Kohonen Self Organizing Map (SOM)-aided Predictions of Aquifer Water Struck Levels in the Merti Aquifer, Northern Kenya:

Meshack OwiraAmimo[1] and Dr K.S.S. Rakesh[2]

[1]*Research Scholar, IIC University of Technology, Cambodia*
[2]*CEO,Gradxs,India*
[1]bmoamimo@gmail.com
[2]kssrakesh@gmail.com

***Abstract*—The aquifer water struck levels in the Merti aquifer were assessed using the Kohonen Self-Organizing Maps algorithm, which employs the Neural Networks. This algorithm mimics the biological sensory and motor neurons. The variable was inferred via predicting the aquifer ground water levels , then subtracting the same from the elevations, measured in meters above the sea levels, of the proposed well point mapped, which is one of the three variables generated using the hand-held GPS. The objective of the study is to help develop a simple prediction model for aquifer well depths in the Merti aquifer, to be used alongside the geoelectrical models generated during geophysical surveys, so as to enhance and modernize groundwater management plan for the Merti Aquifer. Data on well hydraulics for the Isiolo, Garissa and parts of Wajir (south) counties were used to generate the model predictors, employing the Kohonen R package, which clusters and predicts variables using this neural network algorithm developed by Teuvo Kohonen.**

The algorithm was then used to predict the expected groundwater levels of a new area that has not been developed, and this value was subsequently subtracted from the elevation levels, thereby generating water struck levels. The variables employed to achieve this task were longitudes, latitudes, elevation, aquifer depth, resistivity, and the groundwater levels, (gwl) in meters (below ground level) bgl. To predict the gwl of a newly proposed drilling point, the hydrogeological data was run on R platform and models generated inferred and interpreted.The new model was then used to predict the gwl of the new site. Subtracted from elevation of the area, wsl was derived, thusThe study concludes that the neural network SOM mapping algorithm is an accurate predictor of the wsl in the Merti aquifer, as it clusters geological zones bearing the same groundwater levels and aquifer depths together. It should therefore be a useful stochastic hydrological tool for decision making on matters groundwater development in the Merti aquifer.

***Keywords*—cluster, Kohonen Self-Organizing Maps, Merti aquifer, Neural Network, neurons.**

I. INTRODUCTION

The Merti Aquifer is located in eastern Kenya cutting across four counties namely Garissa, Wajir, Isiolo and Marsabit counties. It is also a transboundary aquifer shared between Kenya and Somalia. Climate in the area ranges from humid tropical highlands in the west to semi-arid and arid land in the centre and east. Rainfall across much of the area is less than 300 mm/yr. and is unreliable. Evaporation almost everywhere in the area far exceeds rainfall. Theaverage annual potential evaporation is between 2,100-2,500 mm**.**

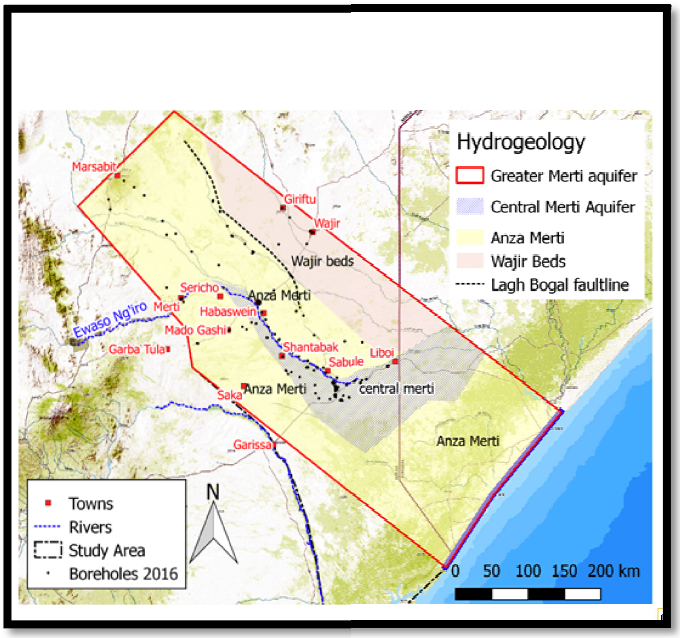


Figure 1: Merti aquifer Study Map , as adapted from The 2017 Merti Aquifer Study Report by Earthwater Consultants.

II. LITERATURE REVIEW

According to study by Han, et al (2016), an accurate groundwater level prediction may be used to enhance water supply essential for growth of agriculture, domestic and

industrial utilizations. The concept of modeling the groundwater successfully for abstraction may be achieved via spatial and temporal methods. The self-organizing Kohonen maps may be useful in this respect as they achieve both predictions of groundwater quality and aquifer hydraulics, as well as clustering the variables for the ease of predicting any single parameter so desired. The maps are self-organizing and generate simple clusters on the account of radii of strength of association between and within study variables used-in this case spatio-temporal. This method employs Neural networks, and was used for the above study in Hexi corridor of the northwest of China. The modeling of the spatial variables influencing groundwater levels in study area established six regions, representing central piezometers and for which sensitivity analysis was undertaken.

The prediction of groundwater resource development and management has been made essential by the hydrological time series modeling, Maiti, et al (2014). Three different modern computing techniques were applied and compared to check their effectiveness. They include; (BNN) Bayesian Neural Networks, (ANN) Artificial Neural Networks, and (ANFIS) Adaptive Neuro-Fuzzy Inference System. All the three were optimized by Scaled Conjugate Gradient (SGC) to predict groundwater level fluctuations. In addition, to compare the robustness of these models, four standard statistical quantitative measures were applied. These quantitative measures include:

a) Pearson's correlation coefficient R
b) Index of agreement (IA)
c) Reduction error
d) Root mean square error

Based on these analyses, it was discovered that the Artificial Neuro-Fuzzy Inference System, ANFIS model was excellent in modeling and predicting noise-free data as compared to the other models. However, the Bayesian Neural Networks, BNN performed better than ANN and ANFIS in the modeling hydrologic time series correlated with huge amount if red noise. Thus, proper care should be considered in order to determine the suitable and best methodology for modeling noisy and complex hydrological time series. The results so obtained may be used to constrain the groundwater fluctuation models, which later on facilitate the implementation and development of a sustainable and effective strategies in groundwater management and planning.

