Geomechanics and Engineering, *Vol. 41, No. 5 (2025) 555-568*
https://doi.org/10.12989/gae.2025.41.5.555

Subsurface characterization for Seoul through optimized geotechnical database based on geospatial interpolation with adaptive grids

Taek-Kyu Chung[1a], Han-Saem Kim*[2], Chang-Guk Sun[3b] and Choong-Ki Chung[1c]

[1]Department of Civil and Environmental Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[2]Department of Civil and Environmental Engineering, Dongguk University, 30, Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea
[3]Earthquake Research Center, Korea Institute of Geoscience and Mineral Resources, 124, Gwahak ro, Yuseong gu, Daejeon 34132, Republic of Korea

(*Received August 8, 2024, Revised May 18, 2025, Accepted May 19, 2025*)

Abstract. In the era of big data, enhancing the reliability of large-scale geotechnical datasets is crucial for accurate subsurface characterization. Although various statistical interpolation techniques have been developed, significant challenges remain in addressing spatial variability and uncertainty inherent in subsurface conditions. This study presents an advanced spatial interpolation framework that integrates carefully preprocessed borehole datasets with adaptive grid-based modeling to improve the precision of subsurface mapping. The borehole data were classified based on elevation discrepancies from a Digital Elevation Model (DEM), temporally segmented by project year and type, and standardized using the 3-sigma rule to minimize outlier-driven distortions. The interpolation process combined kriging with localized averaging strategies and systematically varied grid resolutions to assess performance sensitivity. Leave-one-out cross-validation, using geological layer thickness as the reference metric, demonstrated that finer grids significantly reduced interpolation error near the surface, while deeper layers exhibited increased uncertainty. Notably, the implementation of an adaptive grid system, capable of dynamically adjusting spatial resolution according to data density and terrain complexity, proved essential in mitigating the smoothing effect often associated with kriging. Furthermore, in data-sparse regions, the integration of localized averaging within adaptive cells helped stabilize estimation accuracy. This adaptive approach offers a powerful enhancement to conventional spatial modeling techniques by enabling more faithful representation of geological heterogeneity and by reinforcing the robustness of predictions under variable data availability, ultimately contributing to more informed geotechnical decision-making.

Keywords: boring investigation; digital twin; geospatial interpolation; geotechnical database; outlier analysis; subsurface information

1. Introduction

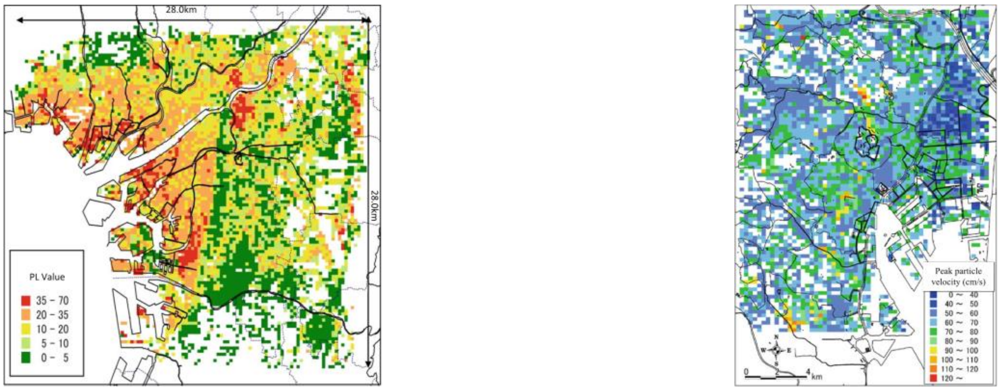
Spatial uncertainty arises from inaccuracies or variability in subsurface conditions at a site, introducing risks and uncertainties. In geotechnical analysis and design, uncertainty is prevalent, particularly in the spatial variability represented by geotechnical investigation data. Understanding spatial uncertainty is crucial, as it often arises from inadequate site investigations (Jaksa *et al.* 2003). Geological heterogeneity and inherent parameter variability in soil and rock units contribute to uncertainty in subsurface conditions (Uzielli *et al.* 2007). This uncertainty can pose considerable risks at construction sites, such as collapses. Therefore, quantifying the spatial uncertainty and comprehensively characterizing the subsurface conditions for borehole investigations are essential. The use of appropriate statistical methods to measure, reduce, or eliminate measurement errors is crucial for mitigating the uncertainty in geotechnical engineering designs (Grasmick 2019, Wang 2018, Sisman *et al.* 2017). Integrating spatial uncertainty knowledge in decision-making optimizes resources and enhances understanding of critical subsurface characteristics, improving construction efficiency (Gangrade *et al.* 2020).

