Arabian Journal of Geosciences (2019) 12: 104
https://doi.org/10.1007/s12517-018-4171-5

ORIGINAL PAPER

Optimizing site-specific geostatistics to improve geotechnical spatial information in Seoul, South Korea

Han-Saem Kim [1] · Hyun-Ki Kim [2]

Received: 5 July 2018 / Accepted: 14 December 2018 / Published online: 4 February 2019

Abstract
Subsurface soil and rock profiles are commonly interpreted from borehole log datasets. These datasets include three-dimensional spatial coordinate information, layer information, and standard penetration test results. More reliable spatial distribution of target physical properties can be obtained from additional testing at locations characterized by outlier observations and geotechnical uncertainties. At a given site, irregular measurements typically differ significantly from bulk measurements or proximal observations. In this study, a process for optimizing site-specific geostatistics, which uses geotechnical spatial information and applies optimum outlier thresholds with a multi-clustering method, is proposed to incorporate site-specific geo-layer uncertainties and identify their geotechnical value. Optimized geostatistical characteristic information for geological strata boundaries was derived and verified based on a sequential procedure applied to representative test areas in Seoul, South Korea.

Keywords Geotechnical spatial uncertainty · Borehole · Geostatistical analysis · Outlier analyses · Clustering

Introduction

The spatial uncertainty in geotechnical properties associated with heterogeneous lithologic compositions has been discussed and reported extensively (Vanmarke 1977; Asaoka and A-Grivas 1982; Degroot and Baecher 1993; Lacasse and Nadim 1997; Phoon and Kulhawy 1999; Chun et al. 2005; Kim et al. 2012). In particular, it is important that geologic strata boundaries and associated geotechnical characteristics are derived for the proposed development site, e.g., foundation, excavation, or embankment. Geotechnical uncertainties decrease geotechnical design reliability, requiring the reduction or removal of measurement errors (over or under the relative design criteria) using appropriate statistical methods (Grubbs 1969; Barnett and Lewis 1994; Zhang et al. 2007; Chandola and Kumar 2009; Orr and Breysse 2008; Phoon 2008). Furthermore, where geotechnical uncertainty exists, focused testing should identify any geological or geotechnical outliers. For a given site, irregular measurements typically differ significantly from other measurements or nearby observations. However, even where sampling is quite regular, spatial correlations can be weak in locations with high geological heterogeneity.

Datasets for geotechnical engineering practices are obtained from a variety of information sources, including in situ investigations, laboratory tests, field monitoring, and computer analyses, which have inherent errors or uncertainties that can be transferred to design parameters. To understand the spatial variability and identify predictable errors, spatial numerical modeling of geotechnical datasets based on appropriate geostatistical or statistical methodologies that consider spatial uncertainties induced by multiple geospatial datasets is necessary. Geostatistics is a branch of spatial statistics with a well-developed mathematical/statistical theoretical framework based on probability theory (Delfiner 1976; Isaaks and Srivastava 1989). It can be regarded as a collection of numerical techniques that characterize spatial relationships. For example, prior work has attempted to obtain the multi-dimensional distribution of soil properties by analyzing a multitude of one-dimensional borehole data using geostatistical techniques (Öztürk and Nasuf 2002).

✉ Han-Saem Kim
adoogen@kigam.re.kr

Hyun-Ki Kim
geotech@kookmin.ac.kr

[1] Earthquake Research Center, Korea Institute of Geoscience and Mineral Resources, 124 Gwahak-ro, Daejeon, Yuseong-gu 305-350, South Korea

[2] Department of Civil and Environmental Engineering, Kookmin University, 77 Jeongneung-ro, Seoul, Seongbuk-gu 136-702, South Korea

Most geostatistical studies do not apply uncertainty measures provided by smoothing methods. Only mapped smoothed rates and interpolated properties are reported and used in such analyses. This has disadvantages because all rate smoothers, including ordinary kriging, cause the loss of local spatial variation risk details (Goovaerts 2006). Therefore, to apply smoothing methods using initial datasets with reduced errors, an optimization framework for site-specific outlier analysis should be developed for geospatial datasets from undermined areas with various spatial correlations. These uncertainties render the geotechnical design less reliable, and measurement errors must be reduced or removed with appropriate statistical methods. Typically, such erratic measurements are numerically distinguishable from the remaining data or deviate significantly from other proximal observations.

Kim et al. (2012) proposed a framework to detect outlier data using statistical analyses, a cross-validation-based method, and generalized extreme-value distribution–based method. Borehole datasets that include the soil depth distribution in regions of central Seoul (South Korea) were assessed to validate the aforementioned methods through a comparison between distribution-based methods and a Moran scatterplot method (Anselin 2004). The results indicated that outlier methods that consider spatial correlations facilitate obtaining more reliable spatial distributions, and with a quantitative evaluation of local reliability. These outlier methods are closely related to clustering methods. Lu et al. (2003) defined a spatial outlier as a spatially referenced object with non-spatial attribute values significantly different from those of its neighborhood. Quantitative methods provide tests to distinguish such spatial outliers from the remaining data in a subcluster. Therefore, to develop geospatial information based on optimized site-specific borehole datasets, conventional outlier analysis and spatial interpolation methodologies should be integrated and optimized while accounting for outlier locations (Yu et al. 2002).

In this study, a framework for optimizing geostatistics for geotechnical spatial information is proposed based on applying a site-specific outlier detection method. The framework uses clustering to incorporate site-specific uncertainties into the geo-layer and geotechnical characteristic values. Sequential optimization phases were designed for the decision-making procedure based on geographic information system (GIS) architecture to develop the geotechnical spatial information grid. Test areas in Seoul, South Korea, were selected to validate the framework. First, the multi-source geo-layer information was archived and standardized to reflect the geostatistical characteristics corresponding to the influence of the geotechnical layer and topological effects. Second, the geostatistics of the geo-layer information were derived using a GIS approach involving geostatistical spatial interpolation and density–topology estimation. Accordingly, a sampling of representative borehole information, considering local site effects, was conducted, and the corresponding test areas were selected. Third, by clustering the borehole information, the geotechnical spatial information for targets in each test area was developed and normalized using geostatistical optimum criteria. The optimum range for zonal clustering was determined based on a cross-validation method using multi-circular zonation. Finally, a cross-validation-based outlier detection method was proposed and applied to each optimized cluster. Here, the relative outlier threshold was determined based on the density and spatial correlations of boreholes in each cluster. The geotechnical spatial information grid was developed without any site-specific outliers. Thus, site classification criteria for optimum outliers at the study site were proposed and validated.

Framework for optimizing site-specific geotechnical spatial information

To determine the geotechnical design variables associated with specific zones, such as redevelopment areas, it is essential to estimate reasonable spatial information from the characteristic geotechnical values using surrounding borehole datasets prior to additional surveying. Therefore, outlier detection for each cluster, within the optimum clustering range, should be conducted to construct a geotechnical spatial information grid using the appropriated geostatistical method. To determine site-specific outliers, considering the distribution pattern and spatial correlations of the borehole datasets, a systematic framework for optimizing geostatistics for geotechnical spatial information is proposed, as illustrated in Fig. 1.

First, multi-source geospatial information, e.g., geotechnical investigation data, geological maps, land cover maps, and other infrastructure information within the same spatial coordinate system, was collected. Then, geo-modeling and reprocessing, using GIS toolsets (Davis et al. 2015), was performed to determine the primary relationships between various geo-datasets corresponding to the overlying and zonal

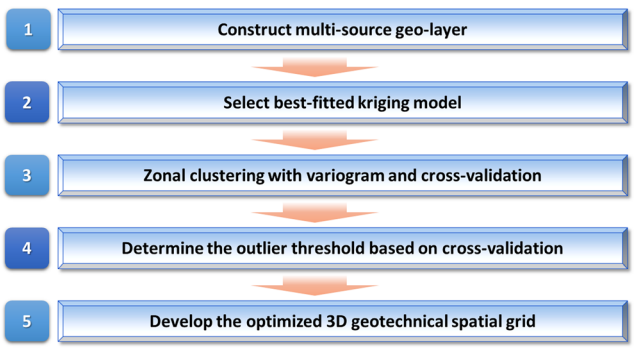


Fig. 1 Schematic procedure for evaluating site-specific geotechnical characteristic values using an outlier analysis based on a GIS platform

characteristics based on observations. Geo-layer datasets with similar geotechnical and geostatistical characteristics were clustered using geostatistical estimates and hot-spot analysis. Local geotechnical datasets were classified using geostatistical density analyses to identify clusters with similar spatial correlations in geo-layer characteristics. The interpolated geo-layer based on kriging, kernel density, and digital elevation models (DEMs) was assigned as the classification criteria for local geo-layer properties. Hot-spot clustering of the classification criteria was conducted to identify representative areas, considering the site classification of the topological surface information based on multiple geo-layers.

Second, to optimize the conditions for random-field assumptions in the kriging methods to incorporate interpolation and zonation, appropriate geostatistical methods and corresponding variogram models were validated and determined using a cross-validation-based verification test (Rue and Follestad 2003). The best-fitting kriging model for depth to bedrock was selected to construct the geotechnical spatial grid. Because there are spatial variations in the bedrock depths, due to heterogeneous geological formations, and spatial density of the borehole measurements, local subsets with constant areas were selected to determine the optimization criteria using site-specific spatial correlations.

Third, the range in optimum circular clustering area for each subset was determined considering the correlation between the radius of the influencing circle and variogram-based effective range for bedrock depths. A trial-and-error approach, by constantly increasing the radius of the circular cluster, was applied to determine the borehole datasets with relatively constant correlations. Subsequently, the outer threshold of each subset was determined using the borehole datasets within the optimized sub-cluster, thus identifying the outlier observations outside the optimum threshold. This is achieved within a more reliable spatial distribution of target physical properties by reducing or removing data that significantly deviate from adjacent observations or results from additional tests performed at the same location. Site-specific outliers were determined for each sub-cluster. The borehole datasets were modified by applying the geostatistical optimum criteria, and a geotechnical spatial grid was developed using these optimized borehole datasets.

Based on the proposed framework (Fig. 1), a previously constructed geospatial database was used as an initial input dataset for the stage-by-stage procedures using a GIS-based multi-layer information platform. To build the database, borehole datasets and geospatial information, including geotechnical engineering, geology, and geomorphology data, were collected and standardized. For a more reliable prediction of geotechnical information in the area of interest, topological surface information was extracted from topographic maps, satellite images, surface geology, and digital elevation models (DEMs) (Kim et al. 2017).

