Viterbi Algorithm-Based Hidden Markov Modeling and the Fuzzy Gaussian Membership Function Regression Mapping of the Lithologic Logs Along the Madogo-Mororo-Maramtu Settlements

[1]Dr. M.O. Amimo, [2]Prof. Dr. K.S.S. Rakesh

[1]Research Scholar (LIUTEBM), [1]Livingstone University of Tourism & Business Management
[2]CEO, Gradxs
[1]bmoamimo@gmail.com, [2]kssrakesh@gmail.com

Abstract*—The present study is an attempt at simplifying the task of supervision and planning for drilling of wells along the Tana Alluvial Aquifer, with respect to the layering sequences of the prevalent stratigraphic columns along the Tana Flow course and even beyond. The climate change dynamics have led to falling well water levels, with some wells ending up as completely dried out, on the account of limited or low laminar flow recharge inputs, arising from the low levels of rainfall being occasioned. It is an attempt at developing simulation models of aquifer drill cuttings using the Hidden Markov Models and also using the Gaussian membership function-based predictive analytics. The basis of using the HMM methods is that the geologic layers have been studied over the years and have a consistent layering sequence defined by the geological Time of deposition of sediments in the Miocene and Jurassic periods. The geo-stratigraphic columns thus behave like a State of Stationary Markov Process, and may thus be modeled using Discrete Transition Probability Matrix (DTPM), and Emission Probabilities (EM) to simulate both observation sequences (clays, fine sandstones and coarse sandstones) and the state sequences (aquifer viability status) of each log. The two stochastic properties, DTPM and EM may be used to predict the state sequences (hidden aquifer status of the drilled logs). The HMM algorithm may be employed to predict the rock lithologic species logged to date (observations), as well as the expected (but hidden) stochastic states, coded as thus: barren, moist and saturated rock units. For the purpose of these HMM simulations, the Viterbi algorithm was used to simulate the expected aquifer status of a geolog assemblage. When the data was fed to an HMM algorithm so developed, so that the clays, coarse sands and fine sands would help infer if the logged material presented a state of barren-ness, moisture, or saturation, for each of the respective observed geologic input, results came out favorably, which matched with the established aquifer viability status of the input (class code of) geologic input material. The logs drilled were sampled and assembled in geologger box. For the ANFIS Neuro-Fuzzy Gaussian models, the data frame of the Tana Alluvial aquifer, complete with longitudes, latitudes, elevations and the logged interval depths of past drilling records and the geology of the material drilled, was indicated. Training and testing data sets were prepared. The new area/site being drilled, and whose GPS coordinates and observed geolog depths are known, had its log depth data fed into the model, which made estimation, with an accuracy of 100.0 percent, of the class of geologic material expected at the log depths specified. The GPS values and the log depths were used to predict the geology of the drill cuttings yet un-drilled. The study concludes that neuro-fuzzy regression algorithm using the frbs R library is a useful stochastic mapping tool for making inferences into the hydro-stratigraphy of the Tana Alluvial Aquifer. The study also vindicates the use of Hidden Markov Models, as a powerful simulation tool, in aiding the researcher make a conclusion, as to how many more meters of depth will be drilled before he obtains a logged lithologic unit, with brighter prospects of a productive aquifer material in the course of drilling progress.

***Key words*—Gaussian Membership Functions, Stochastic Hydrology, Hidden Markov Models, Markov Chains, Transition Probability Matrix, Neuro-Fuzzy, Tana Alluvial Aquifer.**

I. INTRODUCTION

The Tana Alluvial Aquifer project locality is roughly some 0-5 kilometers away from the main River Tana Flow course (highly productive) and also at 6-15 kilometers radius away from Flow course, running all as from Balambala all the way to Ijara/ Masalani areas located both to the west, central, and south of Garissa township. The aquifer benefits over 100000 persons and 200000 animal stocks, on the average, respectively for domestic and livestock use. The Bura sub-county in Tana River county and the subcounties of Garissa such as Balambala, Township, Fafi and Ijara are the localities along which the Tana Alluvial Aquifer runs. The Mororo-Maramtu-Madogo settlements are densely populated and exhibit abstractions of the aquifer on a massive level. There has been a problem with respect to the water supplies in Madogo settlements as well as in the Garissa Township. Consequently, wells are being sunk on trial-and-error basis. Water vendors and local residents of the riparian corridors have been sinking wells for water so that they sell on donkey carts and for domestic uses, respectively. For this reason, there are many dry wells within the aquifer subsets of Tana Alluvial Aquifer in Madogo and Mororo-Maramtu settlements. Consequently, a study was launched to aid the mathematical simulation of layers of saturated, barren and moist aquifer layers, so that if a well of 40m is to be sunk, and the simulation displays/outputs at least three to five layers denoting saturation, then the well is sunk. If the simulations performed using **Hidden Markov Models** suggest barren aquifers or ones that are simply moist (denoting limited recharge flow), then the well drilling

proposed is abandoned. The R library packages known as **markovchain and HMM** were used for this purpose. To complement the simulations, the study employed the Gaussian Fuzzy-based simulation models to predict the geology of sediments of a site yet un-drilled, using the **R package of frbs of the R software.** The ongoing small scale drilling programs shall help alleviate water scarcity in the area, and improve on the overall hygiene status, thus.

Rains have been failing for as many as three consecutive years and recharge of the aquifers is therefore in jeopardy, necessitating the study to act as a climate change intervention tool, in aiding less abstraction of the aquifer. The locals may be close to the river but they fear using the raw water owing to past and recent incidents involving mauling by the marauding crocodiles. The male folk are always by in the farms located by the riverbeds. It is for these reasons that the locals are sinking wells, desperately trying to avoid the scare of crocodile mauling.

II.HYDROGEOLOGY

The Project Area is dominated by thick successions of dark colored Miocene-Pliocene Sediments overlying Carbonates and Siliceous Tertiary sediments. The indigenous acacia shrubs and short, thorny under growths litter the countryside. The Project Area is so flat-lying that it gets easily susceptible to episodes of flooding after heavy rainfall events. Several dry, seasonal streams/rivers are visible in the vicinity. All such dry tributaries drain towards the River Tana flow course. The **Mansa Guda Clay,** coupled with the subsurface Mariakani sandstones, is the main storage parameters in the area, with the former responsible for salinity in the sediments where it dominates the subsurface geology. In fact, the Geo-electrical data generated over the years via vertical electrical soundings indicate less salinized interfaces at sub surface depths **beyond 8m on the account** of the moderate resistivity values recorded in past resistivity mapping work.

The salinity levels arise from the fact that the upper beds are rich in clays – the Kaolin-rich units of alluvial, silty sediments, which have poor hydraulic conductivity and cannot thus channel the precipitation water to trigger dilution effect through interactions in the subsurface hydrology system.

III. PROJECT LOCATION

Location

The project area lies in the Bura county of Tana River county as well as in the Garissa County within the Balambala, Garissa Townships and Fafi subcounties, as well as parts of the Ijara subcounty. Special interest centered on **Kone-gaba, Madogo, Mororo, and Maramtu areas**, on the account of massive abstractions going on. and It is located on the southwestern sides of the main catchments course way. The area is defined by longitudes and latitudes shown in the geophysical curves analyzed, and at an altitude of approximately 60-270m **above sea level.** Oblique dipping sediments litter the terrain alongside some zero-degree dipping units of flood-prone Miocene Pliocene sediments.

Geology and Stratigraphy

The topography is generally flat, and is clayey rich, supporting vegetations that comprise mainly thorny shrubs, undergrowth savannah grass, equatorial weeds typical of the coastal strips and acacia family trees. The geology is defined by **dark to light toned sandy clayey sediments, the Mansa Guda formation,** which overlies the carbonates – namely corallites, aragonitic sediments and calcite. The sandy clayey species are mainly the Mariakani Sandstones. The Jurassic limestone carbonates are fairly fractured and possess water at the shallow depths, though highly mineralized, via the fractures and karstification veins. Water also forms at the contact points between the carbonates 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 in the Jurassic – cretaceous period. Evapo transpiration rates of up to 3,000mm per annum over shadow the annual rains of up to 400mm per annum.

Physiography

The area stands at an average altitude of **60-270m** above sea level within a gently dipping terrain punctuated with several ant hills and flood plains both on the south eastern and north western flanks. The river flows in the northwest-southeastern azimuth.

Evidences abound of jointing and fracturing of the carbonate sediments on the surface, alluding to intense forces of fracturing, carbonation and quaternary tectonic faulting. Much of the south westerly – north easterly directed stress fields helped sculpture the terrain into its present geological state.

Owing to the relatively high fractions of clays in the beds, there is no sufficient time available for maximum catchment input infiltrations into the sub surface zones lying on the adjacent aquifer units in the proposed well sites. This explains the anomalous salinity levels of the boreholes done to great depths in the area.

Drainage

Owing to the relative flat nature of the terrain, there is flood rampancy. The permanent civil structures on the ground to stand the risk of destruction added to 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.

Climate

The project area falls within zone 7 of the classification of climatic/ecological zones of Africa, that is to say arid to semi-arid with temperatures averaging 30 to 34 degrees per day and occasioning evapo transpiration rates of up to 3000mm per annum. The rainfall average falls well below 500mm per year.