In civil and architectural projects, technical and financial risks are often most pronounced on the ground (National Research Council 1984, Institution of Civil Engineers 1991, Littlejohn *et al.* 1994, Whyte 1995, Zumrawi 2014). Improper use of geotechnical investigation data can lead geotechnical engineers to adopt conservative designs to reduce the risk of failure, resulting in increased project costs and designs based on estimates. Geotechnical investigation data inevitably undergoes various uncertainties (Phoon and Kulhawy 1999, Baecher and Christian 2005). Quantifying this uncertainty is essential in geotechnical engineering for reliability and risk assessment, primarily addressed using statistical methods. The quality of field investigation data is influenced by measurement errors, load conditions, and environmental disturbances (Hawkins 1980, Barnett and Lewis 1994, Yuen and Mu

---

*Corresponding author, Ph.D.
E-mail: hansaem@dgu.ac.kr
[a]Ph.D. Candidate
[b]Ph.D.
[c]Professor

ISSN: 2005-307X (Print), 2092-6219 (Online)



(a) (b)

Fig. 1 Nation-wide Electronic Geotechnical Database Systems (Todo *et al*. 2013): (a) Liquefaction potential map, Osaka (Kasugai *et al*. 2009) and (b) Example of calculated surface peak particle velocity, Tokyo (Yasuda and Watanabe 2010)

2012, Zheng *et al*. 2021, Sun *et al*. 2022). Detecting and excluding outliers in the dataset used for quantifying uncertainty in ground parameters is prudent, as their presence can bias results (Rousseeuw 1991, Han and Kamber 2001, Zheng *et al*. 2021, Raheem *et al*. 2021). Kim *et al*. (2012) proposed a framework for outlier detection using statistical analysis, cross-validation-based methods, and generalized extreme value distribution, evaluating its effectiveness in a borehole dataset, including soil depth distribution in Seoul's urban area (Kim *et al.* 2019a). Therefore, for developing geographic spatial information based on an optimized site-specific borehole dataset, integrating existing outlier analysis and spatial interpolation methodologies is crucial, considering outlier locations while optimizing the dataset (Yu *et al*. 2002, Kim *et al*. 2023).

Initially developed for predicting ore grades in mining, geostatistics has been consistently applied across various disciplines, including geology, environmental engineering, and hydrology, for efficiently predicting unknown data (Carr 1995). Spatial interpolation techniques, such as polygon, triangulation, local sample mean, and inverse weighted distance methods, calculate raster data from discrete point data using simple mathematical calculations. Kriging, a prominent geostatistical method, combines mathematical and statistical methods to predict data using a weighted linear combination, effectively reflecting the correlation between data points, which is especially useful for broad-scale spatial interpolation. The types of kriging include simple kriging, ordinary kriging, block kriging, and co-kriging, with ordinary kriging being the most widely used. Since kriging predictions inherently contain errors, evaluating their reliability is crucial. Cross-validation, involving predicting one unknown value while using the rest of the data and comparing predicted values with actual values for each data point, is essential for reliability evaluation (Isaaks and Srivastava 1989). However, the accuracy of error evaluation could be somewhat distorted if the data are unevenly distributed (Choe 2013).

Todo *et al.* (2013) detailed the creation of the nation-wide electronic geotechnical database systems (NEGDS), which construct a baseline model within a 250 m × 250 m area, as shown in Fig. 1. The rationale for choosing this specific size was its capacity to comprehensively represent ground conditions within a specified area. This model draws data from diverse geotechnical information databases and facilitates a standardized assessment of soil properties and depositional environments across different regions of Japan. It provides initial insights into ground conditions, aiding in the preliminary assessment of potential ground hazards, including liquefaction, seismic activity, ground settlement, and slope movement.

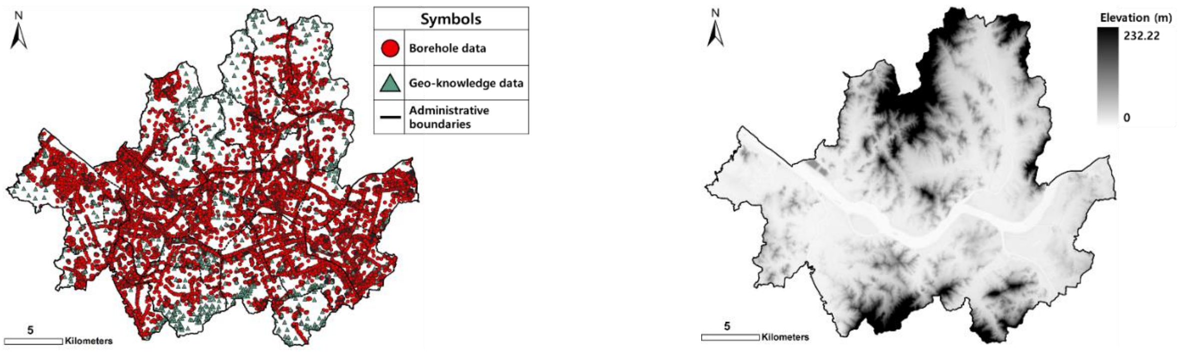
In borehole surveying, inaccuracies can arise from various factors, such as the surveyor's expertise, data entry errors, and mechanical malfunctions. This study addresses these challenges by reconstructing borehole data according to design standards and enhancing accuracy by establishing a refined database employing the 3-sigma rule to eliminate outliers. Using these preprocessed data, our research focused on generating reliable subsurface maps.