According to Daliakopoulos et al (2005), Artificial Neural Network (ANN) are of great significance in water resources forecasting and modeling. The performance of this model in groundwater level forecasting is studied so as to identify and note an optimal ANN architecture that can predict and simulate groundwater level fluctuations for up to 2 years and more. In their research, Messara valley in Greece was studied. The findings were that over the past fifteen years, the groundwater resources have been overexploited thus causing the groundwater level decreasing and deteriorating steadily with time. After the study which involved the use of seven different types of training algorithms and network architectures, it was found that a standard feed-forward neural network which was trained by the Levenberg-Marquardt alogithm predicted accurate results for up to 18 months.

Groundwater levels fluctuation around the world is a very significant component in hydrological research, (Yadav et al, 2017). Some of the main reasons why groundwater levels are fluctuating include soil mismanagement and uncontrolled overexploitation of aquifers, faulty and reckless irrigation practices and the ever increasing rise in water demand. For the groundwater resources to me managed carefully and effectively, it is therefore imperative and significant to have accurate predictions and forecasts of groundwater levels. As a result of the complexity in the nature of the groundwater systems, the creation of computer models and data driven techniques in the field of hydrology has great importance. In order to forecast and predict groundwater levels in the study area, two soft computing techniques were employed, namely, (SVM) Support Vector Machines and (ELM) Extreme Learning Machines. These were conducted for two observation wells located in Canada. The two computing techniques were compared and an eight year monthly dataset from the year 2006 to 2014 was used in the comparative study. The monthly data consisted both meteorological and hydrological parameters such as groundwater level, rainfall, evapotranspiration and temperature. These parameters were then applied in different combinations for multivariate and univariate analysis. After the comparative analysis of the variables the findings were that ELM has better predictive ability and forecasting capacity as compared to SVM for the monthly groundwater level prediction.

According to Sahoo et al (2013), a comparative study was made to determine the potential of (ANN) Artificial Neural Network and (MLR) Multiple Linear Regression soft-computing techniques in the prediction of transient water levels over a groundwater basin. ANN and MLR modeling was carried out in 17 different sites across Japan. The inputs considered in the study were groundwater level, ambient and transient temperature, influential rainfall lags, and eleven seasonal dummy variables. For each of the 17 sites, specific ANN models were established by the use of multi-layer feed-ward neural networks trained with Levenberg-Marquardt backpropagation algorithms. After the analysis, the performance and efficacy of the models

was examined using graphical and statistical indicators. The findings after the comparison were that ANN models had the best goodness-of-fit and showed a better agreement between the observed groundwater levels and the predicted groundwater levels at all the studied sites. These findings were supported by the graphical indicators and the results from the residual analysis. Therefore, it was concluded that MLR is inferior compared to ANN which is the superior technique in the prediction of spatio-temporal distribution of groundwater levels in a basin. However, MLR is a bit advantageous and worth considering as an alternative to ANN as it is a cost effective model and tool in groundwater modeling.

According to Kombo et al (2020), when there is limited data on the quality and the amount of on-site groundwater available, then this makes it difficult and impossible to predict the seasonal groundwater levels. In the present study, a hybrid K-Nearest Neighbour-Random Forest (KNN-RF) algorithm was used in the prediction of groundwater levels variation (L) which belong to an aquifer whose groundwater is close to the surface at around 10m or less. In order to improve the quality of groundwater data, first the time-series smoothing methods are applied. Then, the collective KNN-RF model is run using the hydro-climatic data for the prediction of variations in the groundwater table levels of up to three months and more. Groundwater and climatic data which were collected from eastern Rwanda, were used for the authentication of the model on a rolling window basis. These potential predictors used were precipitation (P), previous day's precipitation [P(t - 1)], daily mean temperature (T), daily maximum solar radiation (S), and groundwater level (L) which, exhibit the highest variation in the groundwater table fluctuations. In addition, the KNN-RF model was shown to be an advanced alternative which can be used instead of SVR, KNN, RF, and ANN models. To conclude with, the results from this study would be useful for the planning and management of groundwater resources.

According to a study by Bretzler, et al (2017), there is an increasing dire need for accurate groundwater level prediction and forecasting in order to ensure effective seasonal water management. In addition, the forecasting models and tools need to be not only effective but also accessible for decision making. In the study, they tested the prophet forecasting procedure to address the challenges. The open source code is based on additive model considering both periodic changes and non-periodic components in a Bayesian framework which has parameters that are easy to interpret. According to the study, the predictions of daily groundwater level data in the area, is affected by excessive pumping close to a tourist complex in the Ramsar wetland area of Donana in Spain, hence compared to other forecasting and prediction methods. The prophet method outperforms many methods in predicting groundwater level. It is a fast and flexible tool preferred by water managers and hydrologists. In addition, the model allows a deep understanding into the influence of each variable of the forecast independently, thus aiding to assess the hydrodynamic response and impacts of external factors such as groundwater pumping.

III. HYDROGEOLOGY

A. Geology and stratigraphy

The geology is mainly comprises the Mio-Pliocene Merti aquifer sediments, and which overlies the carbonates – namely corallites, aragonite sediments and calcite in the upper zones 0-100m bgl. The subsurface geology is dominated by fine, medium and coarse grained sandstones, alongside some silt and gravels. The Miocene clays and sandstones are the major reservoirs of the subsurface water, and are fairly fractured and possess water at the great depths, though fairly, mineralized, via the fractures and some karstification veins. Water also forms at the contact points between the sandstones and the Archaean metamorphic basement units. Groundwater in the upper sediments shall enjoy annual precipitation recharge through direct infiltration, while the deep-seated zones shall be recharged via regional flow aided by the karstification channels and plate tectonics I n the Jurassic – cretaceous period. Evapotranspiration rates of up to 3,000mm per annum over shadow the annual rains of up to 500 mm.

