American Journal of Environmental Engineering 2013, 3(2): 100-106
DOI: 10.5923/j.ajee.20130302.03

Improving the Accuracy of Artificial Intelligence - Based Groundwater Quality Models Using Clustering Technique - A Case Study

Jawad S. Alagha[1], Md Azlin Md Said[1,*], Yunes Mogheir[2]

[1]School of Civil Engineering, Universiti Sains Malaysia (USM), Nibong Tebal, Pulau Pinang, Malaysia
[2]Environmental Engineering Department, Engineering Faculty, Islamic University of Gaza, Palestine

Abstract In this study, the simulation performances of two artificial intelligence (AI) techniques – namely, artificial neural networks (ANNs) and support vector machine (SVM) – for groundwater quality modeling were improved by grouping input data into consistent clusters as a pre-modeling technique. AI techniques were applied to model the concentrations of chloride and nitrate using data from the Gaza coastal aquifer in Palestine, which is a very complex hydro-geological system. Research results indicated that developing separate AI models for each cluster reduced the mean absolute percentage errors (MAPE) of the ANNs' models by 20% and 37% for chloride and nitrate, respectively. Meanwhile, the MAPE of the SVM's models was reduced by 10% and by 13% for chloride and nitrate, respectively. Improving the simulation accuracy of AI techniques would lead to more rational and effective decisions for groundwater management.

Keywords Artificial Neural Networks (ANNs), Chloride, Clustering, Gaza Coastal Aquifer (GCA), Nitrate, Support Vector Machine (SVM)

1. Introduction

In the last decade, artificial intelligence (AI) techniques have become highly popular and widely used in modeling hydrological complicated processes using relatively less cost and effort[1]. The superiority of AI techniques becomes apparent when accurately describing the hydrological process is difficult, and when the available data are insufficient to apply numerical and physical models, which is the case for many groundwater (GW) quality problems[2].

Recognizing their superior capabilities, the uses of various AI techniques, such as artificial neural networks (ANNs) and support vector machine (SVM), in hydrological applications have considerably increased over the previous decade. For example, ANNs have been successfully applied to different GW applications[3-7]. Likewise, the application of SVM has attracted more attention in recent years for modeling both surface water and GW processes[8-10].

Despite the wide strides towards the utilization of AI techniques for GW quality modeling, some areas still require further investigation in this context. Literature revealed that the SVM application for GW modeling is scarcer compared with the growing applications in surface water problems[9]. For example, no study was found to use SVM for modeling the concentration of chloride in GW. Furthermore, with regard to nitrate modeling, none of the earlier studies utilized SVM to estimate nitrate concentration in GW based on the potential influencing variables. As for ANNs, very few applications were found related to model chloride and nitrate concentrations in GW using explanatory variables. In such studies, such as that of Seyam and Moghier[6], the accuracy needs further improvement. On the other hand, studies that were relatively accurate required substantial data input, and utilized sophisticated methods for input calculations; therefore, their applicability could be very limited due to the detailed and accurate data required. An example of such study is that of Almasri and Kaluarachchi[7].

In the field of AI applications for GW quality, the simplification of AI models and their improved accuracy without the need for extra data and effort have become the trend. Thus, research on hybrid models that integrate AI with other techniques is considered to be a promising field, with models being developed that use minimum data, time, and effort[11]. These targeted models could then be effectively utilized to support management decisions related to GW quality.

* Corresponding author:
azlin@eng.usm.my (Md Azlin Md Said)

Published online at http://journal.sapub.org/ajee

This paper aims to improve the simulation performance of ANNs- and SVM-based GW quality models by performing data clustering as a pre-modeling technique. The hybrid systems, composed of AI and clustering techniques, have been applied for modeling both chloride and nitrate concentrations in GW using data from the Gaza coastal aquifer (GCA) in Palestine, which is a very complex aquifer. The improvement of AI simulation performance could in turn lead to more accurate prediction and rational management processes of water resources.

