parameter because of its importance to human life and aquatic organism. The observed value of the two target parameters is presented in Figures 12 and

13. The parameter DO indicates the oxygen level in the water and TDS quantizes the total solids dissolved in the water. From the Figure below, it can be seen that the values of parameter TDS are higher than the MCL limit (500) for drinking water nearly in 99% of the time. On the other hand, the values of parameter DO are higher or lower than the MCL limit (6.5~12) in some days of the year. If the value of DO is continuously out of the range of the MCL limit, it will strongly affect the life of the aquatic organism. Therefore, timely prediction of the water quality is still very important. The water quality dataset collected from the SNWA is split into two parts by 80% and 20% for training and testing purposes, respectively.

Figure 12. The value of parameter TDS from the six water quality monitoring stations in LVW since 2007.



**The Value of Parameter TDS Since 2007**

6000

5000

4000

3000

2000

1000

0

**The Number of Each Sample**



**The Value of Parameter DO Since 2007**

25

20

15

10

5

0

**The Number of Each Sample**

**The Value of TDS (mg/L)**

**The Value of DO (mg/L)**

1

71

141

211

281

351

421

491

561

631

701

771

841

911

981

1051

1121

1191

1261

1331

1401

1471

1541

1611

1681

1751

1821

1891

1961

2031

2101

2171

Figure 13. The value of parameter DO from the six water quality monitoring stations in LVW since 2007.

1

71

141

211

281

351

421

491

561

631

701

771

841

911

981

1051

1121

1191

1261

1331

1401

1471

1541

1611

1681

1751

1821

1891

1961

2031

2101

2171

* + 1. *Experiment Configuration*

Three different kinds of models, ANFIS, the integration of ANFIS with GA (ANFIS-GA), and the integration of ANFIS with PSO (ANFIS-PSO), are evaluated and compared in this study based on the dataset introduced above. A brief introduction of the ANFIS model is given in Chapter 3. All the three models evaluated in this part are developed by the MATLAB toolbox. The ANFIS-GA and ANFIS-PSO models are developed based on the work in [48]. As reviewed

in Section 2.3, the generalized bell-shaped membership function is selected because it has been proven to have better performance than other membership functions in water quality prediction. On the other hand, there are only constant and linear output membership functions available in MATLAB toolbox. The linear function is chosen because of its higher prediction accuracy [30].

* + 1. *Performance Evaluation*

The performance of the three models in the prediction of parameter DO is listed in Table 21. Three evaluation metrics, RMSE, MAPE, and R2, are used to measure the prediction accuracy. It can be seen from Table 21 that ANFIS-PSO has the smallest training and testing error. The result indicates that ANFIS-PSO is the best choice for predicting parameter DO. Meanwhile, the ANFIS-GA model outperforms the ANFIS model with a testing RMSE of 1.248 over 1.263. This proves that evolutionary algorithms can improve the prediction accuracy of the ANFIS model by optimizing the parameters in the membership function. Comparing ANFIS-GA with ANFIS-PSO, it can be concluded that PSO can more accurately optimize the parameters than the GA.

Table 21. The training and testing performance of different models for parameter DO in LVW

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **DO** | | | | | | | |
| **Training** | | | | **Testing** | | | |
| **MSE** | **RMSE** | **MAPE** | **R2** | **MSE** | **RMSE** | **MAPE** | **R2** |
| **ANFIS** | 1.441 | 1.200 | 7.380 | 0.172 | 1.595 | 1.263 | 7.492 | 0.184 |
| **ANFIS-GA** | 1.076 | 1.037 | 7.092 | 0.519 | 1.557 | 1.248 | 7.395 | 0.261 |
| **ANFIS-PSO** | **0.808** | **0.899** | **6.013** | **0.681** | **1.333** | **1.155** | **6.463** | **0.461** |

The experimental result of parameters TDS based on the ANFIS, ANFIS-GA, and ANFIS- PSO is given in Table 22. The prediction result is nearly the same as the parameter DO. The ANFIS-PSO model achieves the best performance among the three models. The testing RMSE error is 85.69 and MAPE is only 3.105%. This means that the ANFIS-PSO model is very

accurate and reliable in TDS prediction. The testing error of the ANFIS-GA model is a bit larger than the ANFIS-PSO model. The ANFIS model has the largest predicting RMSE error which is

100.39. The experiment result of parameter TDS furtherly proves the intelligence algorithms can improve the performance of the ANFIS model and the PSO model outperforms the GA model in optimizing the parameters in the ANFIS model.

