




# Predicted Concentration TSS (Total Suspended Solids) Pollution for Water Quality at the Time: A Case Study of Tan Hiep Station in Dong Nai River

Cong Nhut Nguyen<sup>(✉)</sup> 

Faculty of Information Technology, Nguyen Tat Thanh University,  
Ho Chi Minh City, Vietnam  
ncnhut@ntt.edu.vn

**Abstract.** Water is essential for human life and socio-economic development. Water pollution is a concern for all mankind. In Vietnam, due to the development of factories and industries, water pollution has become more severe, including the Dong Nai River. In this article, the author uses the concentration of TSS at the Tan Hiep control station in Dong Nai River, using the Kriging interpolation method to find the appropriate model and give the results of water pollution prediction. Dong Nai river area over time with high reliability. TSS data were monitored continuously for three months (from the beginning of February 2018 to the end of April 2018), the predicted results using Kriging interpolation with high accuracy with regression coefficient equal to 1,005, the coefficient is 0.859 (the best value is 1), the forecast error is 2.258, the standard error is 0.044. It shows that using the Kriging interpolation method is an effective and suitable solution in mathematical problems with time information.

**Keywords:** Water pollution · Geostatistics · Kriging · Semivariogram

## 1 Introduction

Water pollution is an issue of social concern both in Vietnam in particular and the world in general. Water pollution caused by industrial factories increasingly degrades environments quality, leads to severe problems in health for local inhabitants. The building of water quality monitoring stations is also essential, but also difficult because of expensive installation costs, no good information of selected areas for installation in order to achieve precise results. According to the Center for Monitoring and Analysis Environment (Department of Natural Resources and Environment Binh Duong), automatic water quality monitoring network of Binh Duong province has 4 stations observation including Tan Hiep, Vinh Nguyen, Thu Dau Mot and Tan Uyen. The system continuously monitors daily with monitoring parameters such as TSS, pH, Nitrate, temperature and salinity. With a large area, the province needs to install more new monitoring stations. However, with the rapid development of industrial parks, the problem of environmental pollution, especially water pollution is a hot issue, the scarcity of clean water and polluted water leads to many diseases danger. Therefore, it is necessary to have a

mathematical model to predict whether the quality of water in a certain area is safe to use in the future? By using continuous monitoring data for 3 consecutive months (from February to April 2018, at Tan Hiep station), the author provides an appropriate mathematical model to predict water pollution in the following months.

Currently, there are a number of models predicting water pollution such as models QUAL2K, IPC model, ... Many water quality models were developed over the past years for various types of water bodies. QUAL2E water quality model developed during the earlier stages had many limitations. To overcome those limitations, QUAL2K was developed by Park and Lee in 2002 [1]. Model QUAL2K is a version of the model QUAL2E [2]. This model was developed due to the cooperation between Tufts University and the US Environmental Environment Center (US.EPA). The model is widely used to predict river water quality developments and predict the load of waste into rivers. IPC model developed by the World Bank, the World Health Organization and the US health organization. IPC model assesses river water quality, forecast changes in river water quality, calculates the amount to be cut by each source.

## 2 Materials and Methods

Dong Nai River is the longest inland river in Vietnam, the second largest in the South to the basin, just behind the Mekong River. The Dong Nai river flows through Lam Dong, Dak Nong, Binh Phuoc, Dong Nai, Binh Duong and Ho Chi Minh cities with a length of 586 km (364 miles) and a basin of 38,600 km<sup>2</sup> (14,910 mi<sup>2</sup>). If calculating from the beginning of the Da Dang river source, it is long 586 km. If calculating from the confluence point with Da Nhim river under Pongour waterfall, it is long 487 km. Dong Nai river flows into the East Sea in Can Gio district. 89 data were collected from Tan Hiep automatic water monitoring station on Dong Nai river, continuously monitored daily from February 1<sup>st</sup>, 2018 to April 31, 2018 (see Table 1).

**Table 1.** Pollution data of TSS water at Tan Hiep station.

Time	TSS (mg/l)	Time	TSS (mg/l)	Time	TSS (mg/l)
2/1/2018	21	3/1/2018	19	4/1/2018	24
2/2/2018	19	3/2/2018	16	4/2/2018	26
.....	.....	.....	.....	.....	.....
2/28/2018	15	3/28/2018	23	4/28/2018	30
		3/29/2018	19	4/29/2018	31
		3/30/2018	15	4/30/2018	37
		3/31/2018	20		

TSS parameters (turbidity and suspended solids) are total suspended solids. Usually, it is measured with a turbidity meter (turbidimeter). Turbidity is caused by the interaction between light and suspended solids in water such as sand, clay, algae, microorganisms and organic matter in water. Suspended solids disperse light or absorb them and re-emit them in the manner depending on the size, shape and composition of suspended particles and thus allow application turbidity measuring devices to reflect the change in the type,

size and concentration of the particles in the sample, etc. The author uses a geostatistical method to predict the concentration of TSS water pollution in the next time.

The main tool in geostatistics is the variogram which expresses the spatial dependence between neighbouring observations. The variogram  $\gamma(h)$  can be defined as one-half the variance of the difference between the attribute values at all points separated by has followed [3, 8]

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i + h)]^2 \tag{1}$$

where  $Z(s)$  indicates the magnitude of the variable, and  $N(h)$  is the total number of pairs of attributes that are separated by a distance  $h$ .

Under the second-order stationary conditions [4, 9] one obtains

$$E[Z(s)] = \mu$$

and the covariance

$$Cov[Z(s), Z(s + h)] = E[(Z(s) - \mu)(Z(s + h) - \mu)] = E[Z(s)Z(s + h) - \mu^2] = C(h) \tag{2}$$

Then  $\gamma(h) = \frac{1}{2}E[Z(s) - Z(s + h)]^2 = C(0) - C(h)$

The most commonly used models are spherical, exponential, Gaussian, and pure nugget effect [5, 8]. The adequacy and validity of the developed variogram model is tested satisfactorily by a technique called cross-validation.

Crossing plot of the estimate and the true value shows the correlation coefficient  $r^2$ . The most appropriate variogram was chosen based on the highest correlation coefficient by trial and error procedure.

Kriging technique is an exact interpolation estimator used to find the best linear unbiased estimate. The best linear unbiased estimator must have a minimum variance of estimation error. We used ordinary kriging for spatial and temporal analysis. Ordinary kriging method is mainly applied for datasets without and with a trend.

The general equation of linear kriging estimator is

$$\hat{Z}(s_0) = \sum_{i=1}^n w_i Z(s_i) \tag{3}$$

In order to achieve unbiased estimations in ordinary kriging the following set of equations should be solved simultaneously.

$$\begin{cases} \sum_{i=1}^n w_i \gamma(s_i, s_j) - \lambda = \gamma(s_0, s_i) \\ \sum_{i=1}^n w_i = 1 \end{cases} \tag{4}$$

where  $\hat{Z}(s_0)$  is the kriged value at location  $s_0$ ,  $Z(s_i)$  is the known value at location  $s_i$ ,  $w_i$  is the weight associated with the data,  $\lambda$  is the Lagrange multiplier, and  $\gamma(s_i, s_j)$  is the value of variogram corresponding to a vector with origin in  $s_i$  and extremity in  $s_j$ .

Kriging minimizes the mean squared error of prediction

$$\min \sigma_e^2 = \mathbb{E}[Z(s_0) - \hat{Z}(s_0)]^2$$

For second order stationary process the last equation can be written as

$$\sigma_e^2 = C(0) - 2 \sum_{i=1}^n w_i C(s_0, s_i) + \sum_{i=1}^n \sum_{j=1}^n w_i w_j C(s_i, s_j) \text{ subject to } \sum_{i=1}^n w_i = 1 \quad (5)$$

Therefore the minimization problem can be written as

$$\min \left\{ C(0) - 2 \sum_{i=1}^n w_i C(s_0, s_i) + \sum_{i=1}^n \sum_{j=1}^n w_i w_j C(s_i, s_j) - 2\lambda \left( \sum_{i=1}^n w_i - 1 \right) \right\} \quad (6)$$

where  $\lambda$  is the Lagrange multiplier. After differentiating (6) with respect to  $w_1, w_2, \dots, w_n$ , and  $\lambda$  and set the derivatives equal to zero we find that

$$\sum_{j=1}^n w_j C(s_i, s_j) - C(s_0, s_i) - \lambda = 0, \quad i = 1, 2, \dots, n \text{ and } \sum_{i=1}^n w_i = 1$$

Using matrix notation the previous system of equations can be written as

$$\begin{pmatrix} C(s_1, s_1) & C(s_1, s_2) & \dots & C(s_1, s_n) & 1 \\ C(s_2, s_1) & C(s_2, s_2) & \dots & C(s_2, s_n) & 1 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ C(s_n, s_1) & C(s_n, s_2) & \dots & C(s_n, s_n) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ -\lambda \end{pmatrix} = \begin{pmatrix} C(s_0, s_1) \\ C(s_0, s_2) \\ \vdots \\ C(s_0, s_n) \\ 1 \end{pmatrix}$$

Therefore the weights  $w_1, w_2, \dots, w_n$  and the Lagrange multiplier  $\lambda$  can be obtained by

$$\mathbf{W} = \mathbf{C}^{-1} \mathbf{c}$$

where  $\mathbf{W} = (w_1, w_2, \dots, w_n, -\lambda)$

$$\mathbf{c} = (C(s_0, s_1), C(s_0, s_2), \dots, C(s_0, s_n), 1)'$$

$$C = \begin{cases} C(s_i, s_j), & i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n, \\ 1, & i = n + 1, \quad j = 1, 2, \dots, n, \\ 1, & i = 1, 2, \dots, n, \quad j = n + 1, \\ 0, & i = n + 1, \quad j = n + 1. \end{cases}$$

The GS+ software (version 5.1.1) was used for geostatistical analysis in this study [6].

### 3 Results and Discussions

In order to check the anisotropy of TSS, the conventional approach is to compare variograms in several directions [7]. In this study major angles of 0°, 45°, 90°, and 135° with an angle tolerance of ±45° were used for detecting anisotropy.

Figure 1 shows fitted variogram for spatial analysis of TSS. Gaussian model [Nugget = 6.5 (mg/l); Sill = 64 (mg/l); Range = 95 (mg/l); r<sup>2</sup> = 0.969, and RSS = 101]. It shows the best fitted omnidirectional variogram of water pollution obtained based on cross-validation. Through variogram map of parameter TSS, the model of isotropic is suitable. The variogram values are presented in Table 2.

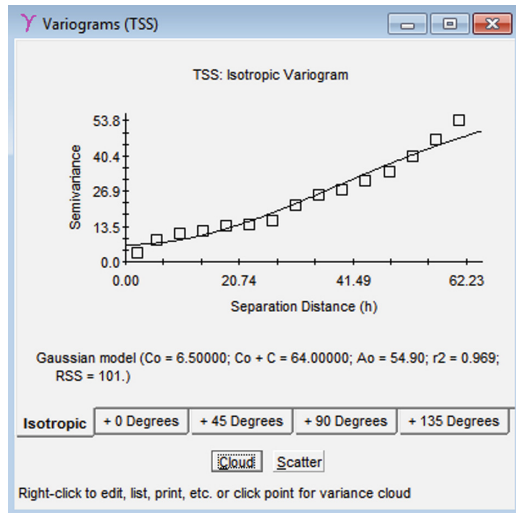


Fig. 1. Model of isotropic variogram for TSS parameters.

Table 2. Variogram values of TSS.

	Nugget	Sill	Range	r <sup>2</sup>	RSS
Linear	0	46.75	60.79	0.951	152
Gaussian	6.5	64	95	0.969	101
Spherical	0.2	61.4	114	0.932	215
Exponential	0.1	61.2	169.2	0.882	410

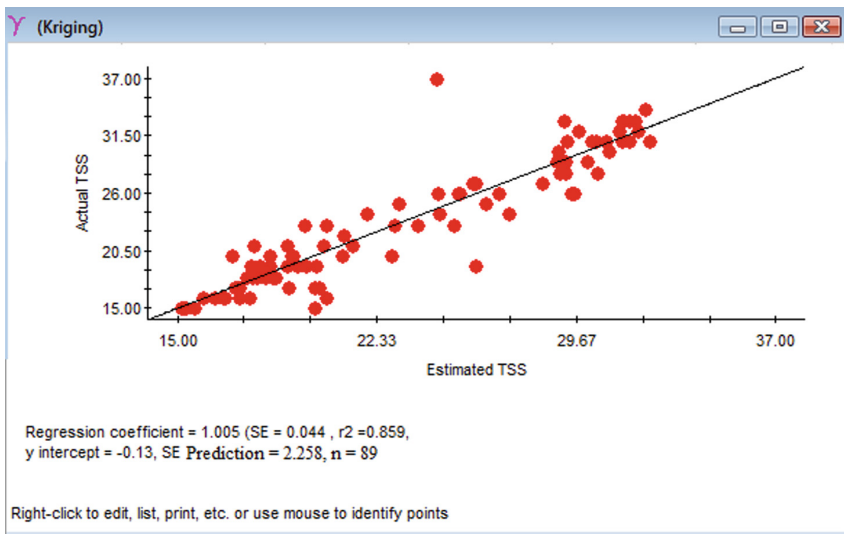
Residual Sums of Squares (RSS) provides an exact measure of how well the model fits the variogram data; the lower the reduced sums of squares, the better the model fits. When GS+ autofits the model, it uses RSS to choose parameters for each of the variogram models by determining the combination of parameter values that minimizes RSS for any given model. The Residual SS displayed in the This Fit box is calculated for the currently defined model.

$r^2$  provides an indication of how well the model fits the variogram data; this value is not as sensitive or robust as the Residual SS value for best-fit calculations; use RSS to judge the effect of changes in model parameters.

Model Testing: The reliable result of model selection using appropriate interpolation is expressed in Table 3 by coefficient of regression, coefficient of correlation and interpolated values, in addition to the error values as the standard error (SE) and the standard error prediction (SE Prediction) [10, 11].

**Table 3.** Testing the model parameters.

Coefficient regression	Coefficient correlation	SE	SE Prediction
1.005	0.859	0.044	2.258



**Fig. 2.** Error testing result of prediction TSS.

Figure 2 shows the results of testing the error between the estimated value and the actual value by interpolation method kriging with isotropic TSS. The regression coefficient is 1.005, the correlation coefficient is close to 0.859 (the best result is 1), the standard error is 0.044 (close to 0) and the forecast error of 2.258 shows the choice of Kriging interpolation model in accordance with the set data in Fig. 3.

Y Cross Validation (Kriging) a)						Y Cross Validation (Kriging) b)					
Record	X-Coordinate	Y-Coordinate	Actual Z	Estimated Z	Error (E-A)	Record	X-Coordinate	Y-Coordinate	Actual Z	Estimated Z	Error (E-A)
1	1.00	1.00	24.00	27.18	3.18	34	34.00	34.00	20.00	19.18	-0.82
2	2.00	2.00	26.00	26.78	0.78	35	35.00	35.00	21.00	20.33	-0.67
3	3.00	3.00	28.00	29.05	1.05	36	36.00	36.00	21.00	21.44	0.44
4	4.00	4.00	33.00	29.21	-3.79	37	37.00	37.00	23.00	20.43	-2.57
5	5.00	5.00	31.00	31.32	0.32	38	38.00	38.00	20.00	21.02	1.02
6	6.00	6.00	32.00	29.71	-2.29	39	39.00	39.00	20.00	19.22	-0.78
7	7.00	7.00	28.00	30.44	2.44	40	40.00	40.00	17.00	20.18	3.18
8	8.00	8.00	29.00	29.23	0.23	41	41.00	41.00	21.00	17.79	-3.21
9	9.00	9.00	29.00	30.03	1.03	42	42.00	42.00	17.00	20.05	3.05
10	10.00	10.00	31.00	30.19	-0.81	43	43.00	43.00	19.00	20.11	1.11
11	11.00	11.00	31.00	31.58	0.58	44	44.00	44.00	22.00	21.09	-0.91
12	12.00	12.00	33.00	31.35	-1.65	45	45.00	45.00	23.00	22.94	-0.06
13	13.00	13.00	32.00	31.92	-0.08	46	46.00	46.00	25.00	23.13	-1.87
14	14.00	14.00	31.00	32.35	1.35	47	47.00	47.00	23.00	25.14	2.14
15	15.00	15.00	33.00	31.36	-1.64	48	48.00	48.00	26.00	24.57	-1.43
16	16.00	16.00	32.00	31.22	-0.78	49	49.00	49.00	26.00	25.36	-0.64
17	17.00	17.00	31.00	29.29	-1.71	50	50.00	50.00	25.00	26.31	1.31
18	18.00	18.00	26.00	29.46	3.46	51	51.00	51.00	27.00	25.86	-1.14
19	19.00	19.00	27.00	28.39	1.39	52	52.00	52.00	27.00	25.94	-1.06
20	20.00	20.00	29.00	28.95	-0.05	53	53.00	53.00	26.00	25.28	-0.72
21	21.00	21.00	30.00	30.83	0.83	54	54.00	54.00	24.00	24.60	0.60
22	22.00	22.00	33.00	31.61	-1.39	55	55.00	55.00	23.00	23.81	0.81
23	23.00	23.00	34.00	32.20	-1.80	56	56.00	56.00	24.00	21.93	-2.07
24	24.00	24.00	33.00	31.81	-1.19	57	57.00	57.00	20.00	22.84	2.84
25	25.00	25.00	31.00	30.39	-0.61	58	58.00	58.00	23.00	19.66	-3.34
26	26.00	26.00	28.00	29.24	1.24	59	59.00	59.00	19.00	19.41	0.41
27	27.00	27.00	26.00	29.55	3.55	60	60.00	60.00	15.00	20.01	5.01
28	28.00	28.00	30.00	26.99	-1.01	61	61.00	61.00	20.00	18.38	-1.62
29	29.00	29.00	31.00	30.74	-0.26	62	62.00	62.00	21.00	19.03	-1.97
30	30.00	30.00	37.00	24.48	-12.52	63	63.00	63.00	19.00	19.72	0.72
31	31.00	31.00	19.00	25.93	6.93	64	64.00	64.00	19.00	19.03	0.03
32	32.00	32.00	16.00	20.48	4.48	65	65.00	65.00	19.00	18.40	-0.60
33	33.00	33.00	17.00	19.07	2.07	66	66.00	66.00	18.00	17.91	-0.09

Y Cross Validation (Kriging) c)					
Record	X-Coordinate	Y-Coordinate	Actual Z	Estimated Z	Error (E-A)
67	67.00	67.00	17.00	17.15	0.15
68	68.00	68.00	16.00	16.70	0.70
69	69.00	69.00	16.00	16.34	0.34
70	70.00	70.00	16.00	16.71	0.71
71	71.00	71.00	17.00	17.11	0.11
72	72.00	72.00	18.00	17.73	-0.27
73	73.00	73.00	19.00	17.71	-1.29
74	74.00	74.00	18.00	17.75	-0.25
75	75.00	75.00	17.00	17.25	0.25
76	76.00	76.00	16.00	17.27	1.27
77	77.00	77.00	17.00	17.18	0.18
78	78.00	78.00	18.00	17.50	-0.50
79	79.00	79.00	18.00	18.22	0.22
80	80.00	80.00	19.00	17.98	-1.02
81	81.00	81.00	18.00	18.47	0.47
82	82.00	82.00	18.00	18.57	0.57
83	83.00	83.00	20.00	16.99	-3.01
84	84.00	84.00	16.00	17.63	1.63
85	85.00	85.00	16.00	15.95	-0.05
86	86.00	86.00	15.00	15.63	0.63
87	87.00	87.00	15.00	15.18	0.18
88	88.00	88.00	15.00	15.10	0.10
89	89.00	89.00	15.00	15.29	0.29

Fig. 3. Cross-Validation (Kriging) (a), (b) và (c) của TSS.

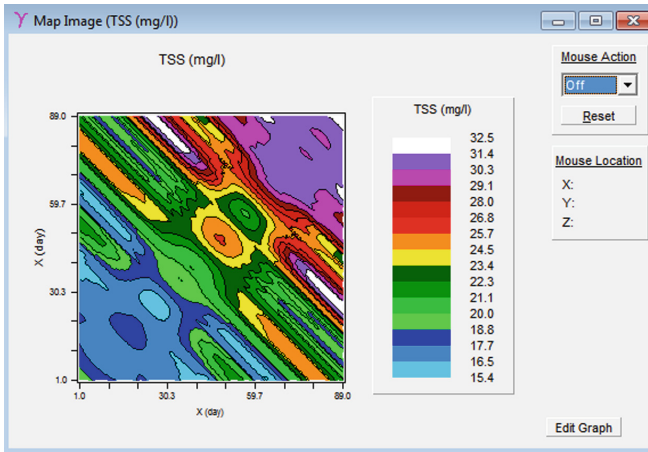


Fig. 4. Kriging interpolation for TSS parameters in 2 dimensions.

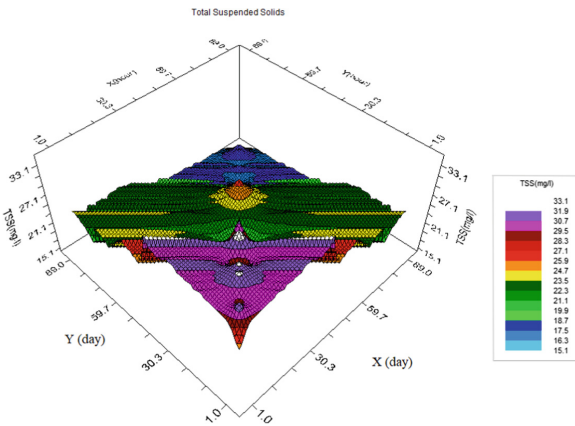


Fig. 5. Kriging interpolation for TSS parameters in 3 dimensions.

From Figs. 4 and 5 we see that from February to April (from 1 to 89 days), in February there is the lowest concentration of TSS and gradually increases to April this is also consistent with the fact because February is the dry season and April is the beginning of the rainy season, so TSS increases. The X-axis and Y-axis represent the number of days (starting from February 1<sup>st</sup>, 2018 to April 30, 2018, which means 89 days). Based on Figs. 4 and 5, we can predict the TSS contamination concentration in next month (May) and offer remedial solutions.

The application of geotechnical methods mentioned to predict the concentration of TSS pollution at Tan Hiep station shows that the forecast results are small errors as shown in Fig. 2. Through this forecast study, methods and tissues are used. Based on interpolation, we can predict the level of TSS pollution levels for the following months without monitoring data, thus suggesting measures to improve and protect the environment.



From the forecast map, we find that the forecast gives the best results in the 89 days period, outside of this time the forecast results may be inaccurate. The more time the pollutant observes, the easier it is to select interpolation models, with higher interpolation results and vice versa. Different colors show different levels of pollution. The lowest level of pollution is blue and the highest is white. Areas of the same color have the same level of pollution. The results of the model still have this error, which may be due to many other factors affecting TSS parameters such as salinity, temperature, nitrate content, water flow ... This is the first article the author uses the method. Kriging interpolation to predict water pollution over time.

## 4 Results and Discussions

The statistical applications for predicting TSS concentrations in rivers at the Tan Hiep monitoring station have resulted in small errors between estimated values and real values (standard errors equal to 0.044 and projected errors reported by 2.258). Since then, the study has shown the effectiveness and rationality with the high reliability of geostatistics to build appropriate predictive models. When building the model, author should pay attention to the error values of the model, the data characteristics of the object. Author also consider the results of the model selection in order to select the most suitable model for the actual data, because separate models provide different accuracy. Therefore, the experience of selecting models also plays a very important role in interpolation results. Finally a comparison of the proposed method with several other methods can be made as follows. Polygon (nearest neighbor) method has advantages such as easy to use, quick calculation in 2D; but also possesses many disadvantages as discontinuous estimates; edge effects/sensitive to boundaries; difficult to realize in 3D. The Triangulation method has advantages as easy to understand, fast calculations in 2D; can be done manually, but few disadvantages are triangulation network is not unique. The use of Delaunay triangles is an effort to work with a "standard" set of triangles, not useful for extrapolation and difficult to implement in 3D. Local sample mean has advantages are easy to understand; easy to calculate in both 2D and 3D and fast; but disadvantages possibly are local neighborhood definition is not unique, location of sample is not used except to define local neighborhood, sensitive to data clustering at data locations. This method does not always return answer valuable. This method is rarely used. Similarly, the inverse distance method are easy to understand and implement, allow changing exponent adds some flexibility to method's adaptation to different estimation problems. This method can handle anisotropy; but disadvantages are difficulties encountered when point to estimate coincides with data point ( $d = 0$ , weight is undefined), susceptible to clustering.

This paper, QUAL2K is not suitable, because QUAL2K has been used to predict pollution on the river section and it has not been applied to the forecast of pollution over time.

**Acknowledgment.** The paper's author expresses his sincere thank to Dr. Man N.V. Minh Department of Mathematics, Faculty of Science, Mahidol University, Thai Lan and Dr. Dung Ta Quoc Faculty of Geology and Petroleum Engineering, Vietnam. Furthermore, author greatly appreciate the anonymous reviewer whose valuable and helpful comments led to significant improvements from the original to the final version of the article.

## References

1. Park, S.S., Lee, Y.S.: A water quality modeling study of the Nakdong River, Korea. *Ecol. Model.* **152**(1), 65–75 (2002)
2. Ashwani, S., Vivek, B., Ratnoji, S., Jayakumar, P., Jainet, P.J.: Application of Qual2K model for prediction of water quality in a selected stretch of Pamba River. *Int. J. Civil Eng. Technol. (IJCIET)* **8**(6), 75–84 (2017)
3. Ahmadi, S.H., Sedghamiz, A.: Geostatistical analysis of spatial and temporal variations of groundwater level. *Environ. Monit. Assess.* **129**, 277–294 (2007)
4. Webster, R., Oliver, M.A.: *Geostatistics for Environmental Scientists*, 2nd edn, pp. 6–8. Wiley, Chichester (2007)
5. Isaaks, E., Srivastava, M.R.: *An Introduction to Applied Geostatistics*. Oxford University Press, New York (1989)
6. Gamma Design Software: *GS+ Geostatistics for the Environmental Science*. Gamma Design Software, LLC, Plainwell (2001). version 5.1.1
7. Goovaerts, P.: *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York (1997)
8. Kitadinis, P.K.: *Introduction to Geostatistics: Applications to Hydrogeology*. Cambridge University Press, Cambridge (2003)
9. Gentile, M., Courbin, F., Meylan, G.: Interpolating point spread function anisotropy. *Astron. Astrophys.* **549**, A1 (2012). manuscript no. psf interpolation
10. Nhut, N.C., Nguyen, M.V.M.: Analyzing incomplete spatial data in air pollution prediction. *J. SE-Asian J. Sci.* **6**(2), 111–133 (2018). ISSN 2286–7724
11. Nhut, N.C.: Metropolitan air pollution prediction: A case study using PM10 data observed in Ho Chi Minh City. *Báo cáo tại Hội nghị Khoa học Công Nghệ, Đại học Hoa Sen* (2016)