Annals of Optimization Theory and Practice

Volume 3, Number 1, 37-46

October 2020

DOI: 10.22121/aotp.2020.243381.1039

Assessment of probability distributions of groundwater quality data in Gwale area, north-western Nigeria

Suleman Ahamad Abubakkar[1]. Singh Vijay Vi[2]✉. Ibrahim Auwalu[3]. Abdullahi Usman Aliyu[4]. Suleiman Suleiman Abubakr[5]

[1] Department of Statistics Kano University, of Science & Technology, Wudil, Kano, Nigeria

[2] Department of Mathematics Yusuf Maitama Sule University, Kano, Nigeria. Formally Northwest University, Kano, Nigeria

[3] Department of Statistics Kano University, of Science& Technology, Wudil, Kano, Nigeria

[4] Department of Statistics Kano University, of Science& Technology, Wudil, Kano, Nigeria

[5] Kano State Ministry of Water Resources, Nigeria

✉ singh_vijayvir@yahoo.com

(Received: **A**ugust 11, 2020 / Accepted: October 8, 2020)

Abstract Groundwater quality plays an important role in human, animal and plant health. Measurements of water quality are random variables which need a probabilistic model. The interest in fitting suitable probability distributions in modeling water quality data remains secure in hydrology and engineering. This paper designed to find the best fitting probability distribution of some calcium concentrations of groundwater collected from 28 sampling sites in Gwale area, Kano state, North-western Nigeria. The parameter estimated for groundwater analyzed using gamma, logistic, lognormal, normal, and Weibull distributions. Best suited distribution selected using the log-likelihood function and the Kolmogorov-Smirnov test, which are the most extensively used goodness-of - fit test measurements. Best fitted distribution has been selected using the log-likelihood function and Kolmogorov-Smirnov test, widely used measures of the goodness-of-fit test. The result indicates that the logistic distribution with the highest log-likelihood value determined to fit the calcium concentrations of groundwater data in the sample area as the most appropriate probability model. Usefulness of probability distribution in the modeling of groundwater quality data and can be used to describe groundwater quality data at any other location. Future reference probability distribution model for calcium concentrations of groundwater data in the study area was also provided.

Keywords Distribution; Groundwater; Logistic; Modeling; Probability

1. Introduction

One of the latest advances in the mathematical analysis of environmental studies is the application of the probability distribution model. Mathematical models are often used as decision support tools to evaluate contamination in groundwater (Cecilia *et al.* 2020). Developing new distribution models certainly helps to easily analyze and interpret the spatial and temporal variability of hydro-morphological as well as groundwater physico-chemical content. It can serve as the reference model for future investigations (Kishore *et al.* 2011).

Measurements of the water quality are random variables that require probabilistic and statistical estimation. Water quality data do not obey the convenient distribution of probability, such as the well-known normal and lognormal distributions on which many conventional statistical methods are based (Lee *et al.* 2001). Unfortunately, preceding studies were insufficient to conclude the suitability of the correct probability distribution functions for modeling water data (Surendran and Tota, 2015). Many efforts have been made to study water quality and rainfall patterns using a probabilistic method including probability plots and frequency distributions. The probability plots are used to determine how well a theoretical distribution describes the sample data. This can be achieved by comparing the curves of the measured values to the theoretical distribution density curves. Conventionally, researchers apply one or more goodness-of-fit tests to see how well a selected distribution fits the dataset (Tahir *et al.* 2016). The choice of distribution to represent any system or quality is verified using the available data (Hahn and Shapiro, 1967).