Table 22. The training and testing performance of different models for parameter TDS in LVW

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Models** | **TDS** | | | | | | | |
| **Training** | | | | **Testing** | | | |
| **MSE** | **RMSE** | **MAPE** | **R2** | **MSE** | **RMSE** | **MAPE** | **R2** |
| **ANFIS** | 11850 | 108.86 | 4.322 | 0.971 | 10078 | 100.39 | 4.456 | 0.978 |
| **ANFIS-GA** | 9796 | 98.98 | 3.479 | 0.977 | 8761 | 93.60 | 3.602 | 0.981 |
| **ANFIS-PSO** | **7766** | **88.13** | **2.643** | **0.981** | **7343** | **85.69** | **3.105** | **0.984** |

The detailed testing result of parameter DO and TDS based on the ANFIS-PSO model is shown in Figures 14 and 15, respectively. The testing result of parameter DO is shown in Figure 14 (a). The black line represents the observed value of parameter DO and the red line is the predicted value based on the ANFIS-PSO model. It can be seen from the upper figure that the prediction value and true value of parameter DO nearly overlap, which demonstrates that the proposed model is reliable. The RMSE and MSE of the testing result are given in Figure 14 (b). A few samples are having abnormal large prediction errors, which are the out-of-range errors occurring in the ANFIS model introduced in Chapter 3. It happens because the data sample size (2218) is still kind of small for  range. There are a few rarely seen water conditions, like a sudden change of factory discharge, weather change, and human behaviors. The out-of-range error can be eliminated when more data samples are collected and the

prediction error can be furtherly reduced. The normal distribution of the testing RMSE is given in Figure 14 (c). The RMSE of parameter DO is less than one in over 93% of the testing cases.

Figure 15 presents the detailed testing result of parameter TDS based on ANFIS-PSO model. The trend is nearly the same with parameter DO. The line of prediction value nearly covers the line of true value in Figure 15 (a), which indicates that the ANFIS-PSO model can accurately formulate the relationship hidden in the dataset. The prediction error of each sample is given in Figure 15 (b). The model is working well except for a few out-of-range errors occurring in the proposed model. The reason for the out-of-range error is as same as that of parameter DO. The histogram of the testing RMSE of TDS is given in Figure 15 (c). It can be seen from the prediction error of normal distribution in Figure 15 (c) that 96% of the testing samples have a prediction error which is less than 200. The RMSE distributions furtherly confirm that the proposed model is reliable in predicting parameters DO and TDS.

