Scientific Review Engineering and Environmental Sciences (2023), 32 (2), 186–197

Sci. Rev. Eng. Env. Sci. (2023), 32 (2) https://srees.sggw.edu.pl

ISSN 1732-9353 (suspended) eISSN 2543-7496 DOI 10.22630/srees.4583

Received: 22.12.2022
Accepted: 27.04.2023

Mus'ab A. Al-TAMIR

Abdullah I. IBRAHIM✉ https://orcid.org/0000-0002-7391-3885

University of Mosul, College of Engineering, Environmental Engineering Department, Iraq

Spatial distribution prediction for the ground water quality in Mosul City (Iraq) using variogram equations

Keywords: GIS, geospatial interpolation, groundwater quality, kriging, semivariogram model

Introduction

Spatially distributed data sets play a crucial role in environmental modeling and monitoring. Geological information system (GIS) has been applied successfully in this field over the last two decades. GIS-aided environmental studies achieved high quality outputs in many aspects, such as water quality policies and modeling, as well as improving controls and assisting confident decisions (Sukkuea & Heednacram, 2022). In addition, integration of GIS-based tools with in-situ measurements have been used extensively to identify and describe spatial and temporal variation of pollutants in surface and groundwater (Xue et al., 2023). The GIS-based tools provide reliable interpolation for specific information even at a regional scale (Kourgialas, Karatzas & Koubouris, 2017). Furthermore, the combination of data visualization and mapping with GIS platform is a powerful tool for extracting information and results interpretation (Fan, Fleischmann, Collischonn, Ames & Rigo, 2015). The GIS platform also has a capability to display and analyze geospatial data as well as to create multiple scenarios for land use (Alqahtany, 2023). Geostatistical analysis is one of the powerful tools embedded in the GIS software. It relies on a combination of several techniques that were employed to

✉abdullah.ibrahim@uomosul.edu.iq

solve problems and predict outputs with respect to spatial distribution of data (Montero, Fernández-Avilés & Mateu, 2015). Thus, the geostatistical analysis is widely used to determine the data magnitude for the positions where measurements are lost or unsampled due to safety restrictions, materials availability, or cost considerations (Sukkuea & Heednacram, 2022). The geostatistical analysis consists of a set of steps as spatial interpolation, kriging interpolation, and data fit and error assessment.

Spatial interpolation

Spatial interpolation is to estimate an unknown data point based on the known or investigated points. The influence of known points is proportional to the distance between points. In other words, the weight of nearby points has higher impact on unknown point than those distant data points. The data at unknown points can be calculated as follows.

$$z_p(x) = \sum_{i=1}^{n} w_i \cdot z(X_i), \tag{1}$$

where $z_p(x)$ donating the value of targeted unknown data point, $z(X_i)$ the value of investigated points S_i at point $i = 1, 2$ through n, and w_i is the weight that assigned to each known point. The summation of weight at each single point has to equal to 1.

Semivariance

Semivariance is a function of the distance between the observed and unknown data points. It measures the average degree of dissimilarity between unsampled sites and nearby data values (Koike et al., 2022). Semivariance is proportional to spatial distribution of data points, the lower semivariance associated to lower distance and vice versa. The optimum semivariogram model simulates natural variability of water quality parameters with least root mean square error (*RMSE*) when compared to real water quality parameters. The semivariogram model that achieves minimum *RMSE* is flexible enough to describe the spatial distribution of the studied parameters. The semivariance (γ_h) calculated according to Eq. (2) is then optimized to find the most representative outputs or surface.

$$\gamma_h = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} \left[z(x_i) + z(x_{i+h}) \right]^2, \tag{2}$$

where $n(h)$ is the number of pairs separated by distance h, $Z(x)$ is the value of z at location x, $z(x + h)$ is the value of z at location $x + h$.