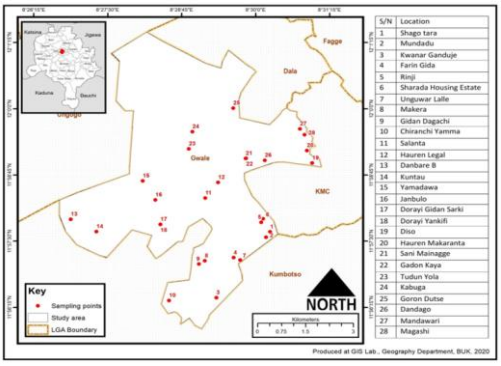
In their study, (Maryam *et al.* 2018) used frequency analysis to predict the likelihood of the most important parameters deteriorating the groundwater quality of the Alashtar region, Iran. It has been revealed that probability values of water quality parameters are plotted for expected return periods to show which location is more sensitive in groundwater quality. To identify the best fit model, six probability distribution models were applied to the Karkheh River annual rainfall data (Machekposhti and Sedghi, 2019). They used the best-fitted distributions to estimate the expected values of rainfall at the rainfall gauging stations. According to (Nwaiwu and Bitrus, 2005), four continuous distributions of probability such as normal, lognormal, gamma, and Weibull distributions were applied to fit the water quality data collected from the water treatment plant in Maiduguri, Nigeria. The analysis of the probability distribution model can be used effectively to extract knowledge about potential environmental effects on water quality and also useful to classify natural groupings within the dataset. Due to uncertainties, these methods are important to avoid misinterpretation of data concerning environmental monitoring. Here, the best-fitted distributions are determined using log-likelihood, AIC, BIC, and K-S tests of goodness-of-fit.

The object of this paper is to evaluate the suitable model of the probability distribution that best models the calcium concentrations of groundwater data obtained from open wells and hand pumps in Gwale local government, Kano state, Nigeria. A total of five probability distributions are applied to the dataset and the log-likelihood, AIC, BIC, and

Kolmogorov-Smirnov measures are used to determine the best fit probability distribution model. The graph of the dataset and estimated probability distribution functions of the competitive fitted models are displayed to see how well the models provide better fits to the dataset graphically.

Study Area

Gwale is a local government area within Kano city, Nigeria. Its headquarters are in the suburb of Gwale. It is situated at a latitude$11^{\circ}58'N$and longitude $8^{\circ}30'E$. The study area has an area of $18km^2$due to the climatic condition of the study area, seven months dry and five months rainy seasons, and the fact that the whole study area lies on the basement rock complex. The large amounts of water required for household and industrial consumptions in the study area depend heavily on groundwater through tube wells, open wells, and hand pumps. The sampling locations and geologic setting of the study area are indicated in Figure. 1 and Figure. 2, respectively.



S/N	Location
1	Shago tara
2	Mundadu
3	Kwanar Ganduje
4	Farin Gida
5	Rinji
6	Sharada Housing Estate
7	Unguwar Lalle
8	Makera
9	Gidan Dagachi
10	Chiranchi Yamma
11	Salanta
12	Hauren Legal
13	Danbare B
14	Kuntau
15	Yamadawa
16	Janbulo
17	Dorayi Gidan Sarki
18	Dorayi Yankifi
19	Diso
20	Hauren Makaranta
21	Sani Mainagge
22	Gadon Kaya
23	Tudun Yola
24	Kabuga
25	Goron Dutse
26	Dandago
27	Mandawari
28	Magashi

Figure 1. Location map of the study area.

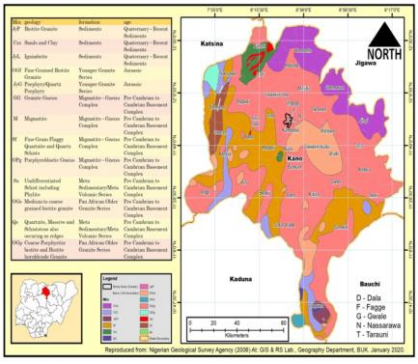


Figure 2. Map of Kano state showing geologic setting of the study area.