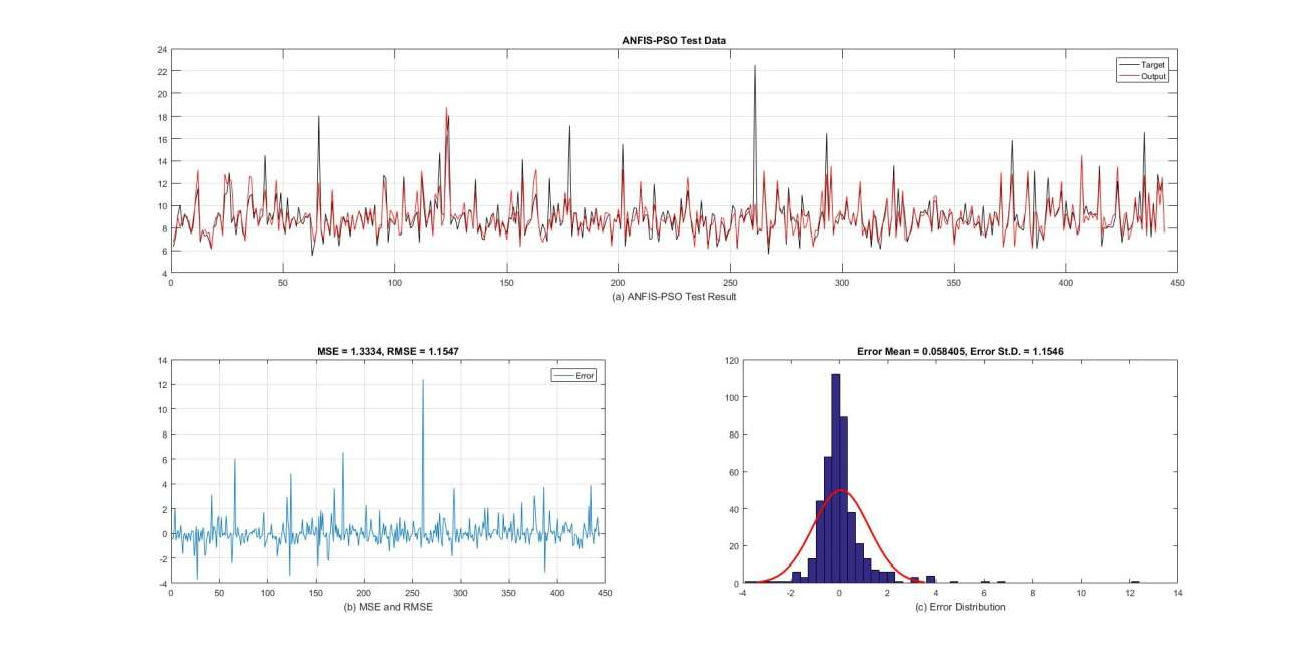


Figure 14. The testing result of parameter DO based on ANFIS-PSO model.

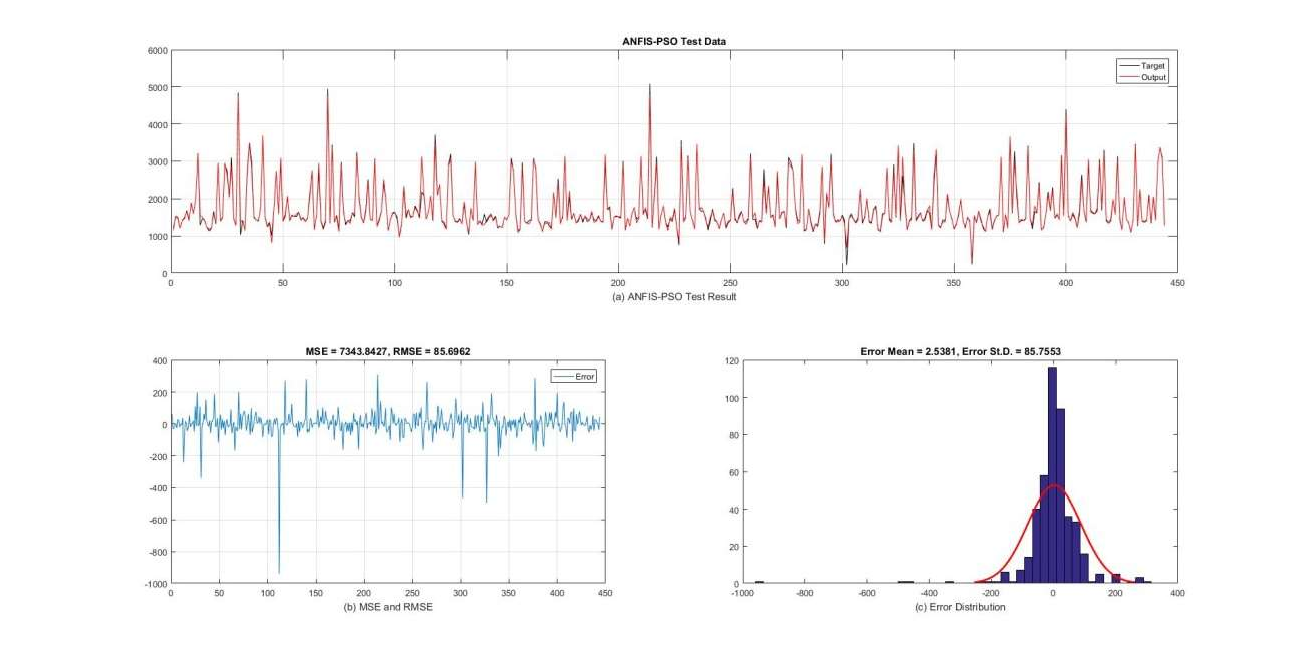


Figure 15. The testing result of parameter TDS based on ANFIS-PSO model.

CHAPTER 7 CONCLUSIONS AND FUTURE WORK

# Conclusions

In this dissertation, the water quality prediction problems are classified into four categories according to the dataset size. Several machine learning methods based on the ANFIS model are proposed to improve the prediction performance for problems falling into Scenarios 2 and 3. First, to eliminate out-of-range errors of the ANFIS model in the testing stage, the stratified sampling strategy is used for mitigating the uneven distribution of training and testing datasets. A general framework of water quality prediction system based on wavelet de-noised ANFIS model using stratified sampling is proposed.

