The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Information Technology
Master of Information Technology

الجامعـــــــة الإســـلاميـة – غـزة
شئون البحث العلمي والدراسات العليا
كليـــــــــــة تكنولوجيا المعلومات
ماجستيـــــر تكنولوجيا المعلومات

Prediction Model Based on Time Series for Chloride and Nitrate Concentration in Municipal Groundwater Wells in Gaza Strip

نموذج توقعي معتمد على السلاسل الزمنية لتركيزات الكلورايد والنترات في الآبار الجوفية البلدية في قطاع غزة

By

Shukri M A Elastal

Supervised by

Dr. Tawfiq Barhoum

Associate Professor – Applied Computer Technology

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Information Technology

July /2017

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Prediction Model Based on Time Series for Chloride and Nitrate Concentration in Municipal Groundwater Wells in Gaza Strip

نموذج توقعي معتمد على السلاسل الزمنية لتركيزات الكلورايد والنترات في الآبار الجوفية البلدية في قطاع غزة

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	شكري محمد الأسطل Shukri M A Elastal	اسم الطالب:
Signature:		التوقيع:
Date:		التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية غزة
The Islamic University of Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي: 1150

الرقم: ج س غ/35/ Ref:

التاريخ: 2017/07/22 Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **شكري محمد عبدالسلام الأسطل** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

نموذج توقعي معتمد على السلاسل الزمنية لتركيزات الكلورايد والنترات في الآبار الجوفية البلدية في قطاع غزة

Prediction Model Based on Time Series for Chloride and Nitrate Concentration in Municipal Groundwater Wells in Gaza Strip

وبعد المناقشة التي تمت اليوم السبت 28 شوال 1438هـ، الموافق 2017/07/22م الساعة الحادية عشر صباحاً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. توفيق سليمان برهوم	مشرفاً و رئيساً
د. إياد محمد الأغا	مناقشاً داخلياً
أ.د. سامي سليم أبو ناصر	مناقشاً خارجياً

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

Abstract

The groundwater in Gaza strip has a high Total Dissolved Solid (TDS), chloride concentration and nitrate concentration. Besides that, the high cost and poor data in Gaza strip are considered as problem for water management.

In the part of the target area, there are many researches for using AI techniques to predict chloride and nitrate in groundwater wells. However, it is not depending on time series and need more data for these wells which not available.

An artificial prediction model for nitrate and chloride concentration in groundwater wells using Data mining techniques in time series is built to improve management of groundwater wells in Gaza strip despite of limitation of data.

In this research, many intelligent techniques as Artificial Neural Network, Support Vector Machine, K-Nearest Neighbours and Linear Regression are applied then determine the optimum one in this case by using Statistics Coefficient of Determination (R^2) and Root Mean Square Error (RMSE). Data collected from Palestinian Water Authority that contains historical reads for groundwater wells in 1974 to 2013 period, then data preprocessed with domain expert contribution. The optimum model has RMSE= 64.68 and R^2 = 0.954 for chloride and 33.33 and 0.828 for nitrate respectively by LR .

Finally, final model evaluation took place using 2014 and 2015 PWA data for the targeted field. The result was RMSE less than 84 and R^2 more than 0.9 for chloride. In additional, RMSE less than 36 and R^2 more than 0.86 for nitrate.

الملخص

تحتوي المياه الجوفية في قطاع غزة على نسبة عالية من المواد الصلبة المذابة ومن تركيز الكلورايد والنترات. إلى جانب ذلك تعتبر التكلفة العالية وقلة البيانات ففي قطاع غزة مشاكل إضافية يعاني منها القطاع في إدارة المياه.

في منطقة الدراسة، هناك العديد من الدراسات التي تتنبأ بنسب الكلورايد والنترات في المياه الجوفية، ومع ذلك فهي لا تعتمد على السلاسل الزمنية، وتحتاج لبيانات إضافية للآبار والتي قد تكون غير متوفرة.

ولتحسين إدارة جودة أبار المياه الجوفية بالرغم من البيانات المحدودة، تم بناء نموذج توقع صناعي للنترات والكلورايد في آبار المياه الجوفية باستخدام تقنيات تنقيب البيانات في السلاسل الزمنية.

وقد تم استخدام التقنيات الذكية في هذا البحث مثل ANNو SVM ،و KNN ،و LR لاختيار أفضل خوارزمية. وتم التقييم باستخدام R^2 و RMSE.

تم تجميع البيانات من سلطة المياه الفلسطينية، والتي تحتوي على قراءات تاريخية للأبار من عام 1974 وحتى عام 2013، وتم معالجة البيانات بالاشتراك مع خبير المجال، وتبين أن نتائج أفضل نموذج كانت كالتالي RMSE = 64.68 و R^2 = 0.954 للكلورايد و 33.33، 0.828 للنترات على التوالي باستخدام LR.

في النهاية تم تقييم النموذج النهائي باستخدام بيانات أعوام 2014 و 2015 والتي تم تحصيلها من سلطة المياه الفلسطينية لنفس الآبار المستهدفة، وكانت النتائج كالتالي RMSE الى أقل من 84 و R^2 أكثر من 0.9 للكلورايد و أقل من 36 وأكبر من 0.86 للنترات على التوالي .

TO MY FAMILY
I DEDICATE THIS WORK

Acknowledgment

First of all, all thanks and appreciations go to Allah for His unlimited blessings and for giving me the strength to complete this work.

Special thanks are due to my supervisor Dr. Tawfiq Barhoom as he was patient and advisable to me.

Also, I like to give special thanks to Palestinian Water Authority (PWA), Dr.Khalil AlAstal, Dr. Jawad AlAgha, Mr. Mahmoud Abdulateef and Mr Ibrahim ElAstal for their corporation in the research.

Finally, my sincere thanks are due to all people who supported me to complete this work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Artificial Neural Networks	ANN
Artificial Intelligence	AI
Boosted Regression Tree	BRT
Case-Based Reasoning	CBR
Chloride	Cl
Classification And Regression Tree algorithm	CART
Data Mining	DM
Decision Tree	DT
Electrical Conductivity	EC
Exploratory Data Analysis	EDA
K - Nearest Neighbor	KNN
Linear Regression	LR
Multinomial Logistic Regression	MLR
Nitrate	NO_3
Ordinal Pairwise Partitioning	OPP
Palestinian Ministry of Health	MOH
Palestinian Water Authority	PWA
Random Forest	RF
Root Mean Square Error	RMSE
Self-Organizing Map	SOM
Statistics Coefficient of Determination	R^2
Support Vector Machine	SVM
Total Dissolved Solid	TDS
World Health Organization	WHO

Chapter 1

Introduction

Chapter 1
Introduction

This chapter consists of six sections. It starts with research motivation then statement problem. The third section is objectives, which consist of two subsections; main objective, and specific objectives.

After that, scope and limitations are illustrated in section 4 and finally, thesis structure.

1.1 Research Motivation

Groundwater in Gaza Strip is not suitable for drinking; it has a high Total Dissolved Solid (TDS), Chloride and Nitrate as shown in Table 1.1.

Table (1.1): Comparison between allowable values and reads for groundwater in Gaza strip (Organization, 2011; Vincent, 2016)

	TDS	Cl	NO_3
Allowable values (WHO)	500	200-600	30
Reads	3000	5000	300

Table 1.1 shows that Groundwater in Gaza Strip has a high TDS, Chloride and Nitrate, which has a bad effect on public health in Gaza. In addition, the lack of financial funding for researches and the lack of data for the water management have also considered as additional problem.

These problems illustrate the needs of adopting new simple, low cost and effective strategies to predict contamination in groundwater by utilize artificial model to manage it.

In this work, developing accurate, simple and cost effective model for nitrate and chloride concentration in groundwater in Gaza Strip using DM techniques in time series is done.

1.2 Statement of Problem

The analytical hydrological models are generally used as groundwater modelling tools worldwide. However, it is time-consuming and expensive. Besides that, the models need a huge amount of detailed and accurate data about hydrological system with understanding of physical process.

These limitations lead up to adopt a totally different approach, in hydrological modelling, such as Data Mining (DM) techniques which have ability to develop an accurate, simple and cost effective model.

1.3 Objectives:

1.3.1 Main Objective

Develop accurate, simple and cost effective model for nitrate and chloride concentration in groundwater in Gaza Strip using DM techniques

1.3.2 Specific Objectives

- Collection and manipulation of data.
- Using statistical and preprocessing techniques to improve data quality.
- Evaluate the performance of DM techniques in hydrological models to choose the optimum one in this case.
- Build the optimum model for this case.
- Evaluation proposed model

1.4 Thesis Importance

The model will help in management of groundwater wells in Gaza strip despite of analytical models limitations.

1.5 Overview of Thesis

This research consists of six chapters. The first one is an introduction that illustrates the problem and the objectives of the research. The second one is theoretical background to illustrate background and theoretical concepts about DM techniques.

Chapter 3 illustrates previous related work in world, regional and local levels and chapter four determines the methodology to develop the model.

Comparing between Models and models results are taken place in chapter five. Finally, conclusion and future work is in chapter six.

Chapter 2

Theoretical Background

Chapter 2

Theoretical Background

In this chapter, background and theoretical concepts, which applied, is present which consist of six sections; the first of them about Data Mining (DM), DM framework and DM techniques. The second one is about outlier detecting. The evaluation methods are appearing in section three. Targeted area and time series analysis and forecasting took place in section four and five, then summary is in the last section.

2.1 Data Mining

This section illustrates theoretical concepts about DM and its frameworks, then illustrate some Artificial Intelligence (AI) algorithms that used in DM.

2.1.1 Data Mining Definition

DM concept is converting data to knowledge. It is defined as an analytical process to extract patterns from data and systematic relationships between variables, then to validate the findings by applying the detected patterns to new subsets of data.(Nisbet, et al., 2009)

DM is relatively less concerned with identifying the specific relations between the involved variables, instead, the focus is on producing a solution that can generate useful predictions.(Nisbet et al., 2009) (Shmueli, et al., 2016)

In other words, the main goal of DM is prediction (Shmueli et al., 2016). Therefore, Data Mining accepts among others a "black box" (Clinchant, Csurka, & Chidlovskii, 2016) approach to data exploration or knowledge discovery without using the traditional Exploratory Data Analysis (EDA) techniques, but by using AI techniques as Artificial Neural Networks (ANN).

2.1.2 Data Mining Framework

Data mining has many frameworks. In American industries, Six Sigma methodology recently has become very common. It supposed a sequence of steps called (DMIAC), which illustrate in Figure 2.1 et al., 2011) (J. F. Chang, 2016)

Define → Measure → Analyze → Improve → Control

Figure (02.1): Six-Sigma Methodology

(J. F. Chang, 2016)

SAS Institute is proposed another framework, called (SEMMA), which is focusing on the technical activities typically involved in DM projects which illustrate in figure 2.2. (Köksal et al., 2011) (J. F. Chang, 2016)

Sample → Explore → Modify → Model → Assess

Figure (02.02) SEMMA

(J. F. Chang, 2016)

All of these frameworks are focused on integrated DM methodology into an organization, extracted information from data, invested information for strategic decision-making by extracting knowledge.

On the other hand, DM deployment process is a repeated sequence of the following steps (Han, et al., 2011):

1. Data preprocessing; it is the first step of DM process; it works to improve the data quality to satisfy the requirements of the intended use. It is improve the data accuracy, completeness, consistency, timeliness, believability, and interpretability. The data preprocessing can be divided to the following steps;
    - Data cleaning; this step works to fill the missing values, identifying outliers and resolve them and correct inconsistencies in the data.