Geology and Hydrology

The study area is underlined by complex rocks in the basement, which underwent a long period of deep clay regolith. The rocks are pre-Cambrian origin and consist of metamorphic and old granites with intruded migmatite, gneiss, and phyllite. Besides, weathering has resulted in a formation group (usually poor aquifers) except where the weathering or fracturing is intensive. The aquifer of the area is a weathered and fractured rock in which groundwater exists under stable conditions. The water table lays at a depth generally less than 20 meters, and the maximum of boreholes rarely exceed 60 meters (Mohammed, 1984). Two hydrological can be identified in the region. The rivers are located in the upland areas, which comprise river Kano and river Challawa. The area received rainfall of over 800mm annually. The temperature varies by warm to hot seasons between November and February. Annual mean temperature ranges from about 21 degrees Celsius to 27 degrees Celsius (Bala *et al*. 2011). The least humidity value is recorded between January and February, while the highest between July and September.

Humidity at 16:00 hours varies from about 20 to 60%, and the highest values are recorded in July and August.

2. Materials and Methods

Water Sampling

Calcium concentrations have been collected from 28 sampling stations in Gwale area in July 2018. The locations were selected at random to ensure equal treatment for representativeness in the area of study. Using the geographical position system, the latitude and longitude of the selected locations were determined on a map. After determining the locations, groundwater samples are taken from hand pumps and open wells, then poured in cleaned plastic containers and put in an iced box before taken to the laboratory according to the standard method (Standard Methods, 2005). The measurements of calcium concentrations expressed in milligram per liter have been analyzed in the federal ministry of water resources laboratory, Kano state, Nigeria.

Probability Distributions of Groundwater Samples

The phenomena whose outcomes cannot be exactly predicted with certainty are termed as uncertain. Events like water quality concentration are uncertainty exists due to many factors, such as social, demographic, political, and hydrological processes (Loucks and Van Beek, 2017). Such events can be adequately characterized by developing appropriate probability distribution models that deal with uncertainties. The developed models can serve as a future reference guide in monitoring water quality data in any study area of interest. Henceforth, the groundwater concentration as a random variable is denoted by letter x in this paper and has been defined in the following probability distribution functions:

The Gamma Distribution

A random variable x has the gamma distribution with two parameters α and φ if its probability distribution function (pdf) is given by

$$f(x) = \frac{1}{\Gamma(\alpha)\varphi^{\alpha}} x^{\alpha-1} e^{\frac{-x}{\varphi}} \qquad 0 \le x \le \infty \qquad (1)$$

where the parameter α is the shape of the distribution and φ is its scale parameter.

The Logistic Distribution

A random variable x has the logistic distribution with two parameters μ and λ if its pdf is given by

$$f(x) = \frac{e^{(x-\mu)/2}}{\lambda(1 + e^{(x-\mu)/2})^2} \qquad -\infty \le x \le \infty \qquad (2)$$

where $\lambda\,\mu$ and are the location and scale parameters of the logistic distribution respectively.

Lognormal Distribution

A random variable x follows the lognormal distribution with two parameters μ and σ if its pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi x^2\sigma^2}} e^{\frac{-1}{2}(\frac{lnx-\mu}{\sigma})^2} \qquad 0 \le x \le \infty \qquad (3)$$

The parameters μ,σ and are the mean and standard deviation of the lognormal distribution, respectively.

Normal Distribution

A random variable X is said to have a normal distribution with two parameters μ and σ if its pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty \le x \le \infty \qquad (4)$$

where the parameter μ is the mean of the normal distribution and σ is its standard deviation.

Weibull Distribution

A random variable X is said to follow a Weibull distribution with two parameters α and β if its probability distribution function (pdf) is given by

$$f(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1} e^{-(\frac{x}{\beta})^{\alpha}} \qquad 0 \le x \le \infty \qquad (5)$$

where α $\beta$and are the shape and scale parameters of the Weibull distribution respectively.

Probability Distribution Tests

These are tests of the hypothesis that assess whether sample data is taken from a population following a predicted distribution of probability. A model is selected to fit a probability distribution on a dataset. Model selection depends on various statistical tests or the experience of observed data. After the model selection, the parameters of the selected distributions are estimated based on the method of parameter estimation. Then the candidate model's performance on the dataset is measured and evaluated by one or more fitness-of-fit tests (Maryam *et al.* 2018). Hence, the selected fitted distribution can reasonably be used as a distributional model for the given dataset.