IV. HYDROLOGY AND STRUCTURAL GEOLOGY

There is ample evidence, both structural and deduced, that the recharge of the Merti aquifer is initiated by flow from hundreds of kilometers away in the Mt Kenya areas which empty waters into the Ewaso Flow course, as it weaves its way across Samburu and Isiolo counties.

This flow shall be from the Mount Kenya areas flowing over to Isiolo, Habaswein, Modogashe, Shant-abak plains, and later into the central axis of Dadaab. The Laghdera flow course is the major control parameter directing both flow and recharge of the subsurface merti waters. Merti Aquifer portions of the regional aquifer is highly concentrated at Dadaab, Madah-gesi, Malailey, Hagadera, Ifo, Kadakso and Kulan areas. This may be inferred from the ground water levels, which rank among the highest-

implying that one did very shallow depths to hit the water tables, compared to the other areas. Owing to the ephemeral states of most of these tributaries, there is no sufficient time available for maximum river bed infiltration into the sub surface zones lying on the adjacent sides of the river course. However there are areas that were exceptionally karstified and fractured within the carbonate beds. These are the zones that store water upon seepage into the sandstone aquifer sediments systems alongside recharging the adjacent sub surface storage systems via the Darcyan flow mechanics.

A. Drainage

Owing to the relative flat nature of the terrain, there is flood rampancy. Nevertheless, no permanent civil structures are on the ground to stand the risk of destruction, other than the occasional loss of lives for both livestock and human persons. Most of the housing units are constructed through shrubs and dry acacia trees locally available, lightening the task of evacuation in the event of impending flood disasters.

V. CLIMATE

The project area falls within zone 7 of the classification ofclimatic/ecological zones of Africa, that is to say arid to semi-arid with temperatures averaging 30 to 34 degrees per day and occasioning evapotranspiration rates of up to 2000-3000mm per annum.

VI. KOHONEN SELF- ORGANIZING FEATURE MAP

The Kohonen Self-Organizing map (SOM) is a neural network trained via the usage of the concept of competitive learning. This Competitive learning refers to a process whereby the competition itself comes up way before the phase of learning. Moreover, this competition dictates that a defined criteria of winning or losing of the element being processed, usually a neuron, be established. After this neuron has been determined, its weight vector would be adjusted according to the learning law employed.

The Kohonen SOIM map generates topologically-similar maps, between input data being used for predictions (namely, longitudes, latitudes, groundwater levels and elevations) and the processing elements of the topological map. The term topologically-similar implies that if the predictor inputs are of similar mathematical traits, then the most active processing elements answering to predictor-inputs are situated in relative proximity to each other ,within the model map so generated. The process of determining this closeness implies using numerical vectors from the dataset provided ordered as appropriate. The weight vectors of the processing elements, which are called into the algorithm, are organized in an ascending to descending order:

$$\mathbf{W_i < W_{i+1}} \text{ for all values of } \mathbf{i} \text{ or } \mathbf{W_{i+1}} \text{ for all values of } \mathbf{i}.$$

In the above line, W_1 represents the weight in the current iteration and W_{1+1} represents the updated weight in the next iteration. The term above is true for 1-D SOM maps, and not the higher dimensional ones. This definition is valid for one-dimensional self-organizing map only. The Kohonen self-organizing map is traditionally presented as a two-dimensional sheet of processing elements. It is observed that each one of the processing elements has its own weight vector, and the learning of SOM algorithm depends on the successful adaptation of these vectors. The processing elements of the algorithm are made competitive, in the process of them going about organizing themselves, or reorganizing themselves-the origin of the term 'self-organizing' process. A specific rule is developed for picking the winner neuron, whose weights are subsequently updated, for use in the next phases of iterations. Primarily, the rule that is employed is set to limit or diminish the Euclidean distances, between the respective input and the weight vectors. The SOM will vary from basic competitive learning, such that instead of adjusting only the weight vector of the winner neurons, weight vectors of neighborhood neurons are simultaneously adjusted.

The size of the neuron neighborhood is to a large extent affected by making the rough ordering of SOM. As each subsequent iteration is effected, this size diminishes, progressively. Finally, only the winner neuron is adjusted, making the process of fine tuning of SOM proceed with relative ease. The use of neighborhood ordering makes topologically-similar neurons to be easily ordered or arranged. Alongside the competitive learning, the process is made non-linear. At the end, the topologically similar clusters represent a certain class attribute defined by the data, so that in our case, the class could be shallow groundwater table levels, or deep and extremely deep groundwater levels. Clustering of variables for classifications and predictions is an important exercise in groundwater hydrology, especially during the process of making decisions under uncertainty, but where there already exists secondary data. The self-organizing map is thus a form of unsupervised learning algorithm, proposed for applications, whereby it is imperative that the topology between input and output spaces is maintained. The Kohonen SOM tool is as much a means of dimensionality reduction, as it is for direct applications, in algorithmic predictions of variables used to assess phenomena in real life context.

Aguilera, P. A., Frenich, A. G., Torres, J. A., Castro, H., Vidal, J. M., & Canton, M. (2001). Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. *Water research*, *35*(17), 4053-4062.

Aguillera et al (2001) used the Kohonen neural network to analyze the nutrient data, comprising mainly ammonia, nitrite, nitrate and phosphate taken from coastal waters in a Spanish tourist area. The activation maps generated were found insufficient for evaluating and predicting the trophic status of coastal waters.