      Each process has many techniques to work; filling the missing values works done by ignoring the tuple, filling in the missing value, using the most probable value to fill in the missing value, etc. Identifying outliers and resolve them done by using regression methods, excluding the outlier values, etc. while correct inconsistencies in the data by using external references, etc.
    - Data integration; this step works to merge data from multiple data sources.

- Data selection; it works to reduced data set to more specific dataset that represent the original data.

  It has many strategies such as; dimensionality reduction which defined as reducing the number of random variables or attributes under consideration, numerosity reduction. It has defined as the techniques to replace the original data volume by alternative set of data just like regression, log-linear models and histograms, and data compression.

- Data transformation; this step work to convert the data into appropriate forms for mining.

It has many strategies such as; attribute construction that constructs and add new attribute from the given set of attributes to help the mining process, aggregation that applies summary operations to the data, normalization, which scale the data to fall in small ranges and discretization that replace the values of numeric attribute by interval labels or conceptual labels.

2. Data mining; it is the essential process where intelligent methods are applied to extract data patterns.

3. Pattern evaluation; it is a step to identify if the pattern is interesting pattern or not. The evaluation process is depending in the ease of human understanding, validity with new dataset with acceptable degree of certainty, potentially useful, and novel.

   In this step, it is very important to have domain expert to have interesting patterns because it is inefficient and unrealistic to generate all possible patterns. This step is considered as an optimization problem.

4. Knowledge presentation; it is the techniques of present and visualizes the mined knowledge to decision makers.

2.1.3 AI Algorithms Used in Data Mining

Data Mining is often considered to be "a blend of statistics, artificial intelligence (AI), and data base research"(Glymour, et al., 1997). In this section, some of AI algorithms, which used in DM will be discussed.

2.1.3.1 Artificial Neural Network (ANN)

ANN is a mathematical model used in machine learning that is inspired by the way biological nervous systems. (Gong, et al., 2016) it is consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation.

In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data.(Schalkoff, 1997)

A feed-forward Neural Network is an ANN where connections between the units do not form a directed cycle. In this network, the information moves in only one direction, forward, from the input nodes to the output nodes and through the hidden nodes. It is important to illustrate that there is no cycles or loops in this network.(Biganzoli, et al., 1998)

Back propagation algorithm is a supervised learning method, which can be divided into two phases: propagation and weight update. These steps are repeated until the performance of the network is good enough(Gevrey, et al., 2003). In back propagation algorithms, the output values are compared with the correct answer to compute the value of some predefined error-function by various techniques (Basheer & Hajmeer, 2000).

The error is then back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is small. (Basheer & Hajmeer, 2000).

2.1.3.2 K - Nearest Neighbor (KNN)

KNN algorithm is based on learning by analogy, that is compare a given test example with training examples that are similar to it. (Larose, 2005)

The training examples are described by n attributes. Each example represents a point in an n-dimensional space. In this way, all of the training

examples are stored in an n-dimensional pattern space. When given an unknown example. KNN algorithm searches the pattern space for the k training examples that are closest to the unknown example. These k training examples are the k "nearest neighbours" of the unknown example. Closeness is defined in terms of a distance metric, such as the Euclidean distance.

The KNN algorithm is one of the simplest machine learning algorithms. (Larose, 2005) The neighbours are taken from a set of examples for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The basic KNN algorithm is composed of two steps: Find the k training examples that are closest to the unseen example. Take the most commonly occurring classification for these k examples. (Larose, 2005)

2.1.3.3 Support Vector Machine (SVM)

SVM is a learning method that exploits prior knowledge of gene-function to identify unknown genes of similar function from expression data.

In addition, it has the ability to identify outliers, and manipulate large feature spaces. (Brown et al., 2000) SVM takes a set of input data and predicts which of the two possible classes comprises for each input, making the SVM a non-probabilistic binary linear classifier.

SVM model is a representation of the examples as points in space which is mapped into separate classes which are divided by a clear gap. New points are then mapped into the same space to determine the point class based on which side of the gap they fall on. (Andrew, 2000) SVM constructs a hyperplane or set of hyperplanes in a high dimensional space, which can be used to classify, regression, or other tasks.

2.1.3.4 Regression

Regression is a learning function that maps a data item to a real-valued prediction variable (Fayyad, et al., 1996). It is also defined as a statistical technique used for numerical prediction that attempts to determine the strength of

the relationship between one dependent variable and a series of other changing variables known as independent variables.(Healy, 2005)

Linear regression (LR) find the relationship between a scalar variable and one or more explanatory variables by fitting a linear equation to observed data.(Montgomery, et al., 2015)

Polynomial regression is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modelled as an nth order polynomial.(Fan & Gijbels, 1996)

2.2 Detecting Outlier

Outliers are considered as noisy points outside a set of defined clusters, however, it is not considered as a noise data.

There are a lot of algorithms that detect outliers using statistics methods, computer science or machine learning.(Wang, 2009) There are three approaches to detect outlier: (Hodge & Austin, 2004)

1. Determine outliers without previous knowledge of the data: it is just like to unsupervised clustering. Statistically, the most remote points in the dataset are flagged as potential outliers.
2. Normality & Abnormality Model: this is considered as a supervised classification and requires training dataset which tagged as normal or abnormal.
3. Only Normality Model: this is considered as a semi-supervised paradigm where only the normal class is taught and the system needs to learn to recognize abnormality.

In this subsection two methods used in this work will be illustrated.

2.2.1 Grubbs' test

Grubbs' test is described, along with methods to handle masking effects related to multiple outliers. A similar analysis can be done with the standard exponential distribution:

$$e^{-X}, \qquad x > 0$$

$$E_n = E(\max\{X_1, X_2, \dots, E_n\})$$

And $\quad c = E_n(\alpha) \quad where\ p(X < c) = (1-\alpha)^{1/n}$

It has a thin tail and plausible outlier values grow slowly with sample size. An exponential with mean μ can be normalized by the substitution

$$\mathbf{Z} = \frac{x}{\mu} \qquad (2.2)$$

It is possible to refine this methodology when the mean μ must be estimated from the sample. (Ghosh & Vogt, 2012) (Hodge & Austin, 2004)

2.2.2 Detect outlier by Standard Deviation

For a normal distribution, outliers can be considered to be points that lie as far as more than three standard deviations from the mean.

If T is a set of values that is truly normally distributed with mean p and standard deviation σ. Def Normal is defined as follows: $t \in T$ is an outlier. (Knorr & Ng, 1997)

$$\text{Iff } \frac{t-\mu}{\sigma} \geq 3 \ or \ \frac{t-\mu}{\sigma} \leq -3 \qquad (2.3)$$

2.3 Measuring Performance

There are many ways to evaluate prediction models; in this subsection two methods will be described

2.3.1 Root Mean Square Error (RMSE)

RMSE is a measure of the differences between values from the predictive model and the actually observed values. It is calculated by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y)}{n}} \qquad (2.4)$$

where y_i is predicted value is computed for n different predictions.(Willmott & Matsuura, 2005)

2.3.2 Coefficient of Determination (R^2)

R^2 indicates the percentage of the variance in the dependent variable that is predictable from the independent variables which computed by

$$\boldsymbol{R^2 = 1 - \frac{SS_{resi}}{SS_{tot}}} \qquad (2.5)$$

Where SS_{tot} is the total sum of squares and SS_{res} is the sum of squares of residuals. (Hill, et.al., 2006)

2.4 Gaza Strip (Targeted area)

Gaza strip is southern part of the Palestinian territory. It is 41 kilometers long, and from 6 to 12 kilometers width. It total area about 365 square kilometers (Al-Khatib & Arafat, 2009). The groundwater is considered the main water resource for more than 1.85 million people live there(Alastal, et. al, 2011). Figure 2.3 shows the location if Gaza strip

2.4.1 Municipality Groundwater Wells Map

Municipality ground water wells are distributed in Gaza strip (about 321) from north to south as shown in figure 2.4

Figure (2.3): Location of Gaza strip

(Al-Khatib & Arafat, 2009)

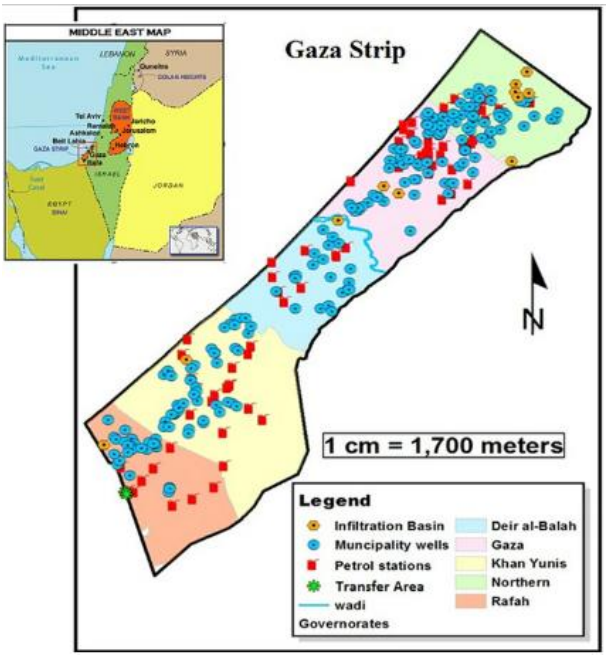


Figure (02.4): Municipal Groundwater Wells Map

(Ghabayen, 2013)

2.4.2 Municipality Groundwater Wells Data Collection Procedure

Data is collected twice per year, the first of them starts in April, and the second cycle starts in October. Ministry Of Health (MOH) is responsible for the collection and analysis of the water samples from the municipal wells in coordination with the municipalities.

MOH analyzes groundwater samples for full chemical parameters including chloride and nitrate. After that, the results are submitting to the Palestinian Water Authority (PWA) for the purpose of assessment and evaluation. (Authority, 2013)

2.5 Time series analysis and forecasting

A time series is a sequence of data points in successive order which regard to differences in time. In other word, it's any variable can be taken change over the time. (De Luca, Zinno, Manunta, Lanari, & Casu, 2017)

Time series analysis comprises methods for analyzing time series data in order to extract meaningful information and knowledge. Prediction in time series is the use of a model to predict future values based on previously observed values. Regression analysis is often used to test theories that the current values of one or more independent time series affect the current value of another time series, which focuses on comparing values of a single time series or multiple dependent time series at different points in time. Interrupted time series analysis is the analysis of interventions on a single time series (Montgomery, 2015)

Window function used to convert time series data into cross section by choosing each X followed values as a row of data. Data classifies as input data and a label. (Chuchro, et al., 2014)

2.6 Summary

In this chapter, theoretical concepts for DM and its artificial techniques are illustrated which used in this work. In addition, this chapter illustrates the statistical methods for detecting outliers and methods to evaluation prediction model which used in this work too.

The description about targeted area, targeted wells and methods that used to collect data is illustrated. Finally, time series analysis and forecasting is took place.

Chapter 3

Related Works

Chapter 3
Related Works

This chapter illustrates and reviews the related work appeared. It discusses the related work in targeted area, related work in other areas and related work just depending on historical data read.

3.1 Modelling water quality in targeted area

There is a limited number of researches about groundwater modelling in Gaza, Al Agha et al. developed chloride concentration model in groundwater wells in Khanyounis city, part of Gaza Strip, using ANN and SVM(Alagha,et al. , 2013). The model depends on data of 22 municipal wells. They classified input data into clusters as a pre-processing technique. The results had a high performance.