For problems in Scenario 2, i.e., medium size dataset and strong correlation between parameters, wavelet transform is integrated with ANFIS model to predict wastewater discharge quality parameters related to salinity levels in Chapter 4. The initial dataset is preprocessed with a statistical stratified sampling strategy. Experimental results show that the proposed model with stratified sampling outperforms MLR, ANNs, ANFIS, EANFIS, and WT-ANFIS models. This demonstrates that the proposed model, with stratified sampling and a general-purpose input parameter selection method, is reliable to model the quality of parameters related to salinity. This model can be applied to reduce the number of parameters monitored to lower the cost associated with monitoring the quality of wastewater discharges.

For problems in Scenario 3, i.e., medium size dataset and weak correlation between parameters, two types of solutions are proposed for datasets with long history monitored at one single station and datasets monitored at multiple stations, respectively. In Chapter 5, for the first type of dataset, the time series analysis method is used to preprocess the water quality dataset to

figure out appropriate input parameters for the ANFIS model. Besides, the FTS model is also applied to this scenario. The experimental results on two water quality datasets show that the FTS model could accurately predict the value of a target parameter when the target parameter has a strong correlation with its historical record, such as the parameter DO from the LVW, and the parameters DO and EC from the BB. On the other hand, when the target parameter has a strong correlation with other parameters, except itself, like parameters EC and TDS from the LVW, the ANFIS-TS model achieved better prediction accuracy over other models. This demonstrates that using the FTS and ANFIS models, integrated with time-series analysis, is an effective and reliable tool to model water quality, even when the correlation between the original parameters is weak.

In Chapter 6, for the second type of dataset, the GA and PSO are used to optimize the parameter in membership function of the ANFIS model. It can be seen from the experimental results that ANFIS-GA and ANFIS-PSO outperform the pure ANFIS model in predicting parameters DO and TDS when the correlation between the input parameters and output parameters is weak even when the time series impact is considered. ANFIS-PSO achieves the best performance in both training and testing data. This proves the ANFIS-PSO is a reliable and robust model to predict water quality.

# Future Work

It can be seen from the experimental results from Chapter 6, both GA and PSO can improve the performance of the ANFIS model. The two algorithms are the most widely used evolutionary algorithms. In the future, more evolution algorithms, like differential evolution and ant colony optimization, will be explored to forecast the water quality. On the other hand, ideas from other data processing methods, like boost learning and weighted timing analysis will be investigated.

More data sources are required to verify the reliability and robustness of the proposed models. So far, the water quality dataset from the LVW collected by Southern Nevada Water Authority and Las Vegas Wash Coordination Committee, and dataset collected from Boulder Basin have been used as the experimental dataset. In the future, more efforts will be made to find more datasets to build a more reliable water quality prediction model.

APPENDIX

The abstract, chapter 1, 2, 3, and 4 are reprinted from Journal of Computers and Electrical Engineering, Zhao Fu, Jiao Cheng, Mei Yang, Jacimaria Batista, and Yingtao Jiang,

Wastewater Discharge Quality Prediction using Stratified Sampling and Wavelet De-Noising ANFIS Model, Vol. 85, pp, 1-15, Jun. 2020.

The abstract, chapter 1, 2, and 5 are reprinted from International Journal of Intelligent Systems and Applications (IJISA), Zhao Fu, Mei Yang, and Jacimaria Batista Using Fuzzy Models and Time Series Analysis to Predict Water Quality, Vol. 12, No. 2, pp. 1-10, Apr. 2020

The abstract, chapter 1, 2, 3, and 4 are reprinted from IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Zhao Fu, Jiao Cheng, Mei Yang, Jacimaria Batista, and Yingtao Jiang Prediction of Industrial Wastewater Quality Parameters Based on Wavelet De-noised ANFIS Model, pp. 301-306, Jan. 2018.

In this dissertation, the ANFIS model is built based on the ANFIS toolbox [23] and hybrid ANFIS model is built based on the framework proposed in [48]. The FTS model is built based on the framework proposed in [49]. The MLR and ANN model are built by Python. The following code segmentions lists the code sketch for the major code blocks.