We employed the kriging methodology to meticulously assess the influence of the grid dimension, which is a pivotal element in ascertaining spatial interpolation precision. Through a comprehensive analysis, we developed an optimized approach for spatial interpolation tailored to urban environments. This innovative method incorporates adaptive grids and synergizes kriging and averaging techniques to enhance accuracy and applicability in urban areas. The adaptive grid system plays a critical role in reducing the smoothing effect inherent to kriging and dynamically adjusts to data density, ensuring more reliable interpolation in both dense and sparse data regions. Finally, we validated the reliability of the proposed method by predicting and analyzing subsurface layer information for Seoul. This comprehensive approach not only improves the accuracy of the data but also offers a robust framework for subsurface mapping in geotechnical studies.

2. Optimization of geotechnical database in Seoul

2.1 Archiving geotechnical data

The geotechnical information portal system (https://www.geoinfo.or.kr/) provided data from 33,867 boreholes drilled in Seoul, encompassing general boring, stratigraphic, and standard penetration test information.



(a) (b)

Fig. 2 Collected data in Seoul: (a) Borehole data (with knowledge-based data) and (b) Digital elevation model (DEM)

Additionally, data from 40,118 boreholes in Seoul were obtained from a public data portal (https://www.data.go.kr/). The two datasets were combined through entity correlation to remove duplicate data. The data with missing geographic coordinates (i.e., x, y, z) or stratigraphic information were removed, resulting in 29,019 boreholes. Due to the urban-centric nature of prior geotechnical investigations, borehole data were predominantly concentrated in developed areas, resulting in a notable deficiency of subsurface information for mountainous regions. To address this limitation, a targeted field investigation campaign was conducted, employing handheld GPS units and manual augers to systematically identify and document geological features unique to mountainous terrains. These included the precise locations of exposed rock outcrops, shallow stratigraphy of cut slopes, and other mountainous terrain characteristics. Furthermore, knowledge-based geological interpretation was enhanced through the integration of expert-reviewed datasets, including supplemental borehole information referenced by Kim *et al.* (2019b) and Sun *et al.* (2008). As a result, 639 additional information records were incorporated, significantly improving the spatial and lithological representation of mountainous zones within the overall geotechnical database. Geographic knowledge-based data were supplemented and reviewed by Kim *et al.* (2019b) and Sun *et al.* (2008), incorporating knowledge-based data from 639 additional boreholes into the dataset.

To simplify and streamline processing speed, we assumed no reversal of strata, postulating that strata, specifically fill soil, sedimentary soil, weathered residual soil, and weathered rock, emerged continuously from the top. When specific stratum information was absent in borehole investigation data, the stratum was treated as 0 m.

The digital elevation model (DEM), obtained from the Korea National Geographic Information Service (https://www.ngii.go.kr/), is a numerical representation of the terrain of man-made structures and trees. With a resolution (cell size) of 5 m×5 m, the DEM serves as a fundamental tool in civil engineering for constructing dams, roads, and railways, as well as in military analysis for selecting suitable radar facility locations.