However, they did not use the data as a time series data. In addition, the model depends on hydrological and physiogeographical data, which is not available in most of wells in Gaza strip (Alagha et al., 2013).

The same authors used the same methodology to build Nitrate concentration model using ANN and SVM(Alagha et al., 2014).

3.2 Modelling water quality in different areas

In Korea, Yooa et al. (Yoo,et al. , 2016) developed a model that can detect patterns of groundwater pollution by used ANN, Decision Tree (DT), Case-Based Reasoning (CBR), and Multinomial Logistic Regression (MLR). The model used hydrogeological and environmental data as input data. DT and rule induction methods gave more accuracy results than others.

In Iran, Barzegar et al. (Barzegar, et al., 2016)developed a model by ANN and used hydrological and physiogeographic data to predict nitrate concentrations.

Also in Iran, Arabgol et al. (Arabgol, et al., 2016) also developed a model using hydrological and physiogeographic data to predict nitrate concentration in groundwater by SVM. They found that SVM model gives fast, reliable, and cost-effective results for groundwater quality evaluation and prediction.

Another study in Iran, Naghibi et al. (Naghibi, et al., 2016) developed a model of groundwater spring potential maps by Boosted Regression Tree (BRT), Classification And Regression Tree algorithm (CART), and Random Forest (RF). The models have acceptable results.

also in Iran, Nourani et al., (Nourani, et al., 2016) used Principal Component Analysis (PCA) model and a Self-Organizing Map (SOM) model to analyse a complex dataset obtained from the river water monitoring stations. The results shows that PCA and SOM are efficient to capture and analyse the behaviour of multivariable, complex, and nonlinear related surface water quality data.

In Spain, Rebolledo et al. (Rebolledo, et al., 2016) used Logic Scoring of Preferences (LSP) model to evaluate groundwater Nitrates from Agricultural Sources. The result shows that LSP map could be used to manage the risk of nitrate.

In China, Wang et al. (Wang et al., 2016) developed a model using cloud model-based approach to water quality assessments. The model realizes the transformation between qualitative concept and quantitative data based on probability, statistics and fuzzy set. Bilateral boundary formula with nonlinear boundary regression is used for parameter estimation. In addition, hybrid entropy-analytic hierarchy process technique is used for calculation of weights and mean of repeated simulations to determine the degree of final certainty. The model is adopted many AI method such as; Scoring Index method, Variable Fuzzy Sets method, Hybrid Fuzzy, Optimal model, and Neural Networks method. The results show that the approach is more representative than other alternative methods and more accurate.

In US, Lee et al. (Lee, et al., 2016) used CART to monitor and predict contaminant degradation. In addition, they found that the groundwater biogeochemical monitoring data should be curated, open-access, up-to-date and comprehensive collection to improve the reliability of the predictive models capabilities.

Another research in US, Asefa et al. (Asefa, et al., 2006) built a model to predict both 6-month ahead annual flow volume and 24-hour ahead hourly stream flow using SVM. The results show that SVM is considered as a promising tool for solving site-specific, real-time water resources management problems.

In India, Oorkavalan et al. (Oorkavalan, et al., 2016) built a model to evaluate the hydro-geochemical quality of shallow groundwater using Multilayer Perceptron Classifier (MLP).The models are built to understand the sources of dissolved ions for evaluate the chemical quality of the groundwater through physico-chemical analysis. The wells are classified as drinking and agricultural wells. In addition, data clustering is used to classify the data based on contamination characteristics of groundwater quality. In Oorkavalan et al. (Oorkavalan, et al., 2016) research, two models are built, the first one using Principal Component Analysis (PCA), while the other one without using it. The results was classification by MLP with PCA is giving better conclusion about the Water Quality Parameters than the MLP method without PCA in terms of correlation coefficient and time consuming to create the model.

3.3 Modelling by using time series

In Iran, Sattari et al. (Sattari, et al., 2016) developed a model using KNN and SVM approaches to predict TDS and electrical conductivity (EC) in water using hydrological data as input. They found SVM is give more accreted results than KNN. The evaluation is took place using RMSE, absolute error and coefficient of determination (R^2). This approach is similar to this research, but targeted area have poor and lack data. In addition, the targeted output is not same.

3.4 Summary

As shown before, researches adopted many tools in hydrological modelling, such as; Artificial Neural Network (ANN), Decision Tree (DT), Case-Based Reasoning (CBR), Multinomial Logistic Regression (MLR), Ordinal Pairwise Partitioning (OPP), Classification and Regression Tree (CART) analysis, Boosted Regression Tree (BRT), Random Forest (RF), Self-Organizing Map (SOM), etc. These tools applied in various applications such as groundwater modelling, Total Dissolved Solid (TDS) modelling, Electrical Conductivity (EC), real-time water resources management, chemical quality of the groundwater, etc. on the other hand, it is hard to compromise these models because the input variables (targeted area, hydrological topography, land uses, etc.) and the outputs (EC, TDS, chemicals concentrations, etc.) are various. Besides that, their

performances are various for the same reasons. In addition, there are limit number of researches about GW modelling in Gaza that focused in specific areas. This situation reflects the needs of groundwater modelling in Gaza Strip

The previous models depend on understanding physical process, in addition to the availability of valid detailed and accurate data about hydrological system, which are not usually available in Gaza strip due to technical, financial and political constraints. So that, the results of these models has unsatisfactory performances (Trichakis, Nikolos, & Karatzas, 2011). In this research, the model is developed to adopt these limitations by different approach in hydrological modelling, such as data mining techniques, which have the ability to develop an accurate, simple and cost effective model.

The previous researchers for the targeted area did not use time series and depended on hydrological and physiogeographical data which make the targeted wells limited (about 22 wells) (Alagha et al., 2014; Alagha et al., 2013), because the database is not valid in the most of wells in Gaza strip and the number of unlicensed wells in Gaza Strip is very huge which illustrated in the next chapter.

Table (3.1): Summary of the related works

Paper Reference	Paper Summary
(Alagha et al., 2013)	Modelling chloride by grouping data into cluster then applying AI techniques
(Alagha et al., 2014)	Modelling nitrate by grouping data into cluster then applying AI techniques
(Yoo et al., 2016)	Model to detect patterns of groundwater pollution by ANN, DT, CBR and MLR
(Barzegar et al., 2016)	A model by ANN and used hydrological and physiographical data to predict nitrate pollution
(Arabgol et al., 2016)	A model used hydrological and physiographical data to predict nitrate concentration in groundwater by SVM
(Naghibi et al., 2016)	A model of groundwater spring potential maps by BRT, CART, and RF with succeeded results
(Nourani et al., 2016)	used PCA and SOM to analyses a complex dataset obtained from the river water monitoring stations
(Rebolledo et al., 2016)	Used LSP to evaluate groundwater nitrates from agricultural sources
(Wang et al., 2016)	A cloud model-based assessment that realize the transformation between qualitative concept and quantitative data, based on probability, statistics and fuzzy set

(Lee et al., 2016)	Used CART to monitor and predict contaminant degradation.
(Asefa et al., 2006)	A model to predict both 6-month ahead annual flow volume and 24-hour ahead hourly stream flow by SVM
(Oorkavalan et al., 2016)	A model to evaluate the hydro-geochemical of shallow groundwater using MLP
(Sattari et al., 2016)	A model using KNN and SVM to predict TDS and EC by used hydrological data as input

Chapter 4

Model Development

Chapter 4

Model Development

This chapter illustrate the proposed model and the research methodology for applied data mining techniques in municipal groundwater wells -Gaza strip as comparative study by using time series to predict chloride and nitrate concentration although limit, irregular, error and lack data.

The main objective of this model is that prediction chloride and nitrate reads in the future by input three previous reads as inputs and predict next. It has done just by depending on historical read for chloride and nitrate without other complicated, missed and costly information as physical and geological data. These reads used as input data and the output is predicted chloride and nitrate reads as shown in Figure 4.1. It can be used as decision support system in groundwater management in Gaza strip. This model evaluated by using 2014 and 2015 reads for 51 wells.

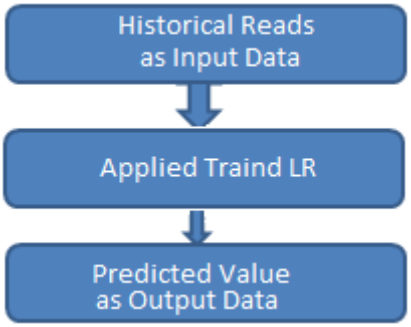


Figure (4.1): Proposed Model

The methodology used starts in data collection, preprocessing, applied DM techniques, evaluation and finally documentation, which is illustrated in figure 4.2

Figure (4.02): Overall Research Methodology

4.1 Model Development Methodology

The proposed model based on methodology to develop chloride and nitrate concentration predication model in groundwater wells outlined in the work Figure 4.3

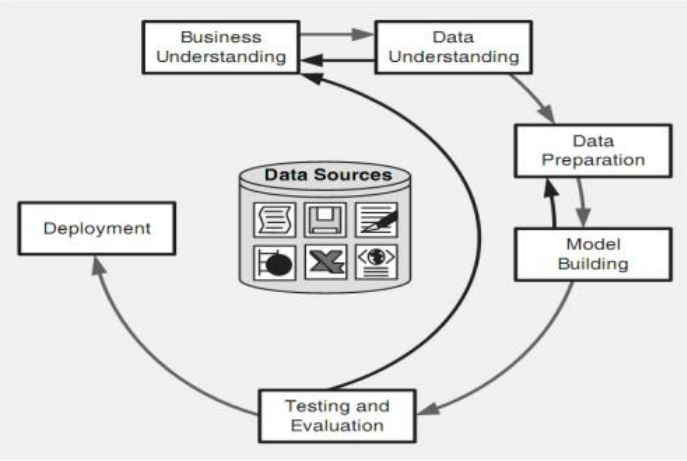


Figure (4.3): CRISP-DM reference model

as in (C. Chang & Lin, 2005; Sharma & Osei-Bryson, 2009)

4.2 Research Methodology

The research methodology is to develop a model for predict chloride and nitrate concentration in groundwater wells by historical read only, as shown in Figure 4.4

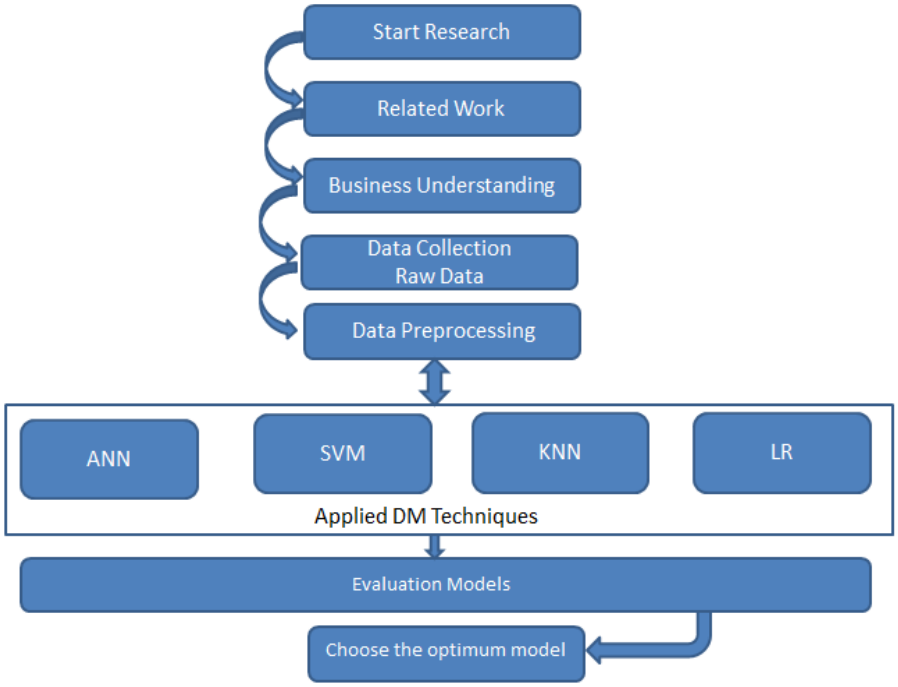


Figure (4.04): Research Methodology

4.2.1 Data Collection

Data collected from Palestinian Water Authority (PWA) as shown in appendix3. It consists of historical read for hydrological parameters as chloride and nitrate from 1974 to 2013 for municipality groundwater wells (321 wells- 4895 reads for Cl and 3793 reads for No_3). In additional, there are some contaminations, water quality parameters and well place are founded. It has missed values and irregular read and missed read that is illustrated in Figure 4.5. In addition, it has incorrect read and outlier. These make trouble in prediction model because there was not dataset that can be used.

