AIMS Environmental Science

AIMS Environmental Science, 12(1): 120–136.
DOI: 10.3934/environsci.2025006
Received: 19 November 2024
Revised: 24 January 2025
Accepted: 06 February 2025
Published: 14 February 2025

https://www.aimspress.com/journal/environmental

Research article

Machine learning and topological kriging for river water quality data interpolation

Rokhana Dwi Bekti[1,*], Kris Suryowati[1], Maria Oktafiana Dedu[1], Eka Sulistyaningsih[2] and Erma Susanti[3]

[1] Department of Statistics, Universitas AKPRIND Indonesia, Yogyakarta 55222, Indonesia
[2] Department of Environmental Engineering, Universitas AKPRIND Indonesia, Yogyakarta 55222, Indonesia
[3] Department of Informatics, Universitas AKPRIND Indonesia, Yogyakarta 55222, Indonesia

* **Correspondence:** Email: rokhana@akprind.ac.id; Tel: +6285711739250.

Abstract: Monitoring of river water quality data is crucial to prevent river water pollution. With limited sampling data, the statistical method of kriging interpolation is indispensable. This method can predict unsampled values based on interconnected surrounding values. Two types of kriging methods that can be applied are Machine Learning (ML) kriging and topological kriging (top-kriging). ML kriging is an extension of ordinary kriging by adding a Super Learning (SL) component. Here, we used SL type Support Vector Regression (SVR). Ordinary Kriging and ML Kriging are based on point values. Top-Kriging is defined as the estimation of streamflow-related variables in ungauged catchments and is based on a non-zero catchment area, not a point value. The three methods were applied in Chemical Oxygen Demand (COD) as water river quality in the Special Region of Yogyakarta (DIY), Indonesia. Based on the Mean Square Error (MSE) and Mean Absolute Error (MAE) comparison, Top kriging provided better accuracy that produced the smallest MSE and MAE. This showed that top kriging is suitable for interpolating data with river flow cases. The interpolation result was that the COD value in the upstream area was low, meaning that the level of organic pollution was minimal. Further downstream, after passing through densely populated residential and industrial areas, the COD values were higher.

Keywords: interpolation; ordinary kriging; machine learning; topological kriging; COD

1. Introduction

Water is a pivotal need for humans and other living things, so it is very necessary to maintain its quality. However, water pollution greatly affects the quality of water. Water pollution has become a serious problem threatening the survival of various ecosystems and people. Industrial and agricultural waste, plastic debris, and hazardous chemicals carelessly dumped into rivers, lakes, and oceans cause severe environmental damage. Polluted water not only destroys the natural habitat of many species but also threatens the quality of water used for daily needs such as drinking, irrigation, and sanitation. This growing water pollution requires concrete action and global cooperation to prevent even greater disasters in the future.

Access to clean water is one of the major goals of the Sustainable Development Agenda (SDGs), particularly Goal 6, which aims to ensure the availability and management of clean water and adequate sanitation. By providing fair and equitable access to safe sanitation and ensuring quality-assured clean water, the SDGs aim to reduce waterborne diseases and improve the quality of life, especially for communities that still lack basic infrastructure. Achieving these goals improves health and supports economic prosperity and a better environment for future generations.

Several water quality parameters can be measured to indicate the good or bad condition of water, each of which provides information about the water's physical, chemical, and biological condition. One parameter often used to evaluate water quality is Chemical Oxygen Demand (COD). The COD level is one of the parameters to determine the quality of river water and its pollution. COD is the need for chemical oxygen to break down all organic matter contained in water. COD shows how much organic matter in water or waste that can be broken down by chemical reactions with strong oxidizers under acidic conditions. Another aspect is that COD is an important parameter in water quality management because it indicates the level of organic pollution [1]. The higher the COD value, the greater the amount of organic matter present and, typically, the higher the potential pollution the sample poses.

Several studies have shown potential risks to health, such as digestive system disorders, hormonal disorders, and immune disorders. Next, Aboyitungiye and Gravitiani [2] found that river pollution in Indonesia significantly impacts human health. Based on data from the Ministry of Environment and Forestry, 73.24% of rivers in 34 provinces are polluted. Using polluted rivers and well water can lead to health problems such as skin disorders, dermatitis, and diarrhea. These are the reasons why monitoring the water quality of rivers is necessary as they continue to experience pollution.

Various qualitative and quantitative methods are used to monitor river water quality. Quantitative methods may use mathematical and statistical approaches. Suphawan and Chaisee [3] predicted water quality indices in the Ping River Basin, Thailand using Gaussian process regression. Moreover, Novianta et al. [4] have used backpropagation artificial neural networks to predict several river water quality parameters to monitor river water quality in the Special Region of Yogyakarta (DIY), Indonesia. Information on water quality will be explored through mapping, namely knowing the distribution pattern of river water quality based on its parameters to know the pollution points from upstream to downstream. Rosyida et al. [5] have created a Web-based Geographic Information System (GIS) for mapping in order to monitor river water quality only at limited points. Additionally, Aneesh and Thomas [6] have mapped the water quality of the Periyar River, Kerala, India using GIS but with limited sample points.

Although there have been many studies on water pollution and its mapping, the data cannot provide information on the entire river because sampling is done only at specific points. This does not describe the water quality of the river as a whole. River water quality data obtained from primary data is generally only at specific points that do not necessarily represent the population of the entire river

flow. This is termed unsampled data. Therefore, the spatial interpolation method is applied to fill in the unsampled data. Spatial interpolation is a method or model to estimate the value of unknown or unmeasured attributes at specific points based on measurements taken at surrounding locations as known sample point values.

Various spatial interpolation models have been developed, with non-geostatistical and geostatistical types, and a combination of both. Raman et al. [7] found that geostatistics has an important role in river water quality analysis. Results are beneficial for researchers and policymakers for sustainable management in downstream areas. The kriging method is a geostatistical method useful for predicting an unmeasured variable. The prediction is based on a weighted average of the region. Types of kriging are ordinary kriging, co-kriging, robust kriging, and universal kriging. Ordinary kriging is the simplest type of kriging, as part of geostatistical methods, and was introduced by Matheron [8]. Chen et al. [9] used it to predict river pollution index numbers in tidal streams. Bekti et al. [10] have also used ordinary kriging to interpolate Peak Ground Analysis data in Banda Aceh. Furthermore, Khan et al. [11] have used universal kriging on spatial interpolation of water quality index data in river and lake water samples. Other researchers that have used ordinary kriging [12–14].

In the developing methods, the kriging algorithm will perform well if it applies machine learning. Machine learning is very effective for handling spatial data and has the potential for broad application in the big data era, especially in interpolating environmental variables [15]. Machine learning modifications to ordinary kriging have been carried out by Erten et al. [16]. The use of machine learning models provides valuable tools for data interpolation. However, they have disadvantages, such as ignoring the samples' spatial proximity and not reproducing the data at their locations. Therefore, machine learning combined with the kriging algorithm through a weighting function based on a kriging variance [16] shows that the combined model yields more accurate estimates than only kriging or machine learning. To prove this, we also applied a combination of both methods.

On the other hand, water quality data in streams is very complex because it relates to the stream's area or length and time [17, 18]. This makes ordinary and universal kriging methods less appropriate to use. One alternative method is the topological kriging (top kriging) method. Top kriging can interpolate variables related to river flow in the catchment area that are not measured. Other advantages of top kriging include that it is the best linear unbiased estimator (BLUE) adapted for the case of flow networks without any additional assumptions. Researchers have applied it to spatially interpolate various variables related to river flow, such as mean annual discharge, flood characteristics, low flow characteristics, concentration, turbidity, and river temperature. Obaid and Mohammed [19] also state that top kriging can predict values over large, irregular areas. Spatial interpolation of river water quality data in Indonesia has not utilized the top kriging method.

The three methods, ordinary kriging, ML kriging, and top-kriging, are applied to interpolate river water quality data in DIY. The water quality parameter used is COD. Our research objective is to obtain the accuracy of the interpolation results of each method. Furthermore, the best interpolation results are presented in a mapping that provides information for monitoring river water quality in DIY.

Our results describe how ordinary kriging, ML kriging, and top kriging work. We also present their similarities and differences and their advantages and disadvantages in the application of data. Providing many alternative methods will contribute to developing kriging methods and illustrate what can be done for the increasingly urgent water pollution problem.

2. Materials and methods

2.1. Data and location

We employed COD as one of the river water quality parameters. COD measures the amount of oxygen required to oxidize organic and inorganic materials in water, making it an important indicator in determining the level of water pollution by chemicals. Secondary data were obtained from the Environment and Forestry Agency, Special Region of Yogyakarta (DIY), Indonesia, in October 2023. The location of the study is about the COD level in DIY. The data location is shown in Figure 1. There were 20 sampling points representing each river area. The detailed data are presented in Table 1.

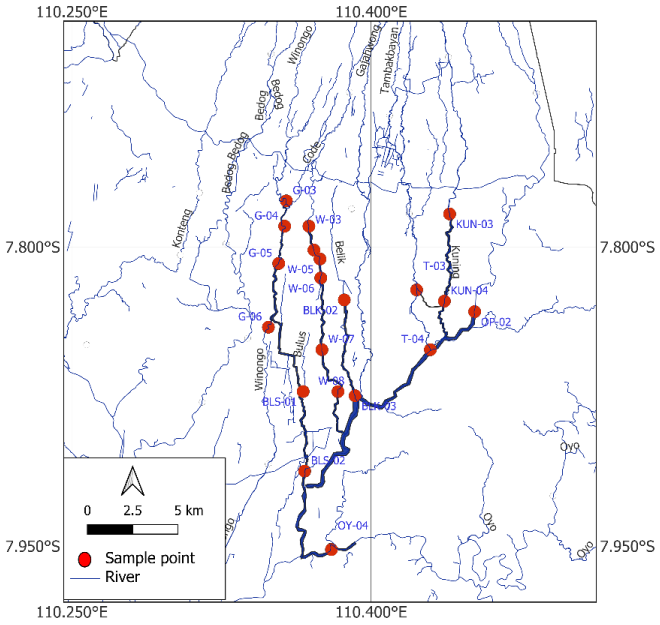

Figure 1. Sample location.

Table 1. River location.

Name code of river	Longitude	Latitude	Name code of river	Longitude	Latitude
OY-04	110.379	-7.953	W-05	110.375	-7.81
BLS-02	110.366	-7.935	W-04	110.372	-7.804
BLK-03	110.385	-7.898	W-03	110.37	-7.795
T-04	110.414	-7.869	T-03	110.428	-7.827
OP-02	110.444	-7.842	BLS-01*	110.366	-7.892
KUN-04	110.436	-7.836	G-06	110.355	-7.839
BLK-02	110.388	-7.848	G-05*	110.351	-7.835
W-08	110.383	-7.883	G-04	110.355	-7.799
W-07	110.379	-7.863	G-03	110.356	-7.783
W-06*	110.375	-7.833	KUN-03	110.439	-7.804

*Note: * Testing data.

In the calculation of ordinary kriging and ML kriging methods, the data was divided into 2, namely 17 training data and 3 testing data. The test data was used to evaluate the comparison between the actual and interpolated values through the Mean Square Error (MSE) and Mean Absolute Error (MAE) values. Moreover, Top Kriging utilized 10 streams, which were splits of the streams that pass through the 20 sample points. A more detailed explanation of this is given in the results section.

2.2. Interpolation method

The interpolation methods used were ordinary kriging, ML ordinary Kriging, and topological kriging. Kriging is a geostatistical method used to estimate the magnitude of the characteristic value at an unsampled location point based on the surrounding sampled point data by considering the spatial correlation. The kriging method is carried out in two stages: The first stage of calculating the value of the variogram or semivariogram and covariance function. Semivariograms include experimental and theoretical semivariograms. The second stage is to estimate the unsampled locations. The steps of interpolation and analysis are shown in Figure 2.

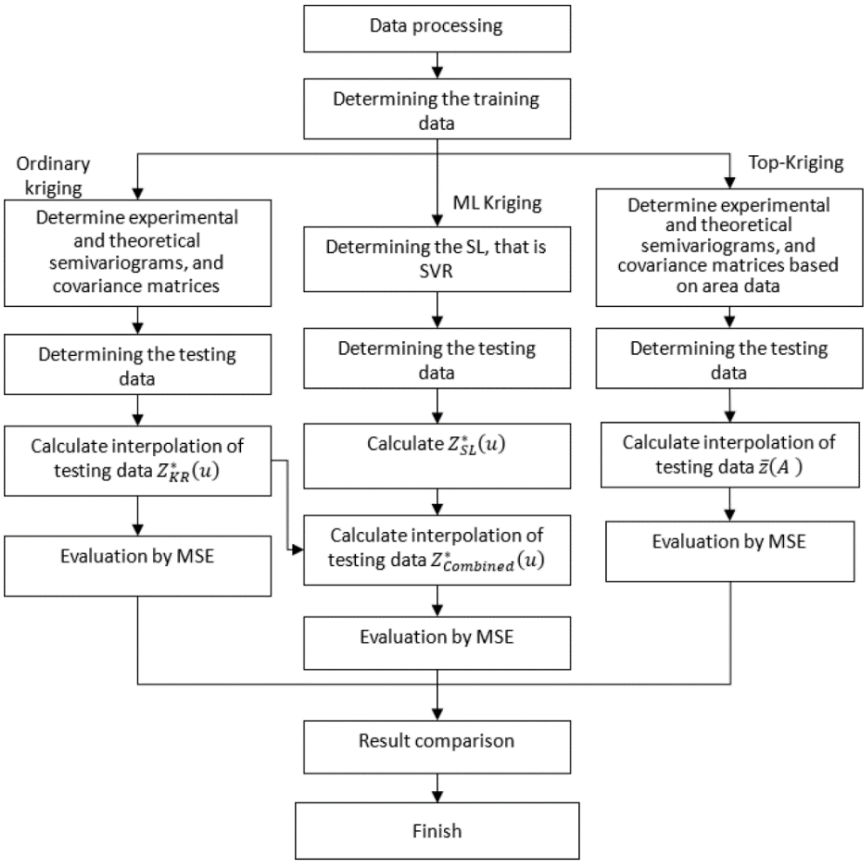

Figure 2. Flowchart of the methodology.

Various types of kriging have been applied. Based on the mean characteristics, kriging can be divided into simple kriging, ordinary kriging, and universal kriging. We used the ordinary kriging type, which assumes that the population mean is unknown, the data is stationary, and does not contain a trend.

Estimator ordinary kriging used formula 1 [20]:

$$Z^*_{KR}(u) = \hat{Z}(X_0) = \sum_{i=1}^{n} \omega_i Z(X_i) \quad (1)$$

where $\sum_{i=1}^{n} \omega_i = 1$, $\hat{Z}(X_0)$ is the interpolation value at location (u), and ω_i is a weight that determines the size of the distance between points. Observation i = 1, 2, ... n, where n is the number of observations that have been sampled or the X variable is known. Value $Z(X_i)$ is the i-th Actual value of variable X.

The value of ω_i obtained from multiplying the covariance matrix as in Eq 2:

$$\omega = C^{-1}D \quad (2)$$

where C is the Covariance Matrix between actual or sampled observations and D is the Covariance Matrix between Actual and interpolated observations. The C matrix elements are obtained from experimental and theoretical semivariograms. One of the theoretical semivariograms used is the Gaussian model with Eq 3.

$$\gamma(h) = C_o + C\left[1 - exp\left(\frac{-(3h)^2}{a^2}\right)\right] \quad (3)$$

We used the combination of Machine Learning and kriging based on the formula in the research of Erten et al. [16], which also applied the Super Learner (SL) model. Super Learner is an ensemble method that combines prediction models to produce more accurate predictions than individual models. Some types of SL that can be applied include Support Vector Regressor (SVR), gradient boosting regressor (GBM), k-neighbors regressor (KNN), Random Forest regressor (RF), and neural network (NN). Here, we used SVR.

The SVR algorithm was introduced as a supervised learning technique [21]. It investigates the relationship between one or more input variables and a target or dependent variable. For a set of training points $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)$, where $x_i \in R_n$ are input values and $y_i \in R$ are target variable, the SVR function is expressed by [22].

$$f(x_i) = w^T(x_i) + b \quad (4)$$

where w and b are coefficients that denote the weight and bias vectors, respectively. To adapt to this research method, this research, the input variables are latitude and longitude coordinates, and the target variable is COD.

The combined ML and kriging algorithm combines the estimates generated by the SL model and estimates obtained from kriging through a weighting function based on a kriging variance. The interpolation result at location *u* is expressed by $Z^*_{Combined}(u)$. The interpolation result with SVR at the *u* location, as one of the SL methods, is denoted by $Z^*_{SL}(u)$ obtained from the function $f(x_i)$ Moreover, the interpolation result of ordinary kriging is denoted by $Z^*_{KR}(u)$. The calculation of $Z^*_{Combined}(u)$ is presented in the following equation:

$$Z^*_{Combined}(u) = w(u).Z^*_{SL}(u) + \big(1 - w(u)\big).Z^*_{KR}(u) \quad (5)$$

$$w(u) = \big(\sigma^2_{KR}(u)\big)^b \quad (6)$$

$$b = b_0 + b_1.Z^*_K R(u) \quad (7)$$

where $w(u)$ is the weight of the SL model, and $\left(1-w(u)\right)$ is the weight of the kriging.

The equations have the rule that when kriging produces an inaccurate estimate of the variance, $\sigma^2_{\mathrm{KR}}(\mathrm{u})=1$, then $\mathrm{Z}^*_{\mathrm{Combined}}(\mathrm{u})=\mathrm{Z}^*_{\mathrm{SL}}(\mathrm{u})$. Conversely, if kriging gives accurate results, $\sigma^2_{\mathrm{KR}}(\mathrm{u})=0$, then $\mathrm{Z}^*_{\mathrm{Combined}}(\mathrm{u})=\mathrm{Z}^*_{\mathrm{KR}}(\mathrm{u})$.

The parameter of b is a linear equation with $\mathrm{b_o}$ and $\mathrm{b_1}$, and $\mathrm{Z}^*_{\mathrm{K}}\mathrm{R}(\mathrm{u})$ is the exponential parameter of the weight. The parameter b must be optimized according to the known data $\mathrm{Z}^*_{\mathrm{SL}}(\mathrm{u})$, $\mathrm{Z}^*_{\mathrm{K}}\mathrm{R}(\mathrm{u})$, and $\sigma^2_{\mathrm{KR}}(\mathrm{u})$.

The topological kriging (top-kriging) algorithm in this study refers to the research algorithms [17,18]. Unlike ordinary and ML kriging, the measurements in Top-Kriging are not point values but are defined over a non-zero catchment area A. In geostatistical terminology, A is the support. As in Formula 1, A point variable z(x) is averaged over an area A. The results of interpolation $\bar{\mathrm{z}}(\mathrm{A})$ are in Eq 8.

$$\bar{z}(A)=\frac{1}{A}\int_A \omega(x)z(x)dx \tag{8}$$

where z(x) is the value of a point in area A and $\omega(\mathrm{x})$ is the weighting.

The detailed steps are as follows:

1) Determine the location of river data as training data and testing data.
2) Calculate the inter-area semivariogram value of the training data.
3) Assuming the existence of a point variogram γ_{p}, the value or the semivariance between two measurements with catchment areas A_1 and A_2 is symbolized as γ_{12} as in Eq 9, with x_1 and x_2 as position vectors within each catchment used for the integration:

$$\begin{aligned}\gamma_{12}&=0.5Var\left(z(A_1)-z(A_2)\right)\\&=\frac{1}{A_1A_2}\int_{A1}\int_{A2}\gamma_p(|x_1-x_2|)dx_1dx_2-0.5\left[\frac{1}{{A_1}^2}\int_{A1}\int_{A1}\gamma_p|x_1-x_2|dx_1\,dx_2\right]\\&+0.5\left[\frac{1}{{A_2}^2}\int_{A1}\int_{A1}\gamma_p|x_1-x_2|dx_1\,dx_2\right]\end{aligned} \tag{9}$$

Moreover, the estimation of the point variogram between two catchments of a pair is

$$\gamma_{obs}(A_1,A_2,h)=\frac{1}{2n(A_1,A_2,h)}\left(\sum_{i=1}^{n(A_1,A_2,h)}[z(x_i)-z(x_i+h)]^2\right) \tag{10}$$

where $\mathrm{h}=|\mathrm{h}|$ is the distance between the centroids of the catchments, and $\mathrm{n}(\mathrm{A}_1,\mathrm{A}_2,\mathrm{h})$ is the number of catchment pairs with areas A_1 and A_2.

A point variogram with the effect of parameter (sill, range, and nugget) is:

$$\gamma_p(h)=ah^b\left(1-e^{-(h/c)^d}\right)+C_{0p} \tag{11}$$

The parameter a is related to the sill of the variogram, c is a correlation length, and b and d define the long and short distance slope of the variogram in a log-log plot, respectively. $\mathrm{C_{0p}}$is a nugget effect. The nugget is calculated by

$$C_0(A_1,A_2)=0.5\left(\frac{C_{0p}}{A_1}\frac{C_{0p}}{A2}\right)-0.5\left(-\frac{2C_{0p}Meas(A_1\cap A_2)}{A_1A_2}\right) \tag{12}$$

4) Calculate the interpolation of the location of the testing data with Eq 8.

Comparison of the interpolation results of the three types of kriging interpolation is done by calculating the Mean Square Error (MSE) and Mean Absolute Error (MAE) accuracy of the testing data,

$$MSE = \frac{\sum_{i=1}^{n}\left(\hat{z}(x_i) - z(x_i)\right)^2}{n} \tag{13}$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{z}(x_i) - z(x_i)|}{n} \tag{14}$$

where $\hat{z}(x)$ is the interpolation results and $z(x)$ the testing data.

3. Results and discussion

The characteristics of COD at the sample points are presented in Figure 3. The minimum value of COD was 3 mg/L, which was owned by 7 sample points and spread across the Winongo, Kuning, and Gadjahwong Rivers. These sample points were located in the northern part of the river or close to the upper reaches of the river on the slopes of Mount Merapi. The highest COD value was 25 mg/L, which was a sample point in the Opak River (No. 5). Furthermore, there was another sample point with COD 21 mg/L in the Tambakbayan River (No. 14). The location with the highest COD levels was in the Opak River, which was polluted. Pollution is caused by the flow of water containing waste that crosses the industry. In addition, pollution can also come from agricultural and household waste.

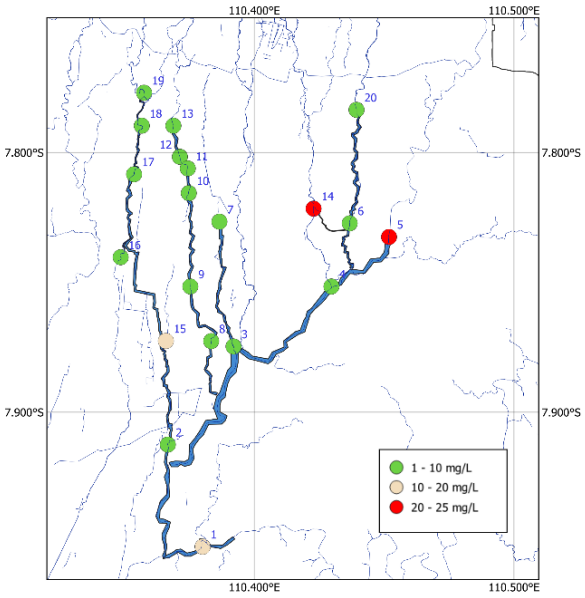

Figure 3. Spatial pattern of COD.

3.1. Interpolation results based on ordinary Kriging

The semivariogram of Ordinary Kriging is presented in Figure 4 and the value is presented in Table 2. The plot shows 14 groups in the interpolation semivariogram plot. Group one consisted of two pairs of points with a distance of 0.006182529 with a semivariance value of 1. Group 14 consisted

of 6 pairs of points that were close to each other with a distance of 0.062022970 with a semivariance value of 50.42. The semivariogram that followed this Gaussian model contained Eq 3.

$$\gamma(h) = 20 + 2\left[1 - exp\left(\frac{-(3h)^2}{0.00497^2}\right)\right] \quad (15)$$

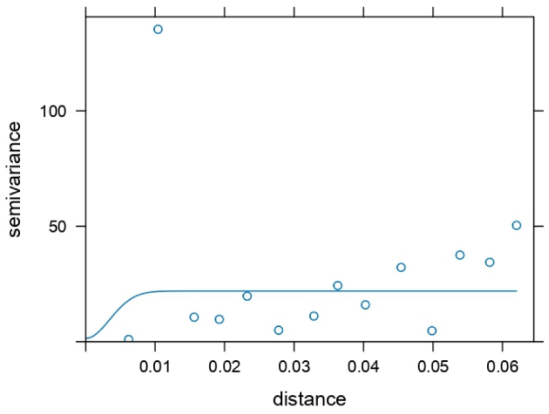

Figure 4. Ordinary kriging semivariogram.

Table 2. Semivariance of ordinary kriging.

No	Number of points	Average distance between points	Semi-variance	No	Number of points	Average distance between points	Semi-variance
1	2	0.006182529	1	8	9	0.036286627	24.33
2	3	0.01042038	135.33	9	11	0.040269201	16
3	4	0.015624882	10.63	10	5	0.04540757	32.2
4	8	0.019242295	9.69	11	4	0.049849159	4.75
5	6	0.023256504	19.75	12	12	0.053872448	37.54
6	2	0.027777759	5	13	7	0.058161632	34.43
7	8	0.032848012	11.13	14	6	0.06202297	50.42

The semivaroigram plot also shows the corresponding model parameters such as sill, range, and nugget. The sill value in the Gaussian model was 20, meaning that the variance value in the Gaussian model will be constant at 20. The range value in the Gaussian model was 0.00497, meaning the distance of the variogram value in the Gaussian model when it reached a sill of 0.00497.

Based on the sill, range, and Gaussian nugget parameters, ordinary kriging produced an interpolation on the test data that gave an MSE of 0.530 and an MAE of 0.696. Interpolation was also performed at 10 other locations, and the results are shown in Table 3.

Table 3. Ordinary kriging interpolation.

No	Lat	Long	COD inter-polation	Variance	No	Lat	Long	COD inter-polation	Variance
1	-7.942	110.371	20.891	1.655	6	-7.849	110.388	7.7678	0.305
2	-7.899	110.384	10.058	0.430	7	-7.832	110.375	3.077	0.308
3	-7.892	110.366	11.000	8.88×10^{-6}	8	-7.828	110.355	2.964	1.755
4	-7.883	110.383	3.000	8.88×10^{-6}	9	-7.827	110.428	21.100	0.001
5	-7.859	110.426	12.263	2.731	10	-7.813	110.438	11.102	1.837

3.2. Interpolation result based on machine learning Kriging

The interpolation calculation in this discussion used Eqs 1–3. $Z^*_{SL}(u)$ was obtained from the Support Vector Regression function. Moreover, $Z^*_{KR}(u)$ is shown in Table 3. The values of b_0 and b_1 are very important in determining the weights $w(u)$. In this research, the values of b_0 and b_1 were determined from 0.00007 to 0.00002, respectively, and then selected based on the smallest MSE value. The following Figure 5 shows the comparison of MSE values between actual data and $Z^*_{Combined}$ at different values of b_0 and b_1. It can be seen that the smaller the values of b_0 and b_1, the smaller the MSE value. Furthermore, if the value is zero, the interpolation result is the same as ordinary kriging, or $Z^*_{SL}(u) = Z^*_{KR}(u)$. This discussion displays the results of COD interpolation with Machine Learning Ordinary Kriging when $b_0 = b_1 = 0.00007$.

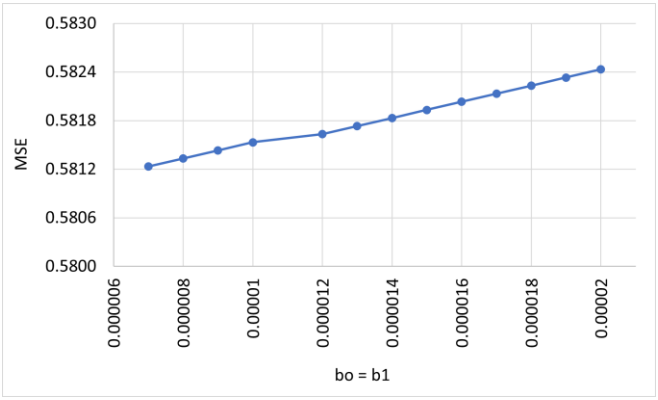

Figure 5. Comparison of b_0 and b_1 against MSE in ML kriging.

ML kriging produces an interpolation on the test data that gives an MSE of 0.581 and an MAE of 0.746. Interpolation was also performed at 10 other locations, and the results are shown in Table 4.

Table 4. ML kriging interpolation.

No	Lat	Long	$Z^*_{SL}(u)$	$Z^*_{Combined}$	No	Lat	Long	$Z^*_{SL}(u)$	$Z^*_{Combined}$
1	-7.942	110.371	9.508	9.508	6	-7.849	110.388	8.044	8.041
2	-7.899	110.384	10.321	10.321	7	-7.832	110.375	3.037	3.037
3	-7.892	110.366	10.937	10.936	8	-7.828	110.355	1.469	1.465
4	-7.883	110.383	3.064	3.064	9	-7.827	110.428	20.936	21.000
5	-7.859	110.426	4.670	3.059	10	-7.813	110.438	6.147	6.151

3.3. Interpolation result based on topological kriging

The topological kriging (top-kriging) included interpolating data based on a non-zero catchment area. Of course, this is different from ordinary and ML kriging. The interpolation step in top kriging is the same as in ordinary kriging, which begins with compiling a semivariogram, determining the parameters of sill, range, and nugget, up to the interpolation. Figure 6 shows an illustration of the area used in top kriging. The training data were 20 river flow areas taken from the sample points in the ordinary kriging method. The 20 areas were also used to interpolate 10 areas. The points in the interpolated area indicate the center point of the area. This data formation references the rtop package in R software.

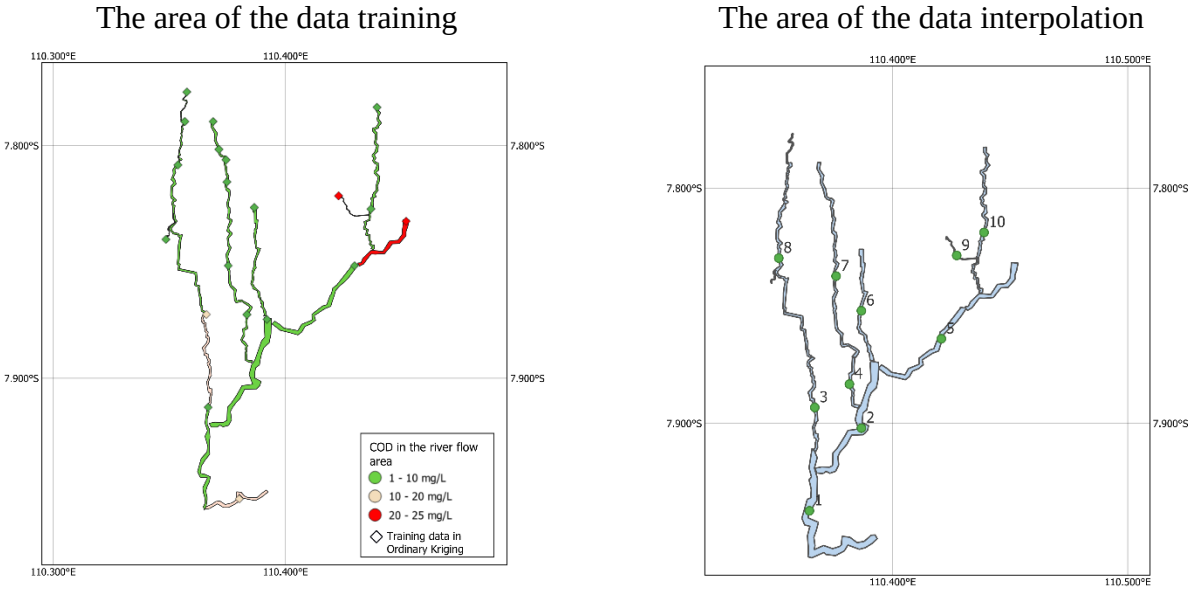

Figure 6. The area used in top kriging.

Semivariograms of topological kriging are presented in Figure 7. This figure shows the point variograms and ordered variograms displayed for different catchment sizes. For example, the solid green line shows the semivariogram for a river area of $(8.4x10^3)$ with $(3.36x10^3)m^2$. Based on this identification, the variogram model used was Exponential, with Nugget 0, Sill 496.41, and Range 12.68 parameters. This model produced a Sum of Squared Errors (SSErr) of 15.301, which indicates the suitability of the model to the data. It also produced an AIC of 0.287.

The interpolation results of Top Kriging are presented in Table 5. The interpolated COD in each area was compared to the average of the training data in each of the corresponding areas. This comparison resulted in an MSE of 0.520 and an MAE of 0.601.

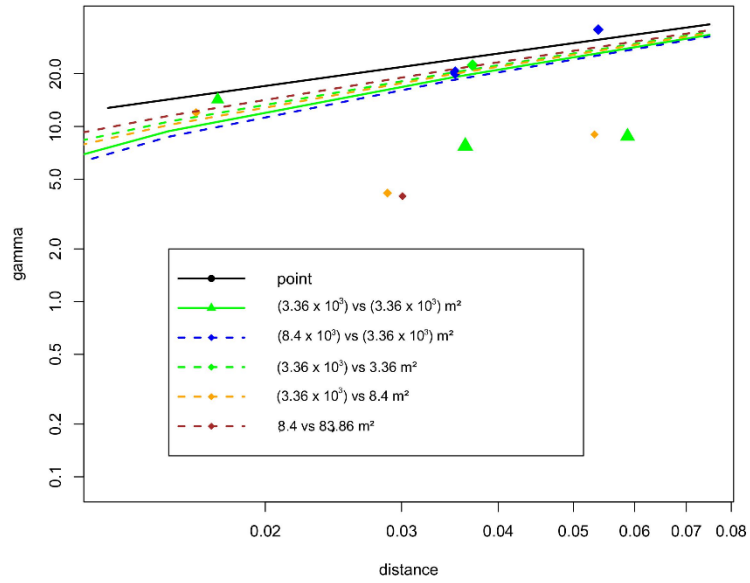

Figure 7. Point variogram and ordered variogram with different catchment sizes.

Table 5. Top kriging interpolation.

Area	Center (Lat, Long)	COD interpolation	Area	Center (Lat, Long)	COD interpolation
1	-7.942, 110.371	21.203	6	-7.849, 110.388	3.150
2	-7.899, 110.384	10.033	7	-7.832, 110.375	10.893
3	-7.892, 110.366	14.587	8	-7.828, 110.355	3.412
4	-7.883, 110.383	7.945	9	-7.827, 110.428	21.088
5	-7.859, 110.426	3.023	10	-7.813, 110.438	5.977

3.4. Comparison of the results

A comparison of the interpolation results of the three methods was done based on the MSE value as shown in Table 6. The MSE and MAE of the ordinary kriging and the ML kriging are based on the location of three points of the test data. Ordinary kriging still provides better performance than ML kriging because ordinary kriging produces smaller MSE and MAE. However, the difference between ordinary and ML kriging is relatively small.

Table 6. The comparison of MSE and MAE.

No	Method	MSE	MAE
1	Ordinary kriging	0.530	0.696
2	ML kriging	0.581	0.746
3	Top kriging	0.520	0.601

Figure 8 shows an illustration of the comparison of ordinary kriging and ML kriging interpolation at 10 other sample points, as shown in Tables 3 and 5. The map also shows 20 training data points. In ordinary kriging, a total of 2 points are predicted to have high COD compared to 8 other points, namely

point 1 at 20.891 mg/L and point 9 at 21.100 mg/L. Both points have COD levels that are relatively the same as those of the nearest training data points. This shows the nature of spatial patterns, where adjacent areas influence each other or have similar characteristics. These characteristics are also present in ML kriging.

However, ML kriging provides different prediction results at some points. For example, point 1 has a COD of 9.08 mg/L, and point 9 has a COD of 21 mg/L, which are lower than the interpolated results of ordinary kriging. However, the interpolated values are nearly equal to the values of the nearest training data points.

On the other hand, MSE and MAE of top kriging are based on ten area locations. The topological kriging method produces a smaller MSE dan MAE value than ordinary kriging and ML kriging, which are 0.520 and 0.601. Area 1 in Oya River with 21,203 mg/L and area 9 in Tambakbayan River with 21,088 mg/L are the areas with the highest COD interpolation results. In comparison with the ordinary and ML kriging results, the COD in region 1 has a value that is relatively the same as the interpolation of the ordinary kriging results at point 1, which is 20,891 mg/L. In other regions, for example, region 8 has an interpolated COD of 3,412 mg/L, which is also relatively equal to the interpolated ordinary and ML kriging results at point 8 and the training data at points 16, 17, 18, and 19, which are located in region 8.

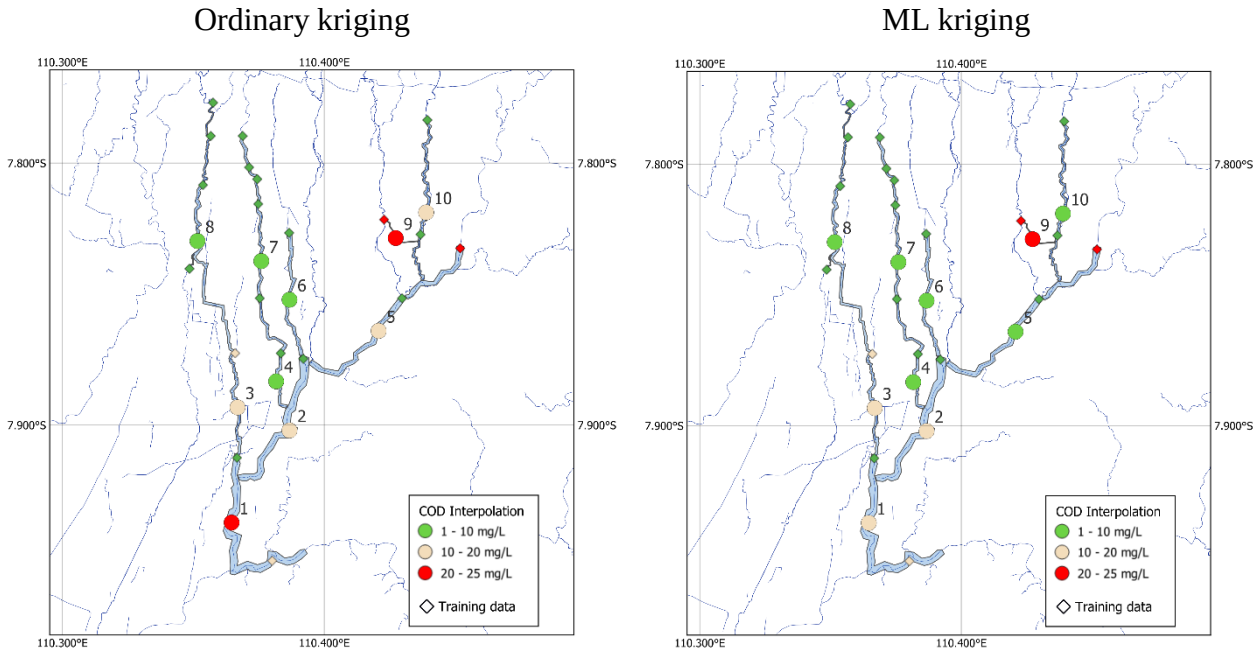

Figure 8. Ordinary and ML kriging interpolation result map.

The visualization of the interpolation results from Figures 8 and 9 show the different forms of interpolation results, where ordinary and ML kriging provide interpolation in the form of point data, while top kriging provides interpolation in the form of river flow areas. The disadvantage of ordinary kriging is that it does not take into account the structure of the river network or the stratified nature of the watershed. As a result, the uncertainty estimates tend to be uniform and do not reflect the intuitive distribution of estimation errors. Top kriging has the advantage of taking into account the information shared by measured and unmeasured watersheds, as well as the stratified structure of the watershed. It provides more accurate and reasonable estimates compared to ordinary kriging. Moreover, ML kriging

has a chance to interpolate well, but it needs to make some modifications. Some things that can be done are choosing the type of Super Learning and optimizing the selection of b_o and b_1.

The results of the interpolation in Figure 9 show that areas with low COD levels (green color) were seen in the north or the upper reaches of the rivers (areas 4, 6, 8, and 10). This identifies that COD levels are good with minimal levels of organic pollution. Furthermore, the southern or downstream part of the river (areas 3, 2, and 1) tends to have high COD levels. In this stream, water quality is predicted to experience a very significant decline, indicating that there is a high level of organic pollution.

For COD interpolation in this study, we did not consider external factors such as river conditions, temperature, water flow, and aquatic biota. However, some other secondary data can describe the things that affect the high and low COD at the same location. Based on the correlation analysis of the training data, COD has a strong relationship with Total Suspended Solid (TSS), pH, Biochemical oxygen demand (BOD), total phosphorus, ammonia, and water discharge. COD will increase as TSS, pH, BOD, Total Phosphorus, and Ammonia increase. Moreover, COD decreases when water flow is high. This can also be explored in future research to see how interpolation can be achieved by considering other factors that may affect COD.

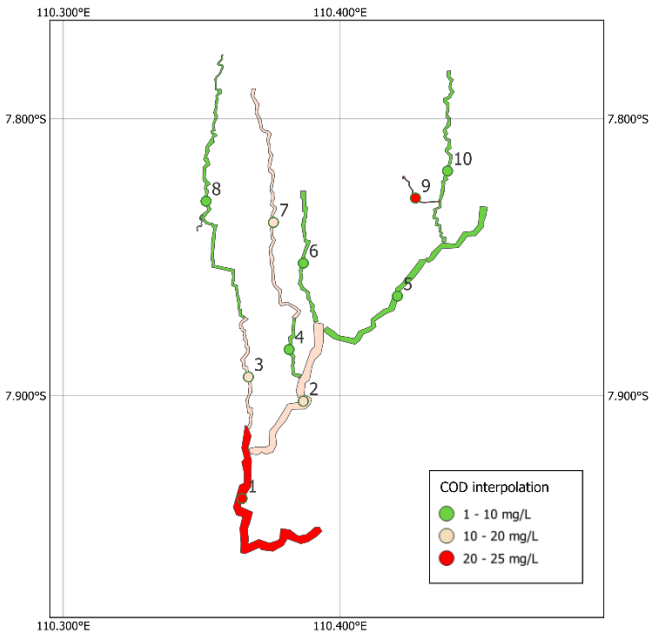

Figure 9. Top kriging interpolation result map.

The prediction that areas 1 and 9 have high COD is also in accordance with the water quality index reported by the concerned agencies. In 2020, the water quality index of the Oya River was 35 and that of the Tambakbayan River was 37.5, which is considered moderately polluted. By 2023, the index was in the same range. River pollution can occur from upstream to downstream. In addition, the dense settlements along the riverbanks also carry domestic waste.

These interpolated results can provide information for pollution prevention in the form of collaboration between the public and government to improve and maintain water quality. Several researchers have also identified the causes of pollution and some remedial strategies. Several causes of river water pollution have also been identified, including waste from slums, settlements, industries, and households. Rohmadi et al. [23] have stated that there is a high correlation between slum areas and

COD in Winongo, Code, and Gajahwong River. This indicates that with the high slum area, the COD levels will be high. However, some efforts have been made to reduce it, namely by development during slum upgrading and improvements in WWTP, drainage channels, and construction of embankments. The results in research of Brontowiyono et al. [24] are that wastewater incredibly contributes to the increase of COD values at Opak River, typically in organic matter from settlement point sources.

4. Conclusions

The three methods were applied in Chemical Oxygen Demand (COD) as water river quality in the Special Region of Yogyakarta (DIY). Ordinary Kriging and ML kriging, whose interpolation basis is data points, provided MSE accuracy of 0.530 and 0.581 and MAE accuracy of 0.696 and 0.746, respectively. This result is different from that in the research of Erten et al. [16], stating that the combined model of ordinary kriging and SL gives more accurate estimates than ordinary kriging. ML kriging will have a chance to interpolate well, but it needs to make some modifications. Some things that can be done are choosing the type of Super Learning and optimizing the selection of b_o and b_1.

Top kriging provided better accuracy, with an MSE of 0.520 and MAE of 0.601. This shows that top kriging is suitable for interpolating data in the case of river flow. However, it should be noted that the data used between ordinary and ML kriging are different. Top kriging has the advantage of taking into account the information shared by measured and unmeasured watersheds, as well as the stratified structure of the watershed.

The interpolation result of top kriging is that the COD value in the upstream area of the river is low, meaning that the level of organic pollution is minimal. The upper reaches of the river are close to the peak of Mount Merapi. Further downstream, after passing through densely populated residential and industrial areas, the COD value is higher. This result has the same characteristics as the results of point interpolation by ordinary and ML kriging. These results have the same characteristics as the results of point interpolation by ordinary and ML kriging, i.e., COD will be high at sample points 1 and 9, which are in areas with high COD also in Top Kriging.

Top-kriging also has the advantage of taking into account the area and integrated nature of the catchment. This method not only provides an estimate of the variable of interest in the unmeasured catchment but also provides an estimate of its uncertainty. This is different from ordinary kriging, which depends only on the centroid distance of the measured and unmeasured catchments. Future research can apply the Top-Kriging method in interpolating various cases. Obaid et al. [19] applied it to risky diseases in Iraq, concluded that top kriging is also better than area-to-point kriging. Then, Archfield et al. [25] did flood prediction and proved that top-kriging is better than Regression kriging. This can also be explored in future research to see how interpolation can be achieved by considering other factors that may affect COD. Thus, combining the concepts of machine learning and top kriging is another possible development.

Use of AI tools declaration

The authors declare they do not use Artificial Intelligence (AI) tools in writing this article.

Acknowledgments

This research was supported by the Directorate of Research and Community Service of the Ministry of Education, Culture, Research, and Technology with contract No:

03/SPP/DP2M/PL/VI/2024 about a Fundamental Research Scheme in 2024. The author also would like to thank DP2M Universitas AKPRIND Indonesia for organizing and facilitating the research.

Rokhana Dwi Bekti: Conceptualization, methodology, writing-original draft, writing-review & editing; Kris Suryowati: Funding acquisition, formal analysis; Maria Oktafiana Dedu: Data curation; Eka Sulistyaningsih: Investigation, resources; Erma Susanti: Software, visualization. All authors have read and approved the final version of the manuscript for publication.

Conflict of interest

The authors declare no conflict of interest.

References

1. Lee J, Lee S, Yu S, et al. (2016) Relationships between water quality parameters in rivers and lakes: BOD 5, COD, NBOPs, and TOC. *Environ Monit Assess* 188: 1–8. https://doi.org/10.1007/s10661-016-5251-1
2. Aboyitungiye JB, Gravitiani E (2021) River pollution and human health risks: Assessment in the locality areas proximity of Bengawan Solo river, Surakarta, Indonesia. *Indonesian J Environ Manage Sust* 5: 13–20. https://doi.org/10.26554/ijems.2021.5.1.13-20
3. Suphawan K, Chaisee K (2021) Gaussian process regression for predicting water quality index: A case study on Ping River basin, Thailand. *AIMS Environ Sci* 8: 268–282. https://doi.org/10.3934/environsci.2021018
4. Novianta MA, Warsito B, Rachmawati S (2024) Monitoring river water quality through predictive modeling using artificial neural networks backpropagation. *AIMS Environ Sci* 11: 649–664. https://doi.org/10.3934/environsci.2024032
5. Rosyida N, Dinira L, Rusydi AN, et al. (2022) Development of web-based geographic information system for water quality monitoring of watershed in Malang. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi* 6: 184–197. https://doi.org/10.29407/intensif.v6i2.17514
6. Aneesh PC, Thomas RM (2024) Assessment and mapping of seasonal variation in water quality of Periyar River, Kerala, India. *J Comput Sci* 17: 1549–3636.
7. Raman RK, Bhor M, Manna RK, et al. (2023) Statistical and geostatistical modelling approach for spatio-temporal assessment of river water quality: A case study from lower stretch of River Ganga. *Environ Dev Sustain* 25: 9963–9989. https://doi.org/10.1007/s10668-022-02472-7
8. Matheron G (1963) Principles of geostatistics. *Econ Geol* 58: 1246–1266.
9. Chen YC, Yeh HC, Wei C (2012). Estimation of river pollution index in a tidal stream using kriging analysis. *Int J Env Res Pub He* 9: 3085–3100. https://doi.org/10.3390/ijerph9093085
10. Bekti RD, Irwansyah E, Kanigoro B, et al. (2018) *Ordinary kriging and spatial autocorrelation identification to predict peak ground acceleration in Banda Aceh City, Indonesia*, In Applied Computational Intelligence and Mathematical Methods: Computational Methods in Systems and Software 2017, Springer International Publishing, 2: 318–325. https://doi.org/10.1007/978-3-319-67621-0_29
11. Khan M, Almazah MM, EIlahi A, et al. (2023). Spatial interpolation of water quality index based on ordinary kriging and Universal kriging. *Geomat Nat Haz Risk* 14: 2190853. https://doi.org/10.1080/19475705.2023.2190853

12. Marthadyanti A, Harisuseso D, Suhartanto E (2024) Mapping of design rainfall distribution at multiple periods using spatial interpolation in Widas Sub-Watershed, Nganjuk Regency. *Jurnal Teknologi dan Rekayasa Sumber Daya Air* 4: 450–459. https://doi.org/10.21776/ub.jtresda.2024.004.01.038
13. Kethireddy SR, Adegoye GA, Tchounwou PB, et al. (2018) The status of geo-environmental health in Mississippi: Application of spatiotemporal statistics to improve health and air quality. *AIMS Environ Sci* 5: 273–293. https://doi.org/10.3934/environsci.2018.4.273
14. Fahimah N, Salami IR, Oginawati K, et al. (2023) The assessment of water quality and human health risk from pollution of chosen heavy metals in the Upstream Citarum River, Indonesia. *J Wate Land Dev* 56: 153–163. https://doi.org/10.24425/jwld.2023.143756
15. Du P, Bai X, Tan K, et al. (2020) Advances of four machine learning methods for spatial data handling: A review. *J Geovis Spat Anal* 4: 1–25. https://doi.org/10.1007/s41651-020-00048-5
16. Erten GE, Yavuz M, Deutsch CV (2022) Combination of machine learning and kriging for spatial estimation of geological attributes. *Nat Resour Res* 31: 191–213. https://doi.org/10.1007/s11053-021-10003-w
17. Skøien JO, Merz R, Blöschl G (2006) Top-kriging-geostatistics on stream networks. *Hydrol Earth Syst Sc* 10: 277–287. https://doi.org/10.5194/hess-10-277-2006
18. Skøien JO, Blöschl G, Western AW (2003) Characteristic space scales and timescales in hydrology. *Water Resour Res* 39. https://doi.org/10.1029/2002WR001736
19. Obaid AN, Mohammed MJ (2020) A comparison of topological kriging and area to point kriging for irregular district area in Iraq. *J Mech Con Math Sci* 15. https://doi.org/10.26782/jmcms.2020.04.00009
20. Fischer MM, Getis A (2010) *Handbook of applied spatial analysis: Software tools, methods and applications*, Berlin: Springer, 125–134.
21. Vapnik V (2013) *The nature of statistical learning theory*, New York: Springer.
22. Balogun AL, Rezaie F, Pham QB, et al. (2021) Spatial prediction of landslide susceptibility in western Serbia using hybrid support vector regression (SVR) with GWO, BAT and COA algorithms. *Geosci Front* 12: 101104. https://doi.org/10.1016/j.gsf.2020.10.009
23. Rohmadi E, Sekine M, Setiawan B (2022) Impact of slum upgrading to river water quality in Yogyakarta City, Indonesia. *Jurnal Teknosains* 12: 85–98.
24. Brontowiyono W, Asmara AA, Jana R, et al. (2022) Land-use impact on water quality of the opak sub-watershed, Yogyakarta, Indonesia. *Sustainability* 14: 4346. https://doi.org/10.3390/su14074346
25. Archfield SA, Pugliese A, Castellarin A, et al. (2013) Topological and canonical kriging for design flood prediction in ungauged catchments: An improvement over a traditional regional regression approach? *Hydrol Earth Syst Sc* 17: 1575–1588. https://doi.org/10.5194/hess-17-1575-2013.s

AIMS Press