2.2 Restoration of borehole data

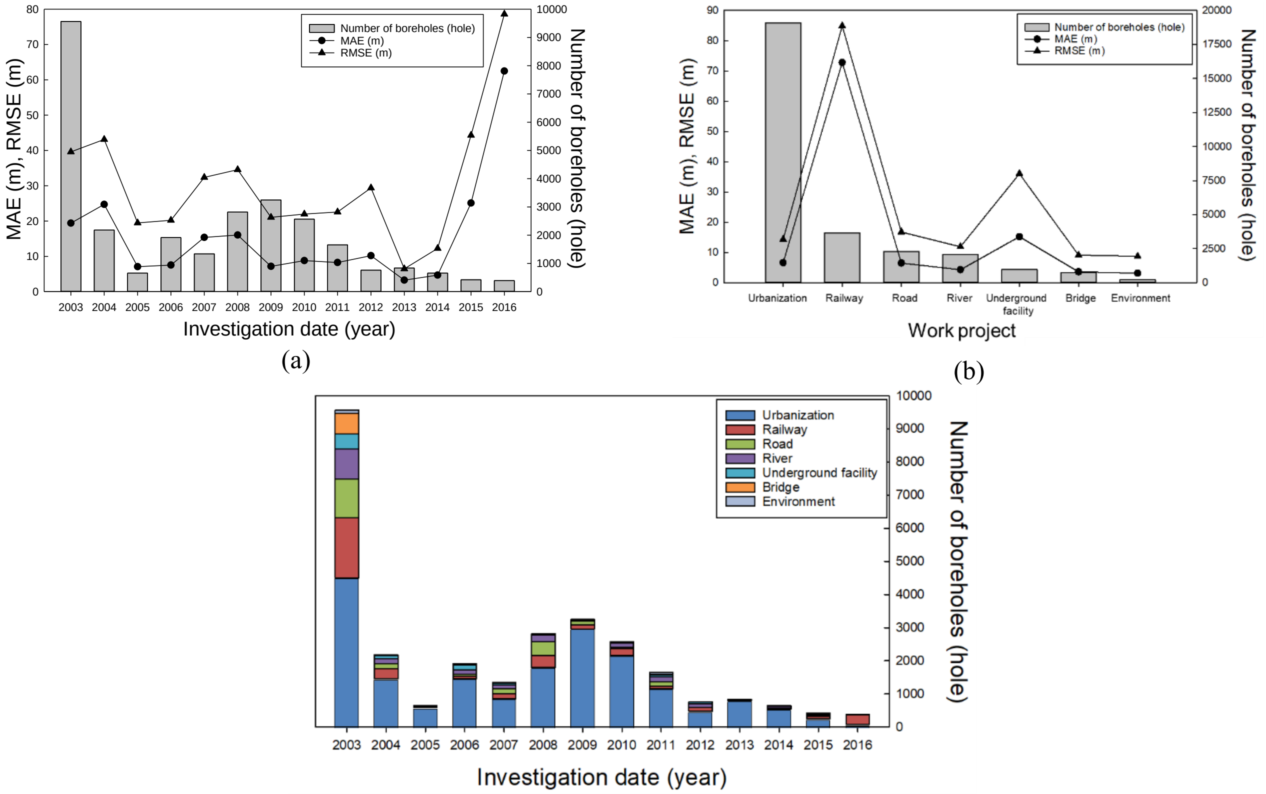
Data serve as fundamental building blocks for generating and utilizing information. Every dataset is susceptible to various quality issues. The adage 'garbage in, garbage out' underscores the common understanding that poor data input results in unreliable data output. Ensuring the dependability of data analytics and applications requires collecting information with high accuracy. Consequently, ensuring data reliability, eliminating errors, and faithfully representing real-world context become our responsibilities, thereby facilitating well-informed decision-making (Man *et al.* 2010, Mans *et al.* 2015, Kilkenny *et al.* 2018). Frequently employed metrics for error assessment include the mean error, mean absolute error (MAE), mean absolute percentage error, mean square error, and root mean squared error (RMSE). The RMSE and MAE are considered some of the most comprehensive indicators of model performance, capturing the average deviation in units between the observed and estimated values (Willmott 1982, Li *et al.* 2008, Denkovski *et al.* 2012).

In this study, errors related to the investigation date and work project were classified into three components for the accuracy assessment. The data were verified using a DEM as a reference point, with the elevation of the DEM set as a constant, and the differences from the borehole data (MAE and RMSE) were calculated for an accuracy assessment. While MAE quantifies discrepancies between paired observations that represent the same phenomenon, the similar concepts of RMSE and standard deviation describe the variance of error for two or more unknowns and the variance of deviation for the population, respectively. Both the borehole and DEM data were entered into ArcGIS, a Geographic Information System (GIS) tool. The 'Point to Multiple Value Extraction' feature in ArcGIS was used to extract DEM elevations at corresponding points, allowing for a comparison with the elevation data from the drilling records (Chung *et al.* 2018).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{2}$$

where, $\hat{Y}_i$ is the observed value, Y_i is the true value, and n is the number of data points.

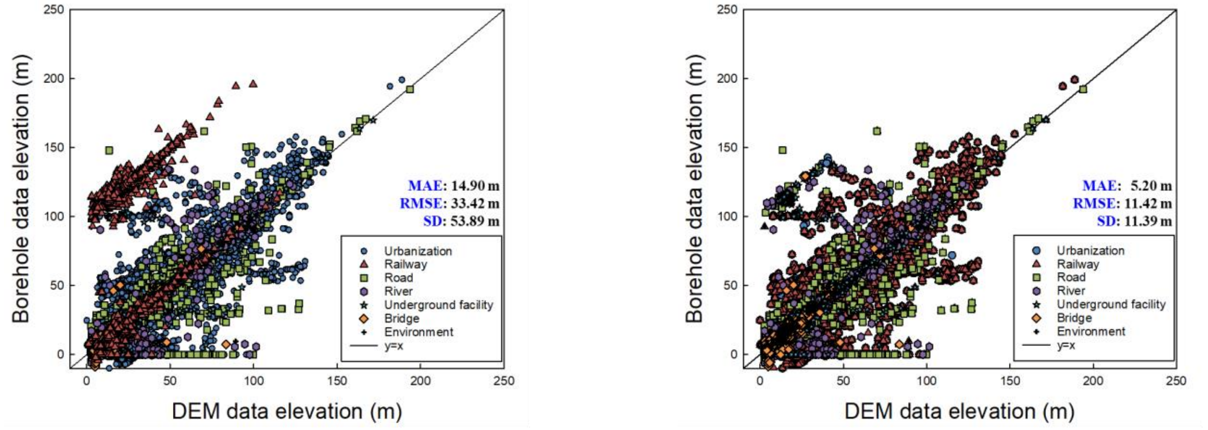


(a)