The previous data collected again from another source by domain Experts as shown in appendix3.

The evaluation data for final model gained from PWA that data of 2014 and 2015.

Well No.	X	Y	Date	EC	TDS	pH	Calcium	Magnesium	Sodium	Potassium	Floride	Chloride	Nitrate	Nitrite	Ammonia	Sulphate	Alkalinity	Hardness
C.79	105349.2878	105095.3091	17/06/1993	2053								329	60					
C.79	105349.2878	105095.3091	22/07/1993	1788								308	28					
C.79	105349.2878	105095.3091	28/10/1993	2053								329						
C.79	105349.2878	105095.3091	28/09/1994	1492								315						
C.79	105349.2878	105095.3091	01/06/1995	1593								329	40					
C.79	105349.2878	105095.3091	23/09/1995	1518								315	55					
C.79	105349.2878	105095.3091	01/10/1995	1772	1181	7.4	83.4	60.1	210	2.3	0.3	413	55			30	287	456
C.79	105349.2878	105095.3091	01/04/1996	1865	1242	7.3	96	58	220	1.6	0.3	434	70			37	278	479
C.79	105349.2878	105095.3091	26/06/1996	1826								385						
C.79	105349.2878	105095.3091	01/04/1997	1894	1176	7.4	149	35	35	35	35	427	35			62.5	290	515
C.79	105349.2878	105095.3091	03/06/1997	2059								384						
C.79.A	105350.281	105094.134	18/04/1999	1894	1176	7.4	149	35				427	35			62.5	290	515
C.79.A	105350.281	105094.134	22/04/1999	2353								463.7	60.9					
C.79.A	105350.281	105094.134	16/04/2000	2427	2427	7.5	145	44	230	2.8		483	57	0.03	0.19	61	299	545
C.79.A	105350.281	105094.134	09/05/2000	322.039474								488.5	93.2					
C.79.A	105350.281	105094.134	15/10/2000	2094	1300	7.11	113	65.4	229.6	2.8		455	68.5		1.94	119.8	347	552
C.79.A	105350.281	105094.134	04/11/2000	2198.118987								483.538	89.175					
C.79.A	105350.281	105094.134	05/11/2000	2181	13522	7	105.7	84.3	240	3.3		478	75.4		1.26	96.2	286.8	611.5
C.79.A	105350.281	105094.134	12/04/2001	2222	1425	7.14	113.4	38.45	298	3.1		473	89.1	0	0.563	67.5	352.9	442
C.79.A	105350.281	105094.134	20/06/2001	2159.223529								473.6	110.7					
C.79.A	105350.281	105094.134	26/09/2001	2184.928571								489.2	98.6					
C.79.A	105350.281	105094.134	23/10/2001	2433	1472	7.33	95.1	78.6	270	2.6		475	97.6	0	0	81.9	393.7	561.8
C.79.A	105350.281	105094.134	01/04/2002	2141	1296	7.1	97.5	73.7	270	3.2	0.9	502.5	106.9	0.01	0.018	68.5	361.2	547.6
C.79.A	105350.281	105094.134	01/10/2002	2453	1521	7.7					0.65	476	105	0	0	80	398	560
C.79.A	105350.281	105094.134	30/03/2003	2350	1457	7.07	99.12	78.41	330	2	1.17	479.7	98	0	0	72	418	570.8
C.79.A	105350.281	105094.134	01/10/2003	2310	1433	7.27	105.4	81.82	290	3.5	1.5	475.7	102.5	0	0	76.56	322.08	600.6
C.137	104987.959	106485.517	22/02/2003	306	204	7.45	30.46	13.07	13.84	4.5		26.94	16.24	0		4.41	100	130

Missed values

Well No.	Date	EC	Chloride	Nitrate
R.306	09/10/2006	1621		255.78
K.19	27/11/2010	1206		172
Safa.Rafah	19/08/2009	4210	986	
Dohair.Ra	19/08/2009	3510	848	

Irregulars read

Well No	Date	EC	TDS	pH	Chloride	Nitrate
A.135	01/04/19	4	2	7	34	
A'Aisha.Bi	28/04/2007	2500	1550	7.15	502	196.4
A'Aisha.Bi	18/11/2007	3100	1922	7.32	774.5	83.06
A'Aisha.Bi	01/10/2008	3900	2418	7.42	910.7	155.2
A'Aisha.Bi	25/05/2009	3860	2393	7.44	910.7	136.2
A'Aisha.Bi	09/05/2010	4160	2579	7.76	977	131
A'Aisha.Bi	31/12/2011	4540	2815	7.18	1055	125
A'Aisha.Bi	31/12/2013	5240	3249	7.48	1306	100.8
A.210	01/11/2010	471	292	7.8	63	36
A.210	31/12/2012	515	319	7.82	64	22.8
A.211	03/05/2010	713	442	7.05	77	52
A.211	01/11/2010	769.9	477	7.69	98	57
A.211	31/12/2011	734	455	7.67	91	49
A.211	31/12/2012	853	529	7.75	114	23.4
A.211	31/12/2013	880	546	8.17	142	52.9
A.231	31/12/2013	1020	632	7.37	99	190.9

Figure (4.05): Row Data Sample and illustrate missed and irregular reads

4.2.2 Data Preprocessing

Data preprocessing is done with domain expert contribution. It is done in many steps as in the following subsections.

4.2.2.1 Irregulars read

The mean of reads in the same year are submitted to unifying time-period as shown in figure 4.5. For example, it's supposed that periodic read each 6 months but many wells are not and some of them have not any read.

4.2.2.2 Missed values and lack data series

There are many wells that don't have complete data series. To a void missed and lack data series as shown in figure 4.5, converting time series into cross sectional data done by windowing data with window size 4, 5 and 6. (Silva, et al., 2017) Table 4.1 illustrates data statistics after windowing process and before second step in detects outlier.

Table (4.01): Data Statistics after Windowing Process

Size of window	Chloride read	Welles	Nitrate read	Wells
4	1860	180	1164	165
5	1645	159	932	140
6	1459	133	736	115

4.2.2.3 Detect outlier

It is done in two steps; the second one used as additional outlier detection in the second experiment.

1. **By grubbs test**

   It is applied on row data with domain experts contribution by using grubbs test (Hodge & Austin, 2004), Then remove outlier reads.

2. **By mean standard deviation**

   It applied outlier mean standard deviation for dataset to improve results for window 4, 5 and 6. Outlier mean standard deviation means that the mean plus or minus three standard deviations consider the threshold and out of range consider outliers then remove them. Table 4.2 shows boundaries to detect outlier. Figure 4.6 showed how it is filtered in Rapidminer.

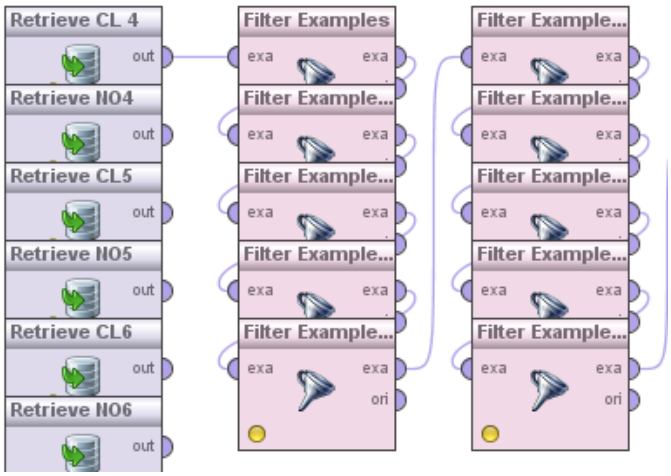


Figure (4.06): Filter Dataset by outlier mean standard deviation

Table (4.2): Detected outlier the 2nd experiment

	Chloride			Nitrate		
	Slandered deviation	Mean	Threshold	Slandered deviation	Mean	Threshold
Window = 4	300.03	356.20	1256.30	87.13	125.57	386.97
	299.93	373.14	1256.01	87.06	373.14	386.77
	299.84	390.44	1255.72	87.03	128.98	386.68
	299.74	409.37	1255.43	87.00	132.50	386.58
Window = 5	382.11	371.20	1517.53	114.81	125.19	469.63
	469.71	400.95	1810.08	116.87	129.24	479.86
	567.66	433.06	2136.06	119.54	132.05	490.70
	682.13	472.34	2563.03	117.38	130.79	482.95
	795.73	516.63	2903.84	117.78	130.94	484.28
Window = 6	229.01	272.51	974.503	85.98	126.25	387.62
	251.88	287.46	1066.53	85.86	129.66	391.51
	382.68	310.88	1478.22	88.21	133.91	399.32
	415.99	330.16	1605.93	88.16	134.68	397.73
	482.03	357.95	1835.60	88.01	133.23	398.32
	549.48	389.48	1886.47	88.83	134.28	381.77

4.2.3 Applied data mining techniques

After preprocessing, the following AI algorithms ANN, SVM, KNN and LR are applied for each window and repeated experiment after redetecting outliers as showed in 4.2.2.3.

Training data set consist of 70% from each dataset randomly and repeated three times for each algorithm to repeat experiments. Figure 4.7 and 4.8 are shows snapshots from Rapidminer for experiments.

In figure 4.8 the additional detecting outlier is added by adding new filters.