Semivariogram models

Semivariogram function fits the spatially autocorrelated paired data points through certain geostatistical procedures. Each set of parameters can fit many empirical models of a semivariogram, such as spherical, circular, quadratic, exponential, rational, tetraspherical, pintaspherical, Gaussian, K-Bessel, J-Bessel, hole effect and stable models (Silva et al., 2023). The selected semivariogram model influences the interpolation of the unknown or missing data. The more the shape of the semivariogram curve is approaching the realistic investigations, the more accurately interpolation will be achieved at virtual or uninvestigated stations (Koike et al., 2022). In other words, the steeper the curve nearby the origin, the higher influence of closed neighbor points on the interpolated surface (Raju, 2016).

According to the regionalized variable theory, two steps need to be followed in the geostatistical analysis to generate a surface. The first step is to construct an empirical semivariogram model with least square error, and then the second step is to produce an interpolation based on the existing known data and the empirical semivariogram model (Grynyshyna-Poliuga, 2019; Sukkuea & Heednacram, 2022).

Kriging interpolation

Kriging interpolation was developed in the 1960s to estimate gold deposited in a rock from a few random core samples (Beana, Sun & Maguire, 2022). Since then, kriging has found its way into other scientific disciplines. Among various spatial interpolation methods, such as trend surface estimation, spline interpolation, and inverse distance weighting, kriging interpolation is the most commonly used in environmental studies (Thanoon, 2018; Yang, Chiu & Yen, 2023). Kriging or Gaussian process regression is an advanced geostatistical approach that uses the semivariogram model to generate unbiased estimation for spatially distributed surface (Sukkuea & Heednacram, 2022). Among many forms of linear and nonlinear kriging, ordinary kriging and lognormal kriging are most commonly used to run interpolation process. Therefore, these kriging procedures were used in this study.

Data fit and error assessment

The collected data were used to generate an empirical semivariogram. Afterward, a reasonable theoretical semivariogram model was fitted based on the least root mean square error (*RMSE*) value (Boroh, Lawou, Mfenjou & Ngounouno, 2022). The *RMSE* was calculated using Eq. (3). The *RMSE* investigate the gap between observed and predicted data points.

$$RMSE = \sqrt{\frac{SSE}{n}}, \tag{3}$$

$$SE = \left\{ \sum_{i=1}^{n} (p_i - o_i)^2 \right\}^{\frac{1}{2}}, \tag{4}$$

where *SSE* is a *t* square root of the sum of squared errors (predicted-observed values) and *n* is the number of pairs (errors).

The *RMSE* is frequently used to evaluate errors in GIS and mapping and many other prediction methods. Thus, it was employed in this study to evaluate outputs of several semivariogram models. A properly fitted semivariogram model was then used to generate autocorrelation to reflect the spatial distribution of the studied variables to produce a mapped surface.

Although the kriging algorithms are used extensively in semivariogram modeling, there is a lack of knowledge about how to optimize this approach to meet water quality modeling requirements (Sukkuea & Heednacram, 2022).

This study aimed to discover the optimum ordinary semivariogram model to predict pH, EC and temperature for groundwater quality for selected wells in the Mosul Province of northern Iraq. The second objective of this study was to explore the pattern of spatial distribution of these parameters over the study area.

Study area and data used

Mosul is the largest city in northern Iraq in terms of population. It is located approximately 400 km North of Baghdad, the capital of Iraq. The latitude coordinate of Mosul City is 36.340000 and the longitude coordinate is 43.130001. According to the latest census, the population of Mosul City is over 3.7 million. Although the Tigris River is the major water supply resource in Mosul, groundwater is also used in some sectors in daily bases for other uses rather than household purposes (Al-Tamir, 2021). Representative groundwater samples were collected from 30 different sites (wells) distributed over different places in Mosul City (Fig. 1). Then the collected samples were analyzed for pH, electrical conductivity (EC), and temperature as a model for groundwater quality in Mosul City.

Al-Tamir, M. A., Ibrahim, A. I. (2023). Spatial distribution prediction for the ground water quality in Mosul City (Iraq) using variogram equations. *Sci. Rev. Eng. Env. Sci.*, *32* (2), 186–197. DOI 10.22630/srees.4583

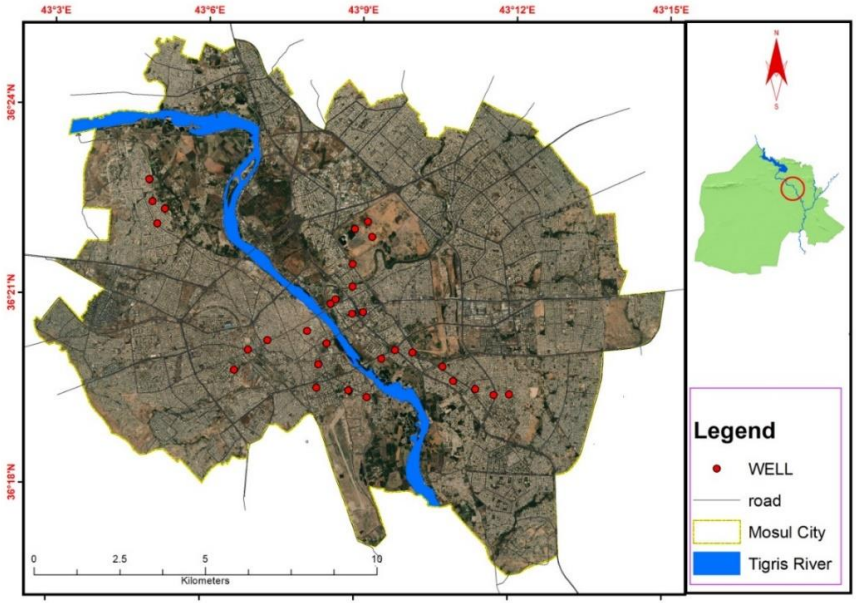

FIGURE 1. Wells distribution map in Mosul City
Source: own work.

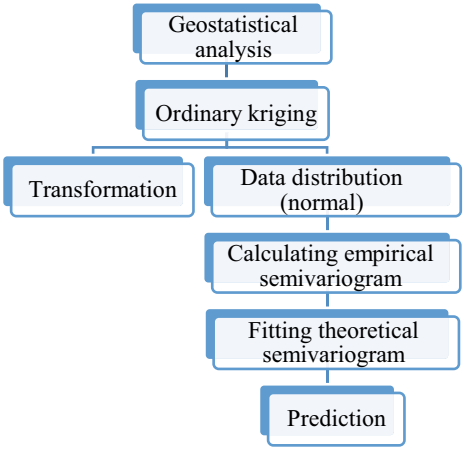

FIGURE 2. Geostatistical analysis steps
Source: own work.

Methodology

The GIS-aided spatial interpolation was applied on the collected data to predict the selected parameters (i.e., pH, EC, and temperature). The geostatistical analysis was applied according to the steps presented in Figure 2.

The normality of the collected data was tested prior to applying the geostatistical analysis. The transformation process was applied, as needed, on data to meet the normal distribution requirements (Fig. 3).

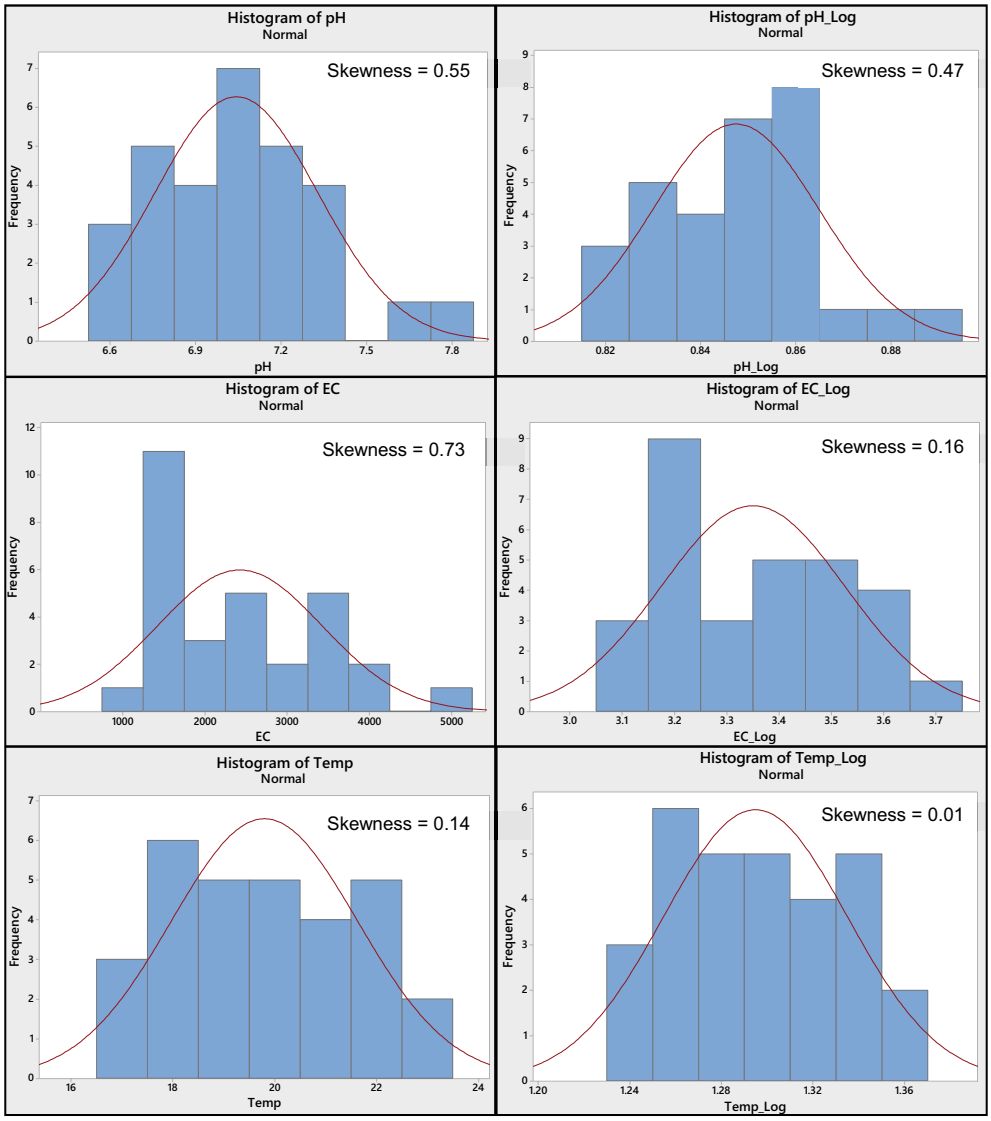

FIGURE 3. Original and transformed normal distribution histogram for each parameter
Source: own work.

The final step was to calculate the empirical semivariogram and compare it with the fitted theoretical semivariogram. The main objective of this procedure was to fit the optimum semivariogram model to represent groundwater quality in sites that have not been sampled. The statistical analysis was performed using Minitab 17 (Minitab 17, 2010).

Al-Tamir, M. A., Ibrahim, A. I. (2023). Spatial distribution prediction for the ground water quality in Mosul City (Iraq) using variogram equations. *Sci. Rev. Eng. Env. Sci.*, *32* (2), 186–197. DOI 10.22630/srees.4583

Results and discussion

Statistical indices

The statistical summary of the collected groundwater quality properties is presented in Table 1. The mean values of pH, EC, and temperature were 7.04, 2426, and 19.8, respectively, while the variance of these parameters was 0.082, 999521, and 3.338, also respectively. Since the variance measures variability of data set from the mean, it shows that the variance of pH and temperature values was close to mean values while the variance of EC was large, which indicates that the collected samples were highly deviated from the mean of the data set. Consequently, this indicates that water quality is impacted by the spatial distribution of wells. This finding is confirmed by the standard deviation test of the same samples (Table 1). Although, the skewness test shows that the data sets of all studied parameters had positive skewness (skewed right), pH and EC were moderately skewed while the temperature data set distribution was fairly symmetrical (Fig. 3). Which means that the distribution of data sets of pH and EC were slightly deviated from the normal distribution. However, the skewness of temperature data set was less than the skewness of other parameters (Table 1). Various functions, such as Log, exponential and square root, were applied on the collected data sets to adjust the normality of distribution. Log was the optimum transformation function because it produced the minimum skewness value (as shown in Table 1). The kurtosis test shows that the distribution of pH and EC data sets was mesokurtic (i.e., kurtosis values between +1 and –1) while the distribution of temperature data set was too flat (platykurtic), (Hair, Hult, Ringle & Sarstedt, 2014).

TABLE 1. The statistical summary for the studied water quality parameters*

Indicator	pH [-]	Electrical conductivity (EC) [-]	Temperature [°C]
Mean	7.04	2 426	19.8
Standard error	0.05	183	0.33
Median	7	2 250	20
Mode	7.2	2 500	18
Standard deviation	0.286	1 000	1.83
Variance	0.082	999 521	3.338
Kurtosis	0.47	–0.23	–1.11
Skewness	0.55	0.73	0.14
Range	1.2	3 850	6
Minimum	6.6	1 150	17
Maximum	7.8	5 000	23
Sum	211.3	72 765	594
Count	30	30	30

*Descriptive statistical analysis indicators were calculated according to common statistical functions.

Source: own work.

Semivariogram model fitting

The empirical semivariogram was computed according to the setting shown in Figure 4. The lag size for pH and temperature was 0.2 for each one, while it sat to 0.4 for EC. The number of lags was 12 for all data sets. The best-fit semivariogram model was selected based on the *RMSE* value (Table 2), the semivariogram model with the least *RMSE* was considered the optimal one. Accordingly, J-Bessel was the best-fit semivariogram model for all investigated parameters because it achieved the minimum *RMSE* as shown in Table 2.

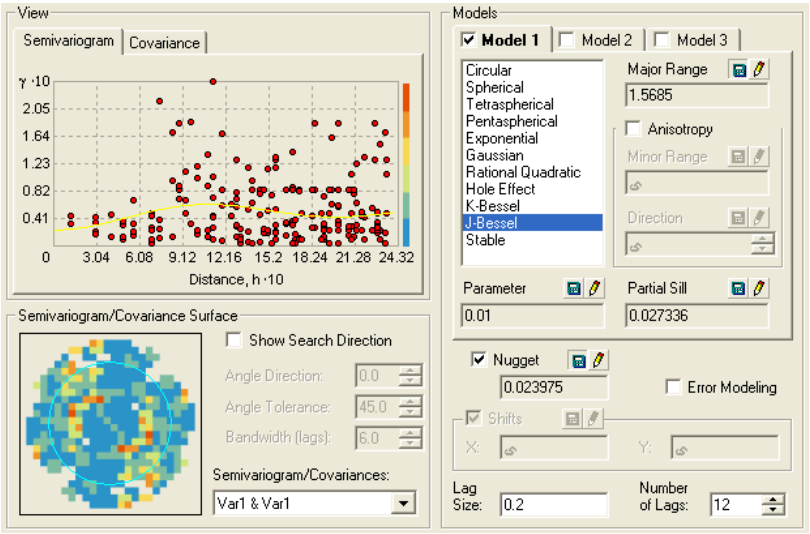

FIGURE 4. Empirical semivariogram settings for each water quality parameter

Source: own work.

TABLE 2. The root mean square error values for the examined models for each studied parameter*

Semivariogram model	pH [-]	Electrical conductivity (EC) [-]	Temperature [°C]
	Lag size: 0.2	Lag size: 0.4	Lag size: 0.2
Circular	0.2248	794.7	1.432
Tetraspherical	0.2261	790.6	1.447
Pentaspherical	0.2262	797.9	1.455
Exponential	0.2263	789.9	1.457
Gaussian	0.2292	886.7	1.554
Rational	0.2237	828	1.474
Quadratic	0.2266	774.6	1.514
Hole effect	0.2239	798.3	1.269
K-Bessel	0.2247	818.3	1.485
J-Bessel	0.2217	740.5	1.209
Stable	0.2237	–	1.492

*For more details refer to Table 1.

Source: own work.

Interpolation maps

Kriging type, output, transformation function, and semivariogram model were set to ordinary kriging, prediction, log, and J-Bessel, respectively. All above settings have been applied to create continuous interpolated surfaces for pH, EC, and temperature in Mosul City (Figs 5–7).

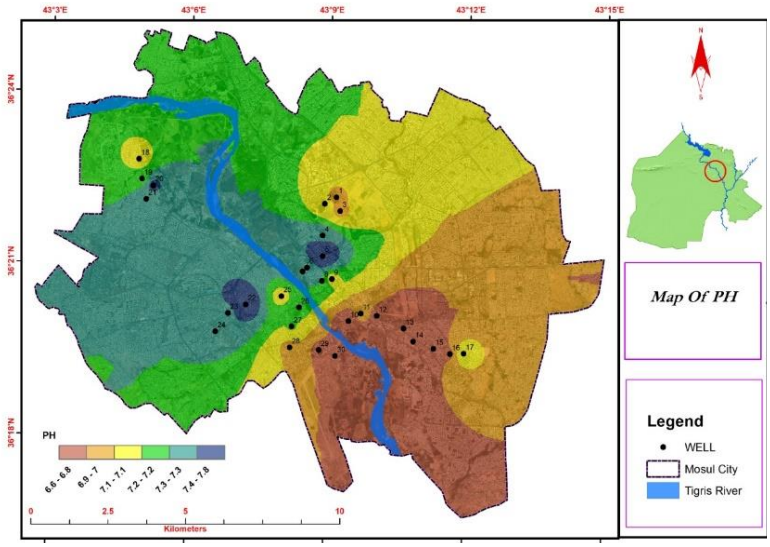

FIGURE 5. Interpolation map of pH in Mosul City
Source: own work.

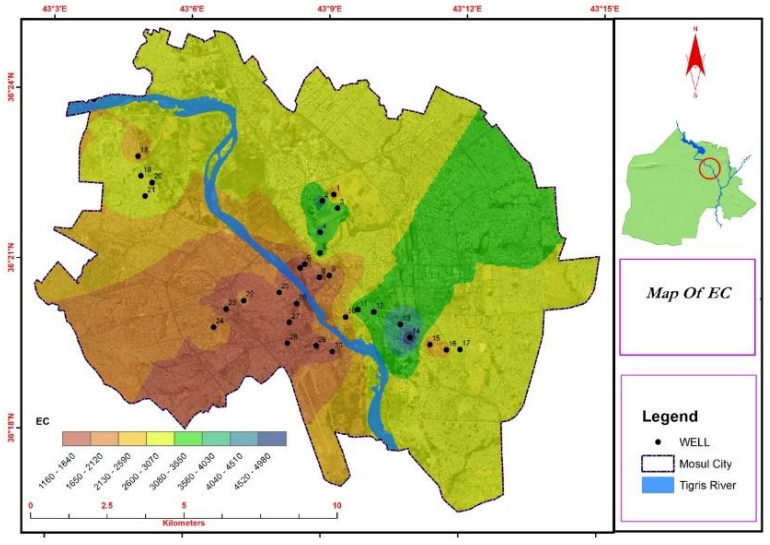

FIGURE 6. Interpolation map of electrical conductivity (EC) in Mosul City
Source: own work.

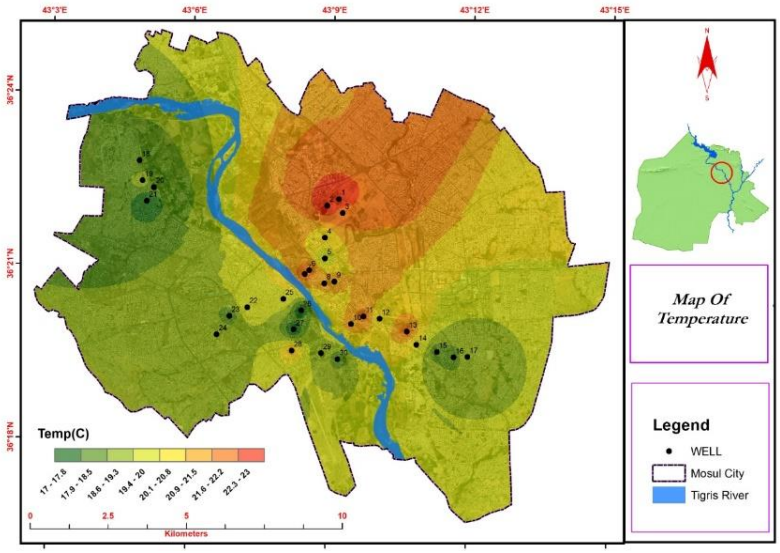

FIGURE 7. Interpolation map of temperature in Mosul City
Source: own work.

The geospatial analysis of the selected groundwater parameters can be considered optimal because it was created using the semivariogram model that has minimum *RMSE* values, as shown in Table 2. The limitation on availability of detailed data due to restricted limitations on drilling of water wells within the city border led to the lack of information regarding groundwater quality. Therefore, the suggested semivariogram model is likely the reliable selection for groundwater spatial modeling in Mosul City and thus it can be used to predict the groundwater parameters in unsampled sites.

Conclusions

Various semivariogram models were applied and optimized to find the best fit model to represent the spatial variability selected groundwater quality parameters in Mosul City. The high restrictions on digging water wells within the city border resulted in small sample size of acquired data sets. Therefore, the skewness test was employed to assess the normality of the data sets distribution. According to that test, natural logarithms were the best fit function to adjust the normality of data sets distribution. The J-Bessel semivariogram model was selected as the opti-

mum model based on the least *RMSE* for all selected groundwater parameters. In spite of the small sample size in terms of the number of wells, interpolation maps for pH, EC, and temperature showed a significant smoothness to represent interpolation surfaces for the studied parameters. Consequently, the adjusted settings are optimal and can be applied successfully for modeling the groundwater quality relaying on a small sample size.

References

Al-Tamir, M. A. (2021). Stability evaluation of Tigris River raw water and treated drinking water from main water treatment plants within Mosul City. *Desalination and Water Treatment*, *226*, 52–61.

Alqahtany, A. (2023). GIS-based assessment of land use for predicting increase in settlements in Al Ahsa Metropolitan Area, Saudi Arabia for the year 2032. *Alexandria Engineering Journal*, *62*, 269–277.

Beana, B., Sun, Y. & Maguire, M. (2022). Interval-valued kriging for geostatistical mapping with imprecise inputs. *International Journal of Approximate Reasoning*, *140*, 31–51.

Boroh, A. W., Lawou, S. K., Mfenjou, M. L. & Ngounouno, I. (2022). Comparison of geostatistical and machine learning models for predicting geochemical concentration of iron: case of the Nkout iron deposit (south Cameroon). *Journal of African Earth Sciences*, *195*, 104662.

Fan, F. M., Fleischmann, A. S., Collischonn, W., Ames, D. P. & Rigo, D. (2015). Large-scale analytical water quality model coupled with GIS for simulation of point sourced pollutant discharges. *Environmental Modelling & Software*, *64*, 58–71.

Grynyshyna-Poliuga, O. (2019). Characteristic of modelling spatial processes using geostatistical analysis. *Advances in Space Research*, *64*, 415–426.

Hair, J. F., Hult, J. G. T. M., Ringle, C. M. & Sarstedt, M. S. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks: SAGE Publications.

Koike, K., Kiriyama, T., Lu, L., Kubo, T., Heriawan, M. N. & Yamada, R. (2022). Incorporation of geological constraints and semivariogram scaling law into geostatistical modeling of metal contents in hydrothermal deposits for improved accuracy. *Journal of Geochemical Exploration*, *233*, 106901.

Kourgialas, N. N., Karatzas, G. P. & Koubouris, G. C. (2017). A GIS policy approach for assessing the effect of fertilizers on the quality of drinking and irrigation water and wellhead protection zones (Crete, Greece). *Journal of Environmental Management*, *189*, 150–159.

Minitab (2010). *Minitab 17 Statistical software*. State College: Minitab.

Montero, J. M., Fernández-Avilés, G. & Mateu, J. (2015). *Spatial and spatio-temporal geostatistical modeling and kriging*. Hoboken: John Wiley & Sons.

Raju, N. J. (2016). *Geostatistical and geospatial approaches for the characterization of natural resources in the environment challenges, processes and strategies*. Berlin: Springer.

Silva, M. V. da, Pandorfi, H., Almeida, G. L. P. de, Silva, R. A. B. da, Morales, K. R. M., Guiselini, C., Santana, T. C., Cangela, G. L. Ch. de, Filho, J. A. D. B., Moraes, A. S., Montenegro, A. A. A. & Oliveira Júnior, J. F. de (2023). Spatial modeling via geostatistics and infrared thermography of the skin temperature of dairy cows in a compost barn system in the Brazilian semiarid region. *Smart Agricultural Technology*, *3*, 100078.

Sukkuea, A. & Heednacram, A. (2022). Prediction on spatial elevation using improved kriging algorithms: An application in environmental management. *Expert Systems with Applications*, *207*, 117971.

Thanoon, S. R. (2018). Application trend surface models with estimation. *Tikrit Journal of Pure Science*, *23* (10), 118–122.

Xue, S., Korna, R., Fan, J., Ke, W., Lou, W., Wang, J. & Zhu, F. (2023). Spatial distribution, environmental risks, and sources of potentially toxic elements in soils from a typical abandoned antimony smelting site. *Journal of Environmental Sciences*, *127*, 780–790.

Yang, J. W., Chiu, S. Y. & Yen, K. C. (2023). Does the realized distribution-based measure dominate particular moments? Evidence from cryptocurrency markets. *Finance Research Letters*, *51*, 103396.

Summary

Spatial distribution prediction for the ground water quality in Mosul City (Iraq) using variogram equations. The GIS-aided spatial interpolation was applied on collected groundwater data to predict selected parameters (i.e., pH, electrical conductivity, and temperature) for the selected water wells distributed over Mosul City in Iraq. A descriptive statistical analysis was conducted on collected samples to explore the statistical indices. The skewness test was also employed to test the distribution of data sets around their mean values. The natural logarithms function achieved least skewness values and thus was applied to transfer data sets in order to adjust normality of the data sets distribution. Among all applied semivariogram models, the J-Bessel semivariogram model was optimal in terms of root mean square error (*RMSE*) values. The average standard errors were 0.2217, 740.5, and 1.209 for pH, EC, and temperature, respectively.