Orak et al (2020) used the Self Organzing maps to assess the drastic change in the water quality of River Ergene in Turkey. It was found that the quality has deteriorated due to both the point and non-point pollution sources. Accordingly, an appropriate assessment of surface water quality was recommended. Water quality classification was calculated separately for each quality parameter in Turkey. An overall assessment of surface water quality is essential for water management. In the study, both the Kohonen self-organizing maps (SOMs) and fuzzy C-means clustering (FCM) methods were employed.

In the present study, SOM maps were used to map aquifer hydrology and predict the groundwater levels that were subsequently used to predict water struck levels.

The whole learning process for the algorithm takes place without supervision as the neural nodes are self-organizing.

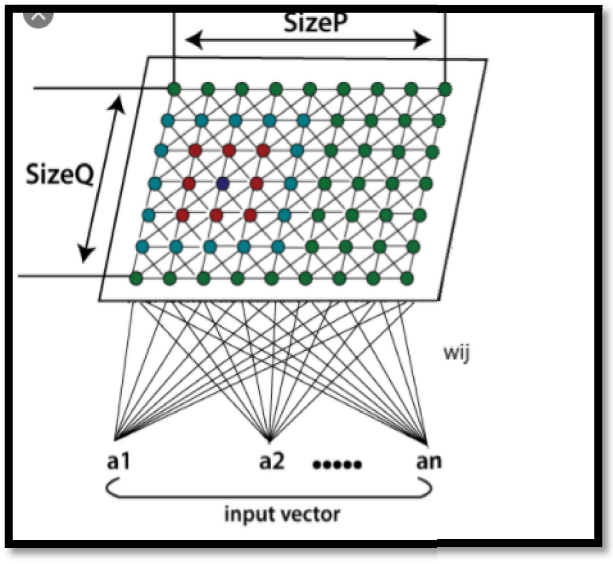


Figure 2: The schematic illustration of the Kohonen Neural Nodes, showing the weights and inputs of the Model.

To the extent that the SOM maps re-train the features in the input dataset and cluster them on the basis of similarity (defined by Euclidean radii), they are known as feature Maps. The high dimensional data comprising dozens or hundreds of variable columns may be easily condensed into a simple 2D model which is a lower-dimensional field, easier to understand.

The SOM model is different from the typical artificial neural network in its architectural and algorithm traits. The structure primarily consists or comprises a single layer linear two-dimensional grid of neurons, rather than the traditional input, output and hidden layers that characterize the simple back-propagation neural networks architecture.

The nodes on the SOM lattice are associated directly to the input vector or array of vectors if the prediction employs multidimensional inputs, but not directly to each other. The implication of this state of affairs is that the nodes do not know the values of their neighborhood neurons, and shall only update the weight of their associations as a function of the given input fed into the model. Moreover, the grid itself is the map that coordinates itself, upon the termination phase of each iteration, again as a function of the input data. A cluster of topological similarity will emerge. Consequently, after this clustering, each node shall have its own coordinate (i.j), which makes it possible for the researcher to compute Euclidean radius, between two nodes by employing the age-old Pythagoras theorem.

It should be noted that the Kohonen Self-Organizing Map employs competitive learning, rather than the error-correction learning method, in order to modify its weights. This means that only an individual node is activated at each cycle, in which the features of an occurrence of the input vector or vectors in the case of many predictor variables shall be introduced to the neural network, since all the nodes present in the architecture shall be competing for the privilege, to respond to the input. In the end, the selected node, also known as the Best Matching Unit or BMU, will be selected according to the similarity generated by the model, to exist between the present input variable values, and all the other nodes in the network architecture.

A. Algorithm:

1. Step number one: This involves generating random weights to be used in the model algorithm. To that effect, each node weight w_ij shall be initialized to a random value.

2. Step number two: In this case, the model for the study has longitudes, elevations, depths and others. The variables shall be picked at random, as a random input vector, X_k.

This X may be the longitude or elevations, just to illustrate how the model was used in the present study.

3. *Step number three:*Repeat steps numbers four and number five all nodes on the SOM map.

4. *Step number four:*This is the most significant phase. It is for computing the radius between the input and weight vectors- Calculate the Euclidean radius between weight vector W_{ij} and the input vector x(t) connected with the first node, where t, i, j =0.

5. *Step number five:*The smallest distance is what is to be looked for in this phase, and it is present in one of the nodes. This is achieved by tracking the node that generates the smallest distance t.

6. *Step number six:*The BMU is then computed for the array of values. This involves computing the overall Best Matching Unit (BMU). It implies picking the node with the smallest distance from all computed values.

A model shall emerge, generated by the rearrangements of the values, which occurred during the iterations. This is the topological neighborhood of BMU in the Kohonen Map generated.it is imperative that one repeats this for all nodes in the BMU neighborhood. The weights are then updated, for the first node in the neighborhood of the BMU by including a fraction of the difference between the input vector x (t) and the weight w (t) of the neuron.
This process is Repeated until the requisite iteration number is attained-the iteration limit of t is set at n, so that t=n.
Summarily, step number 1 represents initialization phase, while the rest represents the training phase.

In our study, these inputs are labeled by their real names.

VII. ACTUAL IMPLEMENTATION

Data was collected from the boreholes of the Merti aquifer and the analysis conducted on the R software plat-form. These wells were distributed such that there were a good number from the central Merti, Proximal Merti and from the distal Merti aquifer portions-which has compromised salinity, unsuitable for domestic purposes. The data was run and generated the following models.

Li et al (2018) generated patterns of water quality variables, which were visualized by the SOM Kohonen models, and similar patterns observed for the variables that highly correlated with each other, suggesting a common source. The clusters so formed led to dimensionality reduction in the water quality parameters and in effect clustered the water quality parameters with similar attributes together.
Hadjisolomou, et al (2018) employed the use of principal components analysis (PCA), cluster analysis, and a self-organizing map (SOM) to evaluate the relationships amongst water quality variables. To achieve this, the study water quality data, sourced from two trans-boundary lakes of North Greece, using the Kohonen self-organizing maps. The Kohonen SOM was employed as it has been considered an advanced and powerful data analysis tool, on the account of its ability to map out the rather complex and nonlinear relationships among multivariate data frames.