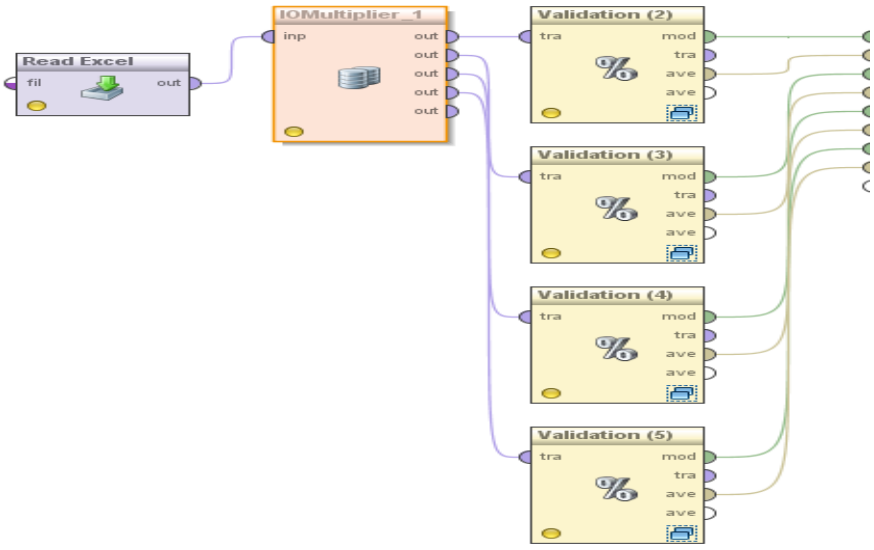


Figure (4.7) The first Experiment: Before Filter Dataset by outlier mean standard deviation

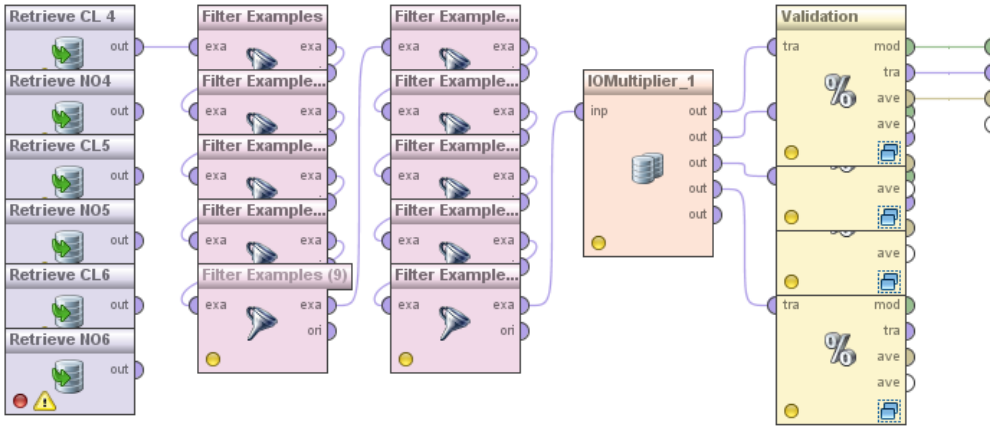


Figure (4.08): The second Experiment: After Filter Dataset by outlier mean standard deviation

4.2.4 Evaluation models

To evaluate models; Statistics Coefficient of Determination (R^2) and Root Mean Square Error (RMSE) are applied. It is done by using 30% from each dataset and repeated them randomly as complement for previous step showed in 4.5 and will be discussed in the next chapter.

Evaluation of the final model is done by using 2014 and 2015 data for 51 wells which have three reads before data gain from PWA (data for 2016 not available).

4.3 Summary

This chapter illustrates the methodology, which applied to develop prediction model for chloride and nitrate in groundwater wells depending on historical read only. The major stages were preprocessing, applied DM techniques and test and evaluation. Four AI algorithms applied as DM techniques and every experiment repeated tree time to make sure from results.

To improve results, feeding back to preprocessing to improve results; by using another way to detect outlier which using outlier mean standard deviation, then applied re-experiment is done.

Evaluation of the final model is done by using 2014 and 2015 data for 51 wells which three reads before. These data gained from PWA.

Chapter 5

Results and Discussion

Chapter 5

Results and Discussion

This chapter illustrates the results and their evaluation as comparison study to determine which DM techniques is the optimal in predicting chloride and nitrate in groundwater wells in Gaza Strip depending on historical read only. This work was done despite lack, missed and irregular data read.

Experiments are done to choose the optimal DM technique for this case in hydrological problem, and the chosen algorithm results is the result of the model. The first experiment is done with detecting outlier by grubbs test.

The second experiment used mean standard deviation to detect outlier as additional step to improve results and evaluated measures calculated by Rapidminer.

For KNN, there are many experiment is done which illustrate in Appendix 1 for first experiment and Appendix 2 for the other.

5.1 The First Experiment Results Presentation and analysis

The first experiment Using ANN, SVM, LR and KNN is done and their result shown in Table5.1 and figures 5.1, 5.2, 5.3 and 5.4.

The results is convergent close, with advantage for KNN for chloride (RMSE=311.228 and R^2=0.895) for window size 4 and LR for Nitrate (RMSE=32.522 and R^2=0.851) for six window size.

Table (5.1): Models evaluation results in the first experiment

	Chloride						Nitrate					
	Window 4		Window 5		Window 6		Window 4		Window 5		Window 6	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
ANN	382.13	0.844	333.905	0.873	429.609	0.732	41.919	0.78	33.991	0.844	32.318	0.854
SVM	564.666	0.828	541.667	0.823	394.916	0.692	40.66	0.786	37.003	0.812	35.038	0.822
LR	379.388	0.846	311.228	0.892	370.278	0.804	40.811	0.793	33.93	0.845	32.522	0.851
KNN	377.182	0.855	334.416	0.895	266.093	0.832	47.782	0.728	37.407	0.808	35.315	0.818

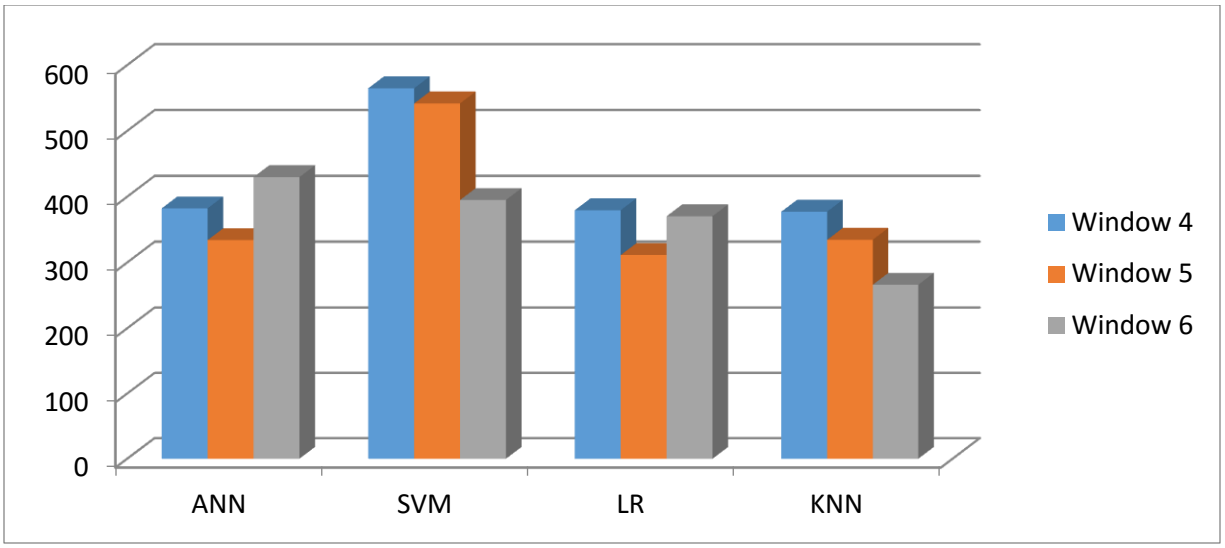


Figure (5.01): the first Experiment results (RMSE) for chloride

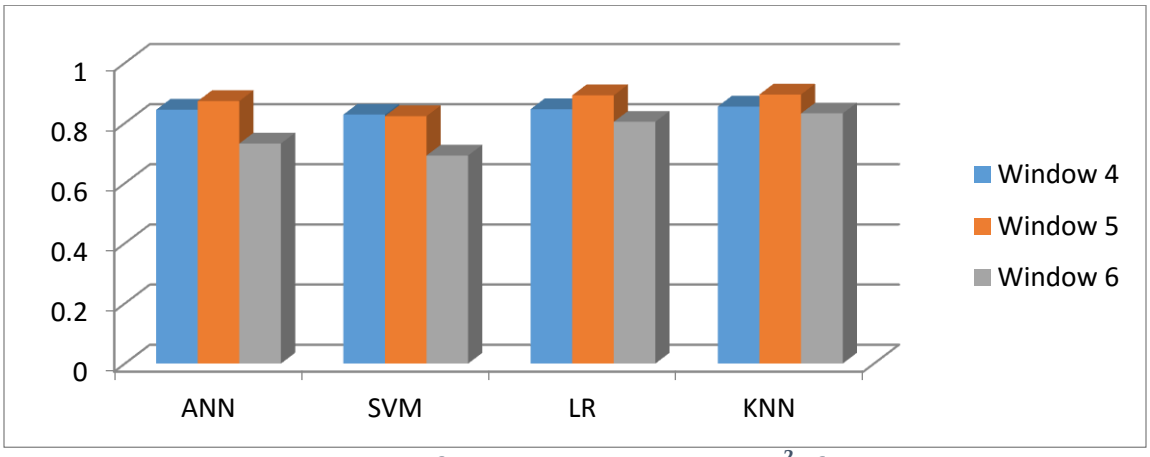


Figure (5.2): The first Experiment results (R^2) for chloride

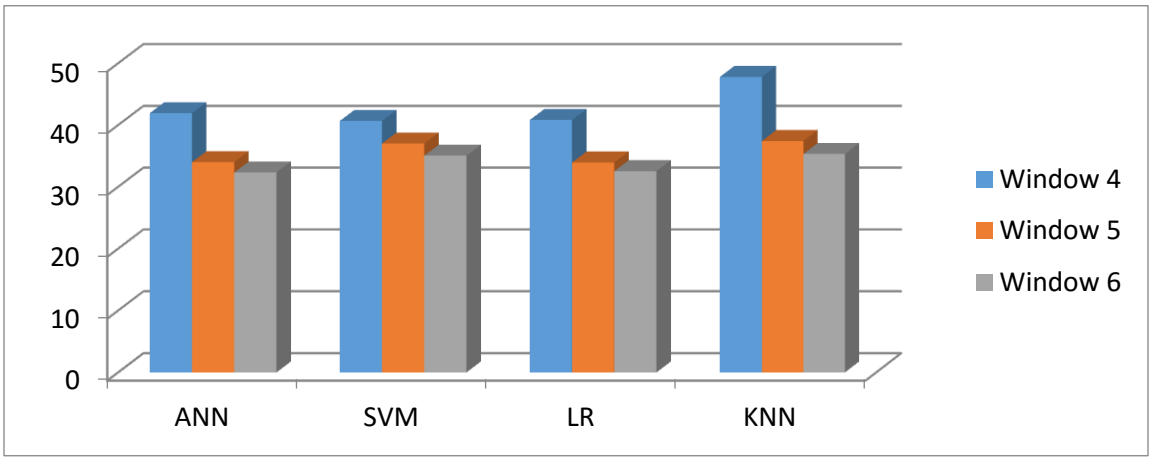


Figure (5.03): The first Experiment results (RMSE) for Nitrate

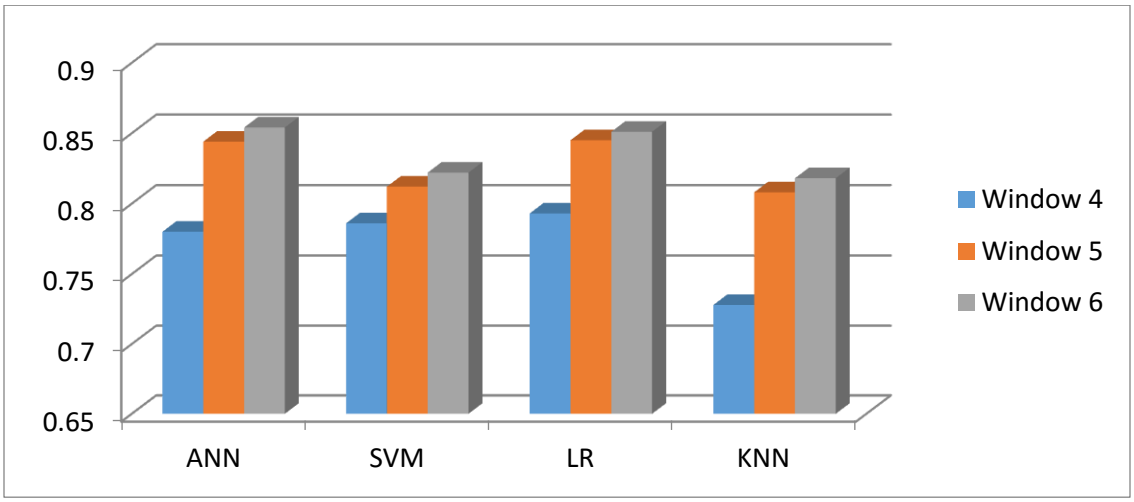


Figure (50.4): The first Experiment results (RMSE) for Nitrate

5.2 The second Experiment Results Presentation and analysis

As trying to improve results in the first experiment, the re-detecting outlier is done by using mean standard deviation. Their results is better than first experiment as shown in table 5.1 and figures 5.1,5.2,5.3 and 5.4.