# calculate correlation relationship import pandas as pd

import numpy as np

data = pd.read\_csv("dataset.csv") data = data.iloc[:, 4:]

corr = np.corrcoef(data.T)

np.savetxt('corr.csv', corr, delimiter=',', fmt='%.3f')

# calculate statistical information import pandas as pd

import numpy as np

data\_median = data.median(axis=0) data\_mean = data.mean(axis=0) data\_std = data.std(axis=0) data\_max = data.max(axis=0) data\_min = data.min(axis=0) data\_var = data.var(axis=0)

dataset\_stat = pd.DataFrame([data\_min.values, data\_max.values, data\_mean.values, data\_median.values, data\_std.values, data\_var.values],

columns=['Chloride', 'EC','DO', 'Fluoride','N','pH','Sulfate', 'T (deg F)','T Water(deg C)', 'TDS','TSS'])

dataset\_stat.to\_csv("dataset\_stat.csv")

# MLR model to predict TDS import pandas as pd

import numpy as np

from sklearn import linear\_model

data\_train = pd.read\_csv("TDS\_rand\_train.csv") data\_test = pd.read\_csv("TDS\_rand\_test.csv")

#data\_train = pd.read\_csv("TDS\_stra\_train.csv") #data\_test = pd.read\_csv("TDS\_stra\_test.csv") train = np.matrix(data\_train.iloc[:, 9:14])

test = np.matrix(data\_test.iloc[:, 9:14]) X\_train = train[:,:4]

y\_train = train[:,4] X\_test = test[:,:4] y\_test = test[:,4]

# create linear regression model

model = linear\_model.LinearRegression() # training model

model.fit(X\_train, y\_train) # verify and test model

skl\_test = model.predict(X\_test)

# use traning data as verifying data skl\_train = model.predict(X\_train) # denormalize

hy\_te = skl\_test \* np.sqrt(data\_var[3]) + data\_mean[3] hy\_trn = skl\_train \* np.sqrt(data\_var[3]) + data\_mean[3]

# evaluate training and testing result from sklearn.metrics import r2\_score train\_r2 = r2\_score(hy\_trn, y\_trn)

print ("Train r2:", train\_r2) test\_r2 = r2\_score(hy\_te, y\_te) print ("Test r2:", test\_r2)

train\_mape = (abs(hy\_trn - y\_trn)/y\_trn).mean(axis=0) print ("Train MAE:", train\_mape)

test\_mape = (abs(hy\_te - y\_te)/y\_te).mean(axis=0) print ("Test MAE:", test\_mape)

train\_rmse = np.sqrt(sum(np.asarray(hy\_trn - y\_trn) \*\* 2) / len(y\_trn)) test\_rmse = np.sqrt(sum(np.asarray(hy\_te - y\_te) \*\* 2) / len(y\_te))

print("Training RMSE: {0} and Testing RMSE: {1}".format(train\_rmse, test\_rmse))

#ANN model to predict TDS

from sklearn.preprocessing import normalize import pandas as pd

import numpy as np

# change the input data folder for different prediction target x = pd.read\_csv("TDS\_trn.csv")

y = pd.read\_csv("TDS\_test.csv")

X\_train = x.iloc[:, 0:3].values

y\_train = x.iloc[:, 3].values.reshape(1774,1)

X\_test = y.iloc[:, 0:3].values

y\_test = y.iloc[:, 3].values.reshape(444,1)

# data standardization

X\_train\_mean = X\_train.mean(axis=0) X\_train\_var = X\_train.var(axis=0)

X\_train = (X\_train - X\_train\_mean) / np.sqrt(X\_train\_var) y\_train\_mean = y\_train.mean(axis=0)

y\_train\_var = y\_train.var(axis=0)

y\_train = (y\_train - y\_train\_mean) / np.sqrt(y\_train\_var) X\_test\_mean = X\_test.mean(axis=0)

X\_test\_var = X\_test.var(axis=0)

X\_test = (X\_test - X\_test\_mean) / np.sqrt(X\_test\_var) y\_test\_mean = y\_test.mean(axis=0)

y\_test\_var = y\_test.var(axis=0)

y\_test = (y\_test - y\_test\_mean) / np.sqrt(y\_test\_var)

# neural network

import tensorflow.compat.v1 as tf tf.disable\_v2\_behavior()

# define the neural network structure

x\_data = tf.placeholder(tf.float32, shape=[None, 3]) y\_target = tf.placeholder(tf.float32, shape=[None, 1]) hidden\_layer\_nodes = 10

