




Applying Geostatistics to Predict Dissolvent Oxygen (DO) in Water on the Rivers in Ho Chi Minh City

Cong Nhut Nguyen^(✉) 

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

Abstract. Geostatistics is briefly concerned with estimation and prediction for spatially continuous phenomena, using data measured at a finite number of spatial locations to estimate values of interest at unmeasured locations. In practice, the costs of installing new observational stations to observe metropolitan water pollution sources, as DO (Dissolvent Oxygen), COD (Chemical Oxygen Demand) and BOD (Biochemical oxygen Demand) concentrations are economically high. In this study, spatial analysis of water pollution of 32 stations monitored during 3 years was carried out. Geostatistics which has been introduced as a management and decision tool by many researchers has been applied to reveal the spatial structure of water pollution fluctuation. In this article, author use the recorded DO concentrations (is the amount of dissolvent oxygen in water required for the respiration of aquatic organisms) at several observational stations on the rivers in Ho Chi Minh City (HCMC), employ the Kriging interpolation method to find suitable models, then predict DO concentrations at some unmeasured stations in the city. Our key contribution is finding good statistical models by several criteria, then fitting those models with high precision. From the data set, author found the best forecast model with the smallest forecast error to predict DO concentration on rivers in Ho Chi Minh City. From there we propose to the authorities to improve areas where DO concentrations exceed permissible levels.

Keywords: Geostatistics · Interpolation · Kriging · Spatial · Variogram

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 HCMC), network quality monitoring water environment of HCMC has 32 stations observation on water in the rivers in HCMC. However, with a large area, the city needs to install more new monitoring stations. The cost to install a new machine costs tens of billions VND, and

the maintenance is also difficult. Therefore, the requirements are based on the remaining monitoring stations using mathematical models based to predict air pollution concentration at some unmeasured stations in the city.

Sources of water pollution are diverse. Many industrial zones, industrial plants and urban areas have discharged untreated wastewater to rivers and lakes which has polluted water sources severely. As a result, water sources in many areas cannot be used. Socio-economic development in each river basin is different and the contribution of pollutants to the environment from different sectors also varies. However, the pressure of waste water mainly comes from industrial and domestic activities. Waste water discharged from industrial establishments and industrial zones exerts the greatest pressure on the surface water environment in the country. Agriculture is the largest user of water, mainly for the irrigation of rice and other water intensive crops. Consequently, waste water discharged by agricultural activities into surface water makes up the largest proportion. Quantity of pollutants from untreated urban waste water. There is an increasing demand for running water in urban areas to meet the need of population growth and the development of urban services. Currently, most cities do not have a treatment system for domestic waste water. In those cities which have this system, the rate of treated waste water is much lower than required. Untreated domestic waste water from residential and tourism areas and discharged by small industrial and handicraft establishments are the major cause of pollution to water sources within cities and their outskirts.

The study area is HCMC in South of Vietnam. It is located between $10^{\circ}10' - 10^{\circ}38'$ northing and $106^{\circ}22' - 106^{\circ}54'$ easting and the area has more than 2096 km^2 (2018). HCMC has more than 9 million people (2018). Figure 1 shows the study area. The city has a tropical climate, specifically a tropical wet and dry climate, with an average humidity of 78–82%. The average temperature is 28°C (82°F) (degrees Fahrenheit).

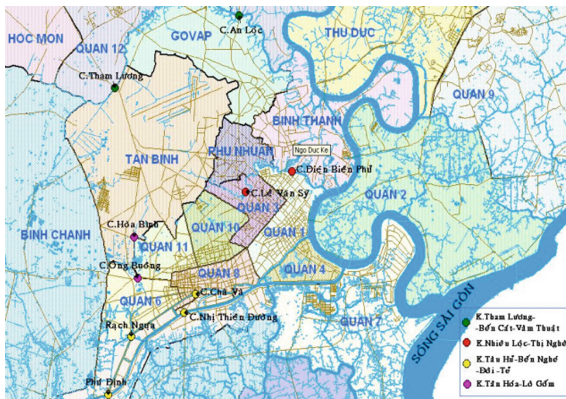


Fig. 1. Location of the study area^a. a. Department of natural resources and environment HCMC.