In the present study, a total of five probability distributions such as Gamma, Logistic, Lognormal, Normal, and Weibull are applied to the calcium concentration measurements to identify the best fit of the probability distribution model(s) using log-likelihood and

Kolmogorov-Smirnov (K-S) tests. The calcium concentrations are measured in the laboratory from the 28 collected groundwater samples in Gwale area of north-western Nigeria. In each case, the parameters of the model are determined by the maximum likelihood test. The measures of goodness-of-fit are computed to compare the fitted model and identify the best fits. The log-likelihood is the value of the calculated log-likelihood function using parameter estimation of maximum likelihood. In this case, the model with the higher log-likelihood values and smaller AIC and BIC values for each selected fitting distribution is known as the best-suited model. Moreover, p-values for each model are computed from the K-S statistic to suggest if the candidate distribution could describe the dataset. The p-values less than 0.05 would indicate that calcium a dataset does not come from the selected distribution with 95% confidence. The statistical measures log-likelihood, AIC, BIC, and K-S are described in details in (Chen and Balakrishnan, 1995). All computations in this study were performed using stat graphics software.

3. Results, Interpretations, and Discussions

The calcium concentrations of the groundwater data are subjected to five selected continuous distributions to identify the best fitting distributions using statistical measures such as log-likelihood, AIC, BIC, and K-S tests. The estimated parameters of the fitted distributions, the log-likelihood, AIC, BIC values, K-S, and p-values are listed in Table 1.

According to results in Table 1, the maximum value of the log-likelihood corresponds to the fitted logistic model among the fitted Gamma, Lognormal, Normal, and Weibull competitive models. Hence, the logistic model gives a suitable fit for the calcium concentration data and, therefore, can be identified as the best-fitted model. However, the best-fitted distribution can be assessed visually by using a frequency histogram and Q-Q plot. The plots of the fitted Gamma, Logistic, Lognormal, Normal, and Weibull density functions are displayed, as shown in Figure. 3 through Figure. 9. These plots have shown that the logistic model closely fits the dataset compared with Gamma, Lognormal, Normal, and Weibull models. Therefore, it can be observed that visual inspection confirms the results given by the numerical measures.

Table 1. Estimated parameters, Log-likelihood, and K-S P-values for calcium concentration data and competing distributions.

Distribution	Gamma	Logistic	Lognormal	Normal	Weibull
Parameter	Shape=9.2134	Mean=15.0597	Mean=2.6532	Mean=15.006	Shape=4.2736
Estimates	Scale=0.6140	Std.dev=4.0270	Std.dev=0.3970	Std.dev=4.063	Scale=16.4138
Log-likelihood	-89.3973	-83.7030	-93.9521	-84.124	-84.6004
AIC	306.3619	297.3786	308.6167	297.6089	306.1793
BIC	309.0263	300.0430	311.2811	300.2733	308.8438
p-value	0.3074	0.6791	0.1162	0.8496	0.8464

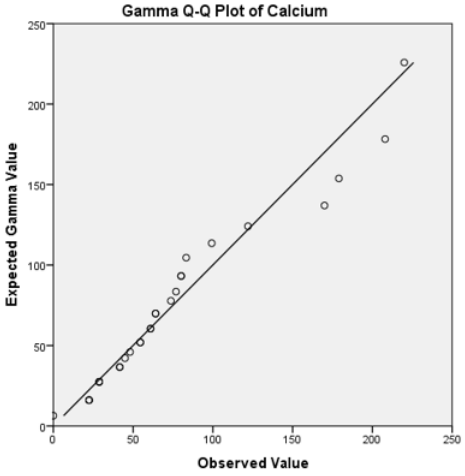


Figure 3. Q-Q plot of calcium concentration and the fitted gamma distribution.