A1 = tf.Variable(tf.random\_normal(shape=[3,hidden\_layer\_nodes]))

b1 = tf.Variable(tf.random\_normal(shape=[hidden\_layer\_nodes])) A2 = tf.Variable(tf.random\_normal(shape=[hidden\_layer\_nodes,1])) b2 = tf.Variable(tf.random\_normal(shape=[1]))

hidden\_output = tf.add(tf.matmul(x\_data, A1), b1) final\_output = tf.add(tf.matmul(hidden\_output, A2), b2)

#define the optimization scheme

loss = tf.sqrt(tf.reduce\_mean(tf.square(y\_target - final\_output))) my\_opt = tf.train.GradientDescentOptimizer(0.01)

train\_step = my\_opt.minimize(loss) init = tf.global\_variables\_initializer() with tf.Session() as sess:

loss\_vec = [] test\_loss = [] sess.run(init)

for i in range(10000): tf.set\_random\_seed(2) np.random.seed(2)

# training and testing

sess.run(train\_step, feed\_dict={x\_data: X\_train, y\_target: y\_train})

temp\_loss, train\_pred = sess.run([loss,final\_output], feed\_dict={x\_data: X\_train, y\_target: y\_train})

loss\_vec.append(np.sqrt(temp\_loss))

test\_temp\_loss, test\_pred = sess.run([loss,final\_output], feed\_dict={x\_data: X\_test, y\_target: y\_test})

test\_loss.append(np.sqrt(test\_temp\_loss))

# store the traing and testing result

train\_pred = train\_pred \* np.sqrt(y\_train\_var) + y\_train\_mean y\_train = y\_train \* np.sqrt(y\_train\_var) + y\_train\_mean y\_test = y\_test \* np.sqrt(y\_test\_var) + y\_test\_mean

test\_pred = test\_pred \* np.sqrt(y\_test\_var) + y\_test\_mean np.savetxt("TDS\_ANN\_train.csv", train\_pred, delimiter=',', fmt='%.8f') np.savetxt("TDS\_ANN\_test.csv", test\_pred, delimiter=',', fmt='%.8f')

# calculate MAPE, R-Square and RMSE from sklearn.metrics import r2\_score

train\_MAPE = (abs(train\_pred - y\_train)/y\_train).mean(axis=0) print ("Train MAE:", train\_MAPE)

test\_MAPE = (abs(test\_pred - y\_test)/y\_test).mean(axis=0) print ("Test MAE:", test\_MAPE)

train\_r2 = r2\_score(y\_train, train\_pred) print ("Train r2:", train\_r2)

test\_r2 = r2\_score(test\_pred, y\_test) print ("Test r2:", test\_r2)

train\_rmse = np.sqrt(sum(np.asarray(train\_pred - y\_train) \*\* 2) / len(train\_pred))

test\_rmse = np.sqrt(sum(np.asarray(test\_pred - y\_test) \*\* 2) / len(test\_pred)) print("Training RMSE: {0} and Testing RMSE: {1}".format(train\_rmse, test\_rmse))

REFERENCES

[1] L. Metcalf and H. P. Eddy, Wastewater Engineering: Treatment and Resource Recovery, 5th Edition. McGraw Hill, Chapter 3, 2014.