A. *Employing Kohonen in Physical hydrogeology*

The data was used to analyze the groundwater levels with a view of using the generated prediction to calculate the value of expected water struck levels, which area actually the depths to the aquifers in the study area. This secondary data was generated from the hydrogeological database of the Northern water works development agency offices in Garissa

```
data=read.csv("kohonenD2.csv",header=T,na.strings="
NA")
```

The data name is known as kohonenD2 and has been saved in the excel format, so that when gwl=shallow, it is a value between 5m and 20m below ground level. When this value is classified as moderate in the dataframe, it ranges between 20 and 60m below ground level. When it is categorized as deep, this implies that the gwl is anywhere between 60m and 200m bgl. The data is called into R from excel, to be analyzed by running the above script in R softyware

A. *library(kohonen)*

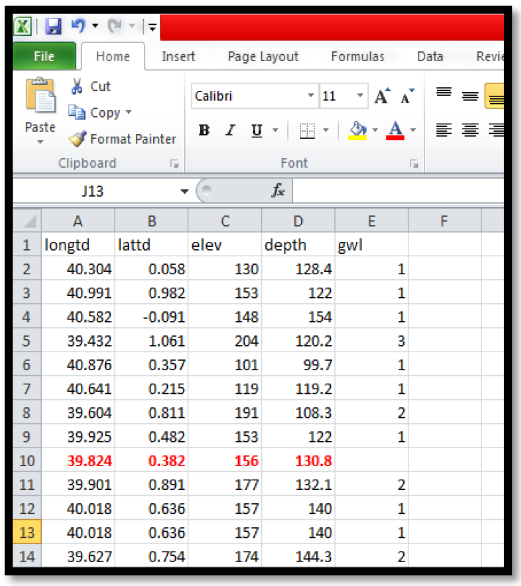
This is the package of R used to generate the prediction models employed in estimating both gwl and the wsl values in the study area. The algorithm will generate the cluster maps as well as the actual prediction of the label class of a row of newly-generated field data of an area hitherto un-mapped.

B. *head(data,6)*

In the R package, the above command helps one get to a glimpse into the first six rows of the dataset, which includes the columns as well.

C. *str(data)*

The above command yields the names of all the variables appearing in the data-frame used that is kohonenD2.csv



	A	B	C	D	E	F
1	longtd	lattd	elev	depth	gwl	
2	40.304	0.058	130	128.4	1	
3	40.991	0.982	153	122	1	
4	40.582	-0.091	148	154	1	
5	39.432	1.061	204	120.2	3	
6	40.876	0.357	101	99.7	1	
7	40.641	0.215	119	119.2	1	
8	39.604	0.811	191	108.3	2	
9	39.925	0.482	153	122	1	
10	39.824	0.382	156	130.8		
11	39.901	0.891	177	132.1	2	
12	40.018	0.636	157	140	1	
13	40.018	0.636	157	140	1	
14	39.627	0.754	174	144.3	2	

Figure 3: The row number 11 appearing in red is deliberately left with no class for the SOM algorithm to classify, later as the modeling progresses.

Nourani et al (2016) used a similar study integrating the concept of co-kriging as spatial estimator and self-organizing map (SOM) as clustering technique to identify spatially homogeneous clusters of groundwater quality data, and allied groundwater hydraulic parameters.

data$gwl=NULL

The above command shows the name of the class item being predicted here, groundwater level, abbreviated as gwl, and which is found under the dataframe presently known as data$gwl, as the name of the dataset was read into R as simply 'data' and as 'kohonenD2'. It is set as NULL.

X=scale(data)

The data then gets scaled so that all the values in the column variable predictors range between 0 and 1.

D. Modeling the SOM predictions now begin as per script hereunder:

```
set.seed(73243)
g=somgrid(xdim=4,ydim=4,topo="rectangular")
map=som(X,grid=g,alpha=c(0.05,0.01),radius=1)
plot(map,type='changes')
```

In the model, a som grid of 4 by 4 in size is generated, to assume a rectangular topography. Alpha is also set to vary between 0.005 and 0.001

E. plot(map)

The above command of plot (map) shall give out the desired predictors in the model as shown hereunder:

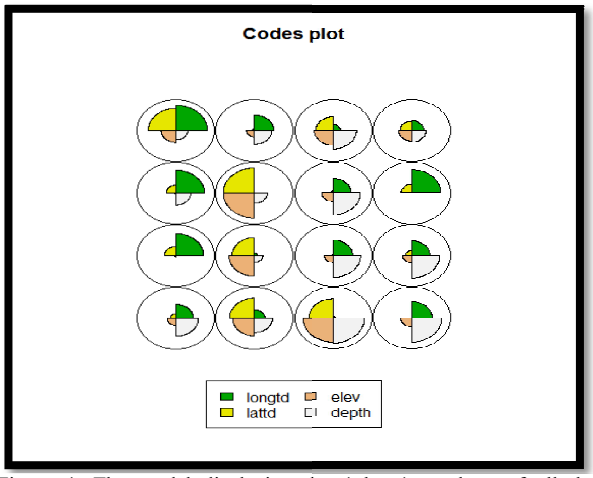


Figure 4: The model displaying the 4 by 4 topology of all the predictor variables in the dataframe.

map$unit.classif

Here the command shall tell the class to which the rows highlighted in the dataframe belongs, and when run, it gives the output indicating that the class of the row lies in one of the 16 pies in the model in figure 3 above.

```
>
> plot(map)
> map$unit.classif
  [1] 14 13  4 10 12  9  6 16 14  2 16 15 15 15 15  6  2  6 16  6  3 16  8 10 14
 [26]  7 16  1 12  3  9  1  9  1  8  1  7  7  9  7  8  7  4 15  9  9 13  6  4 10
 [51] 14 13  4 10 12  9  6 16 14 16 15 15 15 15 15  6  2  6 16  6  3 16  8 10 14
 [76]  7 16  1 12  3  9  1  9  1  8  1  7  7  9  7  8  7  4 15  9  9 13  6  4 10
>
```

Figure 5: The output defining class of each row. For example, 14 is the first reading here.