Geo-spatial analysis and optimization of geotechnical datasets in Seoul, South Korea

Evaluating the test area site descriptions

A geospatial database for Seoul was constructed and applied based on the geodatabase schema in the ArcGIS platform to assess site-specific geospatial distribution patterns, specifically the depth to bedrock and local differences between spatial components in the geospatial database (Fig. 2). The Seoul area was separated into 100-m mesh areas, yielding 225,835 spatial grids (interpolation grids), including the extended area of the Seoul administrative area. Component mesh-unit data were created for each spatial grid (Kim et al. 2017). The authors chose the target study area of the entire territory surrounded by the administrative boundaries of the Seoul metropolitan areas as the administrative region. The target area is the largest urban area in Korea. Geospatial information included borehole datasets, a digital elevation model, digital numerical information (e.g., watershed and administrative boundaries), infrastructure information (e.g., roads, buildings, and pipelines), geological maps, land cover maps, and other geo-proxy-based map information.

Existing borehole data were gathered, and a geo-knowledge-based site visit was conducted across the Seoul area to acquire surface geospatial data, focusing on mountainous or undeveloped areas. Estimates of spatial geotechnical layers across the extended area were collected from a total of ~22,300 existing borehole datasets and ~1700 surface geo-knowledge datasets. To spatially estimate soil layers, the optimized site-specific interpolation method was applied to the Seoul area (53.0 km east–west, 39.0 km north–south). Figure 2 schematically illustrates the procedure for constructing the multi-source geo-layer information in Seoul based on a GIS platform. In addition to the digital elevation model, information for roads, buildings, and pipelines were collected from building registers in Seoul and, using the GIS platform, converted into geospatial datasets based on the coordinate information at each vertex to provide the overall infrastructure. In addition, geophysical survey results were collected and constructed as tomography information with two- or three-dimensional coordinates based on the geodatabase.

Geostatistical characterization of geo-layer information

Generally, there are variations in spatial density of geotechnical datasets because of their collection purpose (Kim et al. 2017). These geotechnical datasets are distributed spatially as linear or circular clusters focused on urban facility sites for engineering projects. Accordingly, there are spatial correlations or patterns of spatial interpolation depending on the

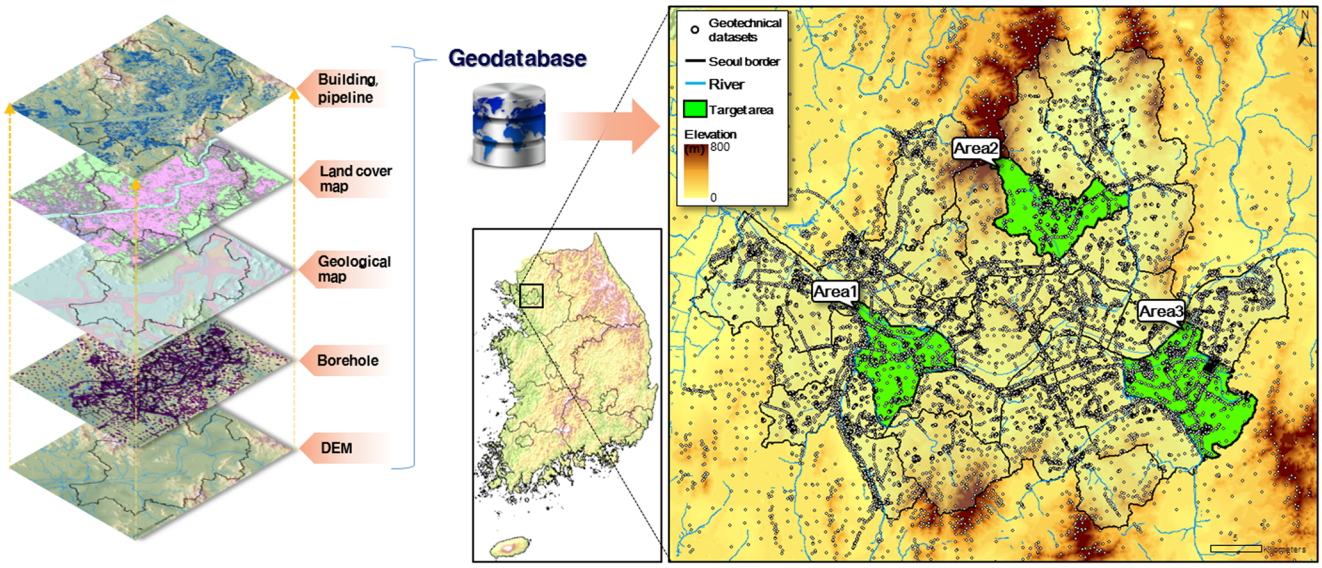


Fig. 2 Schematic procedure for constructing multi-source geo-layer information in Seoul based on a GIS platform

density of a specific cluster in the target area. Thus, geotechnical datasets with similar spatial correlations are grouped together as spatial fields considering the spatial correlations of unit clusters using conventional geostatistical methods; then, they are used to assign sequential analysis categories to determine the appropriate zonation method (Kim et al. 2016). Hence, the spatial correlations or patterns between geotechnical and infrastructure information have to be estimated using a geostatistical density analysis before optimizing the geospatial database and selecting representative areas, which are classified by various geological and topographic properties in the Seoul area.

Kernel density is used to calculate the magnitude per unit area from point or polyline features using a kernel function to fit a smoothly tapered surface to each point. The kernel density is a well-established method used to identify spatial patterns that calculates the density of events around each point, scaled by the distance from the point to each event. The kernel density describes a smooth and continuous surface map of risk targets because a discrete pattern is continuously created by interpolation (Kim et al. 2017). Thus, this method can compensate for any paucity in the data. A general density estimation function is shown as follows:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{K(x - x_i)}{h}, \quad (1)$$

where x_i is the value of the variable x at location i, n signifies the total number of locations, h denotes the bandwidth or smoothing parameter, and K represents the kernel function (Borruso and Schoier 2004).

Using the described spatial analysis method, the geostatistical spatial information was evaluated using borehole datasets and the local DEM, as shown in Fig. 3. The average depth to bedrock was calculated as 18.2 m and was concentrated along the Han River based on the kriging, since the most of boreholes having thick alluvial layer (more than 10 m) were located along the river. Furthermore, a thick soil layer, more than 25 m, was confirmed in the downstream area (west Seoul). The local geotechnical datasets were also classified using the kernel density method to ensure that the geostatistical clusters maintained a similar spatial correlation of geo-layer characteristics. Rather than choosing an arbitrary interval, it is more appropriate to use the mean nearest-neighbor distance for different orders of K, which can be calculated using an ArcGIS toolset. Therefore, the geostatistical estimation was conducted discriminately over large-scale zones with similar spatial correlations based on specific groups (or clusters) considering the kernel density (Fig. 3b). The standard deviation of kernel densities between boreholes and pipelines at each grid was 3.23. To establish a relationship between spatial patterns and topological information, a DEM-based slope was estimated by building a triangular irregular network (TIN). The DEM-based slope was also extracted at borehole locations and connected with geo-layer boundaries to de-trend the effect of sloping layer boundaries and variable layer thicknesses depending on the topographic variability.

Using the kernel function, the optimized hot-spot analysis calculates the Getis–Ord statistic for each feature in a dataset to identify the locations of local clusters for site effect parameters (Gökkaya 2016). The Getis–Ord Gi* of the target feature is included in the analysis and indicates the locations of hot spots (clusters of high values) and cold spots (clusters of low values) in the area (Getis and Ord 1996; Prasannakumar et al. 2011). This method works by inspecting each feature within the context of neighboring features. To be statistically significant, a hot spot or cold spot should have a significantly high or low value and also be surrounded by other features with high or low values.

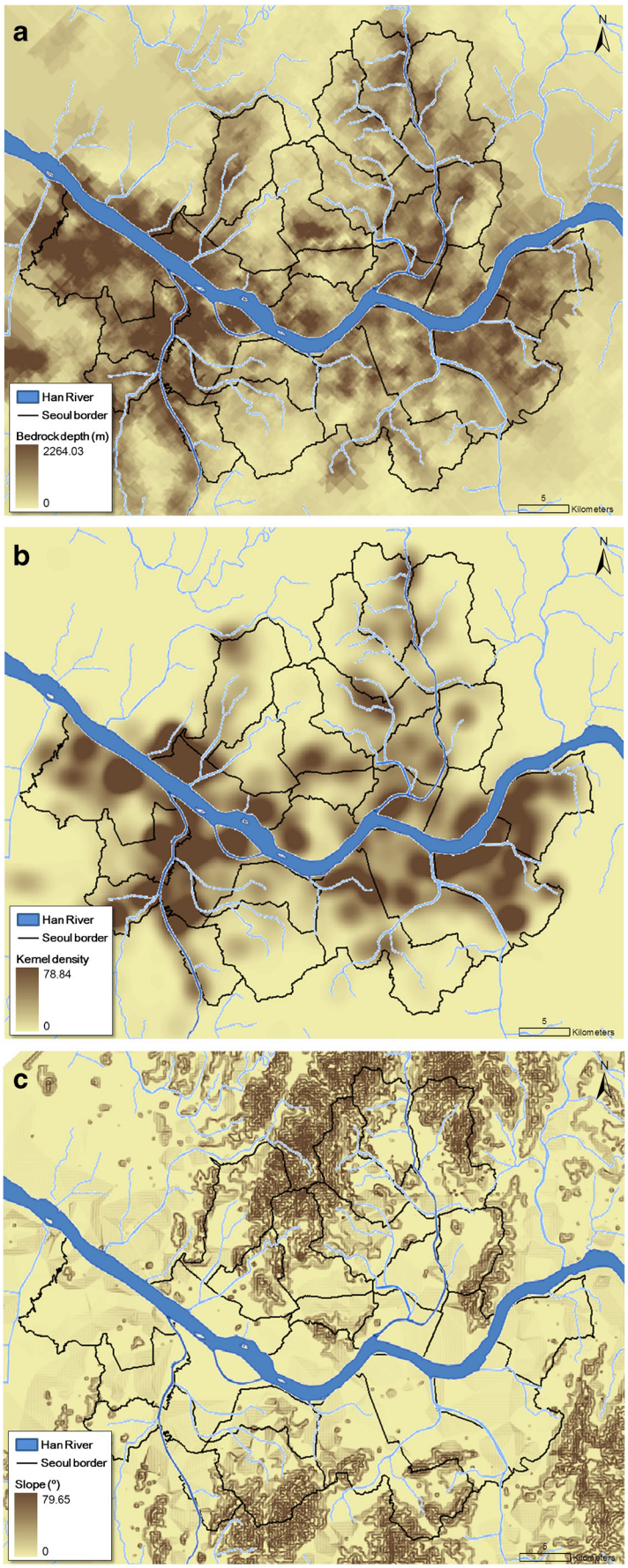


Fig. 3 Variations in geostatistical characteristics of the geotechnical and topology layer. **a** Bedrock depth using ordinary kriging, **b** kernel density of borehole datasets, and **c** slope

The hot-spot analysis method uses geostatistical characteristic values from borehole datasets and DEM data to identify the locations of statistically significant hot spots and cold spots. To evaluate site-specific concentrations of the dual variables, i.e., borehole-based geo-layer boundary and DEM, optimized hot-spot clustering was conducted using the kernel function. As the geo-layers are normally not strictly horizontal layers, the trend of sloping geo-layer boundaries should be removed before the residuals are analyzed by kriging, considering the correlations between slope and geo-layer boundaries. In this study, the local point map of correlations between bedrock depth and land surface information, elevation, slope, geology, and land cover was created using a kernel density estimation map (Fig. 4). Based on confidence distributions, the multivariable clusters were categorized into seven groups: three "cold spots" (with a 90%, 95%, and 99% confidence), three "hot spots" (with a 90%, 95%, and 99% confidence), and one "Not Significant" spot. To identify a strong positive relationship between the density of the depth to bedrock and slope information, a cluster that is related to the greatest depth to bedrock and highest kernel density of land surface characteristics (low slope), respectively, was defined as a "hot spot." In contrast, a "cold spot" was defined as the cluster with the lowest depth to bedrock density and lowest kernel density of land surface characteristics (high slope). The points on the map with a non-critical relationship between land surface information were classified as "Not Significant" spots.

In the Seoul area, a high potential for large depths to bedrock was determined in the western and east–central urban areas. Sub-groups (with a focus on "hot spots") with a lower potential for thick alluvium soil were distributed evenly across the entire area. However, the plains at the mouth of the Han River had a higher density of borehole datasets. Depending on the geotechnical datasets for each group, the construction of 2D geo-layer information was performed separately, based on the appropriate variogram method. Specifically, the variograms for each individual cluster were modeled for the test areas in Seoul following the proposed framework. To de-trend the variable layer thickness depending on the slope, the borehole datasets classified as "cold spots" with a 95% or 99% confidence were removed before the kriging procedure. Although borehole datasets were distributed at the same administrative region, the outlying borehole datasets, which were defined as the "cold spots" with a 95% or 99% confidence, were excluded for selecting the test area based on hot-spot clustering.

From the geostatistical characteristics from the geotechnical and topology layer (Fig. 3) and hot-spot clustering (Fig. 4), zonation information commensurate with the administration borders in Seoul was calculated based on an averaged grid value in each subset (Fig. 5). Consequently, "area 1" and "area 3," having greater bedrock depth and density (i.e., lower slope), were selected as test areas. These areas were classified as "hot spots" with more than 95% confidence. "area 2" was defined with a shallower bedrock depth (generally classified as "cold spot") and density (otherwise higher slope). For "area 1," "area 2," and "area 3," 4532, 2014, and 3854 borehole datasets were constructed, respectively, in the GIS database for site-specific assessment.

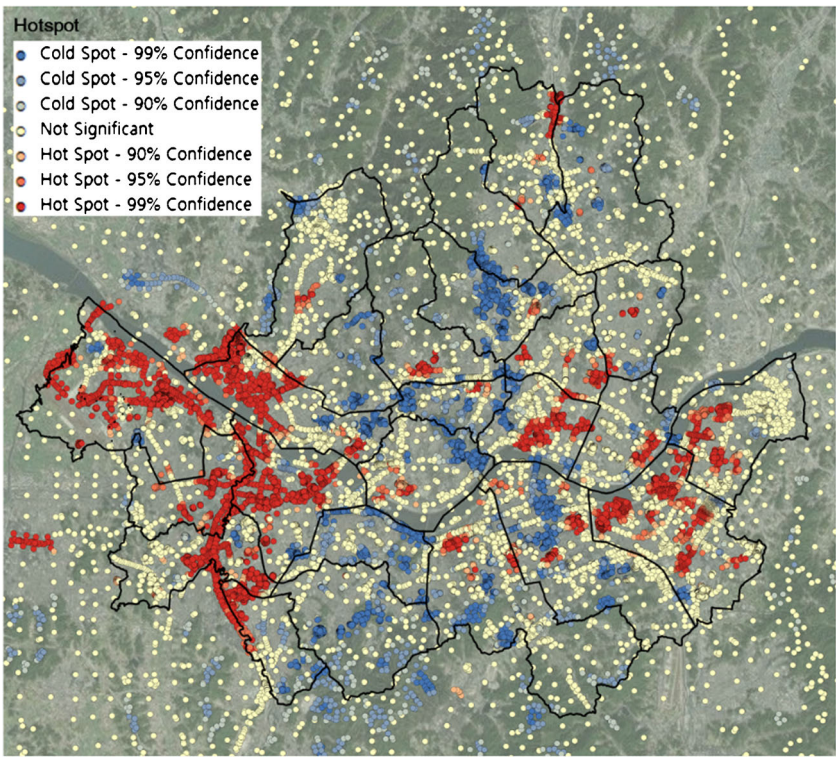


Fig. 4 Hot-spot clustering of geo-layer information considering correlations between the density of bedrock depth and slope

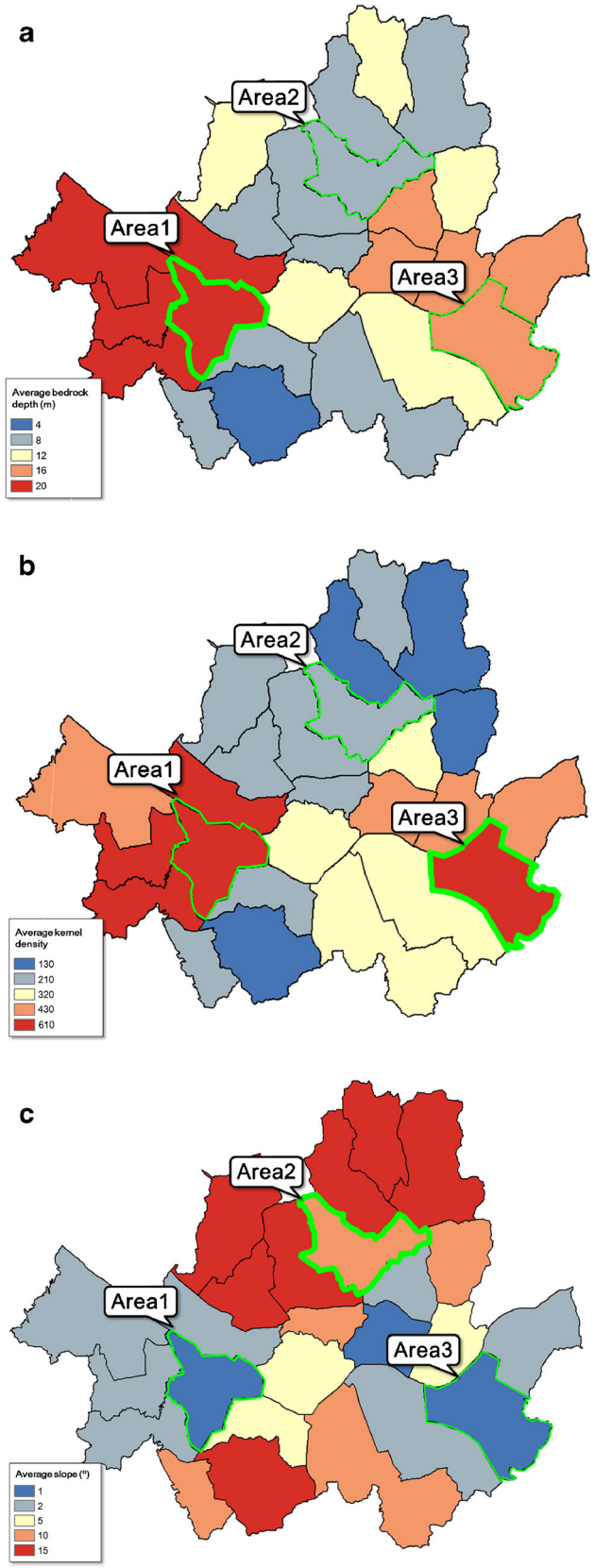


Fig. 5 Zonation information for selecting the test areas according to the administration borders in Seoul. **a** Average bedrock depth, **b** average kernel density, and **c** average slope

Selecting the best-fitting kriging model

Geostatistical interpolation can provide reliable spatial grid information for geo-layers (Kim et al. 2016). However, its efficacy relies on the accuracy of the interpolation method used to define the spatial variability in soil properties (Goovaerts 1998, 1999, 2001). The variogram is a mathematical description of the relationship between the variance of observation pairs and distance separating these observations (h) (Olea 1991). The fitted regression curve minimizes the variance in estimated errors. The variogram model is used to define the weights in the kriging function (Sun and Kim 2016), and the semi-variance is an autocorrelation statistic, defined as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{Z(x_i) - Z(x_i + h)\}^2, \quad (2)$$

where $\gamma(h)$ is the semi-variance for interval distance class or lag interval h, $N(h)$ is the total number of sample couples or observation pairs separated by distance h, $Z(x_i)$ is the measured sample value at point i, and $Z(x_i + h)$ is the measured sample value at point $i + h$ (Isaaks and Srivastava 1989; Azpurua and Dos-Ramos 2016). In this study, we considered the correlated distances within clusters and corresponding weights for the kriging function. The individual variogram model was also validated for every clustered geotechnical dataset using a cross-validation of the best-fitting kriging model.

If there is a lack of data, the large error in the variogram will increase the prediction errors, without it being apparent in the calculated values. Therefore, the results of the proposed step-by-step techniques were verified with separate independent data. The datasets were cross-validated to evaluate the kriging and zonation models and reduce the statistical uncertainty in the borehole data (David 1976; Guarascio et al. 1976; Knudsen and Kim 1978).

Kim et al. (2016) proposed a geostatistical analysis component for the optimum geostatistical estimation of soil conditions using conventional kriging methods. This computer-based framework consists of a step-by-step adaptive optimization technique, with an independent geostatistical method (Deutsch and Journel 1972). To determine the optimum interpolation method, four representative interpolation methods, i.e., the inverse distance method (IDW), simple kriging (SK), ordinary kriging (OK), and empirical Bayesian kriging (EBK), are used in a cross-validation-based verification test. Using the database for each of the three test areas, the site-specific geotechnical spatial datasets were interpolated. Prior to this interpolation, the geostatistical interpolation conditions were optimized and standardized using cross-validation-based verification tests for the geostatistical estimates, considering the site conditions in the study area. An optimum interpolation

method was determined by applying the four aforementioned representative interpolation methods. Consequently, cross-validation-based root mean square errors (RMSEs) were estimated for bedrock depths based on a 100-m grid-cell size in area 1, as shown in Fig. 6.