With the rapid population growth rate, the infrastructure has not yet been fully upgraded, and some people are too aware of environmental protection. So, HCMC is currently facing a huge environmental pollution problem. The status of untreated wastewater flowing directly into the river system is very common. Many production facilities, hospitals and health facilities that do not have a wastewater treatment system are alarming. Figure 2 shows the geographical location of the monitoring stations. The coordinates system used in Fig. 2 is Universal Transverse Mercator (UTM).

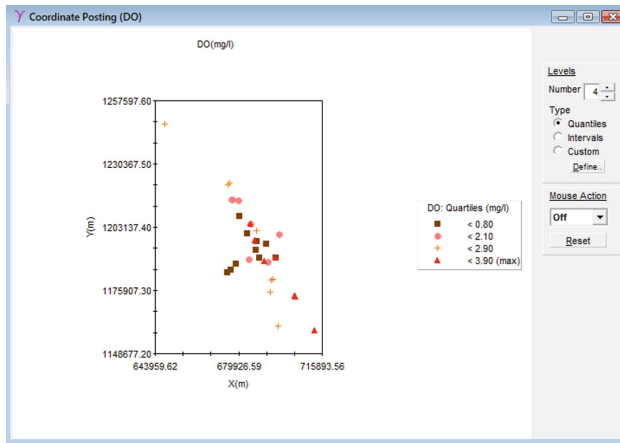


Fig. 2. Map of water quality monitoring stations in HCMC.

2 Materials and Methods

Dissolved oxygen (DO) refers to the level of free, non-compound oxygen present in water or other liquids. It is an important parameter in assessing water quality because of its influence on the organisms living within a body of water. In limnology (the study of lakes), dissolved oxygen is an essential factor second only to water itself. A dissolved oxygen level that is too high or too low can harm aquatic life and affect water quality. The dataset is obtained from monitoring stations in the rivers HCMC with these parameter DO. Figure 2 shows map of water quality monitoring stations in HCMC. DO data of water environment measures 32 stations from 2015 to 2017, (see Table 1). Author applied a geostatistical method to predict concentrations of air pollution at unobserved areas surrounding observed ones.

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 [1, 6]

Table 1. DO data of water quality monitoring stations in HCMC.

Station	x(m)	y(m)	DO (mg/l)
Ba Son	687020.74	1193517.41	0.60
Den Do Apex	692372.50	1188205.59	2.10
Cat Lai Pier	695674.23	1190158.06	3.40
Rach Chiec-Sai Gon	691502.97	1196219.97	0.80
Phu Dinh Port	676558.28	1184762.57	0.20
Binh Khanh Ferry	693943.68	1180318.17	2.80
VCD-Binh Dien Bridge	674736.35	1183824.89	0.20
Tam Thon Hiep	704119.33	1173806.02	3.10
Soai Rap River	693691.06	1175042.20	2.30
Tan Thuan Port	688506.94	1190249.21	0.80
Phu Long Bridge	685004.43	1204724.99	1.50
Hoa Phu Pump station	676867.55	1215207.46	1.80
Bridge Binh Trieu	687447.50	1197076.03	0.70
Lo Gom Bridge	678772.16	1187429.76	0.70
Chu Y Bridge	684059.70	1189290.14	0.90
An Loc Bridge	683576.38	1200370.94	0.70
Cai Stream	697408.50	1200142.76	1.10
Cat Lai	695671.18	1190161.11	0.70
Thi Tinh River	675253.16	1221229.09	2.20
Binh Loi Bridge	686955.30	1197608.09	3.70
Phu My	690858.99	1188710.28	3.00
Rach Tra	680156.73	1207934.77	0.70
Trung An	676079.87	1222198.63	2.20
Phu Cuong	679609.36	1214736.71	1.70
Hao Phu	677250.67	1215117.32	1.30
Phu Long	685004.36	1204737.28	3.20
Tam Thon Hiep	704291.61	1173475.08	3.70
Vam Co	712393.56	1158677.20	3.90
Binh Phuoc	687747.10	1201605.25	2.90
Vam Sat	696879.34	1160493.97	2.50
Nha Be	694496.87	1180871.54	2.20

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

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 [2, 7] 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 [3, 6]. 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. Author 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, λ 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

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

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 λ is the Lagrange multiplier. After differentiating (6) with respect to w_1, w_2, \dots, w_n , and λ 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 λ can be obtained by

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

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

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

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

The GS+ software (version 5.1.1) was used for geostatistical analysis in this study [4].