Table (5.2): Models evaluation results in the second experiment

	Chloride						Nitrate					
	Window 4		Window 5		Window 6		Window 4		Window 5		Window 6	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
ANN	93.866	0.910	138.071	0.717	241.153	0.765	33.523	0.825	38.319	0.805	40.802	0.777
SVM	93.866	0.910	199.086	0.476	362.520	0.462	33.555	0.823	38.943	0.800	41.582	0.769
LR	64.864	0.954	167.438	0.586	236.905	0.771	33.330	0.828	38.015	0.808	40.775	0.777
KNN	87.644	0.924	39.469	0.803	104.027	0.888	35.300	0.797	39.547	0.793	36.382	0.789

The results is convergent close, with advantage for LR for chloride (RMSE=64.864 and R^2=0.954) for window size 4 and LR for Nitrate (RMSE=33.33 and R^2=0.828) for same window size.

The improvement results is succeeded which make less RMSE and higher R^2.

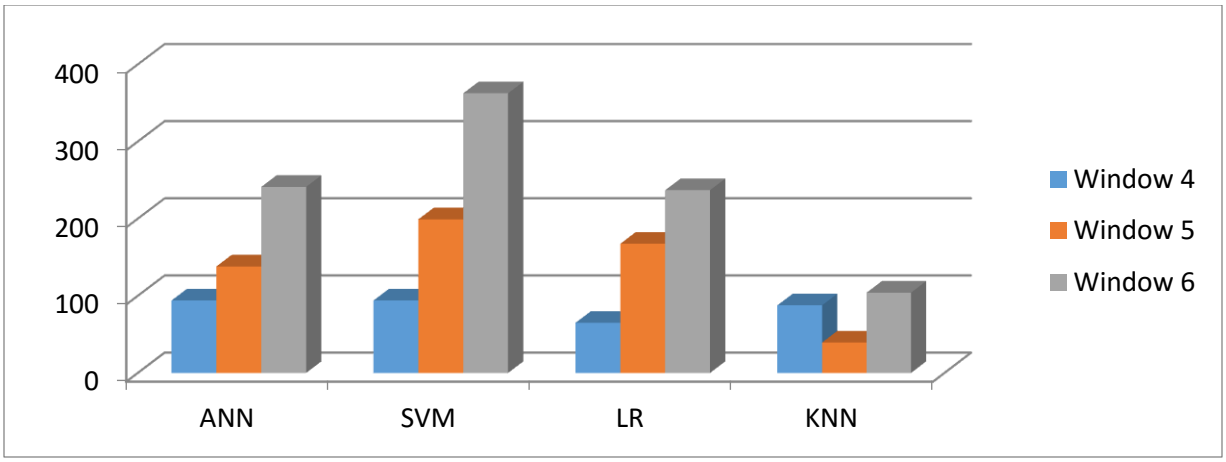


Figure (5.5): The second experiment results (RMSE) for chloride

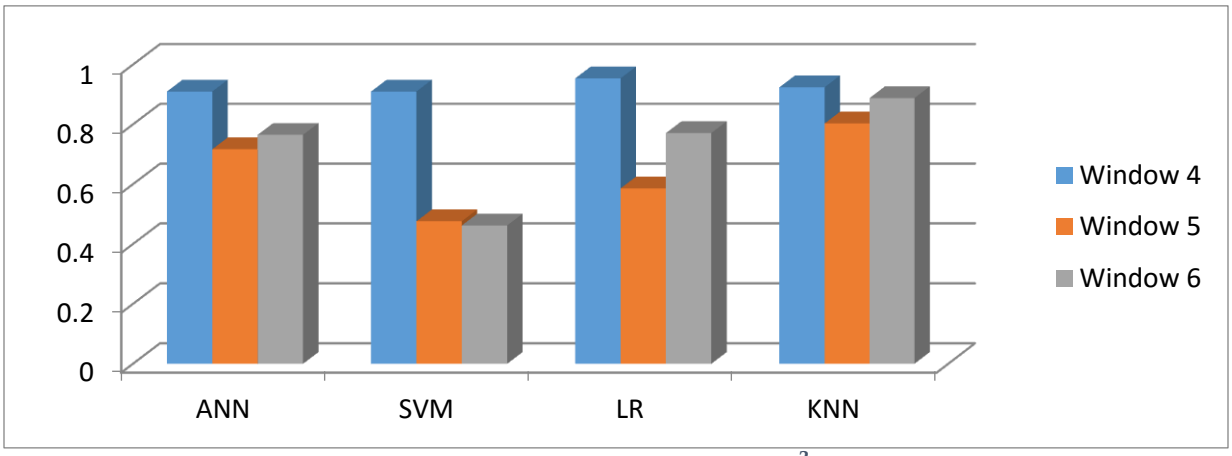


Figure (5.6): The second experiment results (R^2) for chloride

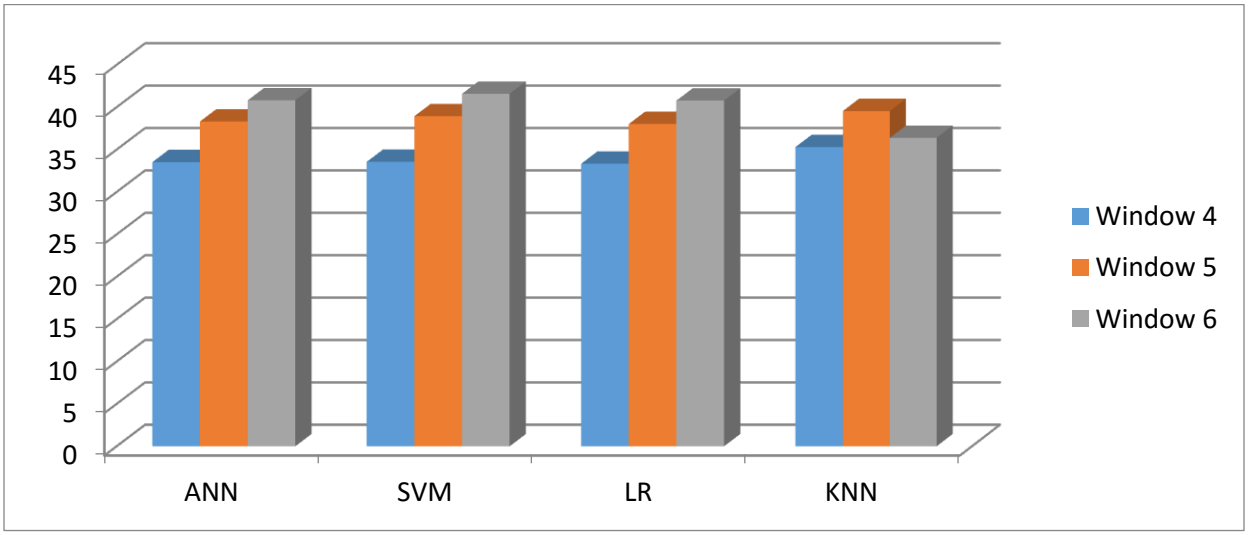


Figure (5.7): The second experiment results (RMSE) for Nitrate

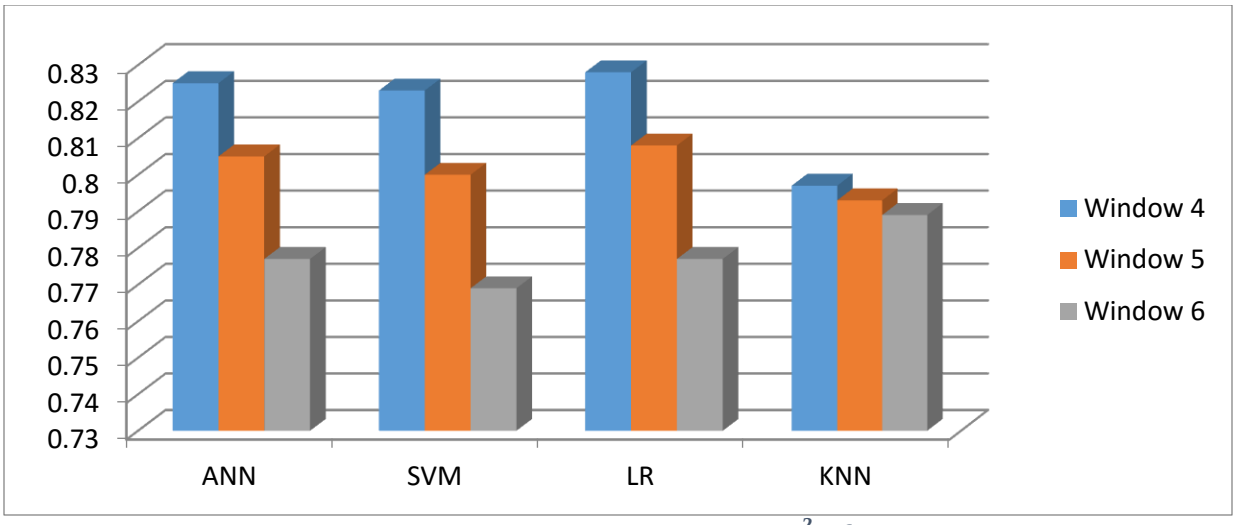


Figure (5.8): The second experiment results (R^2) for Nitrate

5.3 Evaluation results for final model

Evaluation for final model is done by using data 2014 and 2015 for municipal wells in Gaza strip in the same range (1256 for chloride and 387 for Nitrate). It is results as follow:

- RMSE(Chloride 2014) = 82.75 and R2(Chloride 2014) =0.919 (48 wells)
- RMSE(Chloride 2014) = 83.92 and R2(Chloride 2014) =0.917 (48 wells)
- RMSE(Nitrate 2014) =35.835 and R2(Nitrate 2014) = 0.862 (51 wells)
- RMSE(Nitrate 2015) = 32.22 and R2(Nitrate 2015) = 0.898 (51 wells)

5.4 Result presentation and analysis summary

The DM techniques has convergent close with preference for LR. It is not necessary to say that LR is the best but it is optimum for this case despite lack and error read which might change in different area.

The model can predict CL and No3 for groundwater wells in targeted area as output by using last three reads as input. It is evaluated by additional data set in the same range (48 wells).