(b)

(c)

Fig. 3 Classification of borehole data based on characteristics: (a) Investigation date, (b) Work project and (c) Work project by year



(a)

(b)

Fig. 4 Change of error as a result of restoring borehole: (a) Before and (b) After

The available borehole data, mainly comprising borings surveyed before 2010, were surveyed from 2003 to 2016. The accuracy of the surface elevation of the borehole data based on the investigation date is shown in Fig. 3(a). The overall MAE was determined to be 14.90 m, and the RMSE was calculated at 33.42 m. Notably, a pronounced escalation in error magnitudes exists after 2014. From 2014 to 2016, the average error manifested a considerable increase, with an MAE of 25.93 m and an RMSE of 47.70 m, representing an approximate 58% elevation in comparison to the aggregate average. An accuracy examination was conducted concerning individual work projects (Fig. 3(b)) to determine the cause of the increasing error rate post-2014. Work projects were categorized as urbanization, railways, roads, rivers, underground facilities, bridges, and environment. Urbanization includes buildings (apartments, houses, amenities [hospitals, department stores, etc.], and parks), underground facilities (including sewers, telecommunications, and water reclamation centers), and the construction of landfills and waste treatment facilities. Most construction work projects were for urbanization, with railroad-related projects exhibiting the largest error.

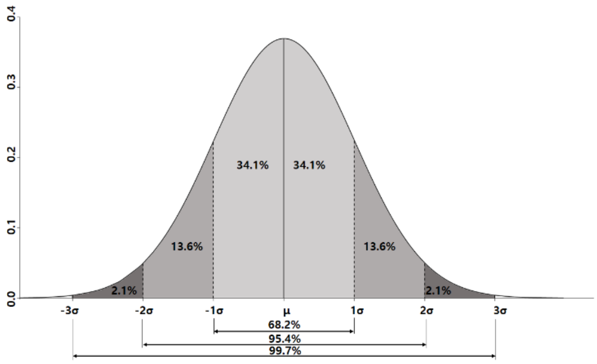


Fig. 5 3-sigma rule (Moivre 2013)

When comparing the borehole data with the surface elevations in the DEM (Fig. 4(a)), the differences were mostly concentrated in the 100 m elevation zone. Specifically, the borehole data for the railway project had errors of at least 100 m. In January 2014, regulations governing general survey work were introduced, specifying that for railway projects, the standard elevation should be determined by adding 100 m to the elevation provided by the National Geographic Information Service (Ministry of Land Transport and Transportation 2015, Korea National Geographic Information 2013). Conversely, surface elevation was restored by subtracting 100 m from the recorded surface elevation. As depicted in Fig. 4(b), the mean error of the borehole data, indicated by the MAE, was 5.20 m, and the RMSE was 11.42 m. This represents a roughly threefold decrease compared to the pre-restoration values. Fig. 3(c) illustrates that the notable surge in borehole data errors post-2014 was closely tied to the implementation of general survey work regulations.

2.3 Outlier analysis of borehole data

Outlier detection is a critical preprocessing step in geotechnical data analysis to enhance reliability and reduce uncertainty. Two widely used methods are the three-sigma rule, based on sample mean and standard deviation (Daw *et al.* 1972, Moivre 2013), and the boxplot rule, which does not assume a specific distribution (Tukey 1977).

The three-sigma rule is highly effective when the sample size is sufficiently large and the data can reasonably be assumed to follow a normal distribution. However, its accuracy may deteriorate in the presence of extreme outliers or small sample sizes, where the influence of anomalous values disproportionately distorts the mean and standard deviation (Leys *et al.* 2013, Lightfoot and O'Connell 2016, Santos 2020, Chung *et al.* 2025). In contrast, the boxplot rule is a simple yet robust method for detecting outliers in distribution-free settings. Nevertheless, it may be biased in skewed datasets, requiring cautious interpretation (Hubert and Vandervieren 2008, Santos 2020, Chung *et al.* 2025).

To assess the normality of the collected borehole data, a histogram-based visual inspection was adopted, which has been previously validated as an intuitive and reliable method (Henderson 2006, Brocker and Ben Bouallegue 2020, Coisnon *et al.* 2021, Chung *et al.* 2025). Where significant deviations from normality were observed, the three-sigma rule was applied to improve data reliability. This rule (Fig. 5) assumes that approximately 68.26%, 95.45%, and 99.73% of observations in a normal distribution fall within ±1, ±2, and ±3 standard deviations from the mean, respectively (Daw *et al.* 1972, Moivre 2013).