 *Journal of Water*, vol. 8, no. 2, pp. 1-15, Jan. 2016

1. A. H. Zare, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," *Journal of Environmental Health Science & Engineering*, vol. 12, no. 1, pp. 1-8, Jan. 2014.
2. L. Li, P. Jiang, H. Guang, L. Dong, G. Wu, and H. Wu " Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China," *Environmental Science and Pollution Research*, vol. 26, no. 19, pp. 19879-19896, Mar. 2019.
3. J.-S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665-685, May. 1993.
4. A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring," *Environmental Science and Pollution Research*, vol. 21, no. 3, pp. 1658 1670, Aug. 2013.
5. J. Wan, M. Huang, Y. Ma, W. Guo, Y. Wang, H. Zhang, W. Li, and X. Sun, "Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system," *Applied Soft Computing*, vol. 11, no. 3, pp. 3238-3246, Apr. 2011.
6. W. Deng, G. Wang, and X. Zhang, "A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting," *Chemometrics and Intelligent Laboratory Systems,* vol. 149, pp. 39-49, Dec. 2015.
7. P. Singh, "Rainfall and financial forecasting using fuzzy time series and neural networks based model," *International Journal of Machine Learning and Cybernetics,vol. 9, no. 3,* pp. 491- 506, May. 2018.
8. A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, and R. N. Sankhua, "Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India," *Modeling Earth Systems and Environment,* vol. 5, no. 3, pp. 951-96, Mar. 2019.
9. A. Sarkar and P. Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique," *Aquatic Procedia,* vol. 4, pp. 1070-1077, 2015.
10. A. P. Piotrowski, M. J. Napiorkowski, J. J. Napiorkowski, and M. Osuch, "Comparing various artificial neural network types for water temperature prediction in rivers," *Journal of Hydrology,* vol. 529, no. 1, pp. 302-315, Oct. 2015.
11. A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Application of artificial neural networks for water quality prediction," *Neural Computing and Applications*, vol. 22, no. 1, pp. 187-201, Apr. 2012.
12. L. Zhang, G. X. Zhang, and R. R. Li, "Water quality analysis and prediction using hybrid time series and neural network models," *Journal of Agricultural Science and Technology*, vol. 18, no. 4, pp. 975-983, Dec. 2015.
13. M. J. Alizadeh and M. R. Kavianpour, "Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean," *Marine Pollution Bulletin,* vol. 98, no. 1, pp. 171-178, 2015.
14. A. G. Mohammad, K. Reza, C. Kwok-Wing, S. Shahaboddin, and T. G. Pezhman,

   -behaved

     *Engineering*

*Applications of Computational Fluid Mechanics*, vol. 12, no. 1, pp. 724-737, Sep. 2018.

1. A. A. Nadiri, S. Shokri, F. T.Tsai, and A. A. Moghaddam, "Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model," *Journal of Cleaner Production*, vol. 180, no. 7, pp. 539-549, Jan. 2018.
2. M. Khadr and M. Elshemy, "Data-driven modeling for water quality prediction case study: The drains system associated with Manzala Lake, Egypt," *Ain Shams Engineering Journal*, pp. 1-9, Sep. 2016.
3. A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," *Journal of King Saud University - Engineering Sciences,* vol. 29, no. 3, pp. 237-243, Jul. 2017.
4. S. Tiwari, R. Babbar, and G. Kaur, "Performance evaluation of two ANFIS models for predicting water quality Index of River Satluj (India)," *Advances in Civil Engineering,* vol. 2018, pp. 1-10, Mar. 2018.
5. S. Shahaboddin, J. N. Ehsan, E. A. Jason, A. M. Azizah, M. Amir, and C. Kwok-wing,

  -day ahead forecasting of chlorophyll-a

 *Engineering Applications of Computational Fluid Mechanics*,

vol. 13, no. 1, pp. 91-101, Dec. 2018.

1. Y. Khan and S. Chai, "Ensemble of ANN and ANFIS for water quality prediction and analysis - a data driven approach," *Journal of Telecommunication, Electronic and Computer Engineering* , vol. 9, no. 2, pp. 117-122, 2017.
2. J.-S. R. Jang, "Frequently Asked Questions - ANFIS in the Fuzzy Logic Toolbox," [Online].

Available: [http://www.cs.nthu.edu.tw/~jang/anfisfaq.htm.](http://www.cs.nthu.edu.tw/~jang/anfisfaq.htm)

1. A. Khashei-Siuki and M. Sarbazi, "Evaluation of ANFIS, ANN, and geostatistical models to spatial distribution of groundwater quality," *Arabian Journal of Geosciences*, vol. 8, no. 2, pp. 903-912, Nov. 2013.
2. N. Q. Sultan, S. Saeed, K. Salar, J. Salar, K. Ozgur, and S.  

monthly pan evaporation using wavelet support vector regression and wavelet artificial neural

 *Engineering Applications of Computational Fluid*

*Mechanics*, vol. 13, no. 1, pp. 177-187, Jan. 2019.

1. A. A. Najah, A. El-Shafie, O. A. Karim and O. Jaafar, "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation," *Neural Computing and Applications*, vol. 21, no. 5, pp. 833-841, Nov. 2010.
2. R. Barzegar, J. Adamowski, and A. A. Moghaddam, "Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 7, pp. 1797 1819, Jan. 2016.
3. Z. Fu, J. Cheng, M. Yang, and J. Batista, "Prediction of industrial wastewater quality parameters based on wavelet-ANFIS model," in *Proc. IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, 2018.

     d ANFIS Combination Model for

 *Revista Tecnica De La Facultad De Ingenieria Universidad*

*Del Zulia*, vol. 39, no. 2, pp.317-328, Dec. 2015.