3 Results and Discussions

In order to check the anisotropy of TSS, the conventional approach is to compare variograms in several directions [5]. In this study major angles of $0^\circ, 45^\circ, 90^\circ$, and 135° with an angle tolerance of $\pm 45^\circ$ were used for detecting anisotropy.

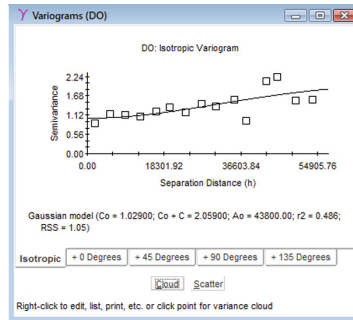


Fig. 3. Fitted variogram for the spatial analysis of parameter DO.

Figure 3 shows fitted variogram for spatial analysis of DO. Gaussian model [Nugget = 0.1 (mg/l); Sill = 2 (mg/l); Range = 75864 (mg/l); $r^2 = 0.486$]. It shows the best fitted omnidirectional variogram of water pollution obtained based on cross-validation. Through variogram map of parameter DO, the model of isotropic is suitable. The variogram values are presented in Table 2.

Table 2. Isotropic variogram values of DO.

	Nugget	Sill	Range	r^2	RSS
Linear	0.9	1.8	53583	0.485	1.05
Gaussian	1	2	75864	0.486	1.05
Spherical	0.22	1.421	4500	0.136	1.77
Exponential	0.883	2.471	201600	0.484	1.06

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).

Table 3. Testing the model parameters.

Coefficient regression	Coefficient correlation	SE	SE Prediction
0.936	0.205	0.336	1.001

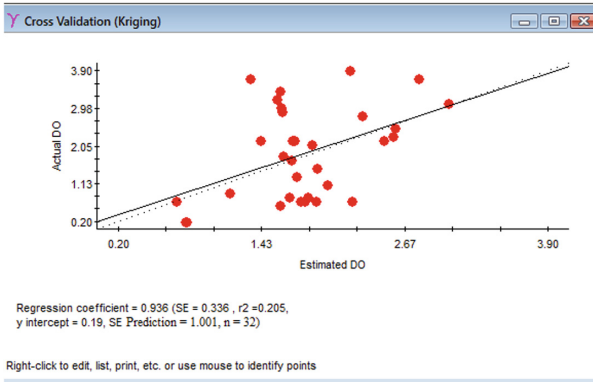


Fig. 4. Error testing result of prediction DO.

Figure 4 shows results of testing of error between real values and the estimated values by the model by cokriging method with isotropic DO. Coefficients of regression are close to 1, where the error values is small (close to 0) indicates that the selected model is a suitable interpolation in Fig. 5.