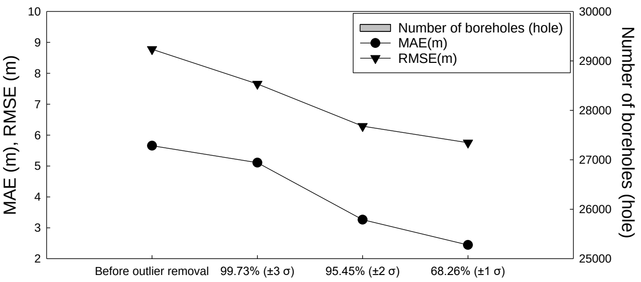


Fig. 6 Rate of outlier removal and degree of error reduction

Errors in borehole data may arise from a variety of sources, including operator inexperience, data entry mistakes, or equipment malfunction. Even after stratifying data by project characteristics and reconstructing it according to relevant design standards, significant anomalies were still observed. Therefore, a data preprocessing procedure was implemented to improve accuracy and minimize errors. Following the approach of Kim *et al.* (2012), which identified the top 10% of samples with excessive deviation as outliers, this study adopted a more conservative strategy by retaining data within the ±1σ (68.26%) interval for estimating surface elevation at each borehole, as illustrated in Fig. 6. This threshold was selected to balance outlier removal with the preservation of valid data, thereby enhancing the reliability of borehole surface elevation estimations.

2.4 Construction of geotechnical information database

In the era of information and big data, information is an immense asset, and the rational analysis of vast amounts of information is crucial for making important decisions. Therefore, proficiency in handling databases is essential for collecting and analyzing information properly. A database system is defined as a computerized record-keeping system. Its overarching purpose is to store information and enable users to retrieve and update it as required (Date 2004). Databases can process and respond to user or program requests in real time, allowing for the continuous modification of data content through operations such as insertion, deletion, updating, and querying.

A Database Management System (DBMS), a software group that manages and operates databases to make them accessible to users and applications, has several advantages, such as real-time data updates, provision for standardized data, reduction of data duplication and inconsistencies, cost savings in system development and maintenance, systematic security against information leaks, and support for database sharing and simultaneous access. As the

Table 1 Results of grid-size-specific spatial interpolation using kriging across Seoul

Multi-scale-grid		RMSE (m)					Interpolative area (%)
		Fill soil	Sedimentary soil	Residual soil	Weathered rock	Average	
Macro-scale (no grid)		2.45	2.55	2.89	3.21	2.77	100
Micro-scale (grid size)	125 m×125 m	1.73	1.78	1.88	2.41	1.95	4.00
	250 m×250 m	1.77	2.00	2.03	2.71	2.11	16.09
	500 m×500 m	1.88	2.01	2.19	2.76	2.21	45.93
	1 km×1 km	1.94	2.11	2.31	2.83	2.30	82.39
	2 km×2 km	1.93	2.28	2.29	3.12	2.41	96.92

volume of data stored on computers rapidly increases and the value of data as an asset becomes more significant, the importance of utilizing a DBMS is becoming increasingly prominent. In this study, PostgreSQL, an open-source database compatible with PostGIS, was employed to map and integrate the spatial information of the GIS and subsurface data. PostGIS is an extension of the PostgreSQL object-relational database system that allows the storage of GIS objects in a database (Toews 2015).

Data modeling is one of the most challenging and crucial tasks in relational Geographic Database Management Systems (Aziz *et al.* 2017). Poor data modeling can fail to meet user requirements, diminish reliability, and result in the inclusion of redundant data in a database (Hong 2012). An Entity-Relationship Diagram (ERD) is a widely utilized technique in data modeling that integrates semantic information about the real world and provides a consolidated view of the data. It is a diagrammatic technique used in the design and description of databases, offering a foundation for integrating various perspectives on data, including network, relational, and entity-set models. The ERD process begins with designers planning entities, their attributes, and the relationships between the entities. Subsequently, logical ERD is transformed into a relational model using scientific rules. This relational model is converted into a physical database based on a software platform for implementation (Chen *et al.* 1976, Hingorani *et al.* 2017).

As shown in Fig. 7, an Entity-Relationship Diagram (ERD) was employed for database modeling in this study, incorporating cleaned data from 26,600 boreholes to enable efficient and scalable data utilization. The database is structured into five relational tables: project, borehole, geolayer, laboratory_test, and field_test. The project relation stores general information about construction sites, borehole includes metadata on borehole locations and conditions, geolayer represents subsurface stratigraphy identified from boring investigations, laboratory_test contains geotechnical parameters obtained through laboratory analysis, and field_test stores field-measured values such as SPT-N and in-situ permeability. To ensure data integrity and reduce redundancy, primary keys (project_id, borehole_id, geolayer_id) were systematically assigned, thereby enhancing the interoperability and scalability of the geotechnical information system.