To estimate the cross-validated residuals based on the kriging and variogram model, an experimental semivariogram was computed and a plausible model was fit. After excluding the measured target values at a given point, the sequential value at each sampling point was estimated using a candidate kriging. The difference between the estimated and measured values at each sampling point was then calculated. For comparison, the RMSE from the cross-validation result was the square root of the average squared distance of a data point from the fitted line, as calculated with the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{3}$$

where y_i and $\hat{y}_i$ are the measured and estimated values of the ith data point, respectively, and n is the total number of data points. As the RMSE approaches 0, the estimate becomes more accurate. The coefficient of variation is the ratio of the RMSE to the mean of the dependent variable (Öztürk and Nasuf 2002). The spatial distribution of the standard deviation of depth to bedrock was also constructed based on the four spatial interpolation methods to estimate the spatial grid information residuals ($\hat{y}_i - y_i$) (Fig. 7).

By calculating the cross-validated RMSE using the residuals at each spatial grid, the OK method was determined to have the lowest RMSE, indicating that this technique was the most optimum geostatistical interpolation method for all target areas (Fig. 8). Additionally, the appropriate variogram model for this method was verified with a cross-validation procedure. To determine the best-fit variogram model, four representative variogram models (exponential model, circular model, spherical model, and Gaussian model) were used in a cross-validation-based verification test. The result showed that the exponential model had the lowest RMSE, indicating that the variogram model was the most appropriate model for all target areas (Fig. 9).

Multi-circular zonal clustering with variogram modeling and cross-validation

To determine the optimum range of a clustering area, accounting for the local variation in borehole data, variogram modeling was conducted for each circular cluster. To determine the need for robust estimation techniques, we fitted three standard theoretical variogram models (exponential, Gaussian, and spherical) to the unevenly spaced experimental variograms. First, each test area was divided into 3 × 3 metrics to define nine subsets with a constant area (Fig. 10). The effective range for each subset of an assumed radius (1, 2, 3, 4, 5, 6, 7, and 8 km) was calculated based on a variogram. Then, the correlation between the radius of the influencing circle and variogram-based effective range for bedrock depth was estimated, and the site-specific circular clustering (optimum subset) for each subset was determined (Fig. 11). The effective variogram range is the lag distance at which the semivariogram reaches a steady value, and is the value at which the variogram levels off. The effective range increases as the clustering boundary addition increases, but not at the same rate in the nine subsets.

The correlation between the radius of the influencing circle and variogram-based effective range indicates the change in spatial correlation (spatial coherency) due to the increasing number of borehole samples in the eight circular clusters. The best model for the exponential variogram was chosen by the minimum sum of squares from the fit, which is a conventional measure for the goodness-of-fit in a least-squares fitting procedure. Similarly, to determine the low variable rate of correlations (Fig. 9), a regression analysis was conducted based on the exponential model. The regression point with the smallest variable for the rate of change (of the correlations), i.e., there are sufficient borehole datasets with constant spatial correlations for spatial interpolation, can be defined as an indicator to determine the constant point of optimum radius for the extended area. Therefore, the optimum radius is assumed as the statistical inflection point of the determined regression line, based on the exponential model. The optimum radii for the nine subsets in area 1 are 0.42, 0.26, 0.53, 0.51, 0.42, 0.62, 1.02, 1.14, and 0.06 km. The geostatistical optimum criteria, including the clustering area for the three test areas, are listed in Table 1. Based on the determined optimum clustering radii, the borehole datasets in each circular cluster were grouped for site-specific outlier analyses and to construct the geotechnical spatial grid (Fig. 10).

Site-specific outlier analysis for multi-clustered borehole datasets

Determining the outlier observations based on optimum thresholds provides a more reliable spatial distribution of the target physical property by reducing or removing data that deviate significantly from proximal observations or by conducting additional tests at the same location. Site-specific outliers were determined for each cluster subset. The relative uncertainties render the geotechnical design less reliable, making it necessary to reduce or remove the "errors" from the measurements (Kulhawy et al. 1972). Typically, erratic measurements from engineering practices are numerically distant from the remaining data or deviate significantly from other proximal observations. Therefore, the possible outliers from the collected borehole datasets are regarded as "re-examinable

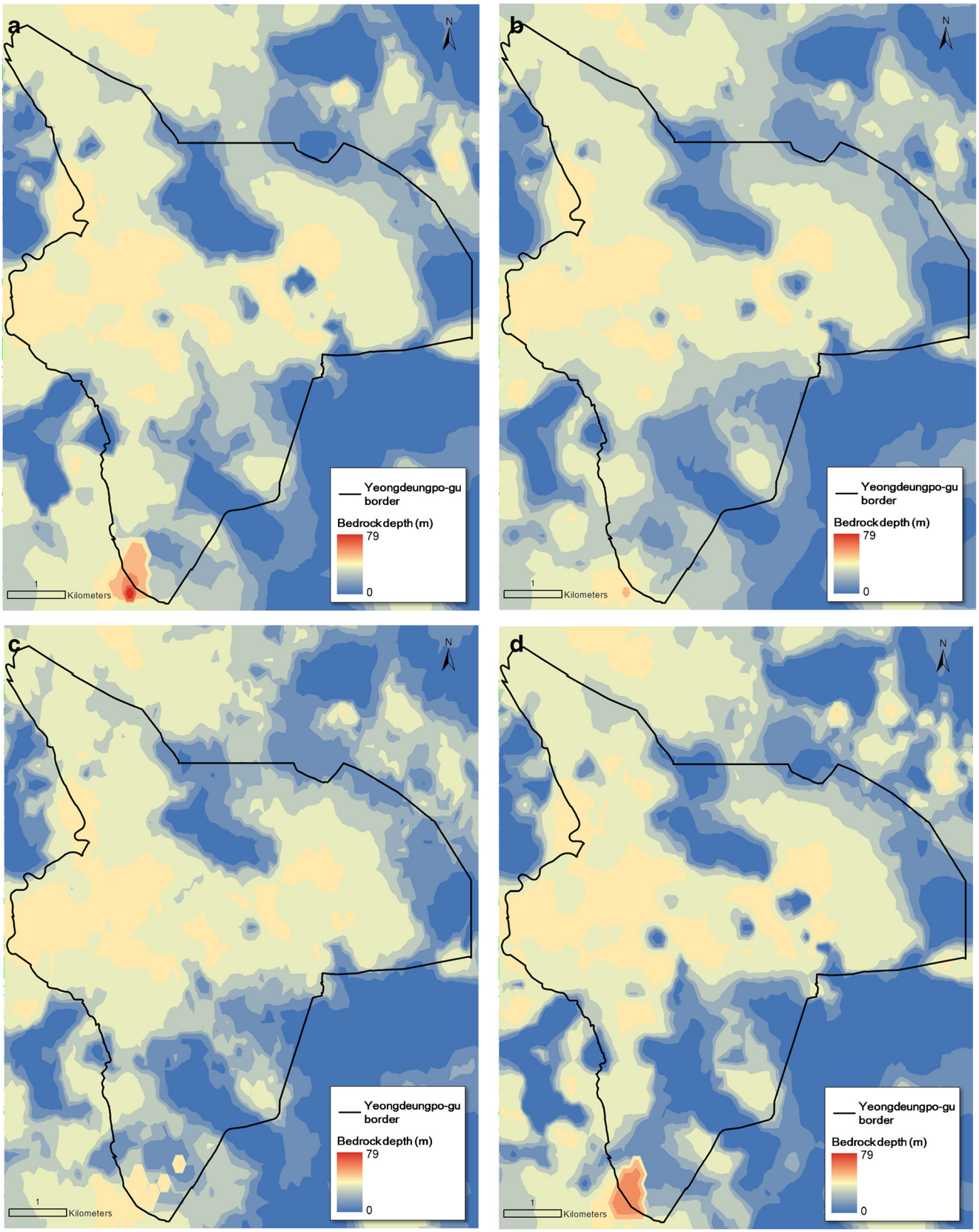


Fig. 6 Example of the spatial distribution of bedrock depth based on the spatial interpolation methods for area 1. **a** Inverse distance method, **b** simple kriging, **c** ordinary kriging, and **d** empirical Bayesian kriging

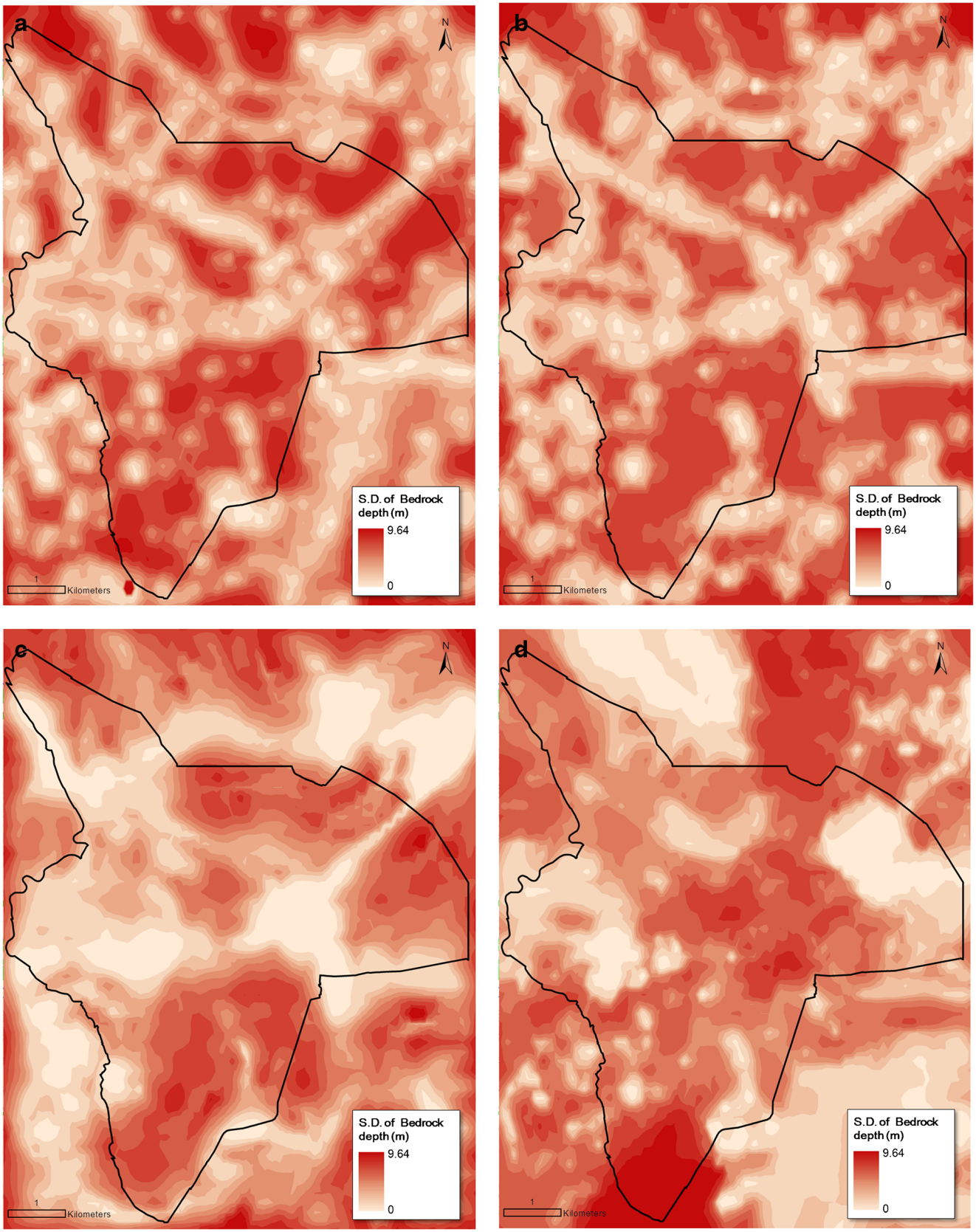
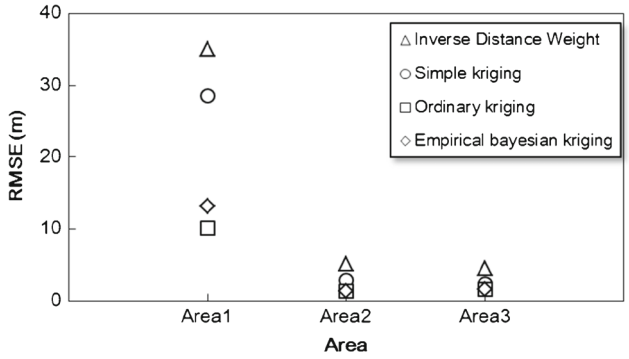


Fig. 7 Spatial distributions of the standard deviations of the bedrock depth for area 1 based on the spatial interpolation methods. **a** IDW, **b** SK, **c** OK, and **d** EBK



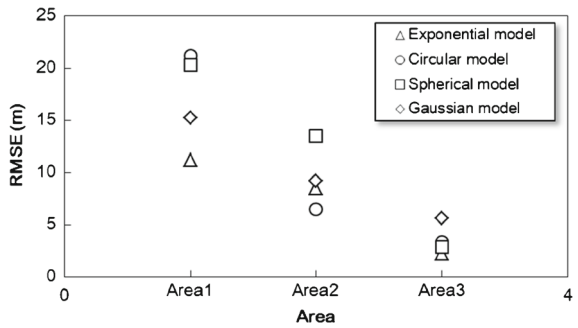
	Area 1	Area 2	Area 3
IDW	35.2	4.8	4.1
SK	28.1	3.5	2.9
OK	10.0	2.1	2.3
EBK	12.6	2.1	2.3

Fig. 8 Cross-validation-based RMSE based on the four spatial interpolation methods in the three test areas

references" to reconstruct geologic strata, and should be preferentially replaced by new borehole data.

To apply the outlier thresholds, which were determined by cross-validation, several basic assumptions should be considered. For example, if the data are normally distributed, with data points being fairly dense and uniformly distributed, fairly good estimates can be implemented irrespective of the interpolation algorithm. The data locations must fall within a few clusters with large gaps, which are non-clustered and non-directional. In addition, almost all interpolation algorithms underestimate high values and overestimate low values, a property inherent to averaging. Under these assumptions, the cross-validation-based outlier threshold determination using kriging provides the best unbiased linear prediction of relative residuals between the observed and estimated values.

According to Kim et al. (2012), the outlier thresholds for all geological layers should be applied equally, such that the



	Area 1	Area 2	Area 3
Exponential	11.3	7.6	2.1
Circular	22.1	6.1	2.9
Spherical	20.3	13.0	2.5
Gaussian	15.1	9.0	5.6

Fig. 9 Cross-validation-based RMSE based on the variogram model using an ordinary kriging technique for the three test areas

number of outliers constitutes 10% of the entire dataset, irrespective of site-specific geotechnical variability. The most suitable site-specific outlier threshold (optimum outlier threshold) can be determined using a trial-and-error approach with cross-validation. In this study, the definition of the optimum outlier threshold incorporates the spatial correlations between borehole datasets following these procedures (Kim et al. 2016). The RMSEs with nine assumed outlier thresholds (0, 5, 10, 15, 20, 25, 30, 35, and 40%) were calculated for each bedrock depth. To determine the site-specific outlier threshold accounting for the spatial correlations between borehole information and land surface conditions, two methods and corresponding criteria for outlier thresholds are proposed.

For a high spatial density of borehole datasets having a high spatial correlation, the possible outliers should be considered conservatively, and relatively larger outlying borehole information should be excluded. For correlations between the RMSE and assumed outlier thresholds, the lowest RMSE point is determined, which indicates the optimum outlier threshold. Alternatively, in the case of a low spatial density of borehole datasets with a low spatial correlation, most borehole data should be conserved. For correlations between the RMSE and assumed outlier thresholds, the point at which the rate of change of correlation is highest is estimated, and two linear regression lines (dashed lines in Fig. 12) are drawn. The intersection point of the two regression lines indicates the optimum outlier threshold. In this study, the criterion for spatial density was set to 200 based on the kernel density.

In area 1, the outlier thresholds for the spatially dense clusters (clusters 1–4), with an estimated kernel density of more than 200, were estimated as 17.8, 13.1, 14.5, and 18.2%, using the lowest RMSE method. In contrast, clusters 5–9, with an estimated kernel density below 200, determined using the highest rate of change for RMSE correlations, resulted in threshold values of 15.2, 15.2, 12.5, 8.5, and 11.1%. Based on the site-specific outlier thresholds, the optimized borehole datasets were modified by removing the outliers from the multi-circular clusters, including the three target administrative zones (Fig. 13). The site-specific outliers within each cluster were determined based on the optimum cluster radius. After removing the outliers, the bedrock depth spatial information was developed based on ordinary kriging with an exponential variogram and the optimized borehole datasets.

Geotechnical spatial grid information in the test areas

Correlations between geostatistical optimum criteria

To predict reliable spatial distribution in a geo-layer, quantitative sub-clustering and site-specific outlier analysis should be conducted based on a trial-and-error approach with sequential

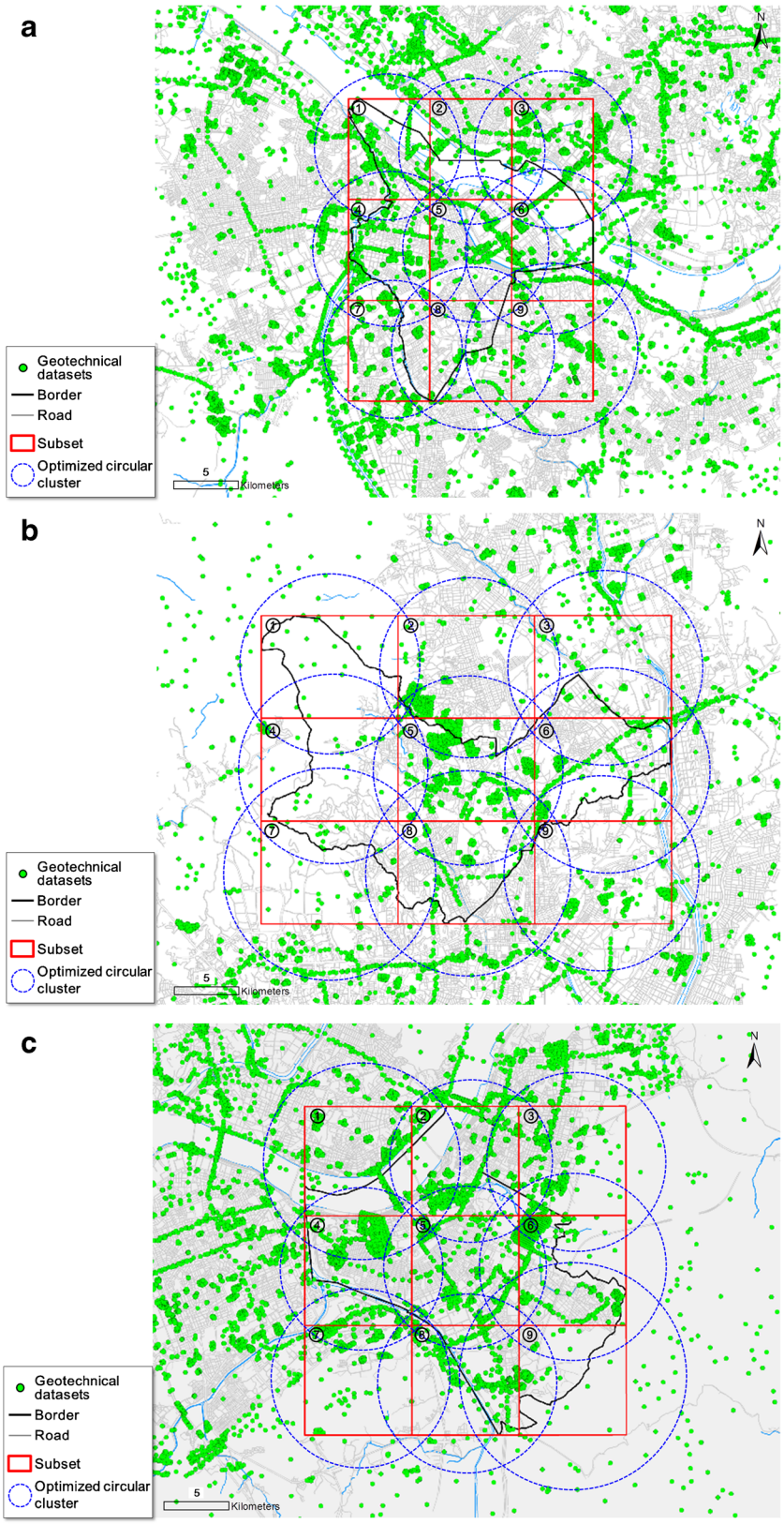


Fig. 10 Optimized circular clusters for subsets in the test areas. **a** Area 1, **b** area 2, and **c** area 3

kriging and cross-validation. Therefore, the optimized borehole datasets, after removing outliers, were applied as initial datasets to develop a geotechnical spatial grid, classified using major geo-layers, fill soil, alluvial soil, and weathered soil,

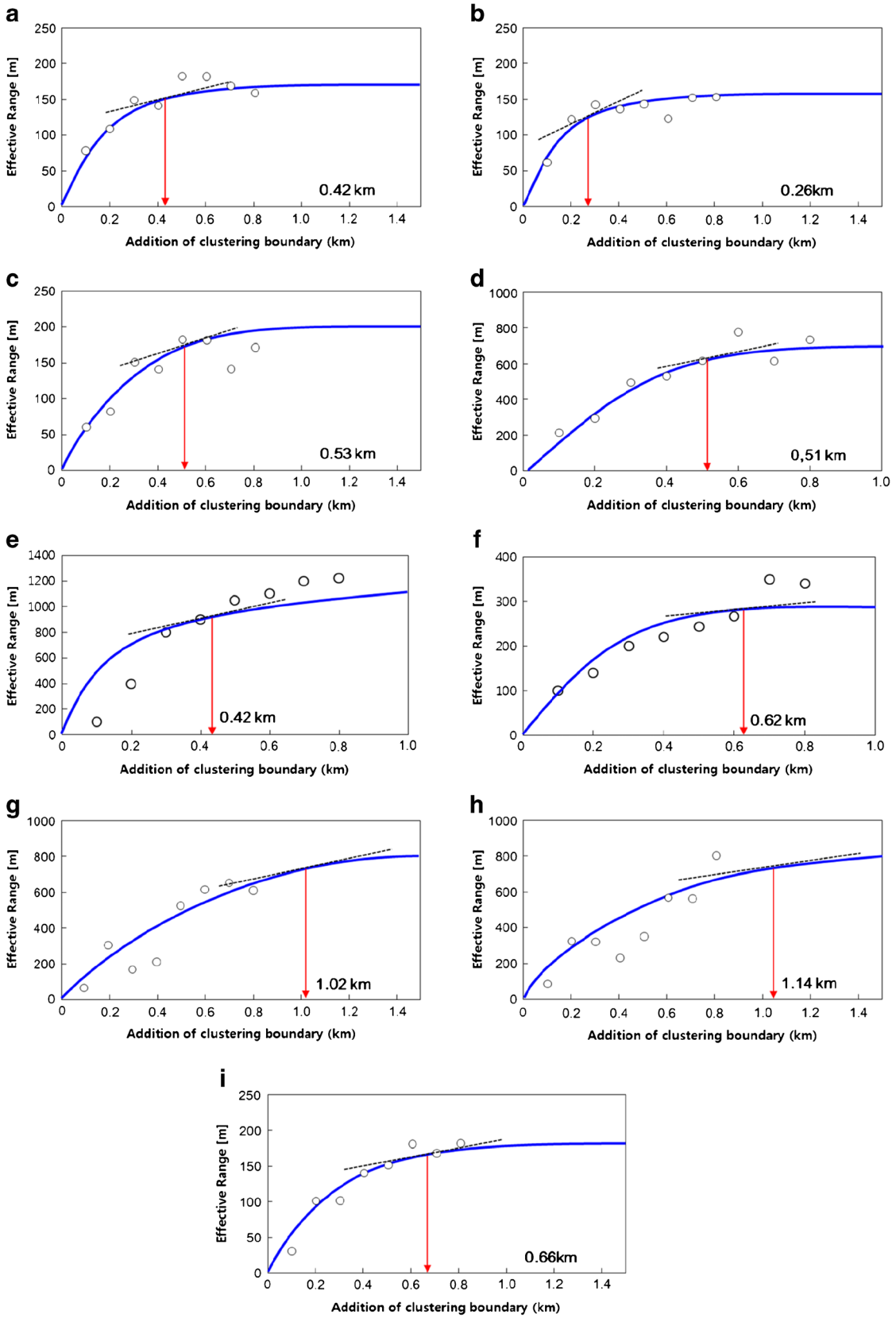


Fig. 11 Correlations between the radius of the influencing circle and variogram-based effective range for bedrock depth to determine the site-specific circular clustering for the nine subsets in area 1. **a** Subset 1, **b** subset 2, **c** subset 3, **d** subset 4, **e** subset 5, **f** subset 6, **g** subset 7, **h** subset 8, and **i** subset 9

Table 1 Geostatistical optimum criteria to construct geotechnical spatial information for the geo-layers in the three test areas, Seoul, South Korea

Test area	Geospatial characteristic criteria	Subset								
		1	2	3	4	5	6	7	8	9
Area 1	Kernel density	248	350	242	220	195	192	129	150	209
	Effective range (m)	521	345	556	549	385	620	702	688	625
	Cluster radius (km)	0.42	0.26	0.53	0.51	0.42	0.62	1.02	1.14	0.66
	Outlier threshold (%)	17.8	13.1	14.5	18.2	15.2	15.2	12.5	8.5	11.1
Area 2	Kernel density	98	146	42	108	121	125	50	49	58
	Effective range (m)	452	593	204	382	395	346	924	892	775
	Cluster radius (km)	0.88	0.69	1.49	0.78	0.82	0.77	1.45	1.88	1.51
	Outlier threshold (%)	16.8	17.2	11.9	15.2	15.9	14.0	12.5	7.0	8.2
Area 3	Kernel density	335	308	274	298	320	185	102	290	255
	Effective range (m)	245	332	252	528	645	752	850	155	204
	Cluster radius (km)	0.35	0.42	0.98	0.33	0.78	1.05	1.22	0.52	0.85
	Outlier threshold (%)	16.2	18.0	14.7	18.2	15.3	8.2	9.6	19.5	13.2

and engineering bedrock. The site-specific geotechnical spatial information that accounts for local uncertainties in the geo-layers and geostatistical optimum criteria, including kernel density, effective range of variogram, cluster radii, and outlier threshold for the nine subsets in each of the three test areas, are summarized in Table 1. Areas 1 and 3, which are flat-lying regions, have dense borehole datasets compared to area 2. However, there are spatial variations in the borehole datasets, with an induced uncertainty observed in the estimated grid-based geo-layer information for every subset in the same unit within each test area.

Spatially dense and correlated subsets can be identified by their kernel density and effective ranges in the variogram. Some subsets, such as subsets 2 and 5 in area 1; subsets 4, 5, and 6 in area 2; and subsets 1, 3, and 8 in area 3, with a higher kernel density and smaller effective range, can be categorized as relatively higher spatially correlated subsets. In these subsets, there were smaller cluster radii and fewer outlier thresholds due to the availability of dense borehole datasets characterized by similar bedrock depths focused at the subset centers. In contrast, subsets 6, 7, 8, and 9 in area 1; subsets 7, 8, and 9 in area 2; and subsets 6 and 7 in area 3 were identified as relatively less spatially correlated subsets. In these subsets, there were larger cluster radii and more outlier thresholds. Therefore, more outlying borehole datasets were determined and removed from among the sub-clustered datasets. Generally, areas 1 and 3, with a greater number of site-specific, dense, and correlated datasets, had smaller cluster radii and fewer outlier thresholds compared to area 2.

However, the sequential procedure for each specific region is a specialization-based, time-consuming process. Therefore, a representative classification for outlier analysis and clustering to construct optimum geotechnical spatial information is recommended based on constructing correlations between geostatistical characteristics. Using the parameters in Table 1, correlations between geostatistical optimum criteria were estimated using a logistical regression model (Fig. 14). The correlations, cluster radius versus outlier threshold and kernel density versus outlier threshold, have R^2 values greater than 70%. If the kernel density of borehole datasets in a target area is acquired using borehole information, the optimum circular cluster radii and outlier thresholds can be determined using the correlations. To establish a reliable and universal classification criterion for Seoul, an additional case study based on the proposed framework should be conducted.

Comparison of the geotechnical spatial grid with the optimization framework

By applying the proposed geostatistical optimum criteria provided in Table 1, geotechnical spatial grids for the three test areas were developed using ordinary kriging with an exponential model based on the optimized borehole datasets after removing the outliers. To validate the reliability and performance of the proposed method, the original datasets, optimized datasets (after removing outliers using a conventional 10% threshold), and only the clustered datasets were additionally assessed using ordinary kriging (Table 2). The standardized residuals for each spatial grid were evaluated with cross-validation-based RMSEs. Consequently, the geotechnical spatial grid information obtained by applying the proposed optimization framework results in the smallest RMSE. Especially, reduction rate (%) in RMSE value for area 1 after excluding outliers, clustering datasets, and optimizing is considerably higher than that for area 2 and area 3. The partial borehole datasets in area 1 have the highest bedrock depth (over 70 m)

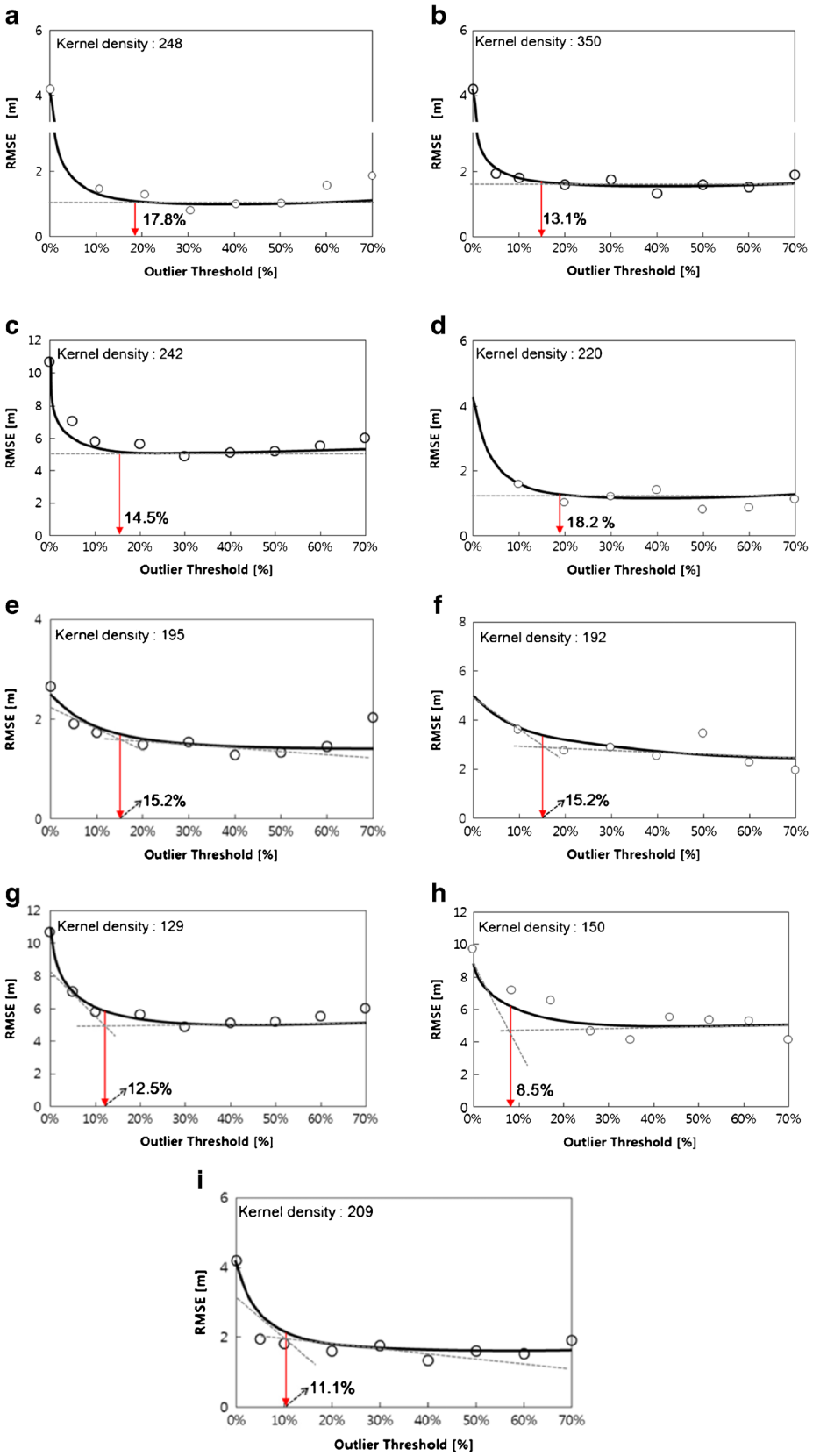


Fig. 12 Examples for determining relative outlier thresholds based on the correlation between sequential cross-validation-based RMSE and the outlier thresholds for optimized circular clusters in area 1. **a** Cluster 1, **b** cluster 2, **c** cluster 3, **d** cluster 4, **e** cluster 5, **f** cluster 6, **g** cluster 7, **h** cluster 8, and **i** cluster 9

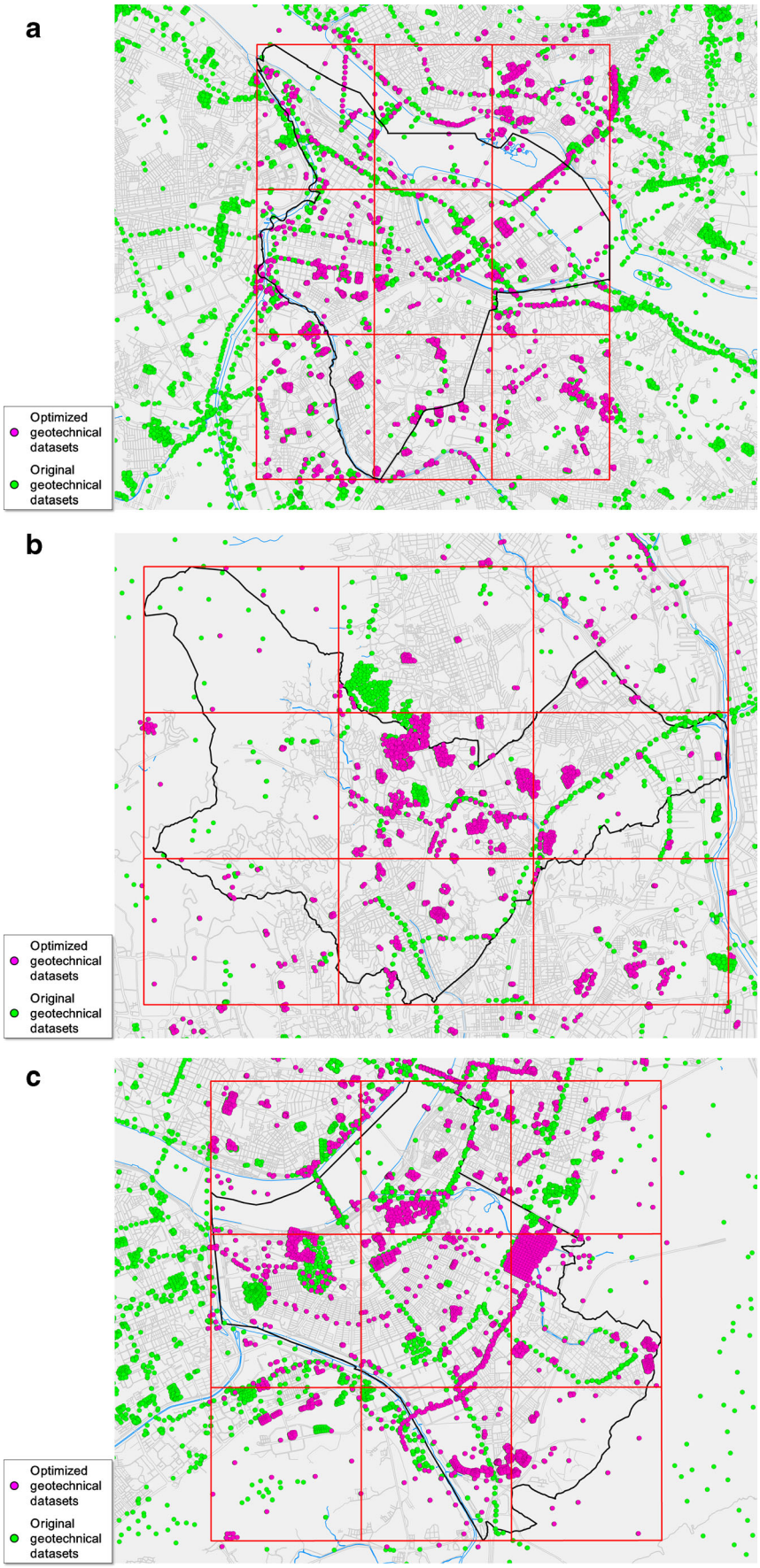


Fig. 13 Optimized and original borehole dataset locations for the nine clusters in the test areas. **a** Area 1, **b** area 2, and **c** area 3

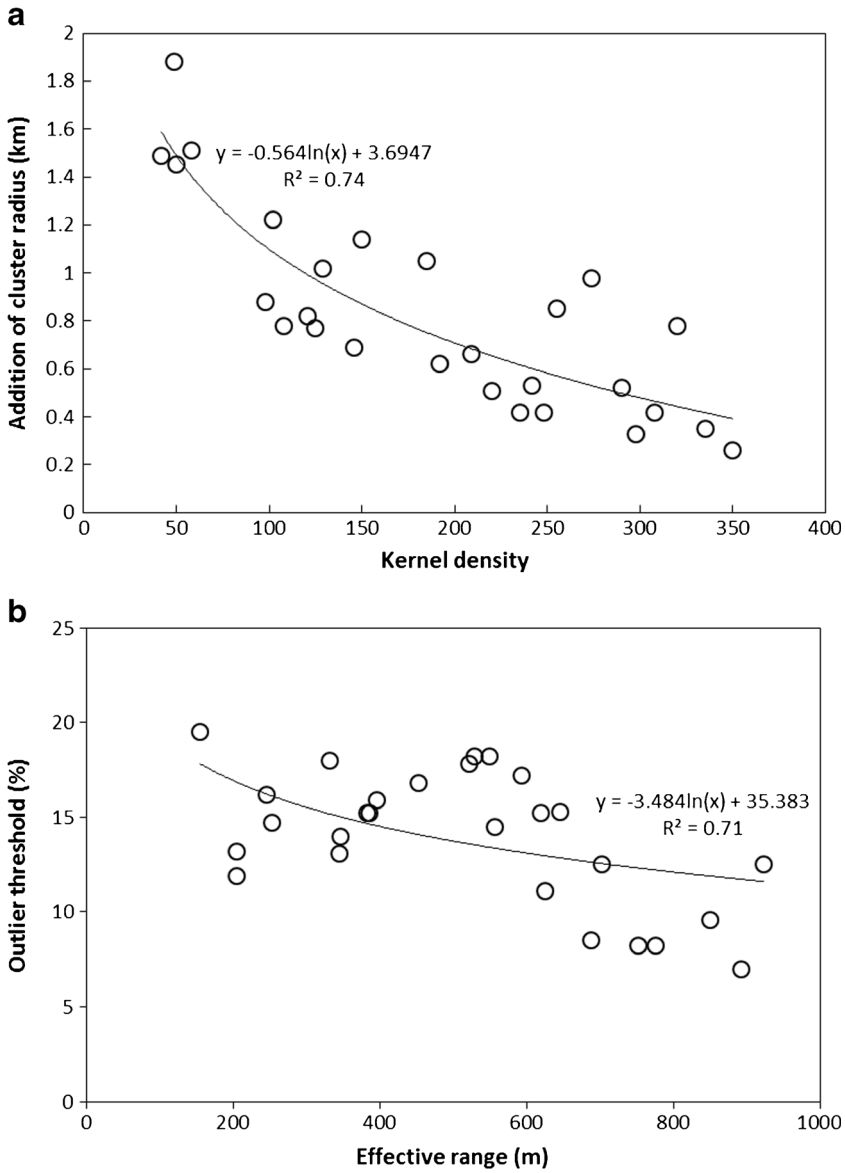
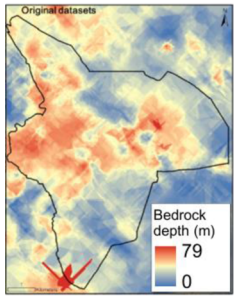
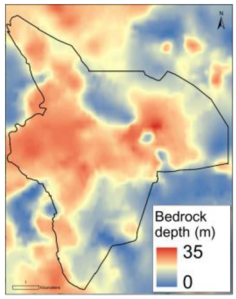
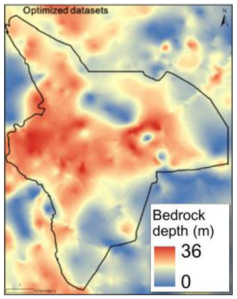
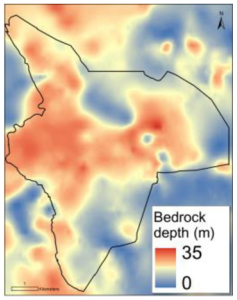
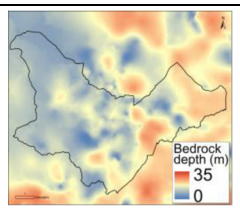
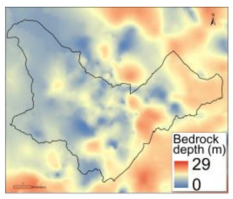
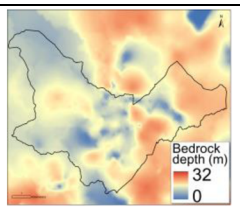
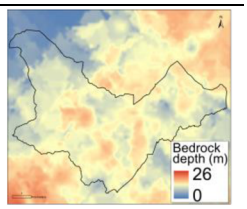
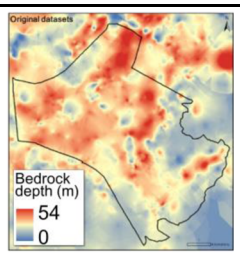
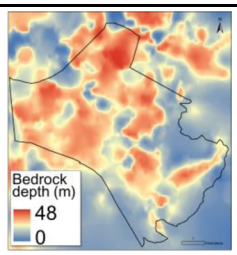
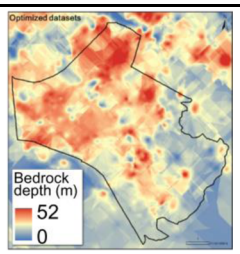
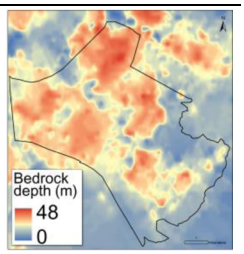


Fig. 14 Correlations of the geostatistical optimum criteria for the construction of geotechnical spatial information for geo-layers. **a** Cluster radius versus kernel density and **b** outlier threshold versus the effective range of the variogram

and the highest spatial density, and were divided by Han River, which induce the differential formation and spatial deviations of subsurface. Accordingly, in northern subsets of Han River, the site-specific outliers having relatively higher bedrock depth (over 45 m) were detected and removed in accordance with 13.1% and 14.5% outlier threshold. Therefore, the proposed method provided more reasonable criteria for developing spatial grid using borehole datasets with spatial uncertainties within the study area.

Additionally, applying the proposed method significantly reduced the quantitative and visible variations in the spatial grid compared to the results from applying the original datasets. The geotechnical characteristic values decreased distortions that produced local outliers in the interpolation results, which can be calculated using optimized borehole datasets. Furthermore, the smoothing effect applying kriging with de-trended datasets, obtained by removing site-specific outliers, produced more improvements than ordinary kriging alone. Generally, ordinary kriging estimates have a serious drawback, well known as the smoothing effect, in which small values are usually overestimated and large values are usually underestimated (Yamamoto 2005). To develop characteristic values for geotechnical design parameters, the depth to bedrock that represents the local target area needs to be considered in the design process for construction projects. Additionally, to verify the influence of each geostatistical criteria corresponding to the multi-scale land surface and spatial correlations of geo-layers, an additional application should be conducted for various regions. The normalized geotechnical characteristic values can then be applied to developing a geotechnical design parameter map.

Table 2 Comparison of geotechnical spatial grids for bedrock depth obtained by applying the original datasets, datasets excluding conventional outliers, clustered datasets, and optimized datasets within clusters of test areas in Seoul.

		Original datasets	Datasets excluding conventional outliers (10%)	Clustered datasets	Optimized datasets within clustering
Area 1	Geotechnical spatial grid				
	RMSE (m)	6.4	3.4	4.6	2.0
Area 2	Geotechnical spatial grid				
	RMSE (m)	8.3	4.8	6.6	4.2
Area 3	Geotechnical spatial grid				
	RMSE (m)	6.3	3.9	4.9	3.3

Conclusions

In this study, site-specific outliers were determined within a proposed systematic framework for optimizing geostatistics for geotechnical spatial information. This process considers the distribution patterns and spatial correlations of borehole datasets and incorporates site-specific uncertainties in geo-layers and geotechnical characteristic values. The sequential optimization was designed as a decision-making procedure based on GIS architecture to develop geotechnical spatial grid information. Test areas were selected for validating the framework in urban Seoul, South Korea.

First, multi-source geospatial information was collected and a geospatial grid was constructed using geo-modeling and reprocessing GIS toolsets. Second, to optimize the conditions from random-field assumptions in kriging methods, and to incorporate interpolation and zonation, appropriate geostatistical methods and corresponding variogram models were determined and validated using a cross-validation-based verification test. Consequently, representative borehole information sampling, considering local site effects in Seoul, was conducted, and the corresponding test areas were identified. Third, by clustering borehole information, the geotechnical spatial information for targets in each test area were developed

and normalized using geostatistical optimum criteria. The optimum range for zonal clustering was determined based on cross-validation using multi-circular zonation. Finally, the outlier threshold for each subset was determined using borehole datasets within the optimized sub-clusters. The relative outlier threshold was also determined based on the density and spatial correlations of boreholes in each cluster. The geotechnical spatial grid information was also developed without any site-specific outliers. Thus, the proposed geospatial characteristic criterion for optimum outliers was validated using the borehole datasets in the Seoul region.

To build a geotechnical database, borehole datasets or other satellite datasets have spatial uncertainties and outliers generated while determining geo-layers and characteristic values. The site-specific criteria for optimization should therefore be considered an improved geostatistical and geotechnical method. Based on the local outlier analysis incorporated within the optimum sub-clustering, the step-wise residuals and accuracy of the analyses were identified and considered parameters in the cross-validation-based optimization. If insufficient borehole datasets are available for application in a target area, proxy-based parameters based on a geospatial database, including topographic maps, satellite images, surface geology, and a digital elevation models, can supplement the process based on this proven case study and normalization.

Funding information The authors wish to express their gratitude for the support from the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2012R1A1A1017659).

References

Anselin L (2004) Exploring spatial data with GeoDaTM: a workbook. Urbana 51(61801):309

Asaoka A, A-Grivas D (1982) Spatial variability of the undrained strength of clays. J Geotech Eng ASCE 108:743–756

Azpurua M, Dos-Ramos K (2016) A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude. Prog Electromagn Res M 14:135–145

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. John Wiley & Sons, New York

Borruso G, Schoier G (2004) Density analysis on large geographical databases. Search for an index of centrality of services at urban scale. Comput Sci Appl 1009–1015

Chandola V, Kumar V (2009) Outlier detection : a survey. ACM Comput Surv 41:1–83

Chun SH, Sun CG, Chung CK (2005) Application of geostatistical method for geo-layer information. J Korean Soc Civil Eng 25:103–115

David M (1976) The practice of kriging. In: Guarascio M, David M, Huijbregts C (eds) Advanced geostatistics in the mining industry. R. Reidel, Boston, pp 31–48

Davis LG, Bean DW, Nyers AJ, Brauner DR (2015) GLIMR: a GIS-based method for the geometric morphometric analysis of artifacts. Lithic Technol 40:199–217

Degroot DJ, Baecher GB (1993) Estimating autocovariance of in-situ soil properties. J Geotech Eng ASCE 119:147–166

Delfiner P (1976) Linear estimation of nonstationary spatial phenomena. In: Guarascio M, David M, Hujibregts C (eds) Advanced geostatistics in the mining industry. R. Reidel, Boston, pp 49–68

Deutsch CV, Journel AG (1972) GSLIB: geostatistical software library and user's guide. Oxford Univ. Press, New York

Getis A, Ord JK (1996) Local spatial statistics: an overview. In: Longley P, Batty M (eds) Spatial analysis: modeling in a GIS environment. Wiley, New York, pp 261–278

Gökkaya K (2016) Geographic analysis of earthquake damage in Turkey between 1900 and 2012. Geomat Nat Haz Risk 7:1–14

Goovaerts P (1998) Geostatistical tools for characterizing the spatial variability of microbiological and physicochemical soil properties. Biol Fertil Soils 27:315–334

Goovaerts P (1999) Geostatistics in soil science: state of the art and perspectives. Geoderma 89:1–45

Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103:3–26

Goovaerts P (2006) Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int J Health Geogr 5(1):52

Grubbs FE (1969) Procedures for detecting outlying observations in samples. Technometrics 11:1–21

Guarascio M, Huybrechts CJ, David M (1976) Advanced geostatistics in the mining industry. In: Proceedings of the NATO Advanced Study Institute held at the Istituto di Geologia Applicata of the University of Rome, p 24

Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York

Kim HS, Kim HK, Shin SY, Chung CK (2012) Application of statistical geo-spatial information technology to soil stratification in the Seoul metropolitan area. Georisk 6:221–228

Kim HS, Chung CK, Kim HK (2016) Geo-spatial data integration for subsurface stratification of dam site with outlier analyses. Env Earth Sci 75:1–10

Kim HS, Sun CG, Cho HI (2017) Geospatial big data-based geostatistical zonation of seismic site effects in Seoul metropolitan area. ISPRS Int J Geo-Inf 6:174

Knudsen H, Kim YC (1978) Application of geostatistics to roll front type uranium deposits. Soc. Mining Eng. AIME, Denver, pp 78–94

Kulhawy FH, Birgisson B, Grigoriu MD (1972) Reliability-based foundation design for transmission line structures (No. EPRI-EL-5507-Vol. 4). Electric Power Research Inst., Palo Alto, CA (United States); Cornell Univ., Ithaca, NY (United States). Geotechnical Engineering Group

Lacasse S, Nadim F (1996) Uncertainties in characterising soil properties. In: Uncertainty in the geologic environment: from theory to practice. ASCE, Madison, WI, pp 49–75

Lu C, Chen D, Kou Y (2003) Algorithms for spatial outlier detection. In: Proc. 3rd IEEE Int. Conf. Data-mining (ICDM'03), Melbourne, FL

Olea R (1991) Geostatistical glossary and multilingual dictionary. Oxford University Press, New York

Orr TL, Breysse D (2008) Eurocode 7 and reliability-based design. In: Phoon K-K (ed) Reliability-based design in geotechnical engineering. Taylor and Francis, Oxon, pp 298–343

Öztürk CA, Nasuf E (2002) Geostatistical assessment of rock zones for tunneling. Tunn Undergr Space Technol 17:275–285

Phoon KK (2008) Reliability-based design in geotechnical engineering: computations and applications. CRC Press

Phoon KK, Kulhawy FH (1999) Characterization of geotechnical variability. Can Geotech J 36:612–624

Prasannakumar V, Vijith H, Charutha R, Geetha N (2011) Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Proc Soc Behav Sci 21:317–325

Rue H, Follestad T (2003) Gaussian markov random field models with applications in spatial statistics (no. NTNU-S-2003-5), SIS-2003-307

Sun CG, Kim HS (2016) Geostatistical assessment for the regional zonation of seismic site effects in a coastal urban area using a GIS framework. Bull Earthq Eng 14:2161–2183

Vanmarke EH (1977) Random vibration approach to soil dynamics. The use of probability in earthquake engineering, ASCE, pp 143–176

Yamamoto JK (2005) Correcting the smoothing effect of ordinary kriging estimates. Math Geol 37(1):69–94

Yu D, Sheikholeslami G, Zhang A (2002) Findout: finding outliers in very large datasets. Knowl Inf Syst 4:387–412

Zhang Y, Meratnia N, Havinga PJM (2007) A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. Technical Report TR-CTIT-07-79. Centre for Telematics and Information Technology. University of Twente, Enschede