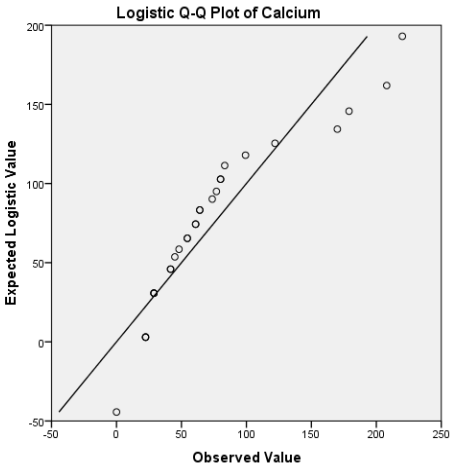


Figure 4. Q-Q plot of calcium concentration and the fitted logistic distribution.

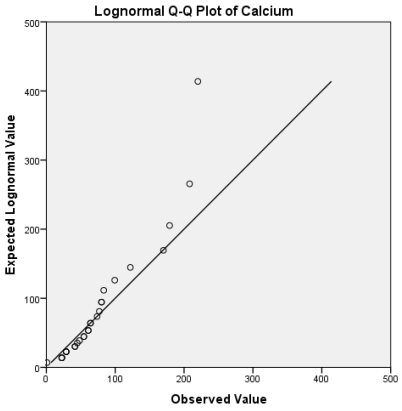


Figure 5. Q-Q plot of calcium concentration and the fitted lognormal distribution.

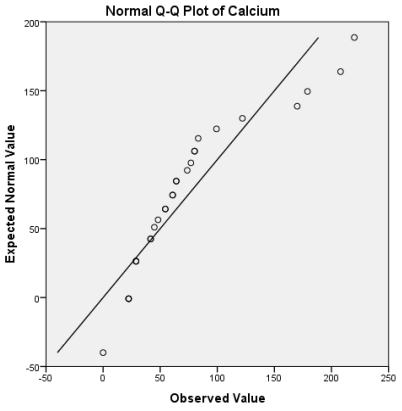


Figure 6. Q-Q plot of calcium concentration and the fitted normal distribution.

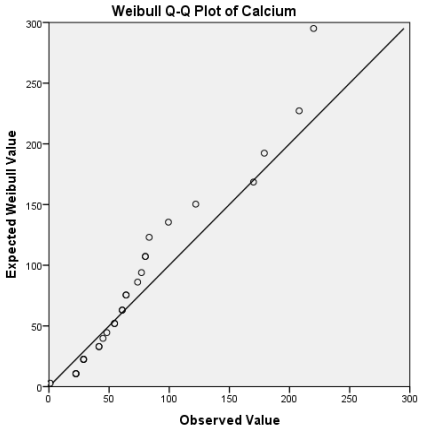


Figure 7. Q-Q plot of calcium concentration and the fitted Weibull distribution.

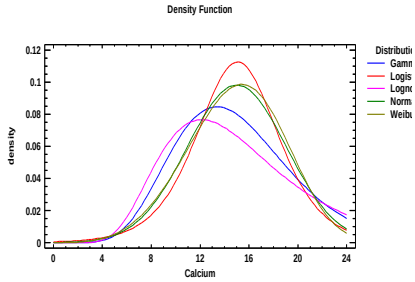


Figure 8. Probability densities of the five fitted distributions.

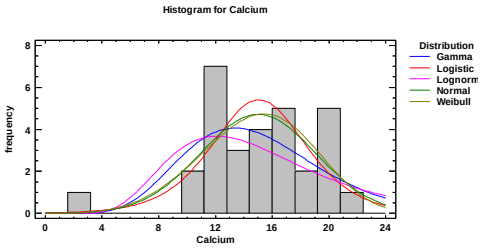


Figure 9. Plot of calcium concentrations and the five fitted distributions.