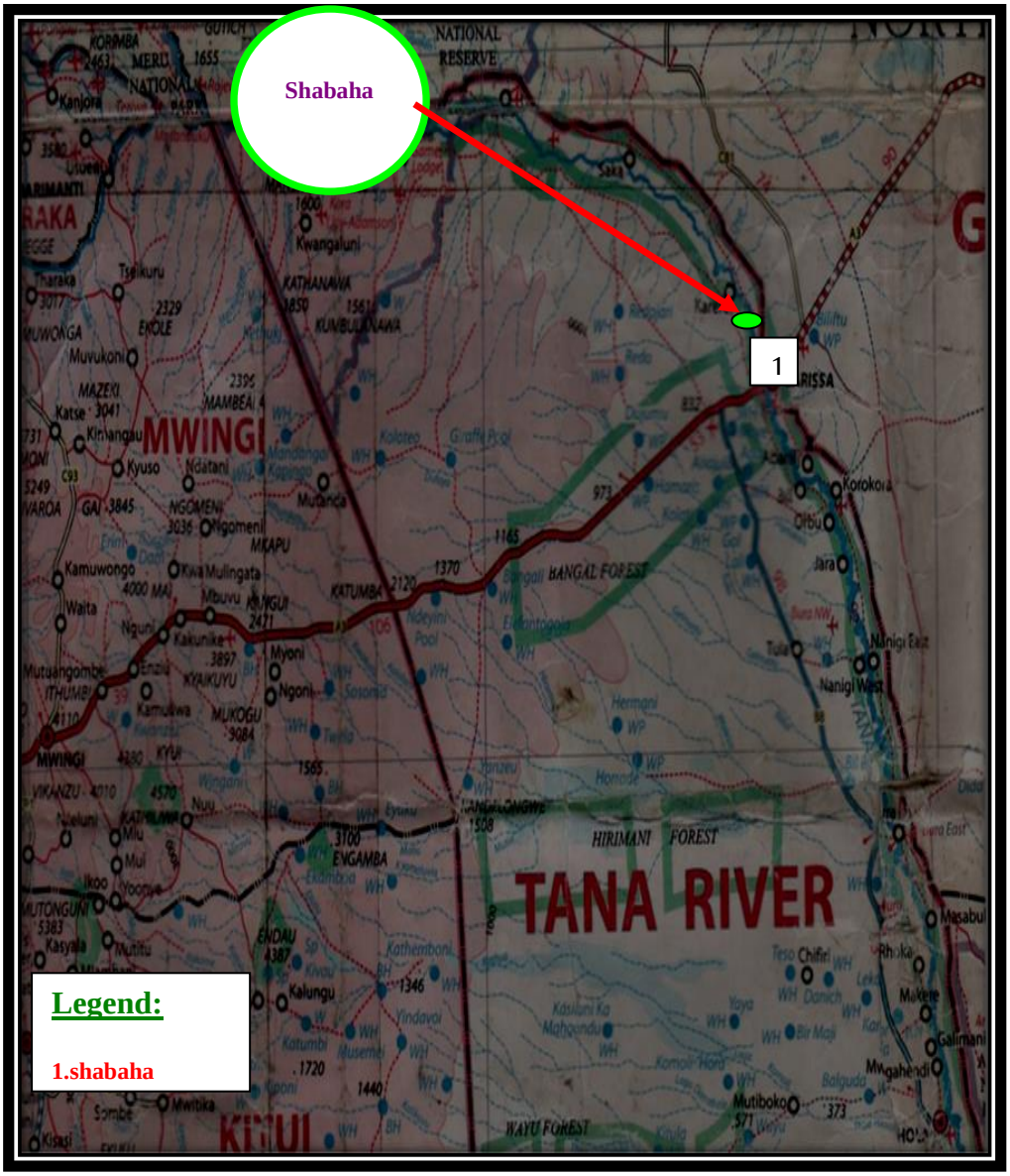


Figure 1: Map showing the location of the study area

IV. LITERATURE REVIEW

Fuzzy Logic Applications in Hydrology

It is not an easy task to make an informed estimate as to the lithologic species characterization of a drill cuttings of a site which is freshly proposed for drilling. However, the existing assemblage of the secondary data of drilling in the Tana Alluvial Aquifer was used to help make predictions of expected aquifer material. This task employed the use of Gaussian Membership Functions in neuro-fuzzy regression. The quest to also estimate whether the said drill cuttings is barren, moist, or saturated (productive) in terms of high-water strike probability was evaluated using the hidden Markov Models by employing the Viterbi algorithms to characterize the aquifers. The Viterbi algorithm helped simulate the expected productivity status of the aquifer during drilling using the mud rotary method for a well that is being drilled and has not been encased, graveled and developed using the poly-phosphatic sodium compounds used to decompose clays, thereby opening aquifers deemed blocked in the course of drilling of the well. The Gaussian membership functions have been found to be effective in modeling the non-linear relationships between the aquifer parameters as well as the lithologic species and aquifer water status of the penetrated material drilled and sampled at intervals of 2m each. The two algorithms provided a useful stochastic tool for mapping the aquifer geology but will be found useful in mapping the lithology of the localities in the neighborhoods bearing similar geology and which on the account of water scarcity but with the physical location being distant, further away from the river flow course, have been proposed for development of surface pans.

The computational AI includes Neural Networks, Genetic Algorithm, Fuzzy systems, and are amongst the many methods that have been used to characterize aquifer hydrology in the context of water quality evaluation in India (Jha, et al, 2020). These methods find appropriate use in that the relationships between the parameters being evaluated for predicting the aquifer lithology happen to be non-linearly related, making the conventional regression techniques obsolete. Among various computational AI methods, Fuzzy Logic (FL) is extensively used to deal with nonlinear relations amongst variables and also comes in handy when modeling phenomena where data scarcity is an impediment.

Fuzzy logic may be combined with other algorithms to aid efficient decision-making with respect to mapping risks probabilities anticipated in dams or large pans. combined with the flood routing results for the Kensiwate reservoir using Monte Carlo simulation, a reservoir overtopping risk model based on right-angle trapezoidal fuzzy numbers was established, and the fuzzy risk index intervals and the corresponding fuzzy risk rate intervals for the Kensiwate reservoir (Yang et al, 2020).

In the present study, fuzzy logic was employed to simply aid in educated guess for the supervising hydrogeologist and the geophysicist who does the investigations for ground water potential in the Tana Alluvial Aquifer, having mapped the salinity risk in previous studies on the area's Hydrogeology.

Fuzzy logic has thus been a useful research tool in mapping stochastic conditions in the environmental, climate change and physical sciences. As an example, Danandeh et al (2020) used a hybrid method of both fuzzy logic and random forests in mapping the classification of droughts in un-gauged catchments so mapped within The Antalya Basin in the nation of Turkey.

The algorithm was used in December 2020 to map the uncertainty which relates to the evapotranspiration dynamics of the subtropical localities of Florida in the United States (Roy et al, 2020).

A good understanding of hydraulic properties of soils and rocks along which water flows is necessary for hydrological studies, artificial recharge of the aquifer, watershed management and agriculture system (Sihag et al 2018). The Sihag study established that the fuzzy logic algorithm performs super-efficiently in rendering high accuracy predictions when combined with neural networks. The study mainly targeted the development of a combined fuzzy -logic and artificial neural network-based models, for use in the approximate mapping of the values of unsaturated hydraulic conductivity of soil.

Again, a study by Sihag et al, (2019) established the power of fuzzy logic Gaussian membership functions in helping characterize the stochastic hydrology and recharge dynamics of a terrain earmarked for agricultural development. The process of Infiltration studied in the research was deemed essential in planning for irrigation-based agriculture and for streamflow, as well as the groundwater recharge, subsurface flow, surface as well as the subsurface water quality and quantity. The study did separate Precipitation is into two subsets for ease of mapping, namely, (i) surface and (ii) subsurface flows via infiltration dynamics. The action of infiltration through soil it was established to be a function of several parameters, ranging from antecedent soil moisture, soil texture and type, soil density, hydraulic conductivity and precipitation levels, to hydrological features deemed stochastic in nature.

The fuzzy models have also been employed to perform simulations of predictive models where data is scarce. As an example, a study in India involving river hydraulics and recharge, runoff simulation models were developed to aid forecasts of runoff for the basin located to the West Godavari district, Andhra Pradesh by utilizing Adaptive Neuro-Fuzzy Inference System, abbreviated as ANFIS models (Kumar et al, 2019)

A study in 2019 in India (Singh et al) characterized the water quality of elected catchments using fuzzy logic, noting the roles played by several parameters in aiding the deterioration or degrading of the water chemistry. The study noted that water quality of most of the rivers in India has been steadily degrading due to increasing numbers of point and non-point sources of pollution. This has all to do with population pressure, socio-economic factors and land-use based pressure on the rivers ecosystem, rendering it vulnerable to pollution impacts. The variables determined thus were:

a) Massive increase in population,
b) rapid urbanization,
c) change in irrigation patterns, and
d) unplanned growth of industries without proper enforcement of environmental standards.

These were established to be the main factors behind deterioration of water quality in the rivers mapped in the study. A study by Baalousha et al (2021) ranks fuzzy-based methods as being superior to the numerical-based DRSTIC techniques in the mapping hydrology of a catchment in Qatar.

Hidden Markov Models Applications in Hydrology

Hidden Markov Models, usually abbreviated as HMM, have been used to characterize the climate change and hydrology of the Nile Basin Catchment systems (Khadr, 2016).

In this study, several homogenous Hidden Markov Models were developed to aid in the scientific predictions of droughts, by employing the variable known as the Standardized Precipitation Index, SPI, at for both the short term as well as the medium-term durations. The study findings were later on validated using the developed models, by comparing the results to the precipitation series observed in twenty two stations situated in the upper Blue Nile River catchment.

Hidden Markov models have been used as a geotechnical and a groundwater exploratory tool in China (Zhao et al, 2020). This is because for one to strike water, the geophysical methods usually strive to identify the weak zones of faulting or weathering, which easily aid groundwater flow within the rock's interstitial settings. Geophysics alone is never absolute in terms of accuracy, and performs only well if complemented by hydrogeology and structural geology. The faults may also be a risk factor with respect to earthquakes or foundational stability of high-rise buildings. It was with this in mind that the Chinese study leveraged on the power of probability-based algorithm of HMM, to help detect faults and the probability of these faults evolving into bigger ones, using the known process geomorphology of the studied terrains.