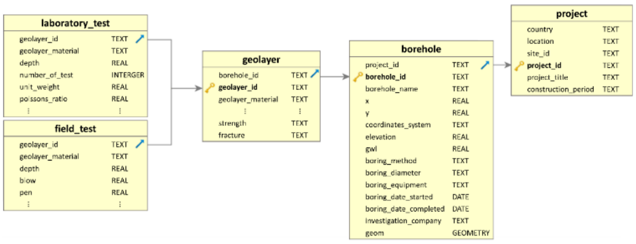


Fig. 7 Entity-relationship diagram in geotechnical information database

Moreover, to preserve the accuracy and applicability of the database over time, this study has established a flexible framework that allows for the seamless integration of additional or updated geotechnical data. Specifically, the system is designed to support: (1) Version-controlled data integration, whereby newly acquired borehole logs, field test results, or laboratory measurements can be ingested with unique timestamps and revision metadata to ensure traceability, (2) Automated validation protocols, which apply predefined quality control criteria to screen and flag anomalous or inconsistent entries prior to acceptance, (3) Dynamic relational mapping, enabling new data to be accurately linked to existing projects, zones, or geological units through foreign key constraints, and (4) Schema extensibility, allowing future incorporation of advanced investigation methods (e.g., CPT, MASW, downhole testing) or real-time sensor data without compromising database structure. Through this expandable and quality-assured architecture, the database can continually evolve to accommodate ongoing geotechnical investigations while maintaining scientific rigor and utility for infrastructure planning, urban development, and seismic risk evaluation.

3. Optimal geospatial interpolation method with adaptive grids in Seoul

3.1 Generation of multi-scale-grid in Seoul

Spatial interpolation, which estimates values at unobserved locations based on data observed in nearby areas, helps fill gaps in the spatial data and enables a more comprehensive understanding of the spatial distribution of the phenomena. Kriging, a prominent spatial interpolation method, estimates values at unobserved locations based on the spatial correlation between data points by considering spatial variability and uncertainty. In contrast, other methods, like inverse distance weighting, assign weights to

Table 2 Standards for borehole investigation (Ministry of land, infrastructure, and transport 2021)

Category		Boring interval
Buildings		Intervals of 30 m to 50 m, depending on the structure's scale
Bridges		One location for each abutment and pier
Box culvert		One borehole per location
Tunnel	Mountain (NATM, TBM)	Intervals of 50 m to 200 m (including entrances and exits); At least one borehole in valley sections/low embankments, with an additional borehole for every 200 m
	Urban areas (Open Cut)	Intervals of 100 m; one borehole per major structure (such as vertical shafts, stations, collecting wells, ventilation shafts)
Slope cutting faces		At least two boreholes per location; an additional borehole for every extension over 100 m; for cutting heights over 20 m, at least two boreholes are required (Trial excavation investigation: 1 to 2 locations)
Fill slope faces	General	Intervals of 500 m (hand auger intervals of 300 m)
	Soft ground	Intervals of 50 m to 100 m (hand auger intervals of 200 m)
Dams		Grid intervals of 20 m to 30 m
Embankments		Intervals of 100 m
Airports		Intervals according to the above categories

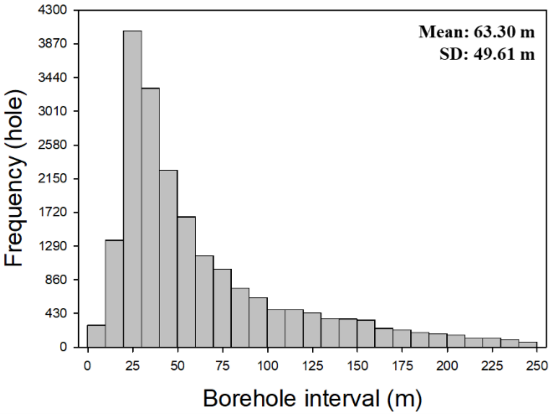


Fig. 8 Borehole interval in Seoul, South Korea

nearby data points based on their distances. Spline interpolation, which fits a smooth curve through observed data points, has also been employed for spatial interpolation. These techniques find application in various fields, such as resource engineering, geotechnical engineering, geology, and environmental engineering, where spatial patterns are crucial for making informed decisions based on available information. In decision-making processes where accurate estimation is essential, reliable spatial interpolation is necessary. Applications that use interpolated values for critical purposes, such as predicting pollutant concentrations or determining optimal infrastructure locations, require the highest interpolation reliability. High-quality spatial interpolation methods ensure that estimated values closely align with actual values, reducing the risk of incorrect decisions based on inaccurate estimates (Chiles *et al.* 2012). The accuracy of kriging varies with the variogram, distribution of data points, and size of the interpolated grid, necessitating caution when modeling the spatial correlation structure (Li *et al*. 2014).