4. Conclusion

Distribution Fitting tests are hypothesis tests that assess whether sample data were drawn from a population following hypothesized probability distributions. Typically, one or more goodness-of-fit tests are applied to actual datasets to determine the efficiency of the candidate fitting distributions based on the results of these statistical measures, and the visual graph can also be used to verify the results of the models selected. In this paper, a total of five probability distributions, Gamma, Logistic, Lognormal, Normal, and Weibull, are applied to calcium concentration data to identify the best-fitted distribution using log-likelihood, AIC, BIC, K-S tests, and visual inspection. According to the results of these statistics and visual inspection, the logistic distribution model has been identified as the best-fitted model than other competing models for the dataset. Hopefully, the use of probability distribution models can draw broader applications in groundwater data fitting.

5. Acknowledgments

The project was fully funded by the Tertiary Education Trust Fund (TETFund) through the Institutional Based Research (IBR) intervention grant 2018. Also, the authors offer their sincere thanks to the editorial team and referee of this journal for their valuable contributions in this paper.

References

1. Bala, A. E., Eduvie, O. M., & Byami, J. (2011). Borehole depth and regolith aquifer hydraulic characteristics of bedrock types in Kano area, Northern Nigeria. *African Journal of Environmental Science and Technology*, *5*(3), 228-237.
2. Chen, G., & Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, *27*(2), 154-161.
3. Hahn Gerrald, J., & Shapiro Samuel, S. (1967). Statistical models in engineering. John Willey & Sons. *Inc. New York.–London–Sydney*, 395.
4. Kishore, K. (2011). Das, Bhanita Das, Bhupen K. Baruah. and Abani K. Misra, Development of New Probability Model with Application in Drinking Water Quality Data. *Adv. Appl. Sci. Res*, *2*(4), 306-313.

5. la Cecilia, D., Porta, G. M., Tang, F. H., Riva, M., & Maggi, F. (2020). Probabilistic indicators for soil and groundwater contamination risk assessment. *Ecological Indicators*, *115*, 106424.
6. Lee, J. Y., Cheon, J. Y., Lee, K. K., Lee, S. Y., & Lee, M. H. (2001). Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons. *Journal of environmental quality*, *30*(5), 1548-1563.
7. Loucks, D. P., & Van Beek, E. (2017). An Introduction to Probability, Statistics, and Uncertainty. In *Water Resource Systems Planning and Management* (pp. 213-300). Springer, Cham.
8. Machekposhti, K. H., & Sedghi, H. (2019). Determination of the Best Fit Probability Distribution for Annual Rainfall in Karkheh River at Iran. *International Journal of Environmental and Ecological Engineering*, *13*(2), 69-75.
9. Maryam, G., Kaveh, O.A., Saed, E., and Vijay, P.S. (2018). Analyzing the groundwater quality parameters using frequency analysis. American Journal of Engineering and Applied Sciences, 11(2), 482-490, doi.org/10.3844/ajeassp.2018.482.490.
10. Mohammed, I. (1984). Hydraulic properties of the Basement Complex and Chad Formation aquifers of Kano State based on test-pumping of selected boreholes. *Unpublished M. Sc. thesis Department of Geology, Ahmadu Bello University, Zaria*.
11. Nwaiwu, E. N., & Bitrus, A. (2005). Fitting probability distributions to component water quality data from a treatment plant. *Global Journal of Environmental Sciences*, *4*(2), 151-154.
12. Standard Methods for the Examination of Water and wastewater, (2005): *21st edn, American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA.*
13. Surendran, S., & Tota-Maharaj, K. (2015). Log logistic distribution to model water demand data. *Procedia Engineering*, *119*(1), 798-802.
14. Tahir, M.H., Adnan, M.H., Cordeiro, G.M., Hamedani, G.G., Mansoor, M., and Zubair, M. (2016): "The Gumbel-Lomex Distribution: Properties and Applications." *Journal of Statistical Theory and Applications. Atlantic Press, Volume 15-1, pp. 61-79*.