Record#	X-Coordinate	Y-Coordinate	Actual Z	Estimated Z	Error (E-A)
1	687020.74	1193517.41	0.60	1.59	0.99
2	692372.50	1186205.59	2.10	1.87	-0.23
3	695874.23	1190156.06	3.40	1.59	-1.81
4	691502.97	1196219.97	0.80	1.82	1.02
5	676556.28	1184762.57	0.20	0.77	0.57
6	693943.68	1180318.17	2.80	2.29	-0.51
7	674736.35	1183824.89	0.20	0.79	0.59
8	704119.33	1173806.02	3.10	3.04	-0.06
9	693691.06	1175042.20	2.30	2.56	0.26
10	685506.94	1190249.21	0.80	1.67	0.87
11	685004.43	1204724.99	1.50	1.91	0.41
12	676867.55	1215207.46	1.80	1.61	-0.19
13	687447.50	1197076.03	0.70	1.76	1.06
14	678772.16	1187429.76	0.70	0.70	-0.00
15	684059.70	1189290.14	0.90	1.16	0.26
16	683576.38	1200370.94	0.70	1.80	1.10
17	697408.50	1200142.76	1.10	1.99	0.89
18	695671.18	1190161.11	0.70	2.21	1.51
19	675253.16	1221229.09	2.20	1.70	-0.50
20	686955.30	1197608.09	3.70	1.33	-2.37
21	690858.99	1188710.28	3.00	1.60	-1.40
22	680156.73	1207934.77	0.70	1.90	1.20
23	676079.87	122198.63	2.20	1.72	-0.48
24	679609.36	1214736.71	1.70	1.69	-0.01
25	677250.67	1215117.32	1.30	1.73	0.43
26	685004.36	1204737.28	3.20	1.56	-1.64
27	704291.61	1173475.06	3.70	2.78	-0.92
28	712393.56	1158677.20	3.90	2.19	-1.71
29	687747.10	1201605.25	2.90	1.61	-1.29
30	696879.34	1160493.97	2.50	2.58	0.08
31	694496.87	1180871.54	2.20	2.48	0.28
32	647959.62	1247597.60	2.20	1.43	-0.77

Fig. 5. Cross-Validation (Kriging) of DO.

From Figs. 6 and 7, we see that, from 2015 to 2017 at Phu Dinh Port and Vam Co Dong - Binh Dien Bridge neighborhood has low pollution levels. Neighborhood of Vam Co have high pollution levels. The Sai Gon river basin extends through many provinces and is strongly impacted by different pollution sources. The major pollution sources derive from industrial activities. The surface water in sections which run

through provinces in the southern key socio-economic area where many industrial zones and towns are located is badly polluted. Saigon River begins to be polluted by organic matter and microorganisms from the Thi Tinh river estuary and the pollution increases in its lower section. The section crossing Ho Chi Minh City is particularly affected by organic matter. The content of BOD₅, COD and bioorganisms all fail to meet the standards set for surface water as a source to supply drinking water.

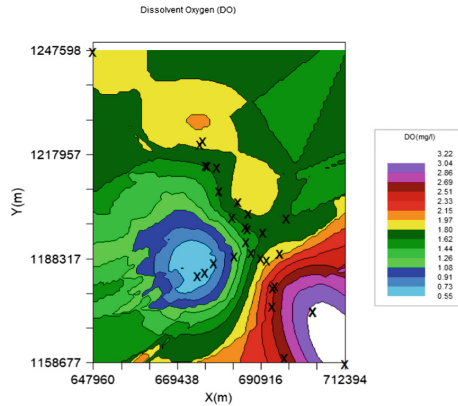


Fig. 6. 2D Cokriging Interpolation Map of DO.

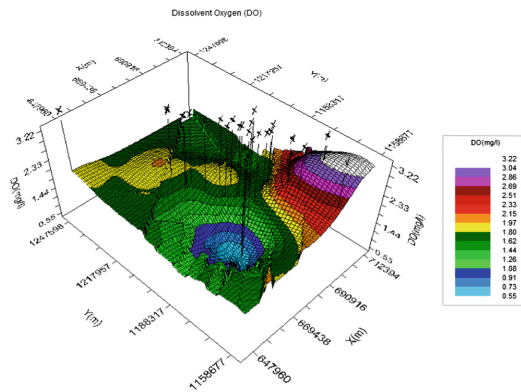


Fig. 7. 3D Cokriging Interpolation Map of DO.

Based on the map, we can also predict the water pollution concentration in the city near the air monitoring locations and to offer solutions to overcome. The mentioned method of applied geostatistics to predict the water pollution concentrations DO on the rivers in HCMC showed that the forecast regions closer together have the forecast deviations as small Fig. 8, meanwhile further areas contribute the higher deviation. Through this forecast case study using spatial interpolation based methods and models, we can predict air pollution levels for regions that have not been installed air

monitoring sites, from which proposed measures to improve the air quality can be taken into account. If the density of monitoring stations is high and the selection of interpolation models is easier, interpolation results have higher reliability and vice versa.

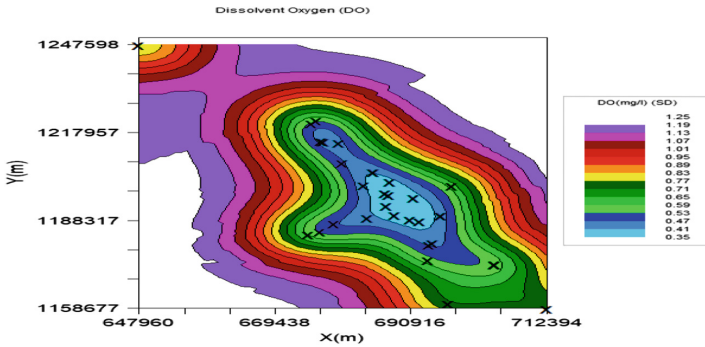


Fig. 8. Estimated error by Cokriging method of DO.

From the forecast maps, we see that, forecast for the region’s best results in areas affected 88921 m, located outside the affected region on the forecast results can be inaccurate. If the density of monitoring stations is high and the selection of interpolation models is easier, interpolation results have higher reliability and vice versa. The middle area represents key outcomes of computation on data. The different colors represent different levels of pollution. The lowest pollution level is blue and the highest is white. Regions having the same color likely are in the same levels of pollution.

Limitations of the article: The forecast results of the model also have errors due to other DO pollutants in the water also have other pollutants such as BOD₅, COD, TSS, ... So in the future, author will study the influence of other pollutants on the DO to reduce the error in the forecast.

4 Conclusion

Geostatistical applications to predict DO concentrations on the rivers in HCMC gave the result with almost no error difference between the estimated values and the real values. Therefrom, the study showed that efficacy and rationality with high reliability of theoretical Geostatistical to building spatial prediction models are suitable. When building the model we should pay attention to the values of the model error, data characteristic of the object. We also looked at the result of the model selection which aimed to choose the most suitable model for real facts, since distinct models provide different accuracies. Therefore, experiencing the selected model also plays a very important role in the 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.

In Vietnam, the modelling methods used the more common, especially in the current conditions of our country. The tangled diffusion model of Berliand and Sutton was used by Anh Pham Thi Viet to assess the environmental status of the atmosphere of Hanoi in 2001 by industrial discharges [8]. In 2014, Yen Doan Thi Hai has used models Meti-lis to calculate the emission of air pollutants from traffic and industrial activities in Thai Nguyen city [9].

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. Ahmadi, S.H., Sedghamiz, A.: Geostatistical analysis of spatial and temporal variations of groundwater level. *Environ. Monit. Assess.* **129**, 277–294 (2007)
2. Webster, R., Oliver, M.A.: *Geostatistics for Environmental Scientists*, 2nd edn, pp. 6–8. Wiley, Chichester (2007)
3. Isaaks, E., Srivastava, M.R.: *An Introduction to Applied Geostatistics*. Oxford University Press, New York (1989)
4. Gamma Design Software: *GS+ Geostatistics for the Environmental Science*. Gamma Design Software, LLC, Plainwell (2001)
5. Goovaerts, P.: *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York (1997)
6. Kitadinis, P.K.: *Introduction to Geostatistics: Applications to Hydrogeology*. Cambridge University Press, Cambridge (2003)
7. Gentile, M., Courbin, F., Meylan, G.: Interpolating point spread function anisotropy. *Astron. Astrophys.* **549**, A1 (2012). manuscript no. psf interpolation
8. Anh, P.T.V.: Application of airborne pollutant emission models in assessing the current state of the air environment in Hanoi area caused by industrial sources. In: 6th Women’s Science Conference, Ha Noi national university, pp. 8–17 (2001)
9. Yen, D.T.H.: Applying the Meti-lis model to calculate the emission of air pollutants from traffic and industrial activities in Thai Nguyen city, orienting to 2020. *J. Sci. Technol.* **106**(6) (2013). Thai Nguyen university