It means that the after computations of neighborhoods distance strengths and the Euclidean rule, the row number 1 is to be found on neuron number 14. Then data row number 51 is also categorized in the same pie.

This way, the one hundred rows in the data frame have been compressed into just a mere 16 pies.

The study next did run the scripts hereunder:

plot(map,type='codes',palette.name=rainbow,main="4 by 4 Mapping of Merti aquifer gwl &wsl")

Running the above lines will generate the model hereunder:

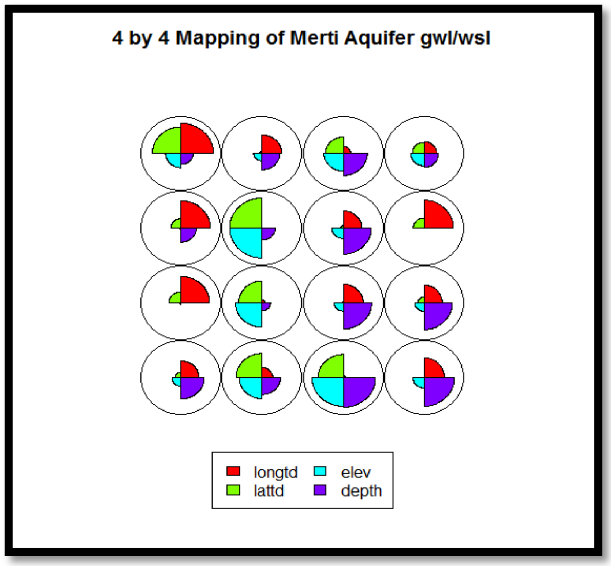


Figure 6: pie models of the SOM map showing predictor variables for gwl.

##The script hereunder was then run

plot(map,type='dist.neighbours')####

The above command generates the model shown hereunder and this is where the rows whose Euclidean distances are closer to one another are compressed into the same winner neurons. There are a total of 16 neurons to capture all the 100 rows of data:

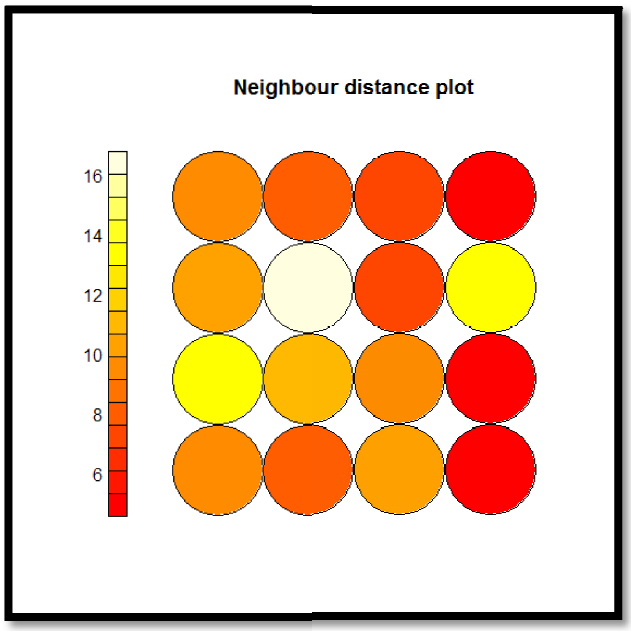


Figure 7: Model shows how many of the rows have been compressed into the pies-in the first row of circles, the final value is extremely red and shows it has six rows or less of data indicated by the color bar scale.

The above manipulations original data is re-read again into R so that a supervised classification is performed.

data=
read.csv("kohonenD2.csv",header=T,na.strings="NA")

##Supervised Self-Organising Map/Data Split

set.seed(12345)

ind=sample(2,nrow(data),replace=T,prob=c(0.7,0.3))

After re-reading data, it is split so that the ration of the training data to the test portions is 70 percent to 30 percent.

train=data[ind==1,]

test=data[ind==2,]

The above lines for 'test' and 'train' will give out the two portions already splits as stated above.

map1=xyf(trainX,classvec2classmat(factor(trainY)),grid=somgrid(7,7,"hexagonal"),rlen=870)

```
plot(map1,type='changes')

plot(map1,type='count',main="gwl predictions using kohonen SOM Maps in  Merti aquifer")
```

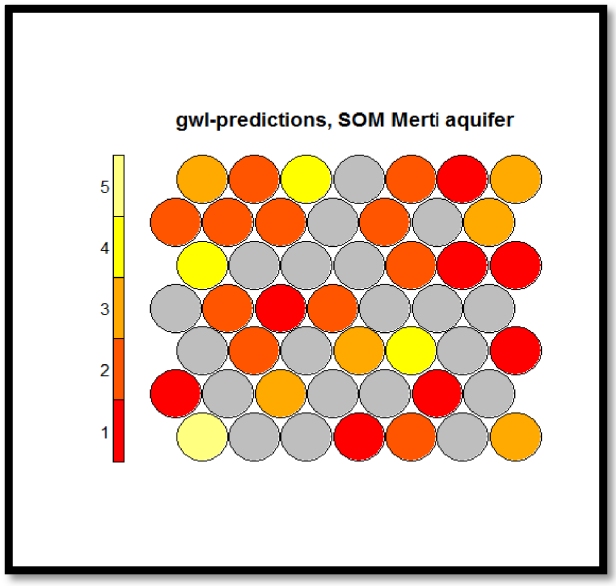


Figure 8: Shows the model predictions generated when a 7 by 7 grid I now used for actual computations after the data is split using the commands shown. thereare a total of 49No grids.