1. -based neural



 *Environmental*

*Monitoring and Assessment*, vol. 191, no. 3, pp. 163-191, Feb. 2019.

1. A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, and F. Sayyahi, "Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River)," *KSCE Journal of Civil Engineering,* vol. 22, no. 7, pp. 2206-2213, Sep. 2017.
2. A. Gholami, H. Bonakdari, I. Ebtehaj, M. Mohammadian, B. Gharabaghi, and S. R.



swarm optimization with ANFIS to predict threshold bank profile shape based on digital laser

 *Journal of the International Measurement Confederation*, vol. 121, pp. 294-

303, Jun. 2018.

1. A. Azad, M. ManOochehri, H. Kashi, S. Farzin, H. Karami, V. Nourani, and J. Shirie,

     -fuzzy inference

  *Journal of Hydrology*, vol. 571 pp. 214-224,

Apr. 2019.

1. Q. Song and B. S. Chissom, "Fuzzy time series and its models," *Fuzzy Sets and Systems,*

vol. 54, no. 3, pp. 269-277, Mar. 1993.

1. S.-M. Chen, "Forecasting enrollments based on fuzzy time series," *Fuzzy Sets and Systems,*

vol. 81, no. 3, pp. 311-319, Aug. 1996.





 *International Journal of Nonlinear Science*, vol. 17, no.3, pp. 234-240, Dec. 2013.

1. S. A. Shanahan and X. Zhou, "Las Vegas wash surface water quality monitoring and assessment plan," *Southern Nevada Water Authority*, Las Vegas, 2011.
2. "Colorado River Basin Salinity Control Forum Website," [Online]. Available: https://[www.coloradoriversalinity.org/.](http://www.coloradoriversalinity.org/)
3. E. Tipton, L. Hedges, M. Vaden-Kiernan, G. Borman, K. Sullivan, and S. Caverly, "Sample selection in randomized experiments: a new method using propensity score stratified sampling," Journal of Research on Educational Effectiveness, vol. 7, no. 1, pp. 114-135, Jan. 2014.
4. M. Sugeno and G. Kang, "Structure identification of fuzzy model," *Fuzzy Sets and Systems*, vol. 28, no. 1, pp. 15-33, Oct. 1988.
5. "Secondary Drinking Water Standards: Guidance for Nuisance Chemicals," [Online].

Available: https://[www.epa.gov/dwstandardsregulations/secondary-drinking-water-standards-](http://www.epa.gov/dwstandardsregulations/secondary-drinking-water-standards-) guidance-nuisance-chemicals.

1. A. K.Venkatesan, S. Ahmad, W. Johnson, J. R. Batista, "Systems dynamic model to forecast salinity load to the Colorado River due to urbanization within the Las Vegas Valley," *Science of the Total Environment of Research on Educational Effectiveness*, vol. 469, no. 13, pp. 2616- 2625, Apr. 2011.
2. S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction," *Environmental Earth Sciences*, vol. 71, no. 7, p. 3147 3160, Aug. 2013.
3. C. Cheng, T. Chen and C. Chiang, "Trend-Weighted Fuzzy Time-Series Model for TAIEX Forecasting," *International Conference on Neural Information Processing*, vol. 3, pp. 469-477, Mar. 2005.
4. J. H. Holland, Adaptation in Natural and Artificial Systems, MA, Cambridge: MIT Press, 1992.
5. R. Hinterding, H. Gielewski,

 in *Proc. 6th International Conference on Genetic Algorithms*, pp. 65 72, 1995.

   ling and prediction of water quality

parameters using a hybrid particle swarm optimization   *International Journal of Environmental Science and Technology*, vol. 16, pp. 4823-4832, Jul. 2018.



Available: [http://yarpiz.com/319/ypfz104-evolutionary-anfis-training.](http://yarpiz.com/319/ypfz104-evolutionary-anfis-training)

[49] P. C. Silva, L. et al. pyFTS: Fuzzy Time Series for Python. Belo Horizonte. 2018. DOI: 10.5281/zenodo.597359. Url: [http://doi.org/10.5281/zenodo.597359.](http://doi.org/10.5281/zenodo.597359)