In the urban landscape of Seoul, a meticulous multiscale grid division approach was employed with the implementation of ordinary kriging for spatial interpolation to ascertain the most effective grid size, as detailed in Table 1. This thorough investigation explored the intricacies of prediction errors and assessed the extent of areas suitable for accurate interpolation depending on the segmentation of grids. Nuanced analysis revealed an escalation in prediction errors when spatial interpolation was performed without the systematic structure of a grid framework, markedly contrasting with the enhanced accuracy observed in areas subjected to a more methodological gridding approach. As the investigation progressed, reducing the grid size considerably enhanced the reliability of the interpolation results. However, this increase in precision inversely affects the scope of the area suitable for spatial interpolation, thus presenting a complex interplay between grid size and interpolation efficacy. To manage the delicate equilibrium between upholding high reliability and achieving computational efficiency, we determined an optimal minimal grid size of 250 m ×250 m as effective. This study proposes an optimized spatial interpolation methodology that synergistically combines the robustness of kriging with a systematic averaging process across varying grid sizes, including 250 m×250 m, 500 m500 m, 1 km×1 km, and 2 km×2 km. This innovative approach was strategically designed to utilize the full potential of spatial interpolation to achieve peak reliability and accuracy in the geospatial analysis of Seoul's urban landscape. Therefore, the proposed methodology demonstrates the intricate balance between precision and practicality, paving the way for more refined and reliable geotechnical assessments in urban planning and development.

In this study, a meticulous examination of borehole data across Seoul focused on the spacing and distribution patterns of boreholes. A detailed interval analysis was conducted using a consistent initial grid size of 250 m, employing clustering techniques around each borehole within a 250 m radius. Our findings (illustrated in Fig. 8) reveal that the average borehole spacing was approximately 63.30 meters, predominantly within the 20–50 m range. This distribution aligns with the prevailing borehole investigation standards (Table 2) set for construction projects, which typically require borehole spacing between

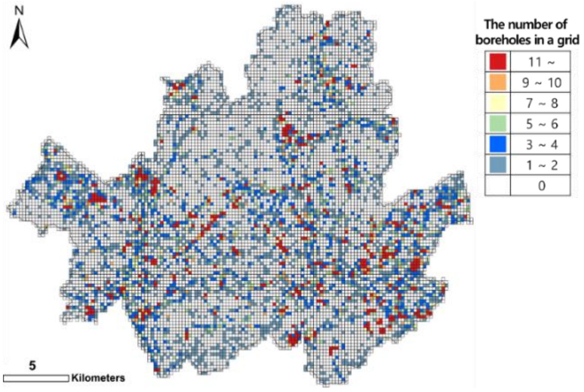


Fig. 9 Borehole distribution in Seoul

30 and 50 m. This pattern corroborates the data classification by project type (Fig. 3), indicating that a significant portion of borehole exploration in Seoul has been conducted primarily for urban development and construction purposes.

As depicted in Fig. 9, spatial distribution of boreholes was calculated based on the number of boreholes within each 250 m×250 m grid, with approximately 41% of these grids containing borehole data. An analysis overlaying this distribution with the DEM of Seoul (shown in Fig. 1(b)) indicated a higher concentration of borehole data at lower altitudes, typically in urban areas. In contrast, higher altitudes, mainly in mountainous regions, exhibited a significant dearth of data. Despite incorporating geographic knowledge-based data, a noticeable insufficiency in the collected data remains.

3.2 Geospatial interpolation method with adaptive grids

The procedural architecture of the proposed framework outlined in Fig. 10 integrates the sophisticated application of two discrete spatial interpolation methodologies. These methodologies have been meticulously developed with a keen emphasis on nuanced complexities and tailored considerations intrinsic to the application of kriging in subsoil geolayer prediction technology. Strategic division of the target area into grids is central to this approach. This division not only enhances the authenticity and alignment with real-world topography but also considerably counters the general smoothing effect, which is a prevalent challenge in extensive kriging applications.

In this context, the Japanese Geotechnical Society's NEGDS harnesses a wide array of subsoil data to formulate a model that standardizes ground conditions over a 250 m × 250 m grid. This model focused on depicting the average ground conditions within this demarcated space, as clarified by Todo *et al.* (2013). In conformance with this methodology, the present study initially adopted a grid dimension of 250 m× 250 m, a decision corroborated by the findings in Section 3.1 regarding the multiscale grid effect. In scenarios where heightened reliability was ascertained, a selective averaging approach reflective of NEGDS practices was employed. Ordinary kriging, a dominant geostatistical technique, was used for spatial interpolation. Thus, the proposed method ensures optimal reliability by segmenting the target area into grids for kriging, deviating from conventional methodologies that aggregate all data concurