Environ Earth Sci (2017) 76:72
DOI 10.1007/s12665-017-6394-2

ORIGINAL ARTICLE

Geostatistical prediction of heavy metal concentrations in stream sediments considering the stream networks

Sung-Min Kim[1] · Yosoon Choi[2] · Huiuk Yi[3] · Hyeong-Dong Park[1]

Received: 8 August 2016 / Accepted: 3 January 2017 / Published online: 10 January 2017

Abstract Heavy metals in mine wastes can considerably influence surrounding surface waters, soils, and human health. To estimate environmental impact, heavy metal concentrations in stream sediments can be utilized because they are indicators of contamination and change negligibly with time. This study proposes a new Kriging method to predict heavy metal concentrations in stream sediments. The proposed methods compensate for the drawbacks of Kriging based on Euclidean distance because they utilize the stream distance for the prediction by analyzing the stream path and networks using digital elevation models. Moreover, the developed method reduces the exaggeration problem in predicting the concentration of an uncontaminated stream segment by considering the catchment basin area in Kriging. Application of these methods to synthetic and real-world datasets proves that they exhibit improvement in terms of overall error reduction, and they provide reasonable predictions at stream junctions, rather than Kriging based on Euclidean distance.

Keywords Geostatistics · Stream sediments · Kriging · Heavy metal concentration · Digital elevation model

✉ Yosoon Choi
yspower7@gmail.com; energy@pknu.ac.kr

[1] Department of Energy Systems Engineering, Seoul National University, Seoul 08826, Republic of Korea

[2] Department of Energy Resources Engineering, Pukyong National University, Busan 48513, Republic of Korea

[3] Korea Institute of Geoscience and Mineral Resources, Daejeon 34132, Republic of Korea

Introduction

Mine leachates and acid mine drainage from mine wastes and tailings can cause environmental contamination in surface waters and soils (Song and Choi 2015). When mine tailings dam fails from a severe rainstorm, the surrounding environments can be greatly influenced. In addition, various harmful metals enriched in these mine wastes will be a serious threat to human health when they move through the ecosystem in biogeochemical cycles (Park et al. 1995; Shamsi et al. 2016). To prevent and mitigate the damage from heavy metal contamination, comprehensive studies on the extents, pathways, distribution patterns, or other characteristics of heavy metals are required (Kim et al. 2012a). However, a field investigation on the effects of mining pollution is difficult to perform because of poor accessibility in steep terrain and extensive range of contamination (Lee and Choi 2016). Sufficient data for analyzing contamination trends in the area of interest are also difficult to acquire because of time and budgetary constraints. Therefore, to predict the pollution degree and extent precisely from the limited data is a significant challenge. Geostatistics, which is a branch of statistics focusing on spatial and spatiotemporal datasets, can be utilized to solve this problem.

Heavy metals in stream sediments are suitable variables to estimate the continuous environmental impacts because they are abundant in contaminated environments and barely change over time (Thornton 1983; Lee et al. 2016). To analyze the heavy metals in stream sediments, the path and flow direction of the stream are essential to be considered (Choi et al. 2011; Choi 2012; Kim et al. 2016). Geographic information systems (GISs) combined with geostatistics can be a useful tool to analyze the characteristics of streams and topography. Furthermore, GIS analysis has the advantage of obtaining data and analyzing extensive areas easily (Choi et al. 2008; Suh et al. 2013). With the growth of computer technologies, various studies on GIS modeling have estimated the pathway and distribution of contaminants (Heathwaite et al. 2005; Yenilmez et al. 2011; Kim et al. 2012b). However, most of these studies have limitations in indicating the real contamination degree and distribution because field survey data were not utilized or were only used to verify the predicted result. On the other hand, geostatistics studies predict the contamination degree using field survey data. Salgueiro et al. (2008) estimated the chemical contamination of stream sediments at the Vale das Gatas mine in Portugal using the geostatistical method. Khalil et al. (2013) assessed the extension and magnitude of soil contamination with toxic elements from abandoned mines in semi-arid areas using geochemical analysis and geostatistics. In addition, many studies predicted soil contamination in cultivated land (Steiger et al. 1996; White et al. 1997; Lin et al. 2001; Liu et al. 2004). Particularly, Kriging, which predicts an unknown value using the weighted linear summation, is widely used and studied in different fields. Although a variety of software platforms provide Kriging tools according to each objective, the aquatic variables could not be predicted reasonably because of their use of Euclidean distance when analyzing the correlationship between samples.

Several studies that predict aquatic variables are represented in Table 1. Smith et al. (1997) suggested a method using spatially referenced regressions of contaminant transport on watershed attributes in regional water-quality assessment. Yuan (2004) used a spatial interpolation to estimate stressor levels in unsampled streams by considering the geology and land use of each sample catchment basin. However, these studies predicted the overall water quality on the area including land because Euclidean distance was used in Kriging application. Dent and Grimm (1999) quantified patterns of nutrient concentration in surface waters of an arid land stream and compared spatial patterns using stream distance. Torgersen et al. (2004) sampled and analyzed multiscale, spatially continuous patterns of stream fishes and physical habitat using the distance between points along the stream path, or the network distance. However, Kriging predictions were not attempted. Curriero (1996), Little et al. (1997), and Rathbun (1998) predicted water quality using Kriging based on stream distance instead of Euclidean distance, but it was not applied to stream networks, and topography was not considered. Skøien et al. (2006) estimated the 100-year flood in stream networks considering the topology of stream networks and nested catchments. However, the method only predicted the variable in the unit of stream segment for the stream networks and was not able to predict the continuous change of the variable in a stream segment. VerHoef et al. (2006) developed spatial statistical models for stream networks that can make predictions at unsampled locations by incorporating flow and stream distance using spatial moving averages. Most of these studies analyzed aquatic variables in a stream or stream networks composed of vector objects. These vector-based methods have difficulties in considering topography, requiring the construction of objects or database to preserve information, such as flow direction and the distance between samples. Therefore, those studies have difficulty in analyzing the continuous change of variables and automating the analysis process. Software development has not been reported.

This study aims to develop the grid-based Kriging algorithm to predict aquatic variables along the stream networks. To compensate for the drawbacks of Kriging based on Euclidean distance, this study utilizes the stream

Table 1 Review of the pattern analysis and prediction for aquatic variables

Authors	Variable	Prediction	Stream distance	Stream network	Catchment properties or flow quantity	Topographical conditions	Software development
Curriero (1996)	Water quality	✔	✔				
Little et al. (1997)	Water quality	✔	✔				
Smith et al. (1997)	Water quality	✔			✔		
Rathbun (1998)	Water quality	✔	✔				
Dent and Grimm (1999)	Water quality		✔				
Torgersen et al. (2004)	Fish abundance		✔	✔			
Yuan (2004)	Water quality	✔			✔		
Skøien et al. (2006)	100-year flood	✔		✔	✔	✔	
VerHoef et al. (2006)	Water quality	✔	✔	✔	✔		
This study	Heavy metal in stream sediment	✔	✔	✔	✔	✔	✔

distance by analyzing the stream path and networks using digital elevation model (DEM). In addition, the area of each sample catchment basin is applied to correlation modeling and Kriging prediction to predict unknown values reasonably at stream junctions. Raster dataset, which is a matrix data structure representing a grid of pixels, is used in this study because of its advantage in analyzing the flow direction, stream networks, catchment basins, and other overlay analyses considering topography. Only the present study and Skøien et al. (2006) in Table 1 considered topography of the study area when predicting variables in the stream by using raster dataset. Furthermore, only the present study developed a software platform designed for automation of aquatic variable prediction. This study presents two case studies of application to a synthetic sample data and stream sediment data of an abandoned mining area in South Korea.

Grid-based Kriging considering the stream distance and catchment basin area

The Kriging method proposed in this study was designed to consider the stream distance and networks in contrast to general Kriging methods, which consider only spatial locations of samples. Figure 1 shows the difference between the stream distance and Euclidean distance in rasterized digital data. In Fig. 1a, the solid line is a stream, and the circle is a sample in streams. A grid of pixels denotes a raster dataset, and the colored pixel X indicates the unknown location to be predicted. Figure 1b shows the rasterized samples and streams. The distances from pixel X to samples A and B are both ($\sqrt{2}$ × pixel size) in raster format (Fig. 1b), and the two samples have same weights in general Kriging. However, it is rational that sample B is more associated with X because sample B is even closer to the pixel X than sample A when considering the stream path.

Figure 2 shows the flowchart for calculating the stream distances between samples and predicting the unknown value using Kriging by considering stream distance (i.e., STD-Kriging in this paper). First, raster format data representing the sample locations and flow direction of the study area are required to calculate the distances between samples. If a sample is located on the stream path, the unit distance in a pixel is accumulated along the flow direction until the next downward pixel meets other samples. There are sample pairs for all samples in Kriging using Euclidean distance (i.e., EUC-Kriging in this paper). For example, ten distance pairs exist excluding duplicates for five samples in Fig. 3a. On the other hand, stream networks should be considered when calculating the stream distance. In Fig. 3b, it is reasonable to calculate the distances between samples 2, 4, and 5 in the Kriging prediction because sample 2 is connected with samples 4 and 5. However, no distance pair between samples 1 and 2 exists because they are unconnected and irrelevant to each other.

After calculating the distances between samples, the empirical variogram should be obtained using the distances and sample values. In this step, an adequate lag distance should be defined because the shape of the empirical variogram depends on the lag distance. The empirical variogram with the lag distance h is defined as follows for observations of z_i at locations x_i ($i = 1, \ldots$) (Cressie, 1985):

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \left|z_i - z_j\right|^2 \tag{1}$$

where $N(h)$ denotes the set of pairs with $|x_i - x_j| = h$, and $|N(h)|$ is the number of pairs in the set. In this study, the approximate distance is used using a tolerance, which is the half of the lag distance, h, to include all data.

Kriging requires valid variogram at every lag distance to predict the unknown value. However, the empirical variogram cannot be computed at every lag distance, and it is not ensured to be valid because of variation in the estimation. Therefore, theoretical variogram models ensuring validity are applied to approximate the empirical variogram (Chiles and Delfiner 2012). Theoretical variogram

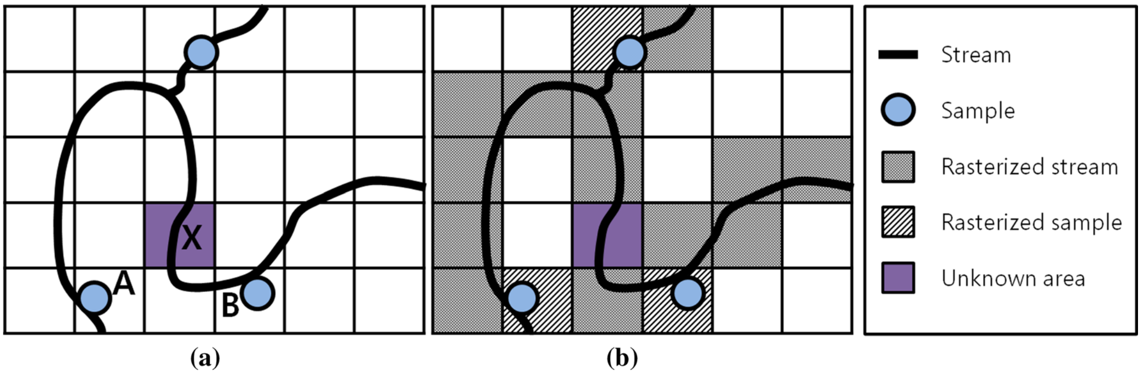


Fig. 1 Rasterization of **a** unknown area, **b** samples and stream paths

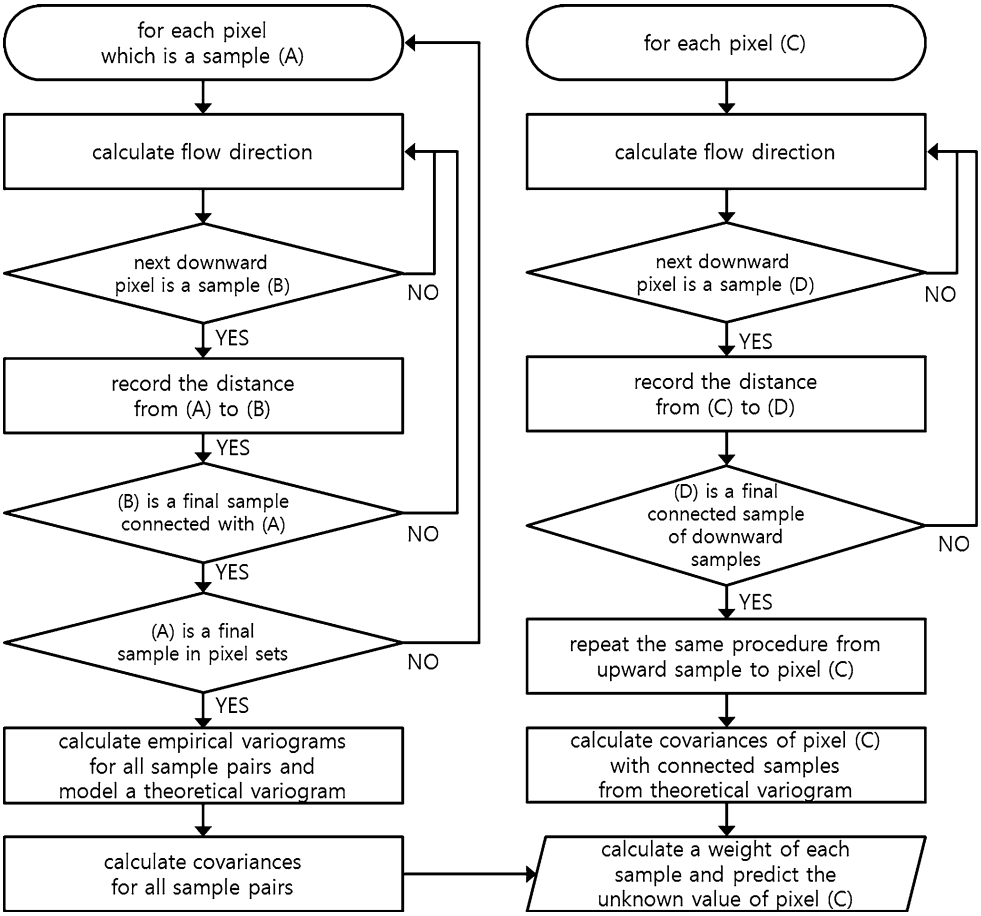


Fig. 2 Flowchart for calculating the stream distances between samples and predicting the unknown value using STD-Kriging

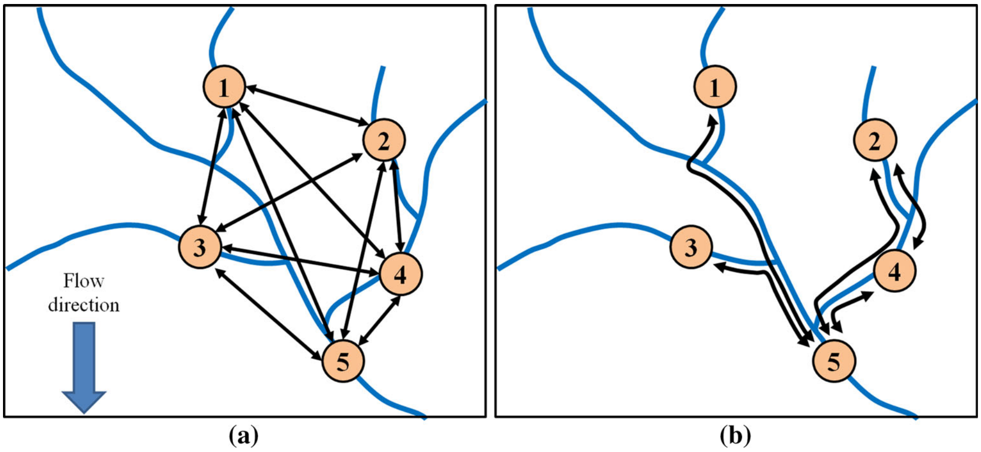


Fig. 3 Euclidean distance (**a**) and stream distance (**b**) between samples in the stream network

modeling is an important step in Kriging because it affects the prediction directly. The final step is to predict the unknown value by calculating weights of samples, which is calculated based on the theoretical variogram according to the distance between the sample and the unknown location to be predicted.

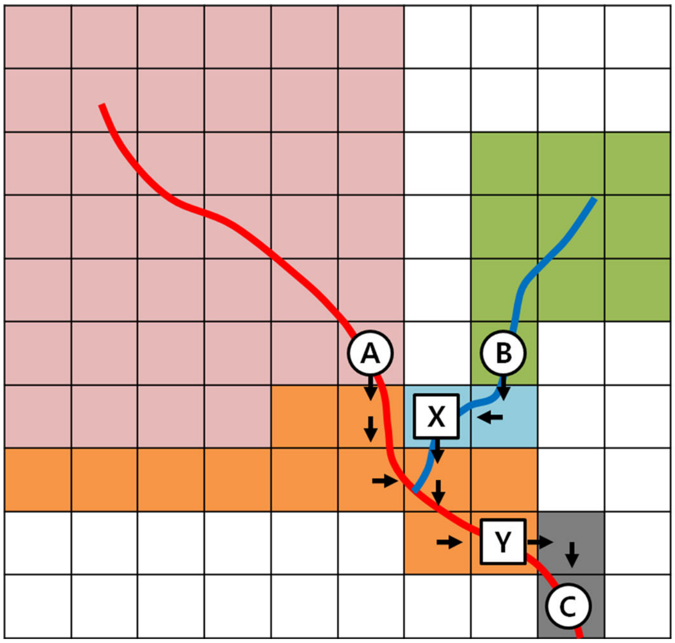
This study proposes a new Kriging method that uses the catchment basin area instead of the distance. A stream sediment sample is a result of constant erosion and sedimentation of introduced materials, which are transported primarily by the flow of rainwater or stream. On the assumption that the surface water flows to the lower area, it is possible to analyze the catchment basin of the sample where the water and materials came from by using GIS-based spatial analysis. The area of the catchment basin of each pixel can also be calculated by using the flow accumulation algorithm (Jenson and Domingue, 1988). If a wide gap of catchment basin area exists between two pixels that are connected to each other, they tend to have different values because many materials are introduced from another area between the two pixels. On the other hand, if a small gap of catchment basin area exists between the two pixels, they tend to have similar values.

Figure 4 shows the advantage of Kriging using the catchment basin area (i.e., CAT-Kriging in this paper) when compared with STD-Kriging. The red line denotes a contaminated stream, the blue line represents an uncontaminated stream, and the arrows indicate the flow direction of the stream pixel. Points A, B, and C are samples with known value, and points X and Y are unknown locations to be predicted. When predicting the value of point X, which is ahead of the stream junction, the stream distances from point X to samples B and C are 2 and 5, respectively. When considering the stream distance, an exaggerated value may be predicted for point X due to the effect of polluted sample C. The differences of the catchment basin area between point X and samples B and C are 2 and 44 (pixels), respectively. CAT-Kriging can provide a more reasonable prediction for point X because the effect of uncontaminated sample B is much larger than the contaminated sample C. When predicting the value of point Y after the stream junction, the stream distances from point Y to samples A, B, and C are 5, 5, and 2, respectively. The closest sample C has the largest weight, and samples A and B have same weights in STD-Kriging. However, it is reasonable that sample A, which is the representative of a larger catchment basin area, has more effect on point Y than sample B because of the assumption that stream sediment is a homogeneous mixture of materials in the catchment basin. In the case of CAT-Kriging, the respective differences of the catchment basin area are 14, 44, and 2 for samples A, B, and C, respectively. Therefore, samples C, A, and B are arranged according to weights on point Y in CAT-Kriging.

Software development to implement the Kriging process

To model the variogram and predict the unknown value using the Kriging method, a new software was implemented. The software was written in Visual Basic 2013 and utilizes the ESRI ASCII grid file as a standard data format, exchangeable with ESRI ArcGIS software.

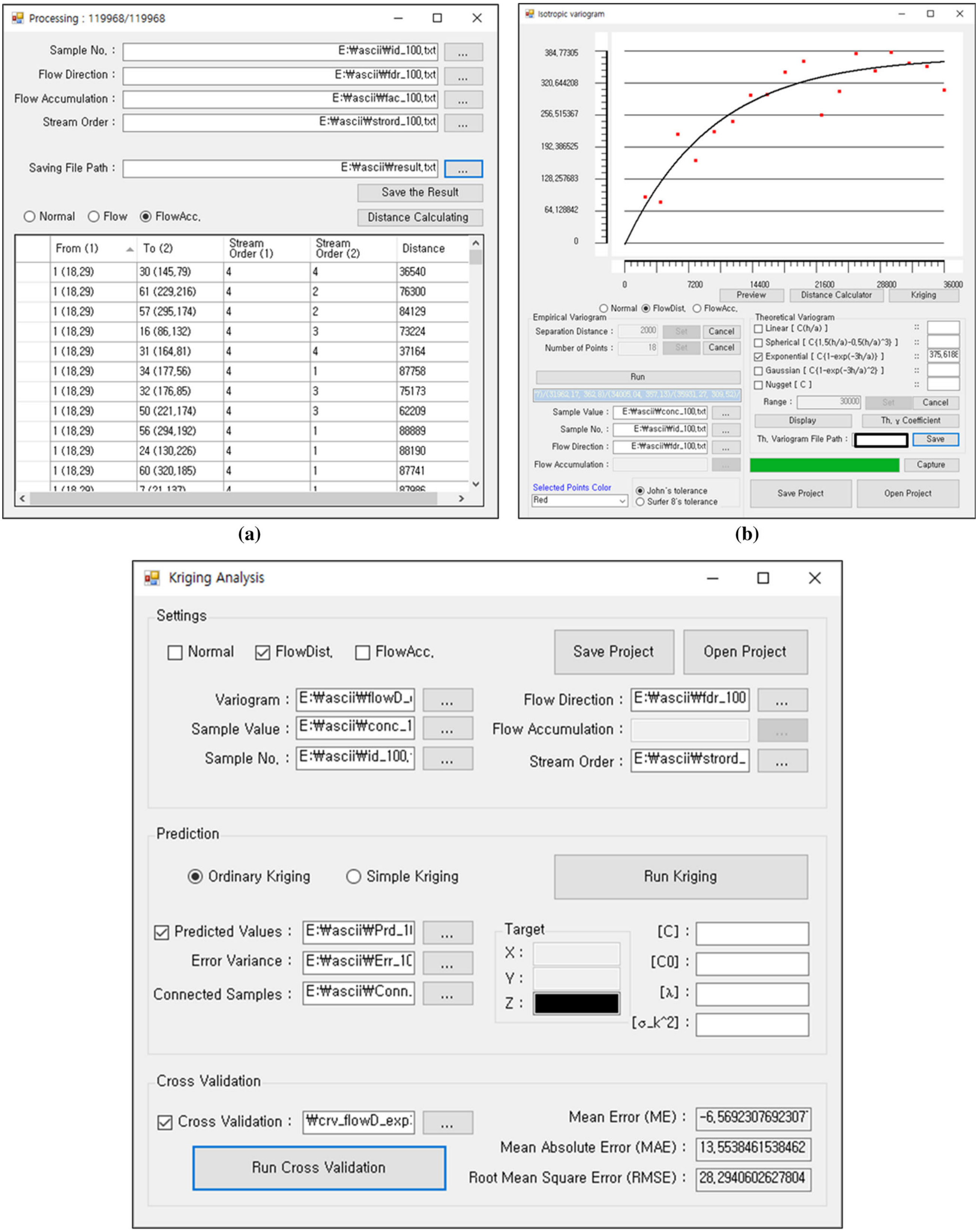
Figure 5a is a module developed for calculating the distance between samples, which provides Euclidean distance, stream distance, and catchment basin area difference as a table. Figure 5b is the variogram-modeling module constructed to calculate the empirical variogram and model the theoretical variogram. Using the raster format input data, such as the sample locations, sample concentrations, and flow direction of the study area, this module calculates the empirical variogram for each lag distance and displays it as points in the graph. To predict the unknown value using Kriging, theoretical variogram models that ensure validity are required to approximate the empirical variogram. In this module, the empirical variogram is approximated by a combination of five widely used theoretical variogram models, such as the linear model, the spherical model, the exponential model,



	Catchment area (Flow accumulation)	Catchment (Color)
A	40	🟥
B	10	🟩
C	56	🟥 🟩 🟦 🟧 ⬛
X	12	🟩 🟦
Y	54	🟥 🟧

		Stream distance	Flow accumulation difference
X	B	2	2
X	C	5	44
Y	A	5	14
Y	B	5	44
Y	C	2	2

Fig. 4 An example showing the effect of the catchment basin area to the chemistry values in the stream network



Processing : 119968/119968

Sample No. : E:₩ascii₩id_100.txt

Flow Direction : E:₩ascii₩fdr_100.txt

Flow Accumulation : E:₩ascii₩fac_100.txt

Stream Order : E:₩ascii₩strord_100.txt

Saving File Path : E:₩ascii₩result.txt

○ Normal ○ Flow ● FlowAcc.

From (1)	To (2)	Stream Order (1)	Stream Order (2)	Distance
1 (18,29)	30 (145,79)	4	4	36540
1 (18,29)	61 (229,216)	4	2	76300
1 (18,29)	57 (295,174)	4	2	84129
1 (18,29)	16 (86,132)	4	3	73224
1 (18,29)	31 (164,81)	4	4	37164
1 (18,29)	34 (177,56)	4	1	87758
1 (18,29)	32 (176,85)	4	3	75173
1 (18,29)	50 (221,174)	4	3	62209
1 (18,29)	56 (294,192)	4	1	88889
1 (18,29)	24 (130,226)	4	1	88190
1 (18,29)	60 (320,185)	4	1	87741
1 (18,29)	7 (21,137)	4	1	87986

(a)

Isotropic variogram

○ Normal ● FlowDist. ○ FlowAcc.

Empirical Variogram

Separation Distance : 2000

Number of Points : 18

?7)/(31962.17, 362.8)/(34005.04, 357.13)/(35931.27, 309.52)/

Sample Value : E:₩ascii₩conc_100.txt

Sample No. : E:₩ascii₩id_100.txt

Flow Direction : E:₩ascii₩fdr_100.txt

Flow Accumulation :

Selected Points Color: Red

● John's tolerance

○ Surfer 8's tolerance

Theoretical Variogram

□ Linear [C(h/a)] ::

□ Spherical [C{1.5(h/a)-0.5(h/a)^3}] ::

☑ Exponential [C{1-exp(-3h/a)}] :: 375.6188

□ Gaussian [C{1-exp(-3h/a)^2}] ::

□ Nugget [C] ::

Range : 30000

Th. Variogram File Path :

(b)

Kriging Analysis

Settings

□ Normal ☑ FlowDist. □ FlowAcc.

Variogram : E:₩ascii₩flowD_

Sample Value : E:₩ascii₩conc_1

Sample No. : E:₩ascii₩id_100.

Flow Direction : E:₩ascii₩fdr_100

Flow Accumulation :

Stream Order : E:₩ascii₩strord_

Prediction

● Ordinary Kriging ○ Simple Kriging

☑ Predicted Values : E:₩ascii₩Prd_1(

Error Variance : E:₩ascii₩Err_1(

Connected Samples : E:₩ascii₩Conn.

Target

X :

Y :

Z :

[C] :

[C0] :

[λ] :

[σ_k^2] :

Cross Validation

☑ Cross Validation : ₩crv_flowD_exp

Mean Error (ME) : -6.5692307692307

Mean Absolute Error (MAE) : 13.5538461538462

Root Mean Square Error (RMSE) : 28.2940602627804

(c)

Fig. 5 Graphical user interface of **a** distance calculator module, **b** variogram-modeling module, and **c** Kriging analysis module

the Gaussian model, and the nugget model. It is important to combine the models to represent the data and empirical variogram appropriately. This module supports modeling the theoretical variogram using weighted least squares regression. The theoretical variogram can be saved as its own format to be used for prediction. Figure 5c is the Kriging analysis module, which predicts the unknown values using GIS layers and the saved variogram file. The predicted result is exported to ESRI ASCII grid file. In addition, this module provides a leave-one-out cross-validation function that uses one observation as the validation set and the remaining observations as the training set. If the number of samples is N, N models are created, and this function is repeated N times. The advantage of this function is that it does not allow for randomness because all the data can be used for training. This function is commonly used and is known as a useful validation method for Kriging (Aelion et al. 2009; Menafoglio et al. 2014). To compare the predicting capabilities quantitatively, mean error (ME), mean absolute error (MAE), and root mean square error (RMSE) are also computed.

Application to synthetic datasets

Synthetic datasets creation

To demonstrate the improvements of the proposed methods, synthetic datasets were used. Synthetic data should be created similar to the principle of real contaminant distribution. In this study, two factors, such as soil erosion and catchment basin derived from real DEM, were considered to reflect the contaminant distribution.

Figure 6a shows a DEM with 100 m resolution and the locations of synthetic stream sediment samples. Figure 6b shows the catchment basins of some samples that are derived from the DEM. Sample 8 is the lowest of the five samples, and the catchment basin is the sum of A, B, C, D, and E. Sample 65 is the highest of the five samples, and the catchment basin is E. The heavy metal concentration of the stream sediment is affected by contaminants in the eroded soil. This study estimated the soil erosion of each pixel using the Universal Soil erosion Equation (USLE) defined as follows:

$$A = R \times K \times L \times S \times C \times P \tag{2}$$

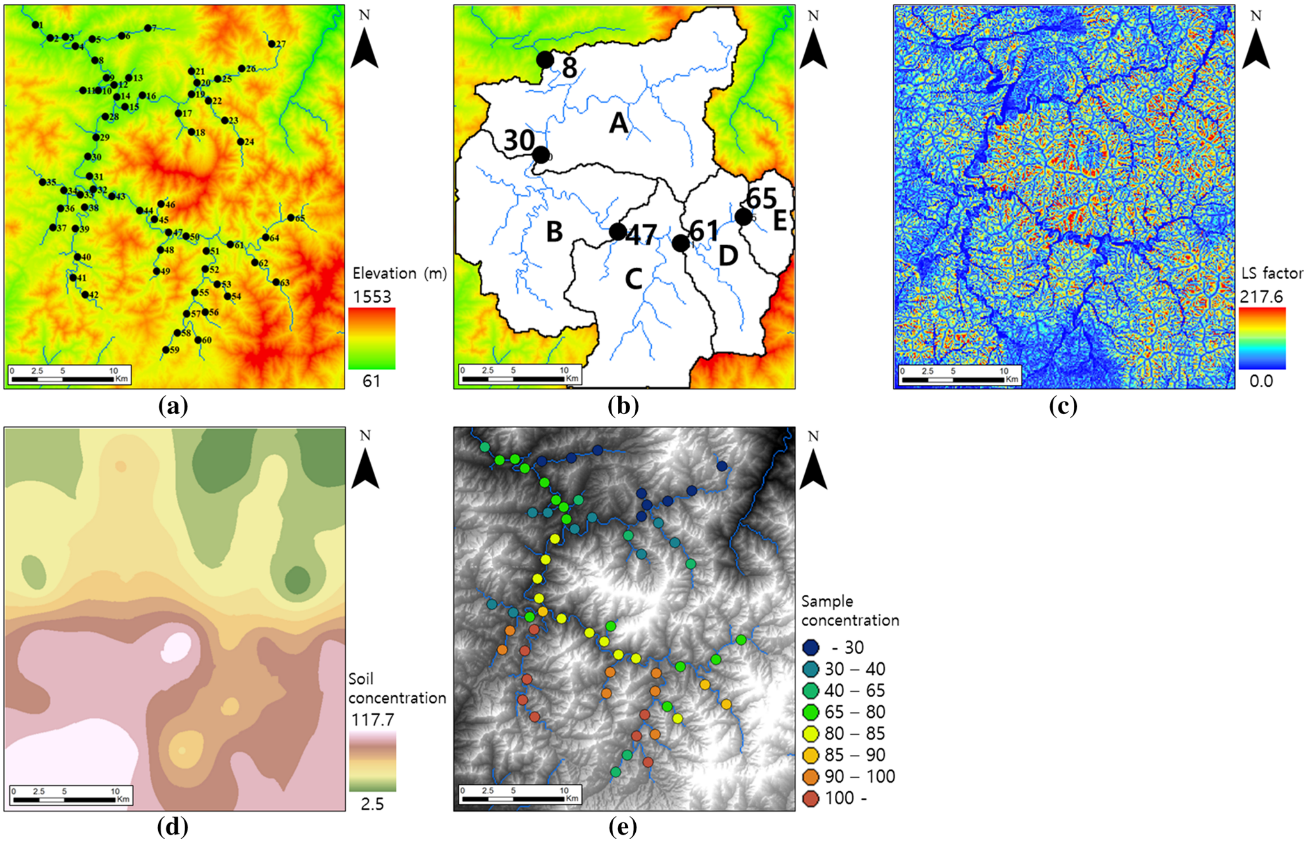


Fig. 6 Synthetic datasets: **a** DEM and samples, **b** catchments of some samples, **c** distribution of LS factor that represents soil erosion, **d** distribution of nonpoint pollution source, and **e** simulated heavy metal concentration of synthetic data

The USLE has been widely used to estimate the average annual soil erosion per unit land area (Wischmeier and Smith 1978) and considers six major factors affecting soil erosion, represented as the rainfall erosivity factor (R), the soil erodibility factor (K), the slope length factor (L), the slope steepness factor (S), the cropping management factor (C), and the supporting conservation practices factor (P). To create synthetic data, this study calculated the L-factor and S-factor as relative soil erosion (Fig. 6c) using GIS-based analyses of slope length and slope gradient (McCool et al. 1987) supposing that other factors are the same for all pixels. Figure 6d shows the synthetic concentrations of soil distributed in the study area as a nonpoint source.

The synthetic concentrations of stream sediment samples were simulated using thematic maps with the following steps:

1. Estimate the amount of heavy metal eroded at each pixel by multiplying the soil erosion (A_i) and concentration (C_i) of each pixel i.
2. Analyze the catchment basin (W_k), and calculate the area (the number of pixels) for each sample k.
3. Define the amount of heavy metal introduced to each sample as $\sum_i^{W_k} (A_i \times C_i)$ on the assumption that the eroded heavy metals are mixed homogeneously.
4. Calculate the concentration of stream sediment sample k by dividing $\sum_i^{W_k} (A_i \times C_i)$ by $\sum_i^{W_k} A_i$, which is the sum of eroded soil in the catchment basin W_k.

Figure 6e shows the concentrations of stream sediment samples derived from these steps. Although this result cannot represent the contamination in real world precisely, synthetic datasets reflecting contamination and dilution could be created.

Prediction for synthetic datasets

The variogram should be modeled before predicting the unknown values using Kriging. Figure 7a–c shows the empirical variograms and the theoretical variograms for the three methods (i.e., Euclidean distance, stream distance, and catchment basin area). The linear model was applied in the same manner, and the ranges of the three models were defined as 25,000 m, 40,000 m, and 800 km^2 (80,000 pixels), based on the shape of the empirical variogram. The range refers to the distance at which the variogram reaches the sill. There is no correlation between two samples.

Figure 8 shows the prediction results using EUC-Kriging (Fig. 8a), STD-Kriging (Fig. 8b), and CAT-Kriging (Fig. 8c). Unlike other methods, EUC-Kriging predicts for all pixels of the study area, but only the predicted values on the stream network were extracted to be compared with other two results. A remarkable difference among the three methods exists at the stream junctions, such as area A and area B. As shown in Fig. 9a, in the case of EUC-Kriging, the predicted values continuously vary regardless of the shape of stream network. On the other hand, the other two methods (Fig. 9b, c) can distinguish both uncontaminated and contaminated streams. The prediction tendencies are similar for STD-Kriging (Fig. 9b) and CAT-Kriging (Fig. 9c). However, at the stream segment marked by a square, CAT-Kriging predicts larger values than STD-Kriging because the stream segment covering sample 57 has larger catchment basin area than the stream segment covering sample 56. Figure 9d–f shows the prediction for area B in Fig. 8a. The stream segment with no observation is predicted using downstream samples, such as sample 28. Predicting these stream segments is challenging task, but it may be generally not contaminated. In the case of STD-Kriging (Fig. 9e), larger values are predicted than CAT-Kriging (Fig. 9f) because of the effect of the adjacent sample 28.

Figure 10 illustrates the predicted concentrations as a graph along the path indicated by pink line. A to M represents junctions where other stream segments meet the path. In the case of EUC-Kriging, the predicted values do not depend on junctions. In other two cases, the predicted values change drastically at junctions. For example, a massive increase at junction C is observed where a contaminated stream joins the path. On the contrary, a huge decrease at junction H occurred where an uncontaminated

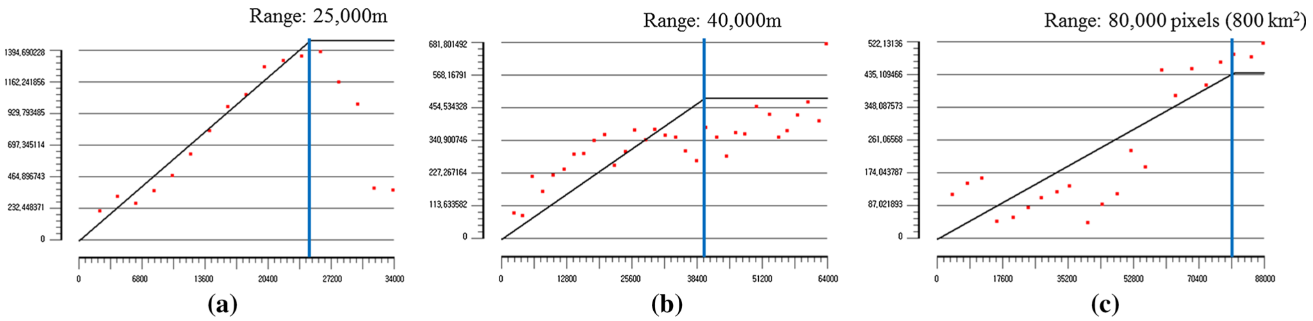


Fig. 7 Empirical variograms and theoretical variograms calculated from **a** Euclidean distance, **b** stream distance, and **c** catchment basin area

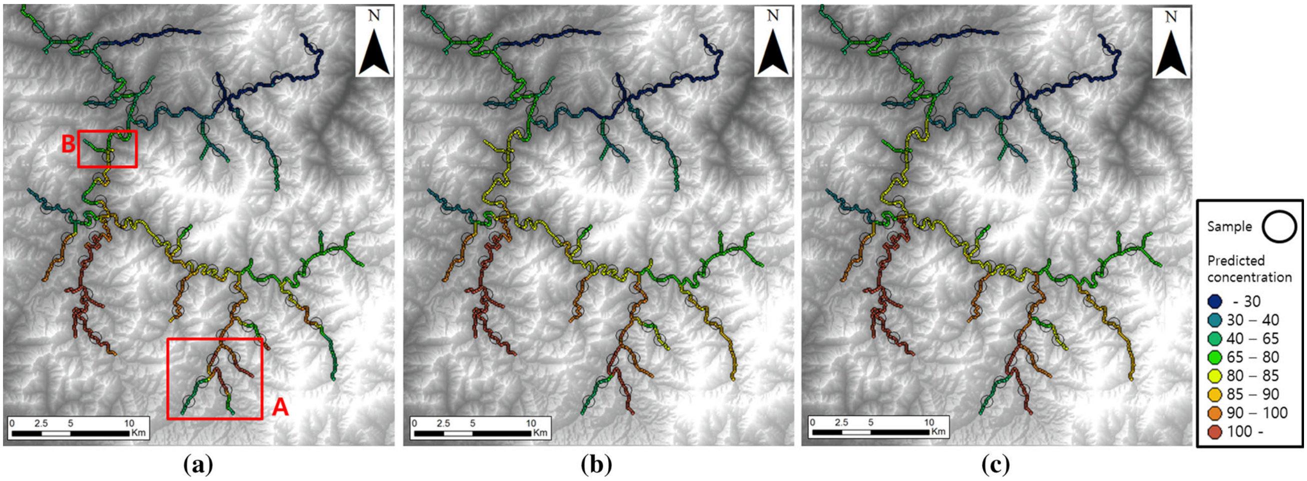


Fig. 8 Prediction of heavy metal concentrations for synthetic data using **a** EUC-Kriging, **b** STD-Kriging, and **c** CAT-Kriging

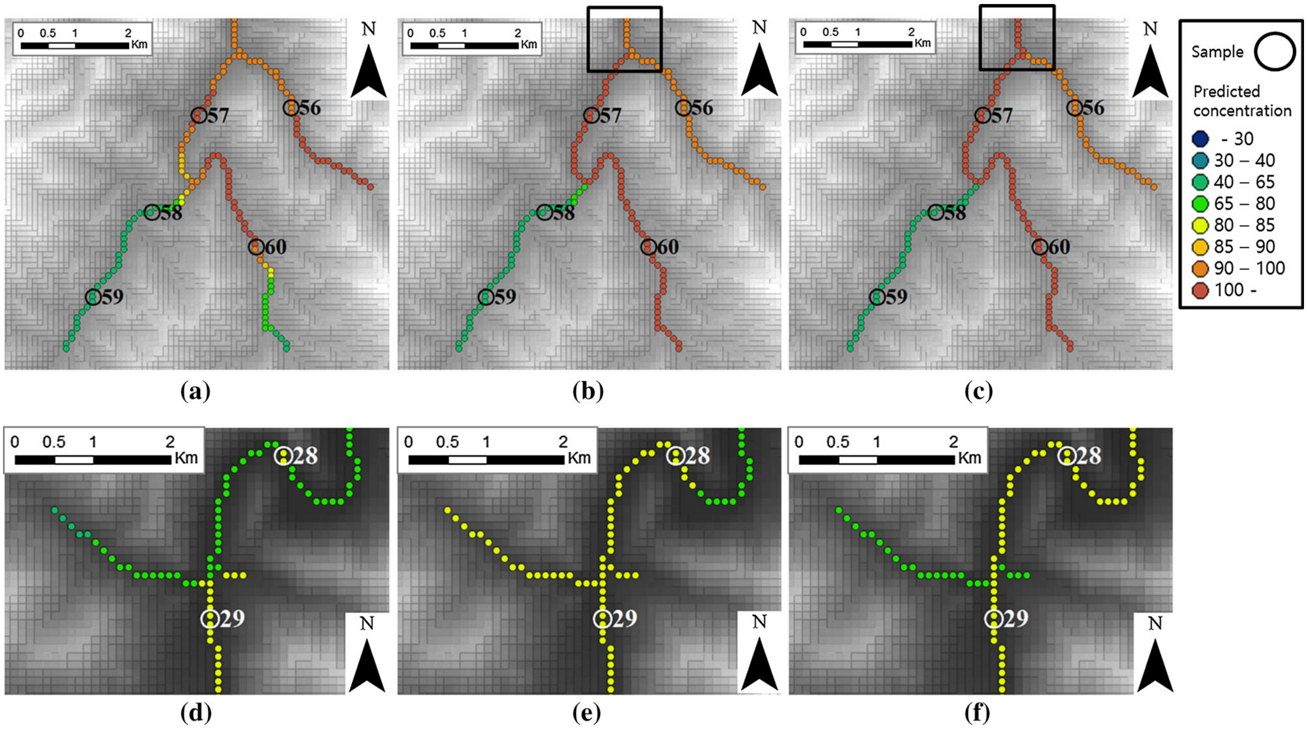


Fig. 9 Prediction of heavy metal concentrations for area A using **a** EUC-Kriging, **b** STD-Kriging, and **c** CAT-Kriging; and for area B using **d** EUC-Kriging, **e** STD-Kriging, and **f** CAT-Kriging

stream with large catchment area joins the path. This trend is significant in CAT-Kriging.

As a result of cross-validation, the ME of EUC-Kriging, STD-Kriging, and CAT-Kriging is −0.2, 3.0, and 1.4, the MAE is 11.5, 6.2, and 5.2, and the RMSE is 15.2, 9.4, and 13.3, respectively. ME indicates bias in prediction, and MAE or RMSE indicates the capability of prediction from the point of overall errors. Compared to MAE, RMSE amplifies large errors because of the square term in the formula. Based on the ME results, the EUC-Kriging result is the most unbiased and the STD-Kriging result is the most biased. Based on the MAE and RMSE results, the prediction capabilities of STD-Kriging and CAT-Kriging are confirmed to have improved in terms of overall error reduction. Even though the MAE of CAT-Kriging is the lowest, the RMSE of CAT-Kriging is larger than that of STD-Kriging owing to the influence of samples 50 and 61, which have large errors. If samples 50 and 61 are excluded, the RMSE of EUC-Kriging, STD-Kriging, and CAT-Kriging is 15.4, 9.4, and 9.3, respectively. Figure 11 shows

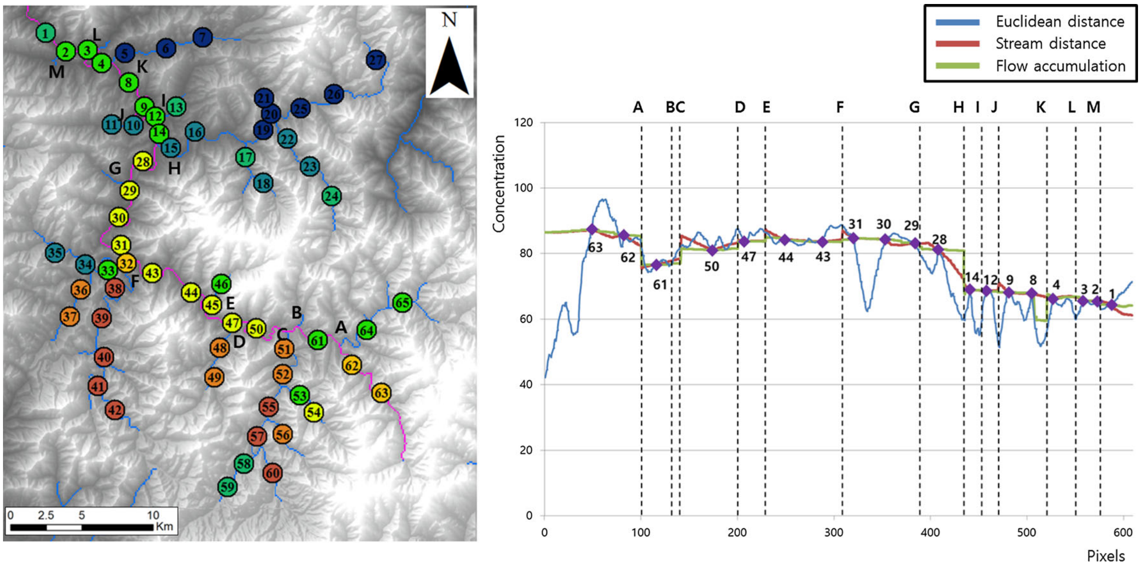


Fig. 10 Graph of the predicted heavy metal concentrations for the stream path represented by a *pink line*. *A* to *M* are stream junctions

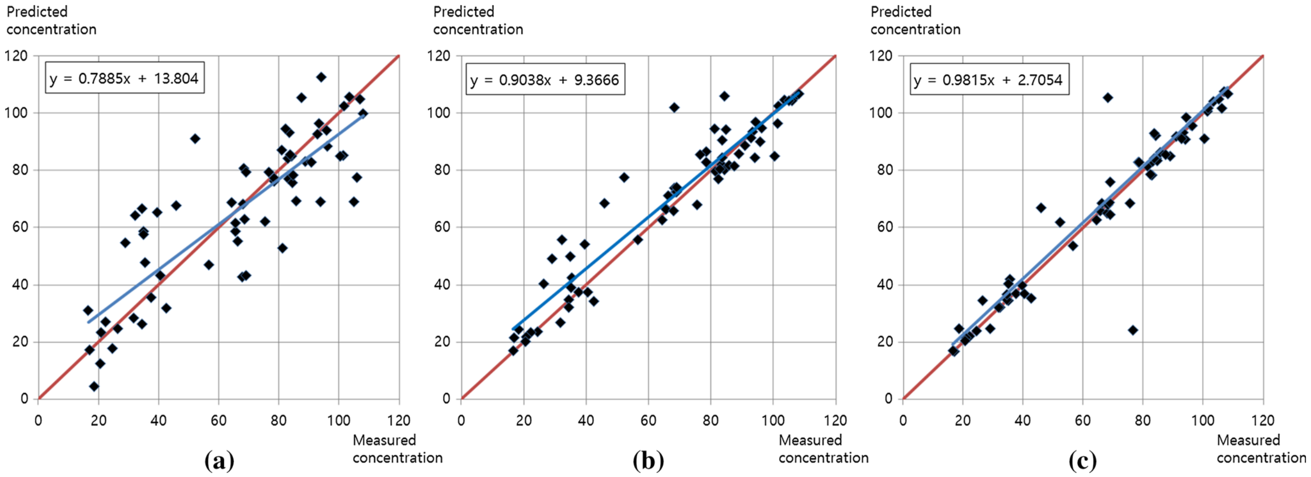


Fig. 11 Original values versus predicted values based on cross-validation for **a** EUC-Kriging, **b** STD-Kriging, and **c** CAT-Kriging. The *red line* indicates the position of one-to-one correspondence. The *blue line* with the equation represents the trend line of points

the relationships between the predicted and the original values. From the comparisons, the predicted values of CAT-Kriging tend to be positioned on the red line, indicating the position of one-to-one correspondence, compared to other methods.

Application to real-world datasets

Study area

For a real-world application, the area of Yeonhwa II mine (37°07′N, 129°08′E), which is located in Samcheok-si, a city in South Korea, was selected as the study area. Two tailings dams (Fig. 12a) hold mine tailings, which include mainly lead (Pb) and zinc (Zn). The average annual rainfall of the study area is 1315.1 mm, and 54.8% of the rainfall is concentrated during summer (July to September), causing surface water contamination by erosion of mine tailings and mine water leaks (MIRECO 2007). Figure 12b shows a DEM (30 m grid spacing) and sampling data of the study area. To generate a DEM, topographical contours (contour interval: 5 m) were extracted from 1:5000 scale topographical maps published by the National Geographic Information Institute of Korea (http://www.ngii.go.kr). A triangulated irregular network surface was created from the topographical contours and converted to a DEM. As seen in a DEM, the western part of the study area is higher than the eastern part, where the main stream is flowing to the East Sea. The circles denote stream sediment samples, and the

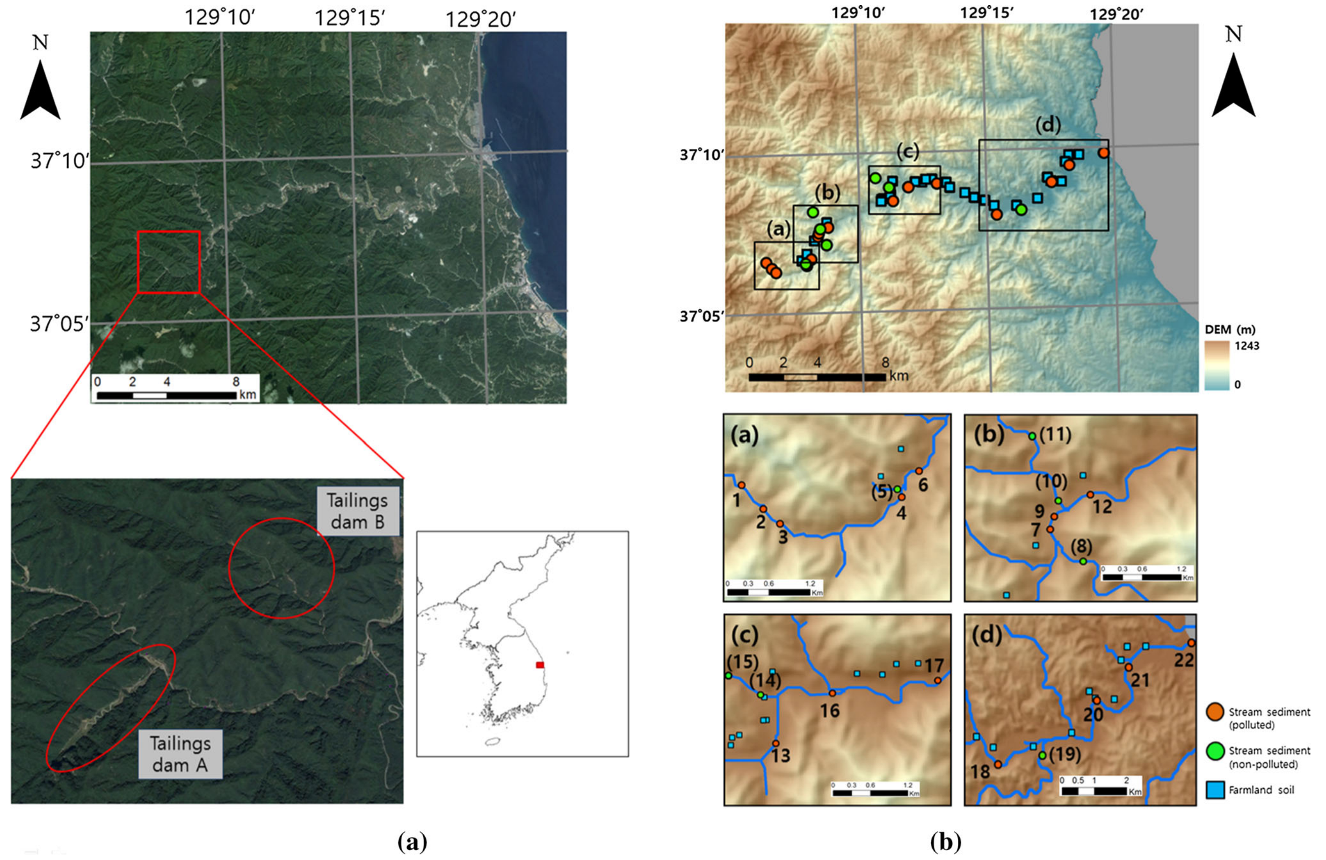


Fig. 12 Location of the study area on the satellite image (Google Earth, *left*), and DEM and locations of samples (*right*). Sample numbers in *brackets* indicate stream sediment samples collected from the uncontaminated stream

squares represent farmland soil samples. The heavy metal concentrations in the stream sediment samples are presented in Table 2. The Zn concentrations of 14 samples out of 22 samples exceed the US standard of sediment removal. The other eight samples were collected from the uncontaminated stream (i.e., not the main stream). The Zn concentrations become smaller on the whole as the distance from Yeonhwa II mine increases. Therefore, Zn is an indicator for pollution in this area. Table 3 shows the heavy metal concentrations in farmland soil samples near a stream. The concentrations of farmland soil are much smaller than those of the stream sediment, which implies that the major cause of contamination in this area is the stream flow. Although farmland soil samples are close to the stream, to consider the stream sediment samples and farmland soil samples as one dataset is irrational. Therefore, the Zn concentration of the stream sediment samples was selected as a variable in this study.

Prediction for study area

To predict the heavy metal concentration, the distances between samples were first calculated (Table 4) based on Euclidean distance, stream distance, and differences of catchment basin area. To make the real-world and synthetic results comparable, the DEM with 30 m resolution was resampled to 100 m resolution, which is the same as that of the synthetic data. It can be observed from the results in Table 4 that the calculated Euclidean distance, stream distance, and differences in catchment basin area are similar for different DEMs. The stream distance is confirmed to be longer than the Euclidean distance for the same sample set. In cases of stream distance and catchment basin area, some samples are not connected with each other. If the relative distances between samples and the relative range are the same for different datasets, the prediction results are the same, according to the characteristic of Kriging. Therefore, the number of pixels can be used as a unit of catchment basin area, instead of km^2 or m^2, regardless of the resolution. The difference of the catchment basin area is noticeably large for some sample pairs despite the short stream distance (e.g., a pair of samples 11 and 12, or the pair of samples 15 and 16). This phenomenon is remarkable where a junction is placed between two samples.

Table 2 Heavy metal concentrations of stream sediment samples (italicized number exceeds the soil contamination warning standards of South Korea, and the bold number exceeds the US standard of sediment removal)

	As	Cd	Cu	Ni	Pb	Zn
Soil contamination countermeasure standards of South Korea	6	0.5	50	40	100	300
Soil contamination warning standards of South Korea	*15*	*4*	*125*	*100*	*300*	*700*
US standard of sediment removal	**93**	**6.7**	**390**		**530**	**960**
Sample ID						
1	*20.5*	**6.7**	5.1	13.6	1.7	**5157.8**
2	9.3	2.8	3.0	8.0	11.6	**5046.3**
3	10.6	1.7	3.3	7.5	20.8	**11,607.3**
4	*21.0*	1.9	3.6	7.1	22.3	**8558.4**
5[a]	*16.1*	3.4	15.7	8.8	6.2	**3090.4**
6	15.0	2.1	4.8	7.2	20.5	**6859.3**
7	14.5	2.0	4.8	7.7	32.4	**11,882.3**
8[a]	1.7	0.1	0.5	8.3	4.9	150.6
9	12.8	1.5	4.5	7.3	19.2	**6684.0**
10[a]	ND	0.0	0.7	11.1	1.1	228.7
11[a]	0.1	0.1	0.6	7.5	2.2	170.3
12	1.5	0.5	1.9	7.4	8.8	**8910.7**
13	7.1	1.4	6.0	7.1	28.5	**4385.2**
14[a]	ND	0.1	0.8	16.1	1.5	161.1
15[a]	0.5	1.5	0.9	9.5	3.5	205.3
16	0.8	0.5	2.2	7.3	9.9	**4346.7**
17	*16.4*	1.9	5.0	7.8	19.5	**3726.2**
18	4.2	0.7	3.5	8.0	18.6	**2380.0**
19[a]	0.5	0.1	0.5	7.2	5.6	108.6
20	7.2	1.1	6.2	15.6	13.1	279.6
21	4.3	0.8	4.3	7.7	21.2	**1955.6**
22	0.4	0.1	1.9	13.7	1.9	307.1

Unit: mg/kg

[a] Sampled from uncontaminated stream

Figure 13a–c shows the empirical and theoretical variograms of the 30-m-resolution DEM for the three methods (Euclidean distance, stream distance, and catchment basin area), and Fig. 13d–f shows the variograms of the 100-m-resolution DEM. Even though there are a few differences between the two resolutions, the variogram shapes are similar. The exponential model combined with the nugget model was identically applied using weighted least squares regression. The ranges of the three models were defined as 14,000 m, 20,000 m, and 140 km^2 for both resolutions. In comparison with the case of synthetic data, the nugget effect is prominent because there is noise in real data and the observed values of the near samples are significantly different in the real world. Kriging prediction was applied to the 100-m-resolution DEM to achieve correspondence with the synthetic data and computational efficiency in analysis.

Figure 14 shows the results of Zn prediction using EUC-Kriging (Fig. 14a), STD-Kriging (Fig. 14b), and CAT-Kriging (Fig. 14c). The predicted values of the western part of the study area are more contaminated than the values of the eastern part in all three results. However, in contrast to the other two methods, EUC-Kriging cannot distinguish the uncontaminated stream segments from the contaminated streams. The stream segment of the western part heading toward the southeast is predicted to have larger values than the other stream segments because the main stream samples connected with the stream segment have large values while the uncontaminated sample directly collected from the stream segment is only one. Therefore, these stream segments need additional sampling to prevent the exaggerated prediction. STD-Kriging tends to be more exaggerated than CAT-Kriging.

Figure 15 illustrates the predicted concentrations as a graph along the main stream path, which is indicated by a pink line. A to M represent the junctions where other stream segments meet the path. In the case of EUC-Kriging and STD-Kriging, there is slight variation in the predicted values except in near samples due to a little variation and large nugget value in the theoretical variogram (Fig. 13a,

Table 3 Heavy metal concentrations of farmland soil samples (italicized number exceeds the soil contamination countermeasure standards of South Korea, and the bold number exceeds the soil contamination warning standards of South Korea)

	As	Cd	Cu	Ni	Pb	Zn
Soil contamination countermeasure standards of South Korea	*6*	*0.5*	*50*	*40*	*100*	*300*
Soil contamination warning standards of South Korea	**15**	**4**	**125**	**100**	**300**	**700**
Sample ID						
1	0.9	0.3	2.0	20.4	27.5	176.6
2	1.3	0.3	1.6	20.0	21.1	183.1
4	0.1	0.1	1.0	21.8	17.6	140.1
5	2.0	0.2	3.8	20.7	8.5	137.5
6	0.5	0.3	3.5	13.3	15.7	127.8
7	1.0	*0.6*	4.8	12.2	15.2	207.8
7-1	0.3	0.1	2.2	15.5	6.9	77.2
7-2	1.1	0.2	1.8	16.1	5.7	116.6
7-3	0.2	0.1	1.5	15.2	4.0	65.2
7-4	0.5	0.2	3.9	12.6	6.2	102.1
8	1.5	0.3	2.0	15.5	1.6	152.8
9	0.7	0.1	1.7	14.4	3.6	110.8
10	4.7	0.1	1.7	16.3	2.4	72.6
11	3.4	0.2	2.0	19.7	2.4	158.7
12	ND	0.1	1.7	16.2	2.8	51.6
13	0.4	0.1	2.0	8.9	7.5	54.1
14	1.1	0.1	2.2	10.8	6.6	83.9
15	1.6	0.2	2.9	10.6	6.2	70.3
16	3.8	*0.5*	4.5	17.8	8.9	204.2
17	1.0	0.2	3.0	13.8	8.7	103.9
18	1.3	*0.6*	5.8	13.8	14.6	213.7
19	0.7	0.1	1.6	9.3	8.1	44.0
20	2.8	0.2	2.6	16.6	3.6	109.8
21	2.4	0.1	2.4	12.5	5.7	99.0
22	1.3	*0.5*	5.5	15.0	13.9	197.5
23	2.0	0.2	2.5	21.7	2.8	121.4
24	1.1	0.2	3.2	23.5	6.2	179.8
25	0.9	0.1	2.7	16.1	4.0	81.7
26	3.3	0.3	6.1	20.9	7.3	179.2
27	1.7	0.1	3.2	21.2	4.9	111.0

Unit: mg/kg

b). However, the predicted values of STD-Kriging change drastically at junctions, unlike EUC-Kriging. The predicted values of CAT-Kriging change significantly at junctions and on the whole.

As a result of cross-validation (Table 5), the number of samples that each method predicted most accurately is 4 for EUC-Kriging, 4 for STD-Kriging, and 14 for CAT-Kriging, respectively. The ME of EUC-Kriging, STD-Kriging, and CAT-Kriging is 37.3, 177.3, and 274.3, MAE is 2954.3, 2337.6, and 2095.3, and RMSE is 3633.0, 3016.7, and 2844.0, respectively. Based on the ME results, the EUC-Kriging result is the most unbiased and the CAT-Kriging result is the most biased. Based on the MAE and RMSE results, the prediction capabilities of STD-Kriging and CAT-Kriging are confirmed to have improved in terms of overall error reduction. Figure 16 shows the relationships between the predicted and the observed values. From the comparisons, CAT-Kriging provides more accurate prediction than the others, particularly in the case of exaggerated predictions of uncontaminated stream segments in Fig. 16b.

Conclusions

In this study, new prediction methods, namely the STD-Kriging and CAT-Kriging, were developed and applied to synthetic and real-world datasets. These methods predicted

Table 4 Euclidean distance, stream distance, and catchment basin area difference between samples

Sample ID (A)	Sample ID (B)	Euclidean distance between (A) and (B) (unit: m)		Stream distance between (A) and (B) (unit: m)		Catchment basin area difference between (A) and (B) (unit: km^2 (pixels))	
		100 m resolution	30 m resolution	100 m resolution	30 m resolution	100 m resolution	30 m resolution
1	2	500.0	488.4	582.8	549.4	3.07 (307)	3.13 (3483)
2	3	360.6	318.9	382.8	344.6	0.48 (48)	0.23 (253)
3	4	1843.9	1841.8	2131.4	2213.1	8.45 (845)	8.87 (9850)
4	5[a]	141.4	134.2	Not connected	Not connected	Not connected	Not connected
5[a]	6	424.3	426.4	482.8	512.1	21.17 (2117)	21.06 (23,396)
6	7	1360.1	1376.4	1690.0	1573.7	53.53 (5353)	53.47 (59,416)
7	8[a]	707.1	700.4	765.7	796.7	30.44 (3044)	30.28 (33,646)
8[a]	9	806.2	823.8	1007.1	1031.5	30.92 (3092)	30.89 (34,327)
9	10[a]	223.6	247.4	Not connected	Not connected	Not connected	Not connected
10[a]	11[a]	1118.0	1075.4	1348.5	1368.8	4.14 (414)	4.57 (5082)
11[a]	12	1272.8	1272.8	1931.4	2003.1	82.68 (8268)	83.06 (92,290)
12	13	4123.1	4117.0	5062.7	5038.6	14.06 (1406)	14.07 (15,630)
13	14[a]	824.6	816.1	Not connected	Not connected	Not connected	Not connected
14[a]	15[a]	943.4	973.5	1007.1	1068.8	1.11 (111)	1.18 (1310)
15[a]	16	1964.7	1986.6	2272.8	2315.5	163.28 (16,328)	155.47 (172,748)
16	17	1711.7	1719.4	2090.0	2156.1	3.56 (356)	3.07 (3412)
17	18	3935.7	3931.0	5045.6	5030.4	23.01 (2301)	22.87 (25,413)
18	19[a]	1529.7	1465.1	Not connected	Not connected	Not connected	Not connected
19[a]	20	2334.5	2399.4	3407.1	3388.2	224.74 (22,474)	224.14 (249,039)
20	21	1486.6	1443.1	2738.5	2564.9	4.62 (462)	4.61 (5118)
21	22	2119.0	2135.1	2572.8	2428.2	10.23 (1023)	9.98 (11,093)

[a] Sampled from uncontaminated stream

the concentrations of stream sediment samples rationally by considering stream networks. In particular, CAT-Kriging reduces the exaggeration problem in predicting the sample of uncontaminated stream segment. As a result of cross-validation, CAT-Kriging obtained the best result for each sample prediction and overall error reduction. This study has several important advantages over existing studies on aquatic variables. First, this study performed a prediction for aquatic variables along the stream network in contrast to studies that apply Kriging on an entire area/region or analyze only the patterns of aquatic variables without prediction. In addition, this study takes flow direction, stream networks, or catchment basin into account by analyzing the topographical conditions based on a DEM. Owing to the raster format basis of the study, the continuous change of variable in a stream segment can be predicted, and there is no necessity for constructing objects or database to preserve information, unlike the vector-based studies. The new software was developed to automate the process in raster format, and it can provide fast analysis by adjusting the variables easily.

A problem exists in predicting the upstream value in the presence of contaminated downstream samples. To reduce the problem, CAT-Kriging was proposed in this study. However, this issue cannot be solved thoroughly if there is no sufficient upstream data in that stream segment. In addition, the upstream value may have some problems in modeling a valid variogram using the catchment basin area instead of the distance. If sufficient samples are available in each stream segment, the application of each variogram and prediction for each stream segment can be a solution for this problem. This study can prioritize environmental hazards and provide useful information for the reclamation of stream networks. Furthermore, this study is universally applicable to abandoned mines, industrial areas, farming areas, and other various environments and can be applied to stream sediments and other aquatic variables. The effectiveness of the prediction is expected to be improved

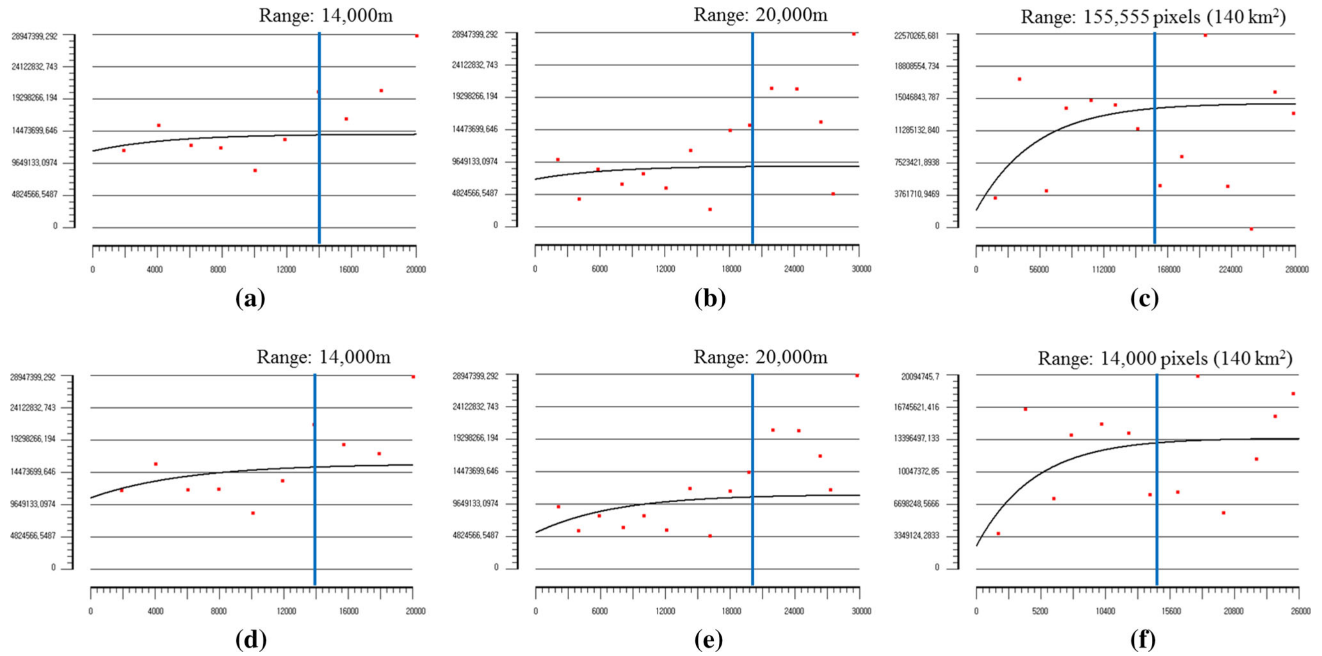
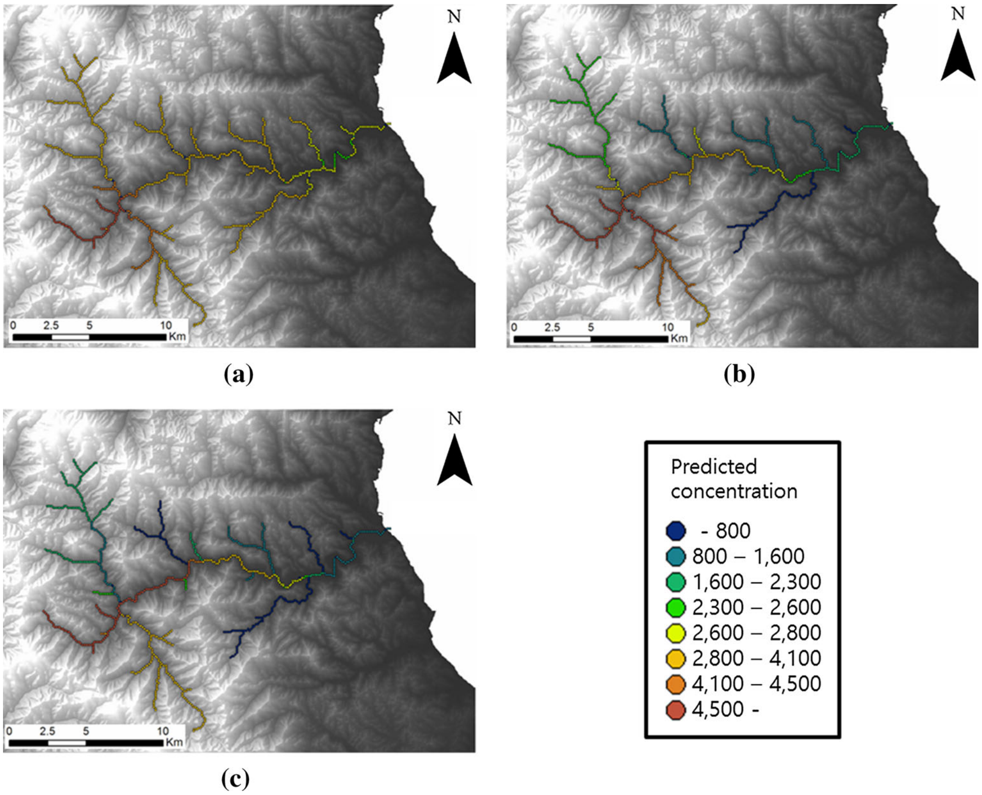
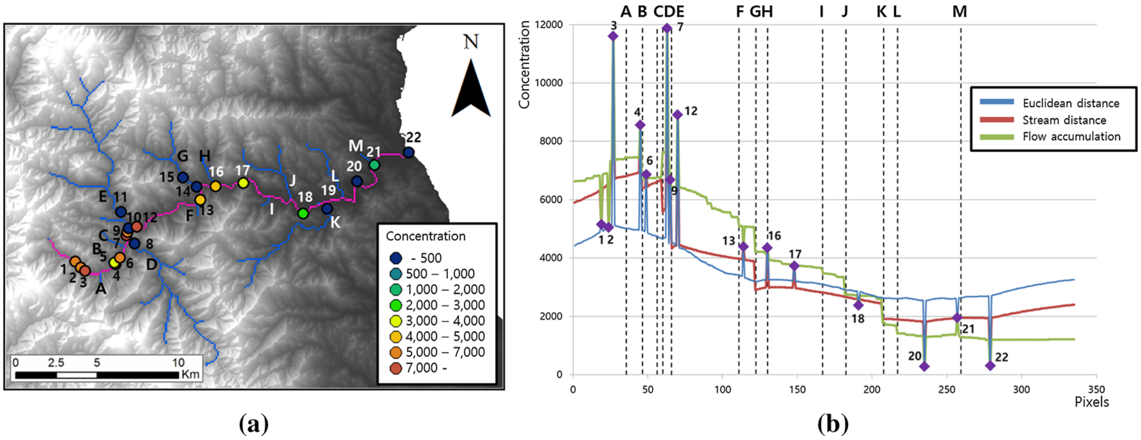


Fig. 13 Empirical and theoretical variograms calculated from **a** Euclidean distance, **b** stream distance, and **c** catchment basin area using 30-m-resolution DEM. Empirical and theoretical variograms calculated from **d** Euclidean distance, **e** stream distance, and **f** catchment basin area using 100-m-resolution DEM

Fig. 14 Prediction of Zn concentrations for the study area using **a** EUC-Kriging, **b** STD-Kriging, and **c** CAT-Kriging





(a)

(b)

Fig. 15 Graph of the predicted Zn concentrations for the study area stream path represented by a *pink line*. *A* to *M* are stream junctions

Table 5 Cross-validation results of EUC-Kriging, STD-Kriging, and CAT-Kriging for the study area (the bold number is the most similar value to the observed value)

Sample ID	Original value	EUC-Kriging	STD-Kriging	CAT-Kriging
1	5157.8	**4868.0**	6593.6	7577.7
2	5046.3	**5036.7**	6834.2	7972.9
3	11,607.3	4222.7	5741.9	**5923.7**
4	8558.4	4529.2	6656.1	**7044.7**
5[a]	3090.4	**5171.4**	6060.1	5931.6
6	6859.3	4680.5	6403.7	**6701.3**
7	11,882.3	3979.9	5094.4	**5421.3**
8[a]	150.6	**5010.2**	6248.5	6584.0
9	6684.0	4405.7	**5697.0**	8064.4
10[a]	228.7	4939.3	3673.3	**1689.7**
11[a]	205.3	4611.7	3372.4	**1615.8**
12	8910.7	3923.8	4178.6	**4231.6**
13	4385.2	3204.9	**3854.0**	5655.0
14[a]	161.1	3594.6	2094.8	**820.2**
15[a]	170.3	3619.2	1986.4	**811.9**
16	4346.7	3099.3	2784.2	**3706.8**
17	3726.2	3126.4	2813.9	**3783.3**
18	2380.0	2943.6	**2636.7**	3033.4
19[a]	108.6	3109.6	811.2	**792.4**
20	279.6	2968.3	2196.6	**1915.7**
21	1955.6	2745.7	**1927.8**	1060.7
22	307.1	3230.2	2439.7	**1899.3**
ME		37.3	177.3	274.3
MAE		2954.3	2337.6	2095.3
RMSE		3633.0	3016.7	2844.0

Unit: mg/kg

[a] Sampled from uncontaminated stream

if the stream networks have a complex and sharp bend. Based on the assumption that contaminants disperse through permanent or transient flow, this study can be applied to the case of soil contamination in semi-arid or arid climates. However, if the effect of rainfall is negligible or the shape of the catchment area is not definite, CAT-

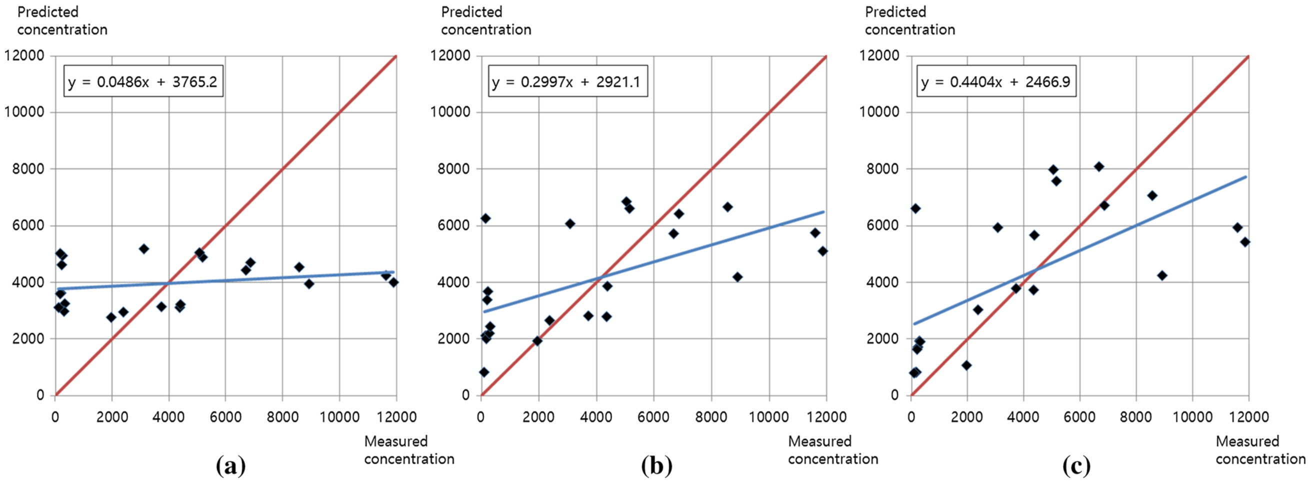


Fig. 16 Observed values versus predicted values based on cross-validation for **a** EUC-Kriging, **b** STD-Kriging, and **c** CAT-Kriging. The *red line* indicates the position of one-to-one correspondence. The *blue line* with the equation represents the trend line of points

Kriging may not provide a good prediction. Therefore, it is necessary to apply this methodology to various environments and conditions.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A1A01061290) and Research Institute of Energy and Resources (RIER), Seoul National University, Korea.

References

Aelion CM, Davis HT, Liu Y, Lawson AB, McDermott S (2009) Validation of Bayesian Kriging of arsenic, chromium, lead and mercury surface soil concentrations based on internode sampling. Environ Sci Technol 43:4432–4438

Chiles J, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, New York, p 734

Choi Y (2012) A new algorithm to calculate weighted flow-accumulation from a DEM by considering surface and underground stormwater infrastructure. Environ Modell Softw 30:81–91

Choi Y, Park HD, Sunwoo C (2008) Flood and gully erosion problems at the Pasir open pit coal mine, Indonesia: a case study of the hydrology using GIS. Bull Eng Geol Environ 67:251–258

Choi Y, Yi H, Park HD (2011) A new algorithm for grid-based hydrologic analysis by incorporating stormwater infrastructure. Comput Geosci 37:1035–1044

Cressie N (1985) Fitting variogram models by weighted least squares. Math Geol 17:563–586

Curriero F (1996) The use of non-euclidean distance in geostatistics. Ph.D. thesis, Kansas State University, USA

Dent CL, Grimm NB (1999) Spatial heterogeneity of stream water nutrient concentrations over successional time. Ecology 80:2283–2298

Heathwaite AL, Quinn PF, Hewett CJM (2005) Modelling and managing critical source areas of diffuse pollution from agricultural land using flow connectivity simulation. J Hydrol 304:446–461

Jenson SK, Domingue JO (1988) Extracting topographic structure from digital elevation data for geographic information system analysis. Photogr Eng Remote Sens 54:1593–1600

Khalil A, Hanich L, Bannari A, Zouhri L, Pourret O, Hakkou R (2013) Assessment of soil contamination around an abandoned mine in a semi-arid environment using geochemistry and geostatistics: pre-work of geochemical process modeling with numerical models. J Geochem Explor 125:117–129

Kim SM, Choi Y, Suh J, Oh S, Park HD, Yoon SH (2012a) Estimation of soil erosion and sediment yield from mine tailing dumps using GIS: a case study at the Samgwang mine, Korea. Geosyst Eng 15:2–9

Kim SM, Choi Y, Suh J, Oh S, Park HD, Yoon SH, Go WR (2012b) ArcMine: a GIS extension to support mine reclamation planning. Comput Geosci 46:84–95

Kim SM, Suh J, Oh S, Son J, Hyun CU, Park HD, Shin SH, Choi Y (2016) Assessing and prioritizing environmental hazards associated with abandoned mines in Gangwon-do, South Korea: the Total Mine Hazards Index. Environ Earth Sci 75:1–14

Lee S, Choi Y (2016) Reviews of unmanned aerial vehicle (drone) technology trends and its applications in the mining industry. Geosyst Eng 19:197–204

Lee H, Choi Y, Suh J, Lee SH (2016) Mapping copper and lead concentrations at abandoned mine areas using element analysis data from ICP–AES and portable XRF instruments: a comparative study. Int J Environ Res Public Health 13:1–15

Lin YP, Chang TK, Teng TP (2001) Characterization of soil lead by comparing sequential Gaussian simulation, simulated annealing simulation and Kriging methods. Environ Geol 41:189–199

Little LS, Edwards D, Porter DE (1997) Kriging in estuaries: as the crow flies, or as the fish swims? J Exp Mar Biol Ecol 213:1–11

Liu XM, Xu JM, Zhang MK, Huang JH, Shi JC, Yu XF (2004) Application of geostatistics and GIS technique to characterize spatial variabilities of bioavailable micronutrients in paddy soils. Environ Geol 46:189–194

McCool DK, Brown LC, Foster GR (1987) Revised slope steepness factor for the universal soil loss equation. Trans ASAE 30:1387–1396

Menafoglio A, Guadagnini A, Secchi P (2014) A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stoch Environ Res Risk Assess 28:1835–1851

MIRECO (2007) A report on the basic and detailed design of project for preventing loss of mine tailings at the area of Yeonhwa II mine, Samcheok-si, Korea. MIRECO, Seoul, p 508 (**in Korean with English abstract**)

Park CY, Park YS, Jeong YJ (1995) Contamination of heavy metals in soil in the Kwangyang mine area. J Korean Soc Miner Energy Resour Eng 32:163–174 (**in Korean with English abstract**)

Rathbun SL (1998) Spatial modeling in irregularly shaped regions: Kriging estuaries. Environmetrics 9:109–129

Salgueiro AR, Avila PF, Pereira HG, Santos Oliveira JM (2008) Geostatistical estimation of chemical contamination in stream sediments: the case study of Vale das Gatas mine (northern Portugal). J Geochem Explor 98:15–21

Shamsi M, Noaparast M, Shafaie SZ, Gharabaghi M (2016) Synergism effect of collectors on copper recovery in flotation of copper smelting slags. Geosyst Eng 19:57–68

Skøien JO, Merz R, Blöschl G (2006) Top-Kriging-geostatistics on stream networks. Hydrol Earth Syst Sci 10:277–287

Smith RA, Schwarz GE, Alexander RB (1997) Regional interpretation of water-quality monitoring data. Water Resour Res 33:2781–2798

Song J, Choi Y (2015) Design of photovoltaic systems to power aerators for natural purification of acid mine drainage. Renew Energy 83:759–766

Steiger B, Webster R, Schulin R, Lehmann R (1996) Mapping heavy metals in polluted soil by disjunctive Kriging. Environ Pollut 94:205–215

Suh J, Choi Y, Park HD, Yoon SH, Go WR (2013) Subsidence hazard assessment at the Samcheok coalfield, South Korea: a case study using GIS. Environ Eng Geosci 19:69–83

Thornton I (1983) Applied environmental geochemistry. Academic Press, San Diego, p 501

Torgersen CE, Gresswell RE, Bateman DS (2004) Pattern detection in stream networks: quantifying spatial variability in fish distribution. In: Proceedings of the second annual international symposium on GIS/spatial analyses in fishery and aquatic sciences, Fishery GIS Research Group, Saitama, Japan

VerHoef JM, Peterson E, Theobald D (2006) Spatial statistical models that use flow and stream distance. Environ Ecol Stat 13:449–464

White JG, Welch RM, Norvell WA (1997) Soil zinc map of USA using geostatistics and geographic information systems. Soil Sci Soc Am J 61:185–194

Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses: a guide to conservation planning (Handbook No. 537). United States Department of Agriculture, Washington, DC, p 58

Yenilmez F, Kyter N, Emil MK, Aksoy A (2011) Evaluation of pollution levels at an abandoned coal mine site in Turkey with the aid of GIS. Int J Coal Geol 86:12–19

Yuan LL (2004) Using spatial interpolation to estimate stressor levels in unsampled streams. Environ Monit Assess 94:23–38