The model proceeds to cluster the rows on the basis of class of ground water levels as thus, so that it generates cluster boundaries.

```
##Cluster Boundaries

par(mfrow = c(1,2))

plot(map1,

type = 'codes',

main = c("Codes X", "Codes Y"))

map1.hc <- cutree(hclust(dist(map1$codes[[2]])), 4)

add.cluster.boundaries(map1, map1.hc)

par(mfrow = c(1,1))

pred$unit.classif

pred$predictions[[2]]
```

F. using computed neuron winner strengths

The model generated indicated that the predicted class which was in row 9 is of class number one. Look at fifth row in all the three tables shown the first one shows 32. This indicates that the pie model in position number 32 represents then class of the gwl being predicted. The third table has a figure of 9 in row 5. This nine is the original row where the unknown class was priori to randomization taking place for prediction. It matches with 1 (the gwl prediction) in table number one. See the middle blue arrow pointing to class predicted

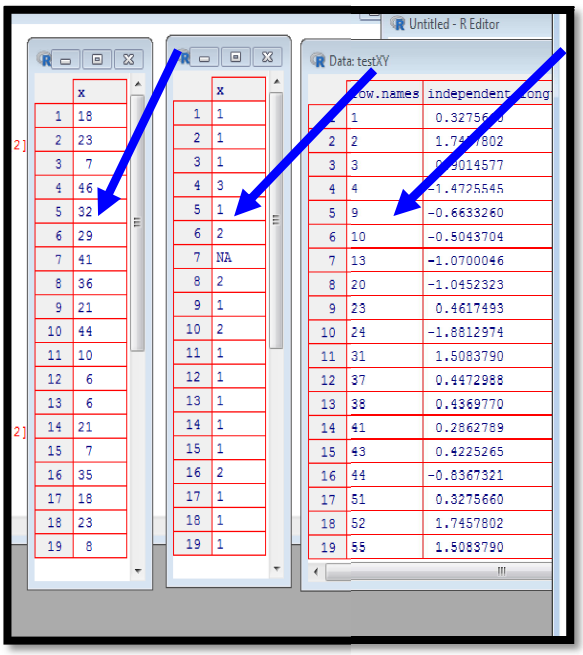


Figure 9: Shows three tables generated for the SOM predictions of class. The class of 1 implies that gwl will range between 5m and 20m bgl.

G. Using the single command line to predict gwl

Running the command datPred also predicts the class as 1 in the study

```
> View(predspredictions[[2]])
>
> View(testXY)
> ## row 11 is data namba 5 now!
> datPred[5,]
[1] 1
Levels: 1 2 3
>
```

Figure 10: Screenshot of the R console showing the output predicted

H. *Using the pie models to predict ground water*

The Predictions hereunder as shown by the pie cluster models indicated that the class of the new site is 1 and this falls in pie number 32. The green color represents 1 as per the key used in the model

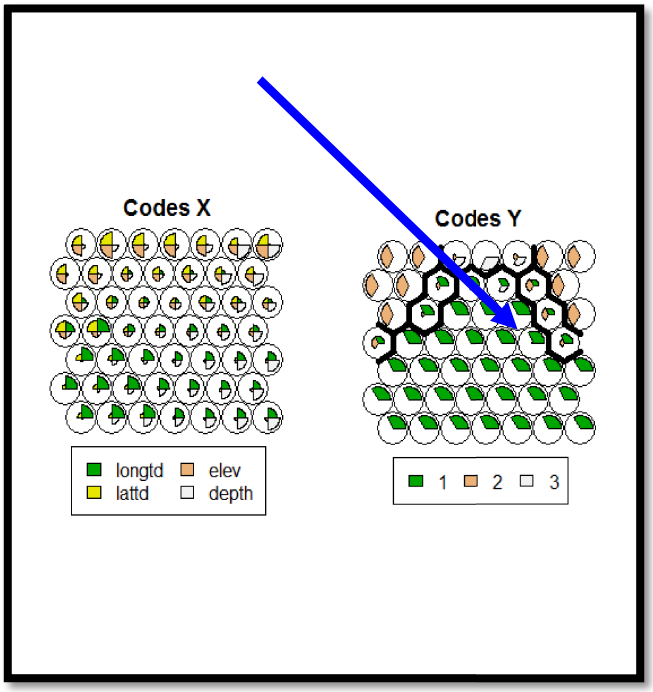


Figure 11: Pie cluster model indicating the class predicted

Predicting the value of aquifer depth for ythe first aquifer in study at area

This shall invite having our data as shown hereunder:

A	B	C	D	E	F	G
ongtd	lattd	elev	depth	gwl		
40.304	0.058	130	128.4	1		
40.991	0.982	153	122	1		
40.582	-0.091	148	154	1		
39.432	1.061	204	120.2	3		
40.876	0.357	101	99.7	1		
40.641	0.215	119	119.2	1		
39.604	0.811	191	108.3	2		
39.925	0.482	153	122	1		
39.824	0.382	156	130.8			
39.901	0.891	177	132.1	2		
40.018	0.636	157	140	1		

Figure 12: The original data table showing the new field data in red color.The class of 1 implies gwl depths of between 5m and 20m.

The elevation here is 156m as measured by GPS. The total depth is 130.8m as per measured by the geo-surveys Terrameter. This implies two things:

a) Minimum wsl depth=elevation- minimum gwl in the range. This is(130.8-5)m=125.8m
b) Maximum wsl depth-elevation-maximum gwl in the range. This gives (130.8-20)m 110.8m
c) The results shows that water struck levels in the aquifer at this point shall be between 110.8m and 125.8m

I. *Validating the model using the kNN algorithms in R*

The neighborhood of a site with groundwater level value was also surveyed with a view of replacing the existing borehole which has the following tabulated details. As usual the historical data was compiled into csv format and called into R. The analysis was then undertaken using the caret library of R. The data already analysed above was also brought into R again to be predicted using the kNN algorithm.