Chapter 6

Conclusions

Chapter 6

Conclusions

6.1 Conclusion

The groundwater management in Gaza strip needs support system to predict chloride and nitrate read in the future despite lack of data, information and fund.

To avoid these problems, this research which depends on DM and its AI techniques took place to predict the results from limited data. A comparison study between many DM and AI approached took place to find the optimum of approach. The research conclude that the LR gave the most reliable and accurate results.

The best result has R^2 =0.954 despite lack, irregular and error data read. After two outliers detecting process, there were few misread which cannot be repaired especially in nitrate, which needs hydrologist to detect it separately as shown in figure 6.1 and figure6.2.

Model evaluation process took place using 2014 and 2015 PWA data for the targeted area. RMSE was less than 84 and 36 and R^2 was more than 0.9 and 0.86 for chloride and nitrate respectively.

The research is independent of physical data, which is not available in Gaza strip. In addition, according to researcher knowledge, it is the first time to used time series in the targeted area.

A1	A2	A3	A4
1076	1004	953	251
2681	4098.5	2008	444.6
953	251	251	838
539.63	206.65	938.8	1424

Figure (6.1): Example of Error read Despite Outlier Detection (Cl)

A1	A2	A3	A4	A5
694	300	155	0	310
266	35	317.5	247.5	355
268	65	355	35	44.5
529	88.7	447.9	458.3	411

Figure (6.2): Example of Error read Despite Outlier Detection (NO_3)

6.2 Future Work

The water quality studies are very important field to be discussed, so the researcher suggests these researches to be done;

- Database correction: the database should be manipulated and corrected using DM preprocessing techniques.
- Repeat the experiment using more DM techniques to have higher accuracy to detect a realistic pattern for groundwater in targeted area.

The Reference List

Al-Khatib, I. A., & Arafat, H. A. (2009). Chemical and microbiological quality of desalinated water, groundwater and rain-fed cisterns in the Gaza strip, Palestine. *Desalination, 249*(3), 1165-1170.

Alagha, J. S., Said, M. A. M., & Mogheir, Y. (2014). Modeling of nitrate concentration in groundwater using artificial intelligence approach—a case study of Gaza coastal aquifer. *Environmental monitoring and assessment, 186*(1), 35-45.

Alagha, J. S., Said, M. A. M., Mogheir, Y., & Seyam, M. (2013). *Modelling of chloride concentration in coastal aquifers using artificial neural network—a case study: Khanyounis Governorate Gaza Strip-Palestine.* Paper presented at the Casp J Appl Sci Res (AWAm International Conference on Civil Engineering and Geohazard Information Zonation).

Alastal, K. M., Alagha, J. S., Abuhabib, A. A., & Ababou, R. (2016). Groundwater Quality Assessment Using Water Quality Index (WQI) Approach: Gaza Coastal Aquifer Case Study. *Journal of Engineering Research and Technology, 2*(1).

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. Knowledge and information systems, 51(2), 339.

Andrew, A. M. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, xiii+ 189 pp., ISBN 0-521-78019-5 (Hbk,£ 27.50): Cambridge Univ Press.

Arabgol, R., Sartaj, M., & Asghari, K. (2016). Predicting Nitrate Concentration and Its Spatial Distribution in Groundwater Resources Using Support Vector Machines (SVMs) Model. *Environmental Modeling & Assessment, 21*(1), 71-82.

Asefa, T., Kemblowski, M., McKee, M., & Khalil, A. (2006). Multi-time scale stream flow predictions: the support vector machines approach. *Journal of Hydrology, 318*(1), 7-16.

Barzegar, R., Moghaddam, A. A., & Baghban, H. (2016). A supervised committee machine artificial intelligent for improving DRASTIC method to assess groundwater contamination risk: a case study from Tabriz plain aquifer, Iran. *Stochastic environmental research and risk assessment, 30*(3), 883-899.

Basheer, I., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods, 43*(1), 3-31.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine, 17*(10), 1169-1186.

Boyd, C. E. (2015). *Water quality: an introduction*: Springer.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., . . . Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences, 97*(1), 262-267.

Chang, C., & Lin, C. (2005). LIBSVM: a library for support vector machines (2001), Software.

Chang, J. F. (2016). *Business process management systems: strategy and implementation*: CRC Press.

Chuchro, M., Lupa, M., Pieta, A., Piórkowski, A., & Lesniak, A. (2014). *A Concept of Time Windows Length Selection in Stream Databases in the Context of Sensor Networks Monitoring.* Paper presented at the ADBIS (2).

Clinchant, S., Csurka, G., & Chidlovskii, B. (2016). *Transductive Adaptation of Black Box Predictions.* Paper presented at the The 54th Annual Meeting of the Association for Computational Linguistics.

Copeland, T. E. (2011). *Drawing a Line in the Sea: The Gaza Flotilla Incident and the Israeli-Palestinian Conflict*: Lexington Books.

da Silva, G. J. C., Neder, H. D., & Santos, H. S. (2017). A lei de Verdoorn-Kaldor-Thirlwall: uma análise empírica. *Revista Econômica do Nordeste, 47*(3), 149-166.

De Luca, C., Zinno, I., Manunta, M., Lanari, R., & Casu, F. (2017). Large areas surface deformation analysis through a cloud computing P-SBAS approach for massive processing of DInSAR time series. Remote sensing of environment.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* (Vol. 66): CRC Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling, 160*(3), 249-264.

Ghosh, D., & Vogt, A. (2012). *Outliers: An evaluation of methodologies.* Paper presented at the Joint Statistical Meetings.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data mining and knowledge discovery, 1*(1), 11-28.

Gong, Y., Zhang, Y., Lan, S., & Wang, H. (2016). A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida. *Water Resources Management, 30*(1), 375-391.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Healy, K. (2005). Book Review: An R and S-PLUS Companion to Applied Regression. *Sociological Methods & Research, 34*(1), 137-140.

Hill, T., Lewicki, P., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*: StatSoft, Inc.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review, 22*(2), 85-126.

Knorr, E. M., & Ng, R. T. (1997). *A Unified Notion of Outliers: Properties and Computation.* Paper presented at the KDD.

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications, 38*(10), 13448-13467.

Larose, D. T. (2005). k - Nearest Neighbor Algorithm. *Discovering Knowledge in Data: An Introduction to Data Mining*, 90-106.

Lee, J., Im, J., Kim, U., & Löffler, F. E. (2016). A data mining approach to predict in situ detoxification potential of chlorinated ethenes. *Environmental science & technology, 50*(10), 5181-5188.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). *Introduction to linear regression analysis*: John Wiley & Sons.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*: John Wiley & Sons.

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental monitoring and assessment, 188*(1), 1-27.

Nisbet, R., Miner, G., & Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*: Academic Press.

Nourani, V., Alami, M. T., & Vousoughi, F. D. (2016). Self-organizing map clustering technique for ANN-based spatiotemporal modeling of groundwater quality parameters. *Journal of Hydroinformatics, 18*(2), 288-309.

Oorkavalan, G., Chidambaram, S. M., Mariappan, V., Kandaswamy, G., & Natarajan, S. (2016). Cluster Analysis to Assess Groundwater Quality in Erode District, Tamil Nadu, India. *Circuits and Systems, 7*(06), 877.

Organization, W. H. (2011). Guidelines for Drinking Water Quality (World Health Organization, Geneva).

Rebolledo, B., Gil, A., Flotats, X., & Sánchez, J. Á. (2016). Assessment of groundwater vulnerability to nitrates from agricultural sources using a GIS-compatible logic multicriteria model. *Journal of environmental management, 171*, 70-80.

Sattari, M. T., RezazadehJoudi, A., & Kusiak, A. (2016). Estimation of Water Quality Parameters With Data-Driven Model. *JOURNAL AWWA, 108*, 4.

Schalkoff, R. J. (1997). *Artificial neural networks*: McGraw-Hill Higher Education.

Sharma, S., & Osei-Bryson, K.-M. (2009). Framework for formal implementation of the business understanding phase of data mining projects. *Expert systems with Applications, 36*(2), 4114-4124.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner*: John Wiley & Sons.

Trichakis, I. C., Nikolos, I. K., & Karatzas, G. (2011). Artificial neural network (ANN) based modeling for karstic groundwater level simulation. *Water Resources Management, 25*(4), 1143-1152.

Vincent, J. (2016). PHYSICO CHEMICAL ANALYSIS OF GROUND WATER NEAR MUNICIPAL SOLID WASTE DUMPING SITES IN ARUMUGANERI, THOOTHUKUDI DISTRICT, TAMILNADU, INDIA. *Green Chemistry & Technology Letters, 2*(1), 35-37.

Wang, C.-H. (2009). Outlier identification and market segmentation using kernel-based clustering techniques. *Expert systems with Applications, 36*(2), 3744-3750.

Wang, D., Liu, D., Ding, H., Singh, V. P., Wang, Y., Zeng, X., . . . Wang, L. (2016). A cloud model-based approach for water quality assessment. *Environmental research, 148*, 24-35.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research, 30*(1), 79-82.

Yoo, K., Shukla, S. K., Ahn, J. J., Oh, K., & Park, J. (2016). Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity. *Journal of Cleaner Production, 122*, 277-286.

Appendices

Appendix 1: Full results for first experiment by using KNN

KNN	Window Size	Chloride	Nitrate
K=1	4	root_mean_squared_error: 455.409 +/- 0.000 squared_correlation: 0.770	root_mean_squared_error: 53.537 +/- 0.000 squared_correlation: 0.654
K=2	4	root_mean_squared_error: 377.182 +/- 0.000 squared_correlation: 0.855	root_mean_squared_error: 109.197 +/- 0.000 squared_correlation: 0.268
K=5	4	root_mean_squared_error: 438.863 +/- 0.000 squared_correlation: 0.816	root_mean_squared_error: 70.571 +/- 0.000 squared_correlation: 0.487
K=10	4	root_mean_squared_error: 484.137 +/- 0.000 squared_correlation: 0.802	root_mean_squared_error: 54.043 +/- 0.000 squared_correlation: 0.637
K=20	4	root_mean_squared_error: 532.791 +/- 0.000 squared_correlation: 0.785	root_mean_squared_error: 49.221 +/- 0.000 squared_correlation: 0.687
K=40	4	root_mean_squared_error: 620.235 +/- 0.000 squared_correlation: 0.712	root_mean_squared_error: 47.215 +/- 0.000 squared_correlation: 0.711
K=60	4	root_mean_squared_error: 668.336 +/- 0.000 squared_correlation: 0.648	root_mean_squared_error: 47.537 +/- 0.000 squared_correlation: 0.713
K=80	4	root_mean_squared_error: 697.839 +/- 0.000 squared_correlation: 0.604	root_mean_squared_error: 47.470 +/- 0.000 squared_correlation: 0.726
K=100	4	root_mean_squared_error: 719.750 +/- 0.000 squared_correlation: 0.567	root_mean_squared_error: 47.782 +/- 0.000 squared_correlation: 0.728
K=1	5	root_mean_squared_error: 339.673 +/- 0.000 squared_correlation: 0.870	root_mean_squared_error: 48.839 +/- 0.000 squared_correlation: 0.694
K=2	5	root_mean_squared_error: 334.416 +/- 0.000 squared_correlation: 0.895	root_mean_squared_error: 44.769 +/- 0.000 squared_correlation: 0.727
K=5	5	root_mean_squared_error: 408.276 +/- 0.000 squared_correlation: 0.842	root_mean_squared_error: 39.810 +/- 0.000 squared_correlation: 0.781
K=10	5	root_mean_squared_error: 441.539 +/- 0.000 squared_correlation: 0.831	root_mean_squared_error: 37.818 +/- 0.000 squared_correlation: 0.803
K=20	5	root_mean_squared_error: 521.472 +/- 0.000	root_mean_squared_error: 37.288 +/- 0.000

		squared_correlation: 0.785	squared_correlation: 0.808
K=40	5	root_mean_squared_error: 609.516 +/- 0.000 squared_correlation: 0.719	root_mean_squared_error: 37.407 +/- 0.000 squared_correlation: 0.808
K=60	5	root_mean_squared_error: 660.842 +/- 0.000 squared_correlation: 0.663	root_mean_squared_error: 38.407 +/- 0.000 squared_correlation: 0.802
K=80	5	root_mean_squared_error: 690.771 +/- 0.000 squared_correlation: 0.620	root_mean_squared_error: 39.525 +/- 0.000 squared_correlation: 0.795
K=100	5	root_mean_squared_error: 713.428 +/- 0.000 squared_correlation: 0.581	root_mean_squared_error: 40.591 +/- 0.000 squared_correlation: 0.788
K=1	6	root_mean_squared_error: 454.883 +/- 0.000 squared_correlation: 0.696	root_mean_squared_error: 49.860 +/- 0.000 squared_correlation: 0.681
K=2	6	root_mean_squared_error: 431.316 +/- 0.000 squared_correlation: 0.693	root_mean_squared_error: 46.713 +/- 0.000 squared_correlation: 0.697
K=5	6	root_mean_squared_error: 335.673 +/- 0.000 squared_correlation: 0.770	root_mean_squared_error: 38.134 +/- 0.000 squared_correlation: 0.789
K=10	6	root_mean_squared_error: 283.983 +/- 0.000 squared_correlation: 0.813	root_mean_squared_error: 36.075 +/- 0.000 squared_correlation: 0.811
K=20	6	root_mean_squared_error: 266.093 +/- 0.000 squared_correlation: 0.832	root_mean_squared_error: 35.524 +/- 0.000 squared_correlation: 0.816
K=40	6	root_mean_squared_error: 304.179 +/- 0.000 squared_correlation: 0.819	root_mean_squared_error: 35.315 +/- 0.000 squared_correlation: 0.818
K=60	6	root_mean_squared_error: 346.694 +/- 0.000 squared_correlation: 0.789	root_mean_squared_error: 36.074 +/- 0.000 squared_correlation: 0.813
K=80	6	root_mean_squared_error: 374.712 +/- 0.000 squared_correlation: 0.758	root_mean_squared_error: 36.800 +/- 0.000 squared_correlation: 0.810
K=100	6	root_mean_squared_error: 396.068 +/- 0.000 squared_correlation: 0.729	root_mean_squared_error: 37.934 +/- 0.000 squared_correlation: 0.804

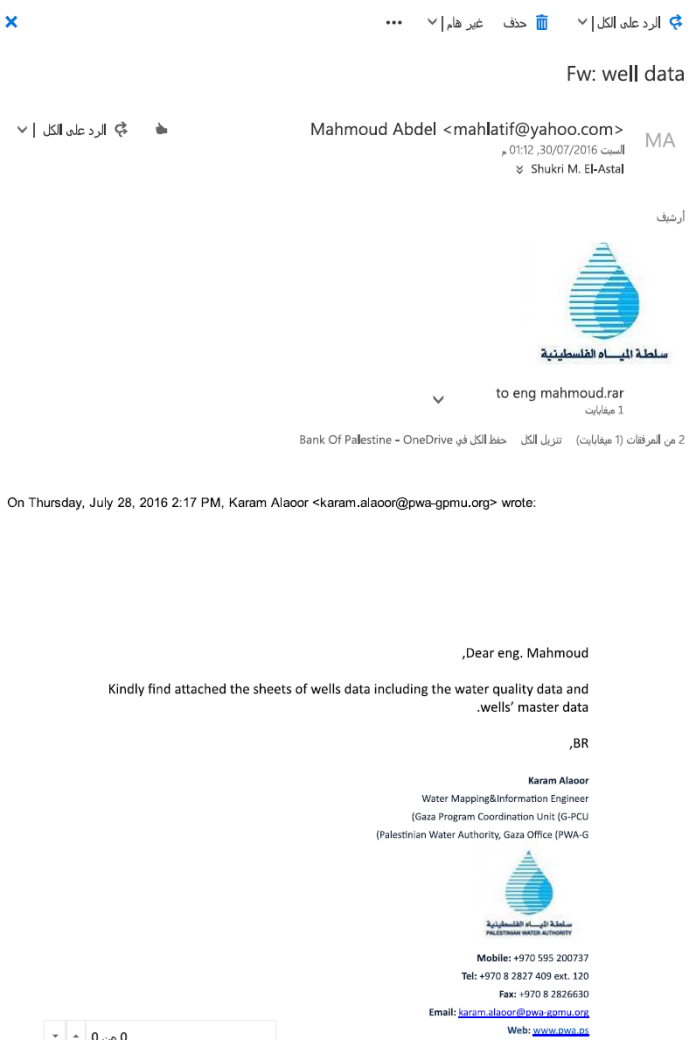
Appendix 2: Full results for second experiment by using KNN

KNN	Window Size	Chloride	Nitrate
K=1	4	root_mean_squared_error: 110.540 +/- 0.000 squared_correlation: 0.874	root_mean_squared_error: 52.557 +/- 0.000 squared_correlation: 0.598
K=2	4	root_mean_squared_error: 99.086 +/- 0.000 squared_correlation: 0.899	root_mean_squared_error: 43.947 +/- 0.000 squared_correlation: 0.691
K=5	4	root_mean_squared_error: 90.056 +/- 0.000 squared_correlation: 0.918	root_mean_squared_error: 38.317 +/- 0.000 squared_correlation: 0.759
K=10	4	root_mean_squared_error: 87.644 +/- 0.000 squared_correlation: 0.924	root_mean_squared_error: 36.461 +/- 0.000 squared_correlation: 0.782
K=20	4	root_mean_squared_error: 89.063 +/- 0.000 squared_correlation: 0.922	root_mean_squared_error: 35.300 +/- 0.000 squared_correlation: 0.797
K=40	4	root_mean_squared_error: 95.240 +/- 0.000 squared_correlation: 0.913	root_mean_squared_error: 35.440 +/- 0.000 squared_correlation: 0.796
K=60	4	root_mean_squared_error: 98.563 +/- 0.000 squared_correlation: 0.907	root_mean_squared_error: 36.427 +/- 0.000 squared_correlation: 0.786
K=80	4	root_mean_squared_error: 102.118 +/- 0.000 squared_correlation: 0.902	root_mean_squared_error: 37.163 +/- 0.000 squared_correlation: 0.780
K=100	4	root_mean_squared_error: 104.719 +/- 0.000 squared_correlation: 0.898	root_mean_squared_error: 38.065 +/- 0.000 squared_correlation: 0.771
K=1	5	root_mean_squared_error: 53.132 +/- 0.000 squared_correlation: 0.664	root_mean_squared_error: 48.460 +/- 0.000 squared_correlation: 0.700
K=2	5	root_mean_squared_error:	root_mean_squared_error:

		47.332 +/- 0.000 squared_correlation: 0.715	46.635 +/- 0.000 squared_correlation: 0.714
K=5	5	root_mean_squared_error: 41.252 +/- 0.000 squared_correlation: 0.782	root_mean_squared_error: 41.594 +/- 0.000 squared_correlation: 0.768
K=10	5	root_mean_squared_error: 40.593 +/- 0.000 squared_correlation: 0.790	root_mean_squared_error: 39.926 +/- 0.000 squared_correlation: 0.787
K=20	5	root_mean_squared_error: 39.469 +/- 0.000 squared_correlation: 0.803	root_mean_squared_error: 39.529 +/- 0.000 squared_correlation: 0.791
K=40	5	root_mean_squared_error: 40.399 +/- 0.000 squared_correlation: 0.796	root_mean_squared_error: 39.547 +/- 0.000 squared_correlation: 0.793
K=60	5	root_mean_squared_error: 41.299 +/- 0.000 squared_correlation: 0.791	root_mean_squared_error: 40.440 +/- 0.000 squared_correlation: 0.788
K=80	5	root_mean_squared_error: 42.774 +/- 0.000 squared_correlation: 0.783	root_mean_squared_error: 41.649 +/- 0.000 squared_correlation: 0.781
K=100	5	root_mean_squared_error: 44.458 +/- 0.000 squared_correlation: 0.770	root_mean_squared_error: 43.136 +/- 0.000 squared_correlation: 0.768
K=1	6	root_mean_squared_error: 113.531 +/- 0.000 squared_correlation: 0.871	root_mean_squared_error: 50.249 +/- 0.000 squared_correlation: 0.615
K=2	6	root_mean_squared_error: 104.027 +/- 0.000 squared_correlation: 0.888	root_mean_squared_error: 39.166 +/- 0.000 squared_correlation: 0.754
K=5	6	root_mean_squared_error: 102.246 +/- 0.000 squared_correlation: 0.896	root_mean_squared_error: 36.538 +/- 0.000 squared_correlation: 0.785
K=10	6	root_mean_squared_error: 110.032 +/- 0.000 squared_correlation: 0.884	root_mean_squared_error: 36.653 +/- 0.000 squared_correlation:

			0.783
K=20	6	root_mean_squared_error: 128.522 +/- 0.000 squared_correlation: 0.842	root_mean_squared_error: 36.382 +/- 0.000 squared_correlation: 0.789
K=40	6	root_mean_squared_error: 140.556 +/- 0.000 squared_correlation: 0.810	root_mean_squared_error: 37.229 +/- 0.000 squared_correlation: 0.785
K=60	6	root_mean_squared_error: 144.597 +/- 0.000 squared_correlation: 0.798	root_mean_squared_error: 38.677 +/- 0.000 squared_correlation: 0.774
K=80	6	root_mean_squared_error: 147.533 +/- 0.000 squared_correlation: 0.792	root_mean_squared_error: 40.083 +/- 0.000 squared_correlation: 0.766
K=100	6	root_mean_squared_error: 149.517 +/- 0.000 squared_correlation: 0.788	root_mean_squared_error: 41.209 +/- 0.000 squared_correlation: 0.759

Appendix 3: Data Sources



الرد على الكل | حذف غير هام |

Fw: well data

Mahmoud Abdel <mahlatif@yahoo.com> MA

السبت 30/07/2016, 01:12 م

Shukri M. El-Astal

أرشيف

سلطة الميـــاه الفلسطينية

to eng mahmoud.rar

1 ميغابايت

2 من المرفقات (1 ميغابايت) تنزيل الكل حفظ الكل في Bank Of Palestine - OneDrive

On Thursday, July 28, 2016 2:17 PM, Karam Alaoor <karam.alaoor@pwa-gpmu.org> wrote:

,Dear eng. Mahmoud

Kindly find attached the sheets of wells data including the water quality data and .wells' master data

,BR

Karam Alaoor
Water Mapping&Information Engineer
(Gaza Program Coordination Unit (G-PCU
(Palestinian Water Authority, Gaza Office (PWA-G

سلطة المياه الفلسطينية PALESTINIAN WATER AUTHORITY

Mobile: +970 595 200737
Tel: +970 8 2827 409 ext. 120
Fax: +970 8 2826630
Email: karam.alaoor@pwa-gpmu.org
Web: www.pwa.ps