In a study of 2019 in China, it was found that better results are attained when the research complemented the traditional algorithms known for use in the Aquifer Contaminant Hydrology mapping with the HMM-based models, listing the use of Support Vector Machines, the Radial Basis Functions, as well as the metropolis-based MCMC algorithms (Xing et al, 2019).

Li et al (2022) studied the water quality dynamics as time-series. The study mapped out water quality parameters that had the characteristics of a time series, and noted that the variations displayed instability and nonlinearity, coupled by complex relationships. This made it difficult to evolve a single standalone predictive data mining model, to characterize and predict the water quality forecasts with a high level of precision and accuracy as envisaged. Subsequently, the study concluded that high accuracy in levels of prediction of water quality parameters was compromised in the use of a single algorithm, necessitating the formulation of a water quality parameter prediction model that combined the Discrete Hidden Markov model abbreviated as the DHMM, and the K-Means Clustering algorithms. The hybrid scheme was then tried on the data of water quality fluctuations with time, in a time-series like formulation. It was found that The model accurately predicted the dissolved oxygen saturation levels as well as the and turbidity levels, prevalent in the marine ranches located within the neighborhoods of the Bohai Rim.

Since the HMM models are primarily probabilistic in nature, they have been validated using the Bayesian models and found to perform to a more or less similar degree of accuracy. The study was performed (Teixeira, et al 2019) consisting of a high-dimensional Bayesian inverse problem and a global sensitivity analysis. The approach was so novel it was deemed a first, as it did mark the first time in karst hydrology, when the active subspace method to find directions in the parameter space that dominate the Bayesian update, from the prior to the posterior distribution in order to effectively reduce the dimension of the problem and for computational efficiency. The space dimensions are a critical factor in understanding the efficiency of a karst system, as it is these very spaces that store water and also are the very ones which admit the flow of water in a recharge-discharge matrix. Moreover, the study HMM opined that the calculated active subspace can be exploited to construct sensitivity metrics on each of the individual parameters and be used to construct a natural model surrogate. This helps understand the karst hydrologic model and aids in quantifying the storage capacity of the material, with respect to its efficiency as an aquifer suitable unit, also known as reservoir capacity. Some of the aquifer subsets mapped in the Tana Alluvial Aquifer belong to the karst hydrogeology, as they contain massive volumes of limestone and bicarbonates.

Climate change hydrogeology is also a discipline considered a ripe candidate for modeling and / or assessment using the hidden Markov models and its accompanying Viterbi algorithm.

A study in Iran by Azimi et al (2020) utilized the HMM algorithm to aid understanding of the long-term climate changes anticipated in the area so mapped. The research generated a steady-state Markov chain model, subsequently used in the prediction of the long-term probability of drought conditions in the study area, with respect to levels of severity expected over time. The study suggested a template for data analysis of drought characterization, and the trends this drought defined over time, for the vast area or geographic extent of the of aquifers mapped on one hand, and the dry bare plains of Iran, on the other hand. The groundwater drought study analyzed a number way greater than twenty-six thousand wells derived from about six hundred meteorological stations. Hidden Markov models were leveraged in mapping subsurface petro-physical properties in a study (Wang et al, 2020). The said petrophysical parameters traditionally differ from one reservoir to the next, and this has an impact on the lithology identification of the geological material, particularly so for the unconventional reservoirs. Consequently, the lithology identification of subsurface reservoirs is a such a difficult task, since there is no one known algorithm that may be regarded as a stand-alone, efficient predictor of the lithology of the reservoirs being mapped. This is further complicated by the mineralogical and petrologic heterogeneities inherent in most such reservoir material. The study noted that machine learning or ML algorithms may be effectively leveraged in this respect to aid speedier and more reliable processing of predictive models to aid decision making. The ML algorithm of Random Forest was combined with HMM, with astounding results in respect of predicting the expected geology/ drill cutting materials geology. By using the Viterbi algorithm to simulate data and then combining the simulated data with well logging data from secondary sources, a reliable prediction of the lithology was observed to occur, subsequently validated using field seismics.

Aguagallo (2021) undertook a study which verified that the HMM algorithm may be used as a standalone time-series algorithm or as a complement to the existing array of time series prediction platforms. In the study, he combined the ARIMA (Autoregressive Integrated Moving Averages)-based models with the HMM models to generate a time series predictive analytical model.

Piho et al (2021) also mapped the groundwater flow path modeling using the probabilistic HMM models and testified to its effectiveness.

V. GAUSSIAN FUZZY MEMBERSHIP FUNCTION PREDICTIVE MODELS

Gaussian Fuzzy mf

Just as bell-shaped membership function, Gaussian Membership (abbreviated as GAUSSIAN in R) bears a smooth curve. As opposed to the other membership functions of the same family of curves, the algorithm uses only two parameters:

i) The variable c for locating center and
ii) The variable σ for determining the width of the curve.
iii) The variable x is the new value whose mf is being determined.
iv) All the three are subsequently expressed as a single mathematical equation, thus:

$$gaussian(x, c, \sigma) = e^{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2}$$

Gaussian membership functions have been verified to be more robust and reliable than the trapezoidal and triangular fuzzy membership functions. The ideal scenarios are that the data given is fuzzified via a process known as fuzzification before being subsequently passed through the fuzzy inference engine, culminating into fuzzified outputs. In the Tana Alluvial Aquifer study, the fuzzy rules were generated using **the frbs** software in R and where the predictions did not yield acceptable levels of accuracy with the sample test data, the labels of number of the defuzzification class were increased progressively until a suitable level of accuracy was attained with the training and test datasets used in the model. The fuzzy inference engine in the study was the Wang-Mendel Algorithm also abbreviated as WM in R language-speak.

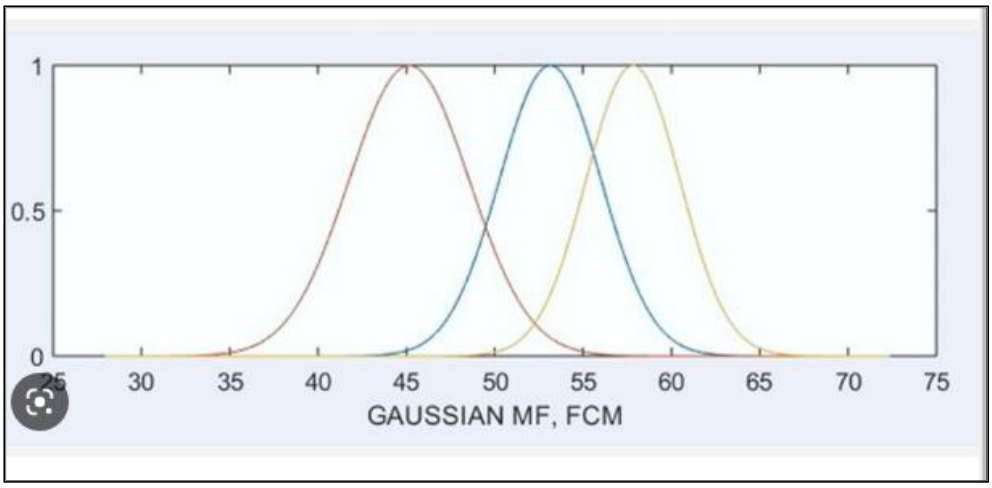


Figure 2: Illustrating Gaussian MF

VI. HIDDEN MARKOV MODELS

Introduction

There is a reason for using markov chains and hidden markov models to model a stochastic process which is sequence dependent or just simply, time dependent. Before one builds a hidden markov model, one has to be able to build a simple markov Model, which is the variable that would subsequently be used to build a Hidden Markov Model.

Examples in real life

One may consider the two examples listed in the tab les hereunder for ease of following the conceptual framework of Markovian process**.**

Table 1: Days in a 3-month duration when town residents are seen to respond to weather as tabulated

WEATHER IDENTITY	OBSERVABLE SIGN
Cloudy Weather	Umbrella
Rainy Weather	Raincoat
Sunny Weather	T-Shirt

Just to explain the tables, there are days when people were seen with umbrellas but it did not necessarily rain, implying the clouds did not yield rainfall. In other day, the umbrellas were carried by persons that had seen clouds and feared rainfall, and indeed, it rained. There are days' people carried raincoats and it did really rain. In other days within the same 90-day period, they carried the raincoats but there wasn't really much rainfall to warrant carrying raincoats. There also instances people wore T-Shirt as the weather was sunny, but later in the day, it did rain.

Markov Models and hidden markov models attempt to answer the questions about the sequences as observed by someone who is only, say, staying in the room, unable to walk outside, maybe for reasons of being ill/indisposed, so that the only way for him to know if it was raining or cloudy or sunny, would be by observing the dress mode and the carrying or umbrella. From the table shown above, it would be misleading for him to conclude that there is a weather of CLOUDY, CLOUDY, SUNNY, RAINS, CLOUDY if he sees a sequence of five CONSECUTIVE days of:

Umbrella, umbrella, T-shirt, raincoat, umbrella.

The sequence may end up as being the following:

CLOUDY, RAINY, RAINY, SUNNY, CLOUDY

The mathematics of probability sequences which relate the observed state to the actual element is the essence of Hidden Markov Modeling.

Now consider a stationary markov Process of layering of Rock sequences in geologic strata comprising lithologic logs, soils, water and lineaments.

Table 2: Rock sequences in the course of drilling a 255m bgl depth borehole in a sedimentary terrain

Log Identity	**Observable Sign (2m Interval of Drill Cuttings)**
Clays and limestone rocks	Barren drill cutting/geolog
Fine sandstones and silts	Wet, low moisture drill cutting/geolog
Coarse sandstone and gravels	Productive drill cutting/ geology of over 1.5 cubic meters per hour

Again, in the above table, one notes that there would be cases whereby productive aquifers occur in the fine sandstones, but obviously at a generally lower frequency than within the sedimentary beds comprising coarse grained sandstone and gravels. The relative probability with which that happens will hele a researcher to predict the expected overserved states of the rock cuttings drilled with respect to its aquifer discharge promise status. It may be interesting to note that the water occurs within the clayey beds with low probability but ends up giving the highest amount of water for the well. The layering of the rocks and soil types in the verticals and horizontal stratigraphic profile is a well-defined sequence, making it ideal for modeling using markov chains computations, and since no time lapse is in consideration here, the process is modeled a stationary markov process.

Explanations of Markov Chains

There are many real-world phenomena which mathematicians refer to as so states and modeling the states form the basis of markov chains.

To start with the Markov Chains is a statistical procedure for modeling a stochastic process, and is defined as thus:

Given a finite set of state $S=\{s_1, s_2,\ldots, s_n\}$ whose cardinality is n. Let $\prod$ be the *initial state distribution* where $\pi_i \in \prod$ represents the probability that the stochastic process begins in state s_i. In other words π_i is the initial probability of state s_i, where the following is true:

$$\sum_{s_i \in S} \pi_i = 1$$

The stochastic process which is modeled gets only one state from S at all-time points. This stochastic process is defined as a finite vector $X=(x_1, x_2,\ldots, x_T)$ whose element x_t is a state at time point t. The process X is called *state stochastic process* and $x_t \in S$ equals some state $s_i \in S$. Note that X is also called *state sequence*. Time point can be in terms of second, minute, hour, day, month, year, or a subjective interval of logging drill cuttings as illustrated above for a stationary markov process.

It is easy to infer that the initial probability $\pi_i = P(x_1=s_i)$ where x_1 is the first state of the stochastic process. The state stochastic process X must meet fully the Markov property, which include, amongst other conditions, that given previous state x_{t-1} of process X, the conditional probability of current state x_t is only dependent on the previous state x_{t-1}, not relevant to any further past state $(x_{t-2}, x_{t-3}\ldots, x_1)$.

This is to state that the present observed state is independent of the preset state but may be explained by the state that was prevalent in the previous sequence. In other words, $P(x_t \mid x_{t-1}, x_{t-2}, x_{t-3},\ldots, x_1) = P(x_t \mid x_{t-1})$,.

One may also note the fact that the term $P(\ .)$ also defines probability, in this derivation example used. A process bearing this kind of property is called first-order Markov process, mathematically speaking.

At some point in the sequence or reference time frame, the process changes to the next state, based PRIMARILY UPON the *transition probability distribution* a_{ij}, which IN ITSELF depends only on the previous state. Consequently, this variable a_{ij} is the probability that the stochastic process changes current state s_i to next state s_j.

This implies that $a_{ij} = P(x_t=s_j \mid x_{t-1}=s_i) = P(x_{t+1}=s_j \mid x_t=s_i)$. The probability of transitioning from any given state to some next state is 1, we have the following: -

$$\forall s_i \in S, \sum_{s_j \in S} a_{ij} = 1$$

All transition probabilities a_{ij} (s) shall now constitute the **transition probability matrix A.** The variable A is an n by n matrix, due to the fact that there exist n distinct states. The matrix A therefore represents state stochastic process X and from the foregoing, it may be further inferred that the initial probability matrix $\prod$ is degradation case of matrix A.

Summarily, a Markov Model, denoted as MM, is the triple $\langle S, A, \prod \rangle$. In typical MM, states are observed directly by users and transition probabilities (A and $\prod$) are unique parameters.

Further explanations on Hidden Markov Models

The Hidden Markov Model, abbreviated here as HMM, is quite similar to MM except that the underlying states become hidden from observer, just as explained in the two tables above, used to explain and simplify the process of **Hidden Markov Model** inference. The HMM adds more output parameters, defined thus as observations. It would be noted that each state (the hidden parameter) has the conditional probability distribution upon such observations.

As explained earlier, the drill cuttings seen maybe clay, with the hidden state being the barren rock or soil with zero to negligible water quantity. The same rock cuttings may be sandstones, with a high-water discharge from underground sources. The use of HMM computations may thus be of immense help in making scientific inferences as to how much water has been attained, way before the drilling unit has assembled a test pumping outfit for testing the aquifer productivity of the well.

The HMM algorithm is thus a helpful scientific / mathematical tool in discovering these hidden parameters (states of water availability in the rock or state of weather being inferred from dressing of passersby) from output parameters (observations - the actual rocks drilled and logged in a logging box or the actual weather indicator like sunny, rainy, or cloudy states), inferred from the stochastic process. The HMM has further properties as thus: -

i) Assuming there was a finite set of possible observations $\Phi = \{\varphi_1, \varphi_2 \ldots, \varphi_m\}$ whose cardinality is m. assume also that there is the second stochastic process which produces *observations* correlating with hidden states. This process is called *observable stochastic process*, a phenomenon explained as a finite vector $O = (o_1, o_2 \ldots, o_T)$, whose element o_t is an observation at time point t.

ii) It is instructive that t $o_t \in \Phi$ equals some φ_k. The process O is often known as *observation sequence*.

iii) Finally, there is a probability distribution of producing a given observation in each state. Let $b_i(k)$ be the probability of observation φ_k when the state stochastic process is in state s_i. This would imply that $b_i(k) = b_i(o_t=\varphi_k) = P(o_t=\varphi_k \mid x_t=s_i$, so that the sum of probabilities of all observations observed in a certain state is 1, will be defined by the equation hereunder:-

$$\forall s_i \in S, \sum_{\theta_k \in \Phi} b_i(k) = 1$$

iv) All probabilities of observations $b_i(k)$ constitute the observation probability matrix B. It is convenient for mathematicians to employ the notation b_{ik} instead of notation $b_i(k)$. Moreover, take note of the fact that B is n by m matrix on the account of these n distinct states and m distinct observations.

v) Summarily, the matrix A represents state stochastic process X, matrix B represents observable stochastic process O. Subsequently, the Hidden Markov Model is the 5-tuple $\Delta = \langle S, \Phi, A, B, \prod \rangle$. It may be also noted that components S, Φ, A, B, and $\prod$ are often called parameters of HMM in which A, B, and $\prod$ are essential parameters.

vi) Return to the weather example. Assume one needs to make a forecast of how weather tomorrow will be like: **sunny, cloudy or rainy**, as one knows only observations about the humidity: *dry*, *dryish*, *damp*, *soggy*. The HMM is totally determined based on its parameters S, Φ, A, B, and $\prod$ according to weather example. We have $S = \{s_1=sunny, s_2=cloudy, s_3=rainy\}$, $\Phi = \{\varphi_1=dry, \varphi_2=dryish, \varphi_3=damp, \varphi_4=soggy\}$. Transition probability matrix A is shown in Table 1.

Table 3: Transition probability matrix A.

		Weather current day (Time point t)		
		sunny	*cloudy*	*rainy*
Weather previous day (Time point $t-1$)	*sunny*	a_{11}=0.50	a_{12}=0.25	a_{13}=0.25
	cloudy	a_{21}=0.30	a_{22}=0.40	a_{23}=0.30
	rainy	a_{31}=0.25	a_{32}=0.25	a_{33}=0.50

From table 3, we have $a_{11}+a_{12}+a_{13}=1$, $a_{21}+a_{22}+a_{23}=1$, $a_{31}+a_{32}+a_{33}=1$.

Initial state distribution specified as uniform distribution is shown in table 2.

Table 4: Uniform initial state distribution $\prod$.

sunny	cloudy	rainy
π_1=0.33	π_2=0.33	π_3=0.33

From table 4, we have $\pi_1+\pi_2+\pi_3=1$.

Observation probability matrix B is shown in table 5.

Table 5: Observation probability matrix B.

		Humidity			
		dry	*dryish*	*damp*	*soggy*
Weather	*sunny*	b_{11}=0.60	b_{12}=0.20	b_{13}=0.15	b_{14}=0.05
	cloudy	b_{21}=0.25	b_{22}=0.25	b_{23}=0.25	b_{24}=0.25
	rainy	b_{31}=0.05	b_{32}=0.10	b_{33}=0.35	b_{34}=0.50

From table 5, we have $b_{11}+b_{12}+b_{13}+b_{14}=1$, $b_{21}+b_{22}+b_{23}+b_{24}=1$, $b_{31}+b_{32}+b_{33}+b_{34}=1.00$

The explanation rendered here IN THE foregoing synthesis of Markov Chains and Hidden Markov Models is necessary for one to follow the data modeling process implemented in software on the Tana alluvial aquifer using the Viterbi algorithms.

VII.DATA ANALYSIS AND FIELD WORK

The field work which was undertaken was complemented with real existing drill cuttings in the repository pf the Ministry of Water, Sanitation and Irrigation as well as those in the Northern Water Works Development Agency. From the logs, a table was prepared with the following: -

i) Longitudes
ii) Latitudes
iii) Elevations
iv) Geology-representing the **OBSERVED** mineralogical characterization of the lithologic unit drilled and logged by the geologist
v) State –representing the **inferred** as well as real aquifer potential status of this drill cutting as inferred and confirmed by geologist during the process of drilling using mud rotary methods, prior to encasement, graveling and development of the borehole.
vi) category1-the coded representation of the geology in (iv) above to aid hmm and allied mathematical computations
vii) category2-the coded representation of the state as shown in (v) above, again to aid computations that helped arrived at predictive and simulation models of hidden Markov models, abbreviated as HMM.

longtd	lattd	elev	logDepth	category1	category2	geology	state
610946	9861112	76	2	1	1	clays/carb	barren
610946	9861112	76	4	3	3	coarse sar	saturated
610946	9861112	76	6	2	2	finesands	moist
610946	9861112	76	8	1	1	clays/carb	barren
610946	9861112	76	10	1	1	clays/carb	barren
610946	9861112	76	12	3	3	coarse sar	saturated
610946	9861112	76	14	3	3	coarse sar	saturated
610946	9861112	76	16	3	3	coarse sar	saturated
610946	9861112	76	18	1	1	clays/carb	barren
610946	9861112	76	20	3	3	coarse sar	saturated
610946	9861112	76	22	3	3	coarse sar	saturated
610946	9861112	76	24	2	3	coarse sar	saturated
610946	9861112	76	26	3	2	coarse sar	saturated
610946	9861112	76	28	3	3	coarse sar	saturated

Figure 3a: The figure above shows the screenshot of the data-frame used for one of the wells whose logs were employed in generating the hidden markov model

The actual model deemed it unnecessary to have the longitudes, latitudes and elevations, minding only the use of geology and state probabilities. The tables were then prepared and the two variable codes as class or categories ranging between 1 and 3 as thus.

Table 6:Geologic Material and Code Used

Geologic material Name	Code used	Comments
Clays/ carbonates	1-lowest percentage of silica material in drill geolog cuttings assembled and analyzed. Grain size is less than 1mm	This is the category with minimum water where it is drilled. However, some clay zones have lots of water. Gypsum, and gypsiferrous limestone is categorized here as well as loams. they are categorized as **BARREN** for the purpose of this study
Fine sands	2- moderate proportions or percentage of silica material in drill geolog cuttings assembled and analysed. Grainsize is between 0.5mm-1mm in diameter	These are fine grained particles of silica with low hydraulic conductivity and may at times be just like the clays in terms of soil mechanics. They have a moderate probability of striking water, though, and are categorized as **MOIST**
Coarse sands	3- highest proportions or percentage of silica material in drill geolog cuttings assembled and analyzed. Grainsize is between 2mm-5mm in diameter	Aquifer material may contain coarse sandstones or gravels but the code used here is just 3.0 and the lithologic term used here to embrace both gravels and sandstones is “coarse sands”. They are categorized as **SATURATED**

Table 7: Aquifer Status and Code Used

Aquifer Status Name (State)	Code used	Comments
barren	1—lowest levels of wetness and moisture content, and hence less probability of striking water. Some logs with this feature have soil mechanics favoring water availability, though. Coarse grained logs were found to be barren in several instances.	This is the category with minimum water where it is drilled in the study area, clay-rich logs dominate this category
moist	2—moderate levels of wetness and moisture content, and hence higher probability of striking aquifer water. Some logs with this feature have soil mechanics favoring barrenness, though. Some clays/ carbonates which were fine grained were found to be moist, in spite of the inferior grain size. Some moist sediments were however found to be barren.	Fine grained sandstone material bear geology which possesses and favor these characteristics.
saturated	3- highest proportions or percentage of silica material in drill geolog cuttings assembled and analyzed. Grainsize is between 2mm-5mm in diameter. Some coarse-grained sediments were still found to be barren, in spite of being saturated	Aquifer material may contain coarse sandstones or gravels in this category

Modeling in Markov Chains rock lithology and aquifer states in Markov Chains

The process was undertaken using the markovchain R library and the visualized output displayed as thus.

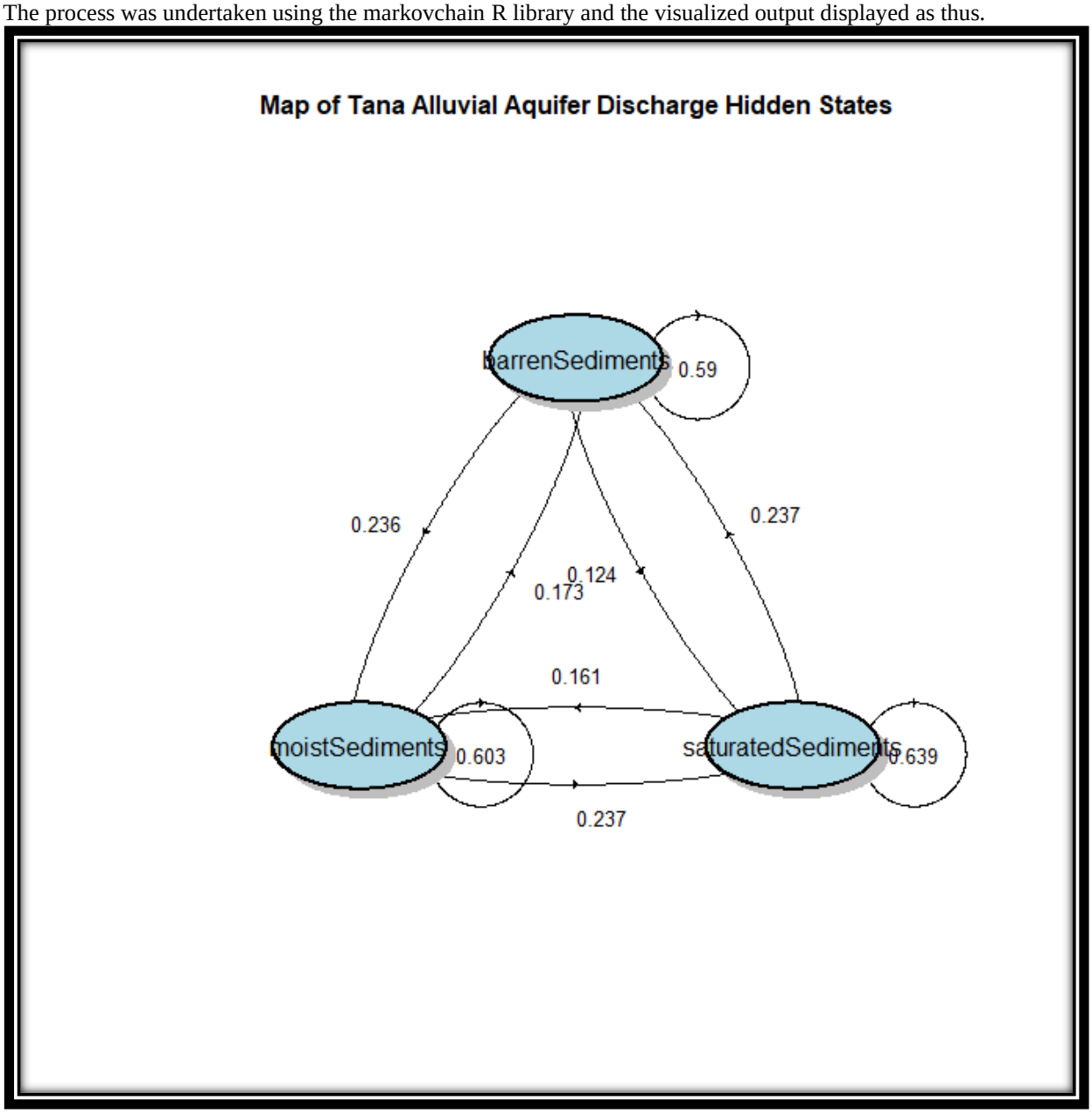


Fig 3b*-The HMM model Map of aquifer state in the geologs within the Tana Alluvial Assemblage*

Explanation –There is a probability of 0.59 that if a drill cutting (log) is barren, then the next cutting at an interval of 2m will also be barren from this markov model. There is also a probability that if a drill cutting saturated with water, it stays so with the next immediate drill cutting with a probability **of 0.64,** whereas if the drill cutting was moist, it stands to remain so with a probability of 0.60 in the immediate next drill cuttings in an on-going drilling episode.

The markov chain algorithm also displayed the following model for aquifer states which are generally hidden and are unknown during drilling phase using mud rotary techniques.

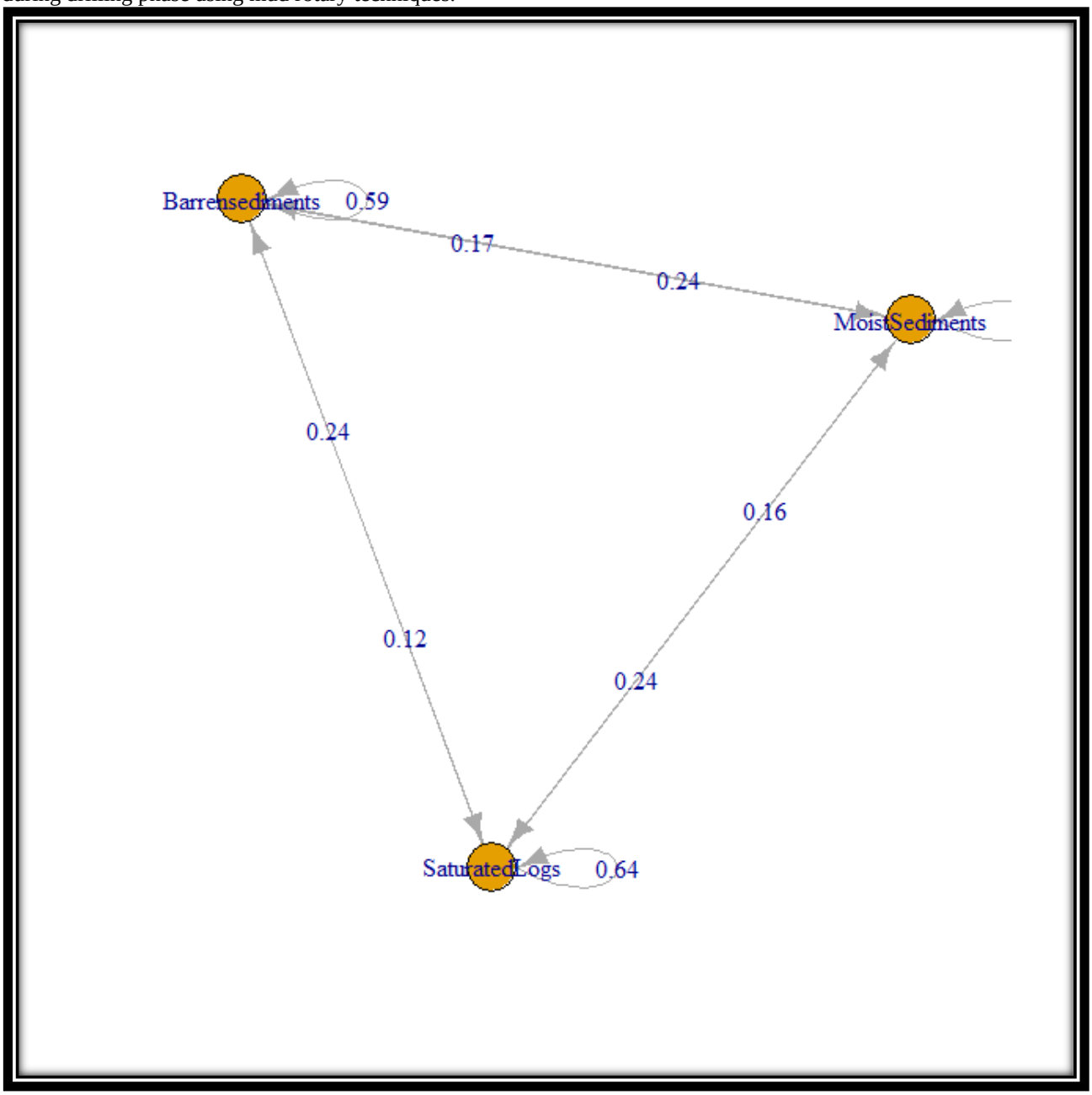


Fig 3c- the Markov chain Model of the aquifer states inferred in the Tana Alluvial aquifer geolog Assemblages.

Explanation –There is a probability of 0.59 that if a drill cutting (log) is barren, then the next cutting at an interval of 2m will also be barren from this markov model. However, there is a probability of 0.17 that the next drill cutting will be a moist sediment, whereas a probability of 0.24 that the next cutting is going to be productive (saturated logs). Conversely, if a material drilled is moist, there is a probability of 0.60 that the next cutting also assumes this (hidden) moist state, and a probability of 0.16 that the next log cutting will be saturated with water, alongside a probability of 0.24 that the next log is going to be barren. Knowing this, it means that a saturated log with water is likely to remain so with a probability of 0.64, with respective probabilities of barren-ness and moisture standing at 0.12 and 0.24.

Modeling using R software

As mentioned earlier, the various categories of 1, 2, and 3 for both geology (observed) and the hidden (state of aquifer productivity or barrenness) were counted in the whole data frame and assembled for use to compute the hmm Model of the Tana Alluvia Aquifer.

```
>
>
> elements <- c("clays","finesand","coarseSand")
> claysStateProb <- c(0.413,0.303,0.284)
> finesandStateProb <- c(0.341, 0.467,0.192)
> coarseSandStateProb <- c(0.171, 0.294, 0.535)
>
```

Figure 4: Table showing probability distribution of the various geologic units.

Generating the emission probability matrix

i) Where clays dominate, there are **41.3% clays, 30.3% fine sands and 28.4% coarse** grained sandstones. These sediments possess the highest probability of delivering a dry well.

ii) Where the fine sands dominate, there are **34.1% clays, 46.7% fine sands and 19.2%** coarse grained sandstones. These sediments possess the most moderate probability of delivering a well with moderate discharge.

iii) Where coarse sandstones dominate, there are **17.1% clays, 29.4% fine sands and 53.5%** coarse grained sandstones. These sediments possess the highest probability of delivering a well with highest discharge.

Running the codes here under gives the output desired.

```
>
>
> #We create an emission probability matrix.
> emissProb <- matrix(c(claysStateProb,finesandStateProb,coarseSandStateProb$
> emissProb
      [,1]  [,2]  [,3]
[1,] 0.413 0.303 0.284
[2,] 0.341 0.467 0.192
[3,] 0.171 0.294 0.535
>
```

Figure 5: The Desired R Output

The next step was to run the codes to generate the transition probability state.

Generating the transition probability matrix

The codes which were run in R to generate the above are as thus:

```
> ####################number twos################
>
> #IF(D58="barren","1",IF(D58="moist", "2","3"))#
> states <- c("barren","moist","saturated")
> barrenProb <- c(0.590,0.173, 0.237)
> moistProb <- c(0.236, 0.603, 0.161)
> saturatedProb <-c(0.124, 0.237, 0.639)
```

i) Where the barren state category dominates there are **59% of getting barren well, 17.3% of getting a moist geolog material and 23.7% saturated high** discharge aquifer, in spite of being of the clays/carbonate geology. These sediments possess the highest probability of delivering a dry well.

ii) Where the moist state category dominates there are **23.6% of getting barren well, 60.3% of getting a moist geolog and 16.1% saturated** high discharge aquifer. These sediments possess the highest probability of delivering a LOW YIELDING WELL.

iii) Where the SATURATED state category dominates there are **12.4% of getting barren well, 23.7% of getting a fine sands aquifer and 63.9% c chance of getting a highly saturated aquifer material log.** These sediments possess the highest probability of delivering a reliable well.

The transmission probability matrix was calculated thus in R:-

```
>
>
>
> #Based on those selection probabilities, we build a transition probability$
> transProb <- matrix(c(barrenProb,moistProb,saturatedProb), 3)
> transProb
      [,1]  [,2]  [,3]
[1,] 0.590 0.236 0.124
[2,] 0.173 0.603 0.237
[3,] 0.237 0.161 0.639
>
```

The study delved into modeling the HMM using these emission and steady state probability matrices in the R software platform. It was from these probability figures that the HMM model for simulating aquifer material was developed using the hmm library in R.

```
>
> hmm <- initHMM(States = states,
+                    Symbols = elements,
+                    transProbs=transProb,
+                   emissionProbs = emissProb)
>
> print(hmm)
```

The output was as thus: -

```
[1] "barren"    "moist"     "saturated"

$Symbols
[1] "clays"      "finesand"   "coarseSand"

$startProbs
   barren     moist saturated
0.3333333 0.3333333 0.3333333

$transProbs
           to
from        barren moist saturated
  barren     0.590 0.236     0.124
  moist      0.173 0.603     0.237
  saturated  0.237 0.161     0.639

$emissionProbs
           symbols
states      clays finesand coarseSand
  barren    0.413    0.303      0.284
  moist     0.341    0.467      0.192
  saturated 0.171    0.294      0.535
```

***Application in Real life*:**

The Simulations of logs and estimating how many of these logs will have water in the Tana Alluvial Aquifer (logging interval=2.0m)

Suppose one is supervising drilling of a borehole of 80m deep within the Tana Alluvial Aquifer, and with mu drilling, the supervisor cannot tell how much water or how close to being dry the well in progress is until after casings, graveling, kalgoning and development flash pumpage. Suppose the supervisor now has the hmm Model developed for the Tana Alluvial Aquifer. All he needs to do is to simulate the first 20 samples (40m drilled so far)

The codes hereunder will accomplish this.

```
> ##simulate twenty logs and estimate how many of them
> ##will be having water in the well being drilled
> simhmm <- simHMM(hmm, 20)
> simulated <- data.frame(state=simhmm$states, element=simhmm$observation)
> print(simulated)
      state    element
```

Figure 6: The R codes for the HMM

The output will be as hereunder:

```
> simulated <- data.frame(state
> print(simulated)
       state    element
1      moist coarseSand
2  saturated   finesand
3  saturated   finesand
4      moist coarseSand
5     barren   finesand
6     barren   finesand
7     barren   finesand
8  saturated   finesand
9  saturated coarseSand
10 saturated coarseSand
11     moist coarseSand
12     moist coarseSand
13 saturated   finesand
14 saturated   finesand
15 saturated coarseSand
16     moist      clays
17     moist      clays
18     moist   finesand
19     moist   finesand
20    barren      clays
```

Figure 7:: R output showing the simulated state

From the outputs, one sees 2,3,8,9,10,13,14,15 as saturated with water, in spite of the varying lithology. Assuming each saturated interface has 1.5 cubic meter per hour discharge capability, these eight logs are capable of transmitting the following discharge into the well:

Saturated Zones Output

=1.50 x8

=12.0

=12.0 cubic meters per hour yield

If one then adds to the moist zones, assuming each moist zone will provide 0.5 cubics per hour, in the simulation model, one sees that the well will have more than 15 cubic meters per hour, assuming one was to stop drilling at the 40m depth mark.

Moist Zones Output

0.5x 7

=3.5

=3.5 cubic meters per hour

Total of the two aquifer states will give 15.5 cubics per hour. The truth is that a 37 m well in Young Muslim Sec School in this aquifer zone has been giving 20 meter cubics. Jarerod Centre well on the way towards Korakora was done to 43 meters and gives 21 cubics per hour. The computations above are thus conservative estimates helpful for the supervising geologist in knowing the depths at which he ought to stop, depending on how much water he needs from the well.

```
>
> print(predState)
     Element     State
1      clays     moist
2 coarseSand     moist
3   finesand     moist
4      clays     moist
5   finesand     moist
6 coarseSand saturated
7      clays saturated
8 coarseSand saturated
```

Figure 8: The R Output showing the Predicted State

Application in real Life No 2:

We are now confronted with a situation whereby the geologist already has the drill cuttings in the course of supervision as opposed to the simulations he did in the first instance. Here is the table of the geolog assemblage he has so far, and he wants to know the aquifer status of the same:

Table 8: Table of Geolog Assemblage

Geolog name	Predicted status for water availability	Depth
clays	unknown	2
Coarse sand	unknown	4
Finesand	unknown	6
clays	unknown	8
Finesand	unknown	10
Coarse sand	unknown	12
clays	unknown	14
Coarse sand	unknown	16
Coarse sand	unknown	18
Finesand	Unknown	20

When we run the following code, it will predict, using the Viterbi algorithm, the sequence of aquifer productivity consistent with the log assemblage input inserted in the codes:

```
> # TRIAL NUMBER ONE
> observations <- c("clays","coarseSand","finesand","clays","finesand","coar$
+ "clays","coarseSand","coarseSand","finesand")
> set.seed(123)
> stateViterbi <- viterbi(hmm, observations)
> predState <- data.frame(Element=observations, State=stateViterbi)
> print(predState)
```

Figure 9: Viterbi Algorithm

The output will be as hereunder:

```
      Element     State
1       clays     moist
2  coarseSand     moist
3    finesand     moist
4       clays     moist
5    finesand     moist
6  coarseSand saturated
7       clays saturated
8  coarseSand saturated
9  coarseSand saturated
10   finesand saturated
```

Figure 10: The Predicted output

final table will thus be as thus:

Table 9: Table of Geolog and the Predicted Status for Water Availability

Geolog name	Predicted status for water availability	Depth
clays	**moist**	2
Coarse sand	**moist**	4
Finesand	**moist**	6
clays	**moist**	8
finesand	**moist**	10
Coarse sand	**saturated**	12
clays	**saturated**	14
Coarse sand	**saturated**	16
Coarse sand	**saturated**	18
Finesand	**saturated**	20

From the foregoing, the supervising geologist is assured of water in the final 10m of drilling, and that is as from 11.0m to 20.0m. therefore, before mud drilling ends in, say, 50m or so, the geologist is already assured of saturated zones which he is assured of getting aquifer water from, even as the drilling progresses.
The HMM is thus an efficient supervision tool if developed for a known aquifer suite.

Assessment of Geology and State of Aquifer Using Fuzzy Logic Gaussian Membership Functions in R.

Cases may arise whereby we have the dataset comprising geologs and allied variables as thus:

Table 10: Geolog Table

	A	B	C	D	E	F	G	H	I	J	K
1	longtd	lattd	elev	logDepth	geology		longtd	lattd	elev	logDepth	geology
2	610946	9861112	76	2	1		536648	9983887	178.0023	6	2
3	610946	9861112	76	4	2		536648	9983887	178.0023	8	2
4	610946	9861112	76	6	3		536648	9983887	178.0023	10	2
5	610946	9861112	76	8	1		536648	9983887	178.0023	12	2
6	610946	9861112	76	10	1		536648	9983887	178.0023	14	2
7	610946	9861112	76	12	2		536648	9983887	178.0023	16	1
8	610946	9861112	76	14	2		536648	9983887	178.0023	18	1
9	610946	9861112	76	16	2		536648	9983887	178.0023	20	3
10	610946	9861112	76	18	1		536648	9983887	178.0023	22	3
11	610946	9861112	76	20	2		536648	9983887	178.0023	24	2
12	610946	9861112	76	22	2		536648	9983887	178.0023	26	3
13	610946	9861112	76	24	2						
14	610946	9861112	76	26	2						
15	610946	9861112	76	28	2						
16	610946	9861112	76	30	3						
17											

The present study is one such case whereby we **have excel sheet 508 rows of geolog** data defining the various logging depths and respective longitudes and latitudes:

Beginning from here:

	A	B	C	D	E
1	longtd	lattd	elev	logDepth	geology
2	610946	9861112	76	2	1
3	610946	9861112	76	4	2
4	610946	9861112	76	6	3
5	610946	9861112	76	8	1
6	610946	9861112	76	10	1
7	610946	9861112	76	12	2
8	610946	9861112	76	14	2
9	610946	9861112	76	16	2
10	610946	9861112	76	18	1
11	610946	9861112	76	20	2
12	610946	9861112	76	22	2
13	610946	9861112	76	24	2
14	610946	9861112	76	26	2
15	610946	9861112	76	28	2
16	610946	9861112	76	30	3
17					

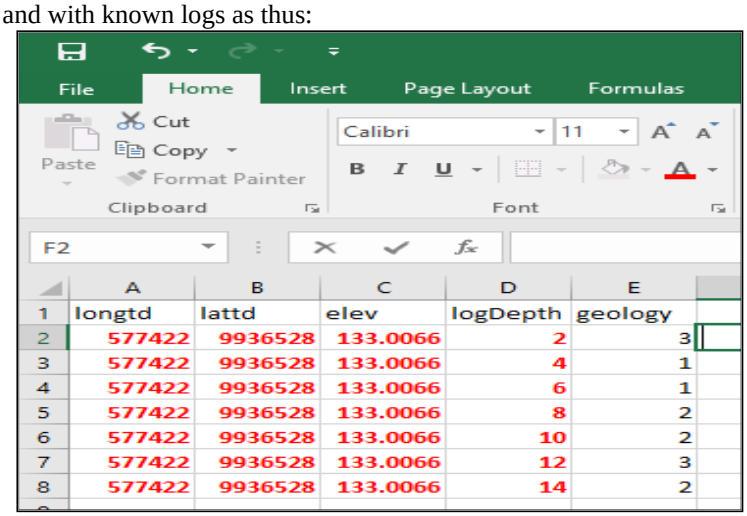
All the way to here:

499	longtd	lattd	elev	logDepth	geology
500	536502	9983924	180.0045	28	2
501	536502	9983924	180.0045	30	3
502	536502	9983924	180.0045	32	1
503	536502	9983924	180.0045	34	1
504	536502	9983924	180.0045	36	2
505	536502	9983924	180.0045	38	3
506	536502	9983924	180.0045	42	1
507	536502	9983924	180.0045	44	2
508	536502	9983924	180.0045	46	2
509	536502	9983924	180.0045	46	2

Suppose we are undertaking a drilling at a place with known longitudes and latitudes as well as sample drill cuttings as well as the respective geologging depths expected. Do we have a way of estimating the geology of the material at respective depths expected to be sampled, way even before the drill cutting has been attained? This is useful in three ways:

a) During the planning phase for drilling
b) During the actual drilling, wherein the drilling supervisor is able to estimate/ predict the lithology of the geologic material expected
c) One is able to estimate how many slotted and plain casings may be purchased in advance before drilling

Fuzzy logic predictive power may be leveraged here, with the test data being estimated. First we have the know details of a well which was mapped and drilled, and with known logs as thus:



	A	B	C	D	E
1	longtd	lattd	elev	logDepth	geology
2	577422	9936528	133.0066	2	3
3	577422	9936528	133.0066	4	1
4	577422	9936528	133.0066	6	1
5	577422	9936528	133.0066	8	2
6	577422	9936528	133.0066	10	2
7	577422	9936528	133.0066	12	3
8	577422	9936528	133.0066	14	2

In the table above, 1.0 represents clays and carbonate based lithologic logs, 2 represents fine grained sandstones whereas 3.0 represents the coarse-grained sandstones.

Now comes the part the geologist is assigned as a real application test of fuzzy logic. He has a table of data being drilled and is unsure of the lithologic sequence expected in the next ten samples, with intervals of 2.0m each. See the table below:

	longtd	lattd	elev	logDepth
2	578955	9937507	159	84
3	578955	9937507	159	86
4	578955	9937507	159	88
5	578955	9937507	159	90
6	578955	9937507	159	92
7	578955	9937507	159	94
8	578955	9937507	159	96
9	578955	9937507	159	98
10	578955	9937507	159	100
11	578955	9937507	159	102

The drilling of the well at this location has proceeded up to 84m so far, but depths as from 0-84m have logs which are non-promising in terms of giving more water. The geologist wishes to know whether he has any chance of striking any more water down the stratigraphic column as from 84m-102m bgl, so that he makes a decision on whether or not to proceed with the drilling.

Now he has to take the old data with known geolog class in the Tana Alluvial aquifer, and combine it with this site with as yet unknown geologs, which he wishes to predict. This is because the known site has coordinates that have already been trained or tested with the neuro-fuzzy model in the past, to generate a stochastic mapping tool. The new data-frame will look like this:

	longtd	lattd	elev	logDepth
2	578955	9937507	159	84
3	578955	9937507	159	86
4	578955	9937507	159	88
5	578955	9937507	159	90
6	578955	9937507	159	92
7	578955	9937507	159	94
8	578955	9937507	159	96
9	578955	9937507	159	98
10	578955	9937507	159	100
11	578955	9937507	159	102
12	577422	9936528	133.0066	2
13	577422	9936528	133.0066	4
14	577422	9936528	133.0066	6
15	577422	9936528	133.0066	8
16	577422	9936528	133.0066	10
17	577422	9936528	133.0066	12
18	577422	9936528	133.0066	14

Figure 11: New Test Data Frame with both the new dataset in rows 1 to 10, and also the old data with known logs from rows 11 to 17

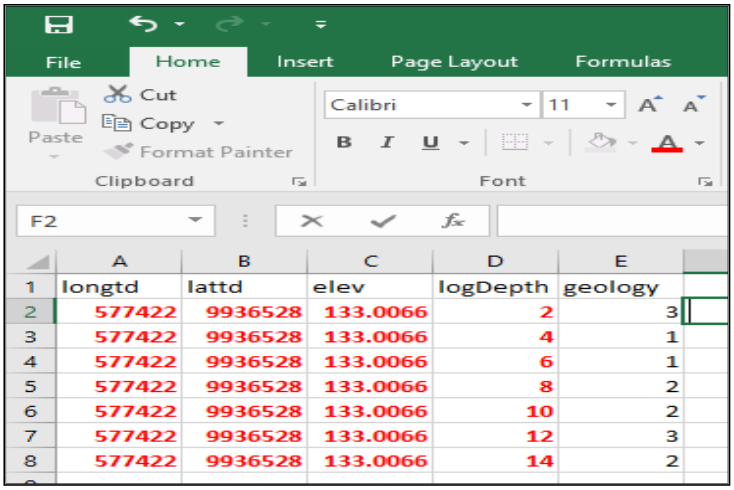
The prediction runs with GAUSSIAM mf in R using 135 labels of fuzzification labels generate a prediction as follows:

```
 [1,] 3.000000
 [2,] 1.000000
 [3,] 2.000000
 [4,] 3.000000
 [5,] 1.000017
 [6,] 1.000000
 [7,] 2.000000
 [8,] 2.000000
 [9,] 1.000000
[10,] 3.000000
[11,] 3.000000
[12,] 1.000000
[13,] 1.000000
[14,] 2.000000
[15,] 2.000000
[16,] 3.000000
[17,] 2.000000
>
```

Now pick values in Rows number 11 to 17 and note down the values for comparison with the old data table with known logs.

```
[11,] 3.000000
[12,] 1.000000
[13,] 1.000000
[14,] 2.000000
[15,] 2.000000
[16,] 3.000000
[17,] 2.000000
>
```

Read the geology column in the excel screenshot. It is observed that it is a hundred percent prediction accuracy for the geology class of the old known well. If this is so, then this same model may be now used to correctly predict the true lithology of the unknown rows from 1 to 10. Therefore, the unknown rows predicted are also presumed to be correct.



	A	B	C	D	E
1	longtd	lattd	elev	logDepth	geology
2	577422	9936528	133.0066	2	3
3	577422	9936528	133.0066	4	1
4	577422	9936528	133.0066	6	1
5	577422	9936528	133.0066	8	2
6	577422	9936528	133.0066	10	2
7	577422	9936528	133.0066	12	3
8	577422	9936528	133.0066	14	2

The expected class prediction of lithologic logs are thus as follows:

```
 [1,] 3.000000
 [2,] 1.000000
 [3,] 2.000000
 [4,] 3.000000
 [5,] 1.000017
 [6,] 1.000000
 [7,] 2.000000
 [8,] 2.000000
 [9,] 1.000000
[10,] 3.000000
```

The geologist thus expected coarse grained sandstones which are synonymous with aquifers that are productive in 84m, 90m, and at 100m. The advice he would give is that the driller continues with drilling to 102m.

	kumbi				
3	log-Inter	logDepth	geology	state	category1	category2
4	0.0-2.0	2	finesands	barren	1	1
5	2.0-4.0	4	coarse sands	moist	2	2
6	4-6.0	6	clays/carbonate	saturated	3	3
7	6.0-8.0	8	finesands	saturated	1	3
8	8.0-10.0	10	finesands	saturated	1	3
9	10.0-12.(	12	coarse sands	saturated	2	3
10	12.0-14.(	14	coarse sands	saturated	2	3
11	14.0-16.(	16	coarse sands	moist	2	2
12	16.0-18.(	18	finesands	moist	1	2
13	18.0-20.(	20	coarse sands	moist	2	2
14	20.0-22.(	22	coarse sands	saturated	2	3
15	22.0-24.(	24	coarse sands	saturated	2	3
16	24.0-26.(	26	coarse sands	saturated	2	3
17	26.0-28.(	28	coarse sands	saturated	2	3
18	28.0-30.(	30	clays/carbonate	saturated	3	3
19	30.0-32.(	32	clays/carbonate	barren	3	1
20	32.0-34.(	34	clays/carbonate	barren	3	1
21	34.0-36.(	36	finesands	barren	1	1
22	36.0-38.(	38	coarse sands	barren	2	1
23	40.0-42.(	42	coarse sands	saturated	2	3
24	42.0-44.(	44	coarse sands	moist	2	2
25	44.0-46.(	46	coarse sands	saturated	2	3
26	46.0-48.(	48	coarse sands	saturated	2	3
27	48.0-50.(	50	clays/carbonate	saturated	3	3
28	50.0-52.(	52	clays/carbonate	barren	3	1
29	52.0-54.(	54	clays/carbonate	barren	3	1
30	54.0-56.(	56	clays/carbonate	moist	3	2
31	56.0-58.(	58	finesands	moist	1	2
32	58.0-60.(	60	coarse sands	moist	2	2
33	60.0-62.(	62	coarse sands	saturated	2	3
34	62.0-64.(	64	coarse sands	saturated	2	3

Figure 12: Sample Extract of the datasheet with drill cutting geolog as analyzed by The geologist who supervised the drilling in the study area

VIII. CONCLUSION AND RECOMMENDATIONS

The study avails a new tool for hydrogeologists and water engineers alike, for decision making under clouds of uncertainty, namely, simulation of number of layers with promise of aquifer water, as well as the prediction of identity of the geologic layer units before actual drilling takes place at a site earmarked for drilling, using algebraic **Neuro-Fuzzy Modeling and the Hidden Markov Models**. THE FOLLOWING ARE FURTHER ADVANTAGES:

i) The simulations will be weighted against known discharge values for specific lithologies and estimation of discharge appropriately assigned, thus. As an example, coarse grained sandstone aquifers have a value of 1 to 1.5 cubic meters per hour discharge. A simulation of six coarse grained sandstones layers/2m-interval logs is an indirect estimate of at least 9 cubic meters per hour, for a well being drilled but as yet un-developed and un-tested.

ii) It is a tool which may be used to help avoid wastages in drilling of dry aquifers, by abandoning a proposed drilling site of a well, if the number of productive layers simulated shows limited promise of viable aquifers, saving the aquifer from unnecessary depletion, via having so many wells located near each other, on a trial-and-error basis, which is currently taking place, within the Tana Alluvial Aquifer.

iii) Noting that the HMM helps determine probable identity of lithologic logs, it is a useful tool in helping estimate the number of slotted and plain casings required for drilling.

REFERENCES

[1] Azimi, S., Hassannayebi, E., Boroun, M., & Tahmoures, M. (2020). Probabilistic analysis of long-term climate drought using steady-state Markov chain approach. *Water Resources Management*, *34*(15), 4703-4724.

[2] Baalousha, H. M., Tawabini, B., & Seers, T. D. (2021). Fuzzy or Non-Fuzzy? A comparison between fuzzy logic-based vulnerability mapping and DRASTIC approach using a numerical model. A Case Study from Qatar. *Water*, *13*(9), 1288.

[3] Chaglla Aguagallo, D. K. (2021). *Forecasting groundwater level recession patterns through ARIMA and Hidden Markov Model: A comparative study* (Bachelor's thesis, Universidad de Investigación de Tecnología Experimental Yachay).

[4] Danandeh Mehr, A., Tur, R., Çalışkan, C., & Tas, E. (2020). A novel fuzzy random forest model for meteorological drought classification and prediction in ungauged catchments. *Pure and Applied Geophysics*, *177*(12), 5993-6006.

[5] Hameed, I. A. (2011). Using Gaussian membership functions for improving the reliability and robustness of students' evaluation systems. *Expert systems with Applications*, *38*(6), 7135-7142.

[6] Jha, M. K., Shekhar, A., & Jenifer, M. A. (2020). Assessing groundwater quality for drinking water supply using hybrid fuzzy-GIS-based water quality index. *Water Research*, *179*, 115867.

[7] Khadr, M. (2016). Forecasting of meteorological drought using Hidden Markov Model (case study: The upper Blue Nile river basin, Ethiopia). *Ain Shams Engineering Journal*, *7*(1), 47-56.

[8] Kumar, A., & Tripathi, V. K. (2019). Adaptive neuro fuzzy inference system for runoff modelling—A case study. *Int. J. Curr. Microbiol. App. Sci*, *8*(4), 2054-2061.

[9] Piho, L., & Kruusmaa, M. (2021). Subsurface Flow Path Modeling from Inertial Measurement Unit Sensor Data Using Infinite Hidden Markov Models. *IEEE Sensors Journal*, *22*(1), 621-630.

[10] Roy, D. K., Barzegar, R., Quilty, J., & Adamowski, J. (2020). Using ensembles of adaptive neuro-fuzzy inference system and optimization algorithms to predict reference evapotranspiration in subtropical climatic zones. *Journal of Hydrology*, *591*, 125509.

[11] Sihag, P. (2018). Prediction of unsaturated hydraulic conductivity using fuzzy logic and artificial neural network. *Modeling Earth Systems and Environment*, *4*(1), 189-198.

[12] Singh, A. P., Dhadse, K., & Ahalawat, J. (2019). Managing water quality of a river using an integrated geographically weighted regression technique with fuzzy decision-making model. *Environmental Monitoring and Assessment*, *191*(6), 1-17.

[13] Teixeira Parente, M., Bittner, D., Mattis, S. A., Chiogna, G., & Wohlmuth, B. (2019). Bayesian calibration and sensitivity analysis for a karst aquifer model using active subspaces. *Water Resources Research*, *55*(8), 7086-7107.

[14] Wang, P., Chen**, X., Wang, B., Li, J., & Dai, H. (2020). An improved** method for lithology identification based on a hidden Markov model and random forests. *Geophysics*, *85*(6), IM27-IM36.

[15] Xing, Z., Qu, R., Zhao, Y., Fu, Q., Ji, Y., & Lu, W. (2019). Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model. *Journal of Hydrology*, *572*, 501-516.

[16] Yang, K., Chen, F., He, C., Zhang, Z., & Long, A. (2020). Fuzzy risk analysis of dam overtopping from snowmelt floods in the nonstationarity case of the Manas River catchment, China. *Natural Hazards*, *104*(1), 27-49.

[17] Zhao, W., Shi, T., & Wang, L. (2020). Fault diagnosis and prognosis of bearing based on hidden Markov model with multi-features. *Applied Mathematics and Nonlinear Sciences*, *5*(1), 71-84.