1. *Working of the k-Nearest Neighbor Method:*

To show that the SOM algorithm is effective in classification and predictions of the groundwater hydrology parameters of the merti aquifer, the kNN algorithm, which a sister algorithm to the SOM Kohonen, was also used to model the same gwl data. The K-nearest neighbor algorithm primarily operates via forming a majority voting Rule, between the k most-similar instances to a given "newly-introduced" variable observation. The concept of similarity is defined, according to a specified distance metric, between two selected data points. The

popular one in use has been the Euclidean distance method. The x and y in the formula hereunder refer to separate variable data points separated by the distance being measured by this method, which operates on the Pythagorean Theorem.

The other algorithms in common use for kNN include the Manhattan distance, Minkowski distance as well as the Hamming distance algorithms. In order to perform classification using categorical variable predictors, the hamming distance method has been deemed most convenient.

TABLE 1

SNo	longtd	Lattd	elevation	depth	gwl
1	39.432	1.061	204	120.2	3
39.824	39.824	0.382	156	130.8	1

The table was analyzed minus the gwl as thus:

TABLE 2

SNo	longtd	lattd	elevation	depth
1	39.432	1.061	204	120.2
39.824	39.824	0.382	156	130.8

The data was then analyzed using the K-nearest neighbor algorithm in R and generated the predictions that more or less tallied with that of the SOM Maps. When the data was called into R, renamed as 'kohonenD33i', the following a script was used to make kNN predictions:

```
df1=read.csv("kohonenD33i.csv",header=T,na.strings="NA")

df1

head(df1)


##install.packages("caret")


library(caret)


# k-Nearest Neighbors (KNN)
set.seed(101)

mymodel3 <- train(gwl~., data=df1)

head(mymodel3)

##now bring in ur new field data==========say row number 1
new.well2 <- data.frame(longtd=c(39.432),lattd=c(1.061),
 elev=c(204),depth=c(120.2))
```

Figure 13: Script used to make the kNN predictions

The algorithm was run and predicted as follows:

```
> ##now bring in data from zone THREE of gwl 97m bgl
> new.wellTHREE <- data.frame(longtd=c(39.432),lattd=c(1.061),
+  elev=c(204),depth=c(120.2))
>
> predTHREE=predict(mymodel3,new.wellTHREE)
>
> predTHREE
       1
2.842133
>
> ##now bring predict data from zone THREE with gwl of 17m bgl
> new.wellONE <- data.frame(longtd=c(39.824),lattd=c(0.382),
+  elev=c(156),depth=c(130.8))
>
>
> predONE=predict(mymodel3,new.wellONE)
>
> predONE
       1
1.094167
>
```

Figure 14: Output of the kNN predictions

The data from zone with high values of gwl of 97 was predicted as 2.842, which is 3.0 to the nearest whole number, representing class three of gwl level values ranging between 60 to 200m in the model. The data taken from the zone one which was already predicted as category 1.0 yielded a value of 1.094 which is approximately 1, meaning gwl ranged between 5m and 20m bgl values

VIII. CONCLUSIONS FROM STUDY

The study has employed the SOM Kohonen map models to generate predictions for water struck levels in the study area where the general depths are uncertain, so that they are inferred with reasonable accuracy of prediction, way before drilling. The efficacy of the model is confirmed by generating the same predictions using the kNN algorithm.

The SOM neural network is thus a useful clustering, classification and prediction tool ideal for use in stochastic hydrology.

REFERENCES

[1] Aguilera, P. A., Frenich, A. G., Torres, J. A., Castro, H., Vidal, J. M., & Canton, M. (2001). Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Water research, 35(17), 4053-4062.

[2] Bretzler, A., Lalanne, F., Nikiema, J., Podgorski, J., Pfenninger, N., Berg, M., &Schirmer, M. (2017). Groundwater arsenic contamination in Burkina Faso, West Africa: predicting and verifying regions at risk. *Science of the Total Environment*, *584*, 958-970Han, J. C.,

[3] Daliakopoulos, I. N., Coulibaly, P., &Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of hydrology*, *309*(1-4), 229-240.

[4] Hadjisolomou, E., Stefanidis, K., Papatheodorou, G., &Papastergiadou, E. (2018). Assessment of the eutrophication-related environmental parameters in two Mediterranean lakes by integrating statistical techniques and self-organizing maps. International journal of environmental research and public health, 15(3), 547.

[5] Huang, Y., Li, Z., Zhao, C., Cheng, G., & Huang, P. (2016). Groundwater level prediction using a SOM-aided stepwise cluster inference model. *Journal of environmental management*, *182*, 308-321.

[6] Kombo, O. H., Kumaran, S., Sheikh, Y. H., Bovim, A., &Jayavel, K. (2020). Long-Term Groundwater Level Prediction Model Based on Hybrid KNN-RF Technique. *Hydrology*, *7*(3), 59.

[7] Li, T., Sun, G., Yang, C., Liang, K., Ma, S., & Huang, L. (2018). Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes. Science of the Total Environment, 628, 1446-1459.

[8] Maiti, S., & Tiwari, R. K. (2014). A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction. *Environmental earth sciences*, *71*(7), 3147-3160.

[9] Merti Aquifer Study Report , Report No. 28/2017, Paper Page

[10] Nourani, V., Alami, M. T., &Vousoughi, F. D. (2016). Hybrid of SOM-clustering method and wavelet-ANFIS approach to model and infill missing groundwater level data. Journal of Hydrologic Engineering, 21(9), 05016018.

[11] Orak, E., Akkoyunlu, A., & Can, Z. S. (2020). Assessment of water quality classes using self-organizing map and fuzzy C-means clustering methods in Ergene River, Turkey. Environmental Monitoring and Assessment, 192(10), 1-10.

[12] Sahoo, S., &Jha, M. K. (2013). Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. *Hydrogeology Journal*, *21*(8), 1865-1887.

[13] Yadav, B., Ch, S., Mathur, S., &Adamowski, J. (2017). Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. *Journal of Water and Land Development*, (32), 103-112.