2. Materials and Methods

2.1. Study Area

The Gaza Strip (GS) area is located at the eastern coast of the Mediterranean Sea (Figure 1 (b)). It is one of the most densely populated areas in the world with an average density of more than 4300 inhabitants/km^2; and it is expected that the population density will exceed 5835 inhabitants/km^2 in 2020[12]. GS is administratively divided into five governorates, among which Khanyounis governorate which is the study area as shown in Figure 1(c), has the largest area of about 112 km^2 with a total population of about 300,000 inhabitants[13]. GS is an extreme model on how unstable political environment, disastrous economic situation, decaying environmental conditions and unplanned human activities are combined together to further deteriorate the GW quality[14].

Gaza coastal aquifer (GCA) as shown in Figure 1(a) is the only natural source of water for different purposes in GS. According to UNCT[12] the GW situation in GCA is deteriorating and it could become unusable as early as 2016. GCA suffers from two water quality problems which are the high concentrations of chloride and nitrate[15]. Where, less than 5% of GS municipal water wells meet world health organization (WHO) chloride standards. Moreover chloride concentration in many wells of Khanyounis governorate reached 10 times more than WHO standards[16]. The main sources of the elevated chloride concentration in Khanyounis governorate are seawater intrusion, extensive exploitation, saline water flux from the neighboring eastern Eocene aquifer, and salty water lenses exist in many locations at deeper layers[17]. Likewise, the average nitrate concentration in Khanyounis governorate wells is 191 mg/l which is almost 4 times WHO standards for nitrate[18]. The main sources of the high nitrate levels are disposing of untreated wastewater into the aquifer through cesspits and septic tanks[15]. Additionally agricultural activities where thousands of tons of animal manure and synthetic fertilizers that exceed crop demands are usually applied resulting in leaching the excess nitrogen load into the aquifer[19].

2.2. Data Collection

Two separate groups of models were developed using potential influencing variables: the first one was related to chloride, while the second, to nitrate. The primary step for modeling water quality using AI techniques is to develop an input-output response matrix between the inputs (potential influencing variables) and outputs (concentration of chemical parameters). Based on the availability of monitoring data, 22 wells that constitute 80% of the municipal wells in Khanyounis governorate were used to develop the AI models. Periodic water quality analyses for municipal wells are usually performed twice a year, in spring (May) and in autumn (November). Chloride and nitrate monitoring data in the case study wells and all associated variables from 2000 to 2010 were collected from the database of the related institutions.

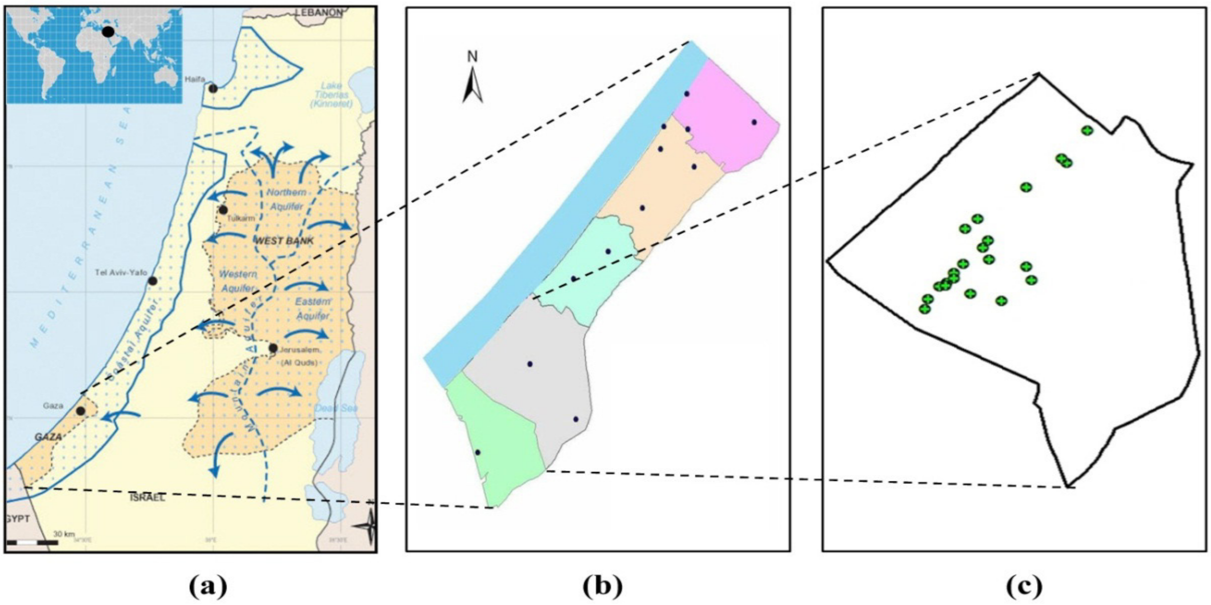

Figure 1. (a) Gaza Strip and Gaza coastal aquifer layout; (b); Gaza Strip Map; and (c) Location of municipal wells in Khanyounis governorate

The Thiessen polygons technique was used to delineate the influence area of each well, in which all calculations related to input variables were based on. Thiessen polygons is a simple, well-known, and widely used technique that has been used for various hydrological applications[20, 21].

To delineate the wells' influence area, each municipal and agricultural well in Khanyounis governorate (about 1100 agricultural wells) was plotted on the map. Subsequently, Thiessen polygons were created using ArcMap10. Afterwards, the polygons belonging to the case study municipal wells, (22 wells), were used for further analyses. To account for the effects of land use land cover (LULC) on GW quality, three aerial photos of the study area for the years 1999, 2003, and 2007 were analyzed using ERDAS IMAGINE 11 and ArcGIS 10 software. The entire area of each Thiessen polygon was grouped into three LCLU categories (built up, open, and agricultural areas). The values for the LULC recharge coefficient of each category were obtained from previous studies[22]. Additionally, the soil recharge coefficient for the study area, which depends on soil type and texture in the well's area, was also considered based on the basic GS soil type classification maps[23]. All GW recharge sources, including rainfall, leakage from water distribution networks, areas without sewers, and return flow from irrigation were considered in calculating the total recharge from each LCLU category.

The potential input variables for chloride model included variables such as the previous chloride concentration monitoring record (Cl_o), which was measured 6 months prior; recharge from each LCLU category; cumulative abstraction from each well for the past 6 months; distance to the Khanyounis center, which accounts for the effects of both seawater intrusion and lateral flow from the adjacent eastern aquifer; aquifer or sub-aquifer thickness; and well screen depth. The potential input variables for nitrate model included many variables, such as the previous nitrate concentration monitoring record (NO_{3o}); cumulative abstraction from each well for the past 6 months; the total recharge from surface to aquifer inside each Thiessen polygon during the past 6 months from built up areas, open areas, and agricultural areas; the estimated surface nitrogen load (N-load) during the past 6 months from built up areas and agricultural areas; overall N-load from each Thiessen polygon by all LULC categories; as well as the multiplication of GW recharge and N-load in built up and agricultural areas.

Both ANNs and SVM models were applied for the 22 wells for each GW parameter (NO_3 and Cl) as one group; hereafter this group will be termed as *an un-clustered model*. Then, k-means clustering technique was applied for clustering the 22 wells according to their similarity with respect to a number of chemical parameters. Afterward, AI models were separately applied on each cluster. Then these separated models were assembled together forming *an aggregated clustered model*. AI modeling and clustering have been performed using Statistica7 and Microsoft Office Excel 2010 softwares. Two performance evaluation criteria were used for models' evaluation. These criteria were the mean absolute error (MAE), and the mean average percentage error (MAPE). The formula to calculate the error indicators are:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \qquad (1)$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{O_i - P_i}{O_i}\right| X100\% \qquad (2)$$

Where: n = number of data pairs (observations); O_i = the i_{th} observed value; P_i = the i_{th} predicted value.

3. Results and Discussion

3.1. Clustering of Monitoring Wells

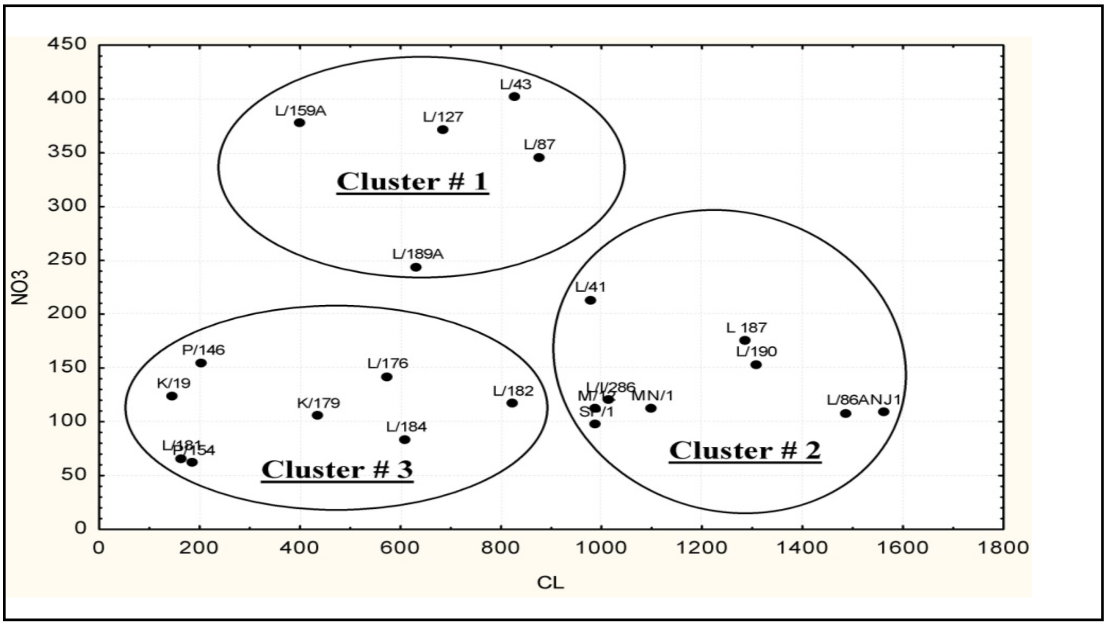

Figure 2. The concentrations of both Cl and NO_3 in case study wells in 2007 for each well's clusters

The 22 case study wells were clustered based on the characteristics of their water quality by using the k-means clustering technique. Figure 2 depicts the concentrations of Cl and NO_3 in the case study wells in 2007, which are grouped into three clusters.

Well cluster #1 is characterized by relatively low chloride and high nitrate concentrations as compared with the overall mean. In all cluster #1 wells, built-up areas characterized by high population comprise the dominant land use (except for 189A

For well cluster #2, chloride concentration is relatively high, whereas nitrate concentration is relatively low compared with the overall mean. These wells differ in location, and are characterized by mixed land use. Finally, the concentrations of all chemical parameters including chloride and nitrate in cluster #3 wells are relatively low. Open areas associated with agricultural activities constitute the primary land use category of the well's areas.

3.2. Modeling of Chloride Concentration

For ANNs' models, the architecture that delivered the best results for both un-clustered and aggregated clustered models is the multi-layer perceptron feed forward neural network with one hidden layer. The Levenberg-Marquardt technique provided the best results as a training algorithm. On the other hand, different SVM models were evaluated and optimized until the best performance was achieved. Radial basis function was used as a Kernel function.

Table 1 presents the results of the best ANNs and SVM models for chloride. Based on the results, the model's performance evaluation criteria for the aggregated clustered model were better compared to the un-clustered model, thus indicating the positive effect of well clustering on AI performance. However, the improvement that resulted from clustering in the SVM model was less than that of ANNs.

The clustering-induced improvement could be attributed to the fact that the clustering divides the wells into groups that possess a high degree of common characteristics. Accordingly, when separately applying AI technique for each cluster, the model can easily grasp the common variables that affect the output. Moreover, the influence (weight) of each variable on the model's output is almost the same for all wells at the same cluster.

The effect of clustering could be more obvious by investigating the input variables and their order in each separate model, as shown in Table 2. The five input variables of the un-clustered model ordered according to their weights are: Cl_o, overall recharge, municipal abstraction, distance to Khanyounis center, and bottom screen depth. Meanwhile, for the cluster #1 model, LULC recharge coefficient replaced the distance to Khanyounis center, because almost all cluster #1 wells have the same distance to Khanyounis center; thus, this variable is insignificant for this cluster. Furthermore, built-up areas are the dominant land use for this cluster, and thus the recharge capacity basically depends on the LCLU recharge coefficient. Additionally, the bottom screen depth of the wells in cluster #1 was noted to have the second most significant influence. This observation may be related to the deeper saline water lenses that exist in the locations of cluster #1 wells. Likewise, the input variables of cluster #2 wells are the same as those of the un-clustered model, but possessed different relative weights. These wells have relatively high chloride concentration as compared with the overall mean, and are spatially scattered over the study area. The ranking of the input variables for this cluster shows the three main sources of elevated chloride concentration, which are seawater intrusion, lateral flow, and the effect of saline lenses. The first two sources are expressed by the distance to Khanyounis center, whereas the bottom screen depth indicated the effect of saline lenses. Finally, the input variables of cluster #3 model comprised only Cl_o, municipal abstraction, and overall recharge, which are the three common input variables for all models. The chloride concentrations in these wells are relatively low, and the wells' areas are characterized by high GW recharge. Other variables are insignificant because these wells have relatively the same distance to Khanyounis center, and almost all have relatively small bottom screen depths. The results of the best ANNs' chloride models in this study are indicated higher accuracy than the results obtained by Seyam and Mogheir[6], who developed an ANNs-based model for simulating Cl in GCA. Their result for MAPE was 14%.

Table 1. Modelling results of both un-clustered and aggregated clustered ANNs' and SVM's models for chloride

Model	MAE	MAPE %
ANNs Models		
Un-clustered Model	19.0	4.5
Aggregated Clustered Model	15.1	3.7
% Improvement	25.7	20.5
SVM Models		
Un-clustered model	19.3	4.6
Aggregated clustered model	17.4	4.1
% Improvement	10.7	10.8

Table 2. Ranking the influence of input variables for un-clustered and clustered AI chloride models

Model	Input Variables					
	Cl_o	Overall Recharge	Municipal Abstraction	Distance to KYC	Bottom Screen Depth	LULCRC
Un-clustered	1	2	3	4	5	-
Cluster # 1	1	4	5	-	2	3
Cluster # 2	1	5	4	2	3	-
Cluster # 3	1	3	2	-	-	-

3.3. Modeling of Nitrate Concentration

The architecture that delivered the best results for both un-clustered and aggregated clustered ANNs models was the multi-layer perceptron feed forward neural network with one hidden layer. In addition, the back-propagation and Levenberg-Marquardt training algorithms resulted in the best performance for the various models. For SVM, the different models were evaluated and optimized until the best performance was achieved. Sigmoid and radial basis function were used as a Kernel function.

Table 3 presents the results of the best ANN and SVM models for nitrate. The aggregated clustered model has a lower error compared with the un-clustered model for both ANNs and SVM. The improvement of the nitrate model due clustering technique before AI modeling is due to the same reason as that for chloride; that is, the uniform characteristics of the wells falling under the same group, which consequently facilitates the ability of the model to identify the common influencing variables. Table 4 presents the ranking of the input variables for both un-clustered and clustered nitrate models. $NO3_o$ produced the largest effect on the nitrate concentration for both clustered and un-clustered models. For the un-clustered models, clusters #1 and #2, the second, third, and fourth influencing input variables were recharge and N-load from built up areas, recharge from open areas, and recharge and N-load from agricultural areas respectively. However for cluster #3, recharge and N-load from built up areas was not included in the model. Ranking the input variables for each developed model demonstrated the effect of LULC on GW nitrate concentration. For instance, recharge and N-load from built up areas was insignificant in cluster #3 wells because the average built up area around the wells of this cluster was less than 5%. Therefore, the effect of the built up areas was limited. On the other hand, the area of the three LULC categories for un-clustered, cluster #1, and cluster #2 models were considerable. Therefore, all categories are significant input variables in the models.

The results of nitrate modeling are comparable with those of other similar studies. For example, Almasri and Kaluarachchi[7] modeled nitrate concentration in the Sumas-Blaine aquifer of Washington, United States using ANNs and achieved a 6.7% MAPE. They utilized highly accurate maps with 21 LULC categories. Moreover, they used a relatively sophisticated method in identifying the well's buffer zone. By contrast, the present study used low quality aerial photos in classifying the entire study area into three LULC categories, using an easy method for the well's buffer zone delineation. The results of this research are relatively more accurate than those obtained by Almahallawi et al.[24], wherein the model's MAPE was 8.43%.

With regard to the comparison between ANNs and SVM, the research results are largely consistent with those obtained by Dixon[10], who used both techniques to differentiate between the wells that were contaminated and uncontaminated by nitrate. He reported that ANNs outperformed the SVM, especially on training data. Nevertheless, the results of both techniques were comparable for the test data set.

The accuracy of the nitrate models (MAPE = 7.0%) was less than that of chloride model (MAPE = 3.7%). This result may be attributed to the high complexity of the GW contamination by nitrate. The nitrate simulation results are mainly affected by LULC categories, which were obtained through an analysis of aerial photos, and the quality of the aerial photos played a crucial role. Moreover, the calculations of input variables were based on an estimation of the average N-load and recharge for each LULC, which are not always accurate. Other variables, such as bacterial role in nitrification and de-nitrification processes, may likewise affect simulation accuracy. On the other hand, the values for most of the input variables for the chloride model were specific and accurate, including the abstraction quantities, distance to Khanyounis center, as well as the depth of the well's bottom screen.

Table 3. Modelling results of both un-clustered and aggregated clustered ANNs' and SVM's models for nitrate

Model	MAE	MAPE %
	ANNs Models	
Un-clustered Model	11.9	11.2
Aggregated Clustered Model	8.7	7.0
% Improvement	26.4	37.3
	SVM Models	
Un-clustered model	9.2	8.3
Aggregated clustered model	9.0	7.1
% Improvement	2	13.7

Table 4. Ranking of the input variables of nitrate model for un-clustered and clustered models

Model	Input Variables			
	$NO3_o$	RNBA	RNAA	ROA
Un-clustered	1	2	4	3
Cluster #I	1	2	4	3
Cluster #2	1	2	4	3
Cluster #3	1	-	3	2

4. Conclusions

Assessment of the effect of the wells' clustering technique in improving the simulation performance of AI-based GW quality models in complex aquifers as conducted in this study indicated that the clustered models outperformed the un-clustered models. This result indicates the effectiveness of wells' clustering as a pre-modeling technique on AI models' performance, especially for ANNs. AI models for each distinct cluster captured the input-output relationships more accurately due to the similarity of the characteristics of wells grouped under the same cluster. The improvement of the AI models due to data clustering is obvious, even though the clustered models have less data sets compared with the un-clustered model, which adversely affected the model's performance. Consequently, clustering the sampling points and stations before applying AI techniques particularly for heterogeneous systems is highly recommended. However, the number of clusters must be kept to a minimum, such that data scarcity associated with clustering does not affect the model's performance.

Introducing a comprehensive and periodic GW monitoring system is never an easy task, owing to various financial and technical constraints in many regions of the world, particularly in developing countries. Thus, accurately modeling the most sensitive and dominant GW quality parameters using cost-effective techniques that rely on few monitoring data presents a highly advantageous opportunity, as setting rational GW management strategies depend on the availability of accurate, applicable, and reliable simulation models. Therefore, the importance of the present study stems from the growing need to improve the accuracy of the GW quality model, without requiring additional data and effort. The clustering technique does not require extra data, apart from routine monitoring data. Consequently, the accurate, simple, and applicable AI models developed in this study can be applied in setting appropriate strategies and making rational decisions related to GW management.

REFERENCES

[1] Chen, S.H., A.J. Jakeman, and J.P. Norton, Artificial intelligence techniques: an introduction to their use for modelling environmental systems. Mathematics and computers in simulation, 2008. 78(2): p. 379-400.

[2] Trichakis, I.C., I.K. Nikolos, and G. Karatzas, Artificial neural network (ANN) based modeling for karstic groundwater level simulation. Water resources management, 2011. 25(4): p. 1143-1152.

[3] Schulze, F., et al., Applications of artificial neural networks in integrated water management: fiction or future? Water science and technology: a journal of the International Association on Water Pollution Research, 2005. 52(9): p. 21.

[4] Yesilnacar, M.I. and E. Sahinkaya, Artificial neural network prediction of sulfate and SAR in an unconfined aquifer in southeastern Turkey. Environmental Earth Sciences, 2012: p. 1-9.

[5] Nourani, V., Conjugation of Artificial Neural Network and Geostatistics Approaches for Groundwater Modeling. Recent Researches in Environmental and Geological Sciences. 2012. ISBN: 978-1-61804-110-4

[6] Seyam, M. and Y. Mogheir, Application of Artificial Neural Networks Model as Analytical Tool for Groundwater Salinity. Journal of Environmental Protection, 2011. 2: p. 56-71.

[7] Almasri, M.N. and J.J. Kaluarachchi, Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. Environmental Modelling & Software, 2005. 20(7): p. 851-871.

[8] Behzad, M., K. Asghari, and E.A. Coppola Jr, Comparative Study of SVMs and ANNs in Aquifer Water Level Prediction. Journal of Computing in Civil Engineering, 2010. 24: p. 408.

[9] Yoon, H., et al., A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. Journal of Hydrology, 2010. 396(1): p. 128-138.

[10] Dixon, B., A case study using support vector machines, neural networks and logistic regression in a GIS to identify wells contaminated with nitrate-N. Hydrogeology Journal, 2009. 17(6): p. 1507-1520.

[11] Chau, K., A review on integration of artificial intelligence into water quality modelling. Marine pollution bulletin, 2006. 52(7): p. 726-733.

[12] UNCT, Gaza in 2020 A liveable place? 2012, Office of the United Nations Special Coordinator for the Middle East Peace Process (UNSCO) - A report by the United Nations Country Team in the occupied Palestinian territory: Jerusalem.

[13] PCBS, Statistical Yearbook of Palestine 2012, Palestinian Central Bureau of Statistics 2011: Ramalla- Palestine.

[14] Shomar, B., Groundwater contaminations and health perspectives in developing world case study: Gaza Strip. Environmental Geochemistry and Health, 2011. 33(2): p. 189-202.

[15] Almasri, M.N. and S.M.S. Ghabayen, Analysis of nitrate contamination of Gaza coastal aquifer, Palestine. Journal of Hydrologic Engineering, 2008. 13: p. 132.

[16] Shomar, B., S. Fkher, and A. Yahya, Assessment of groundwater quality in the Gaza Strip, Palestine using GIS

Mapping. Journal of Water Resource and Protection, 2010. 2(2): p. 93-104.

[17] Yakirevich, A., et al., Simulation of seawater intrusion into the Khan Yunis area of the Gaza Strip coastal aquifer. Hydrogeology Journal, 1998. 6(4): p. 549-559.

[18] Shomar, B., K. Osenbruck, and A. Yahya, Elevated nitrate levels in the groundwater of the Gaza Strip: Distribution and sources. Science of the Total Environment, 2008. 398(1-3): p. 164-174.

[19] Baalousha, H., Analysis of nitrate occurrence and distribution in groundwater in the Gaza Strip using major ion chemistry. Global NEST J, 2008. 10: p. 337-349.

[20] Bekesi, G., M. McGuire, and D. Moiler, Groundwater allocation using a groundwater level response management method—Gnangara groundwater system, Western Australia. Water resources management, 2009. 23(9): p. 1665-1683.

[21] Coulibaly, M. and S. Becker, Spatial interpolation of annual precipitation in South Africa-comparison and evaluation of methods. Water International, 2007. 32(3): p. 494-502.

[22] Hamdan, S.M., U. Troeger, and A. Nassar, Stormwater Availability in the Gaza Strip, Palestine. International Journal of Environment and Health, 2007. 1(4): p. 580-594.

[23] Metcalf and Eddy, Coastal Aquifer Management Plan (CAMP). Final model report (task 7). . 2000, USAID.

[24] Al-Mahallawi, K., et al., Using of neural networks for the prediction of nitrate groundwater contamination in rural and agricultural areas. Environmental Earth Sciences, 2012. 65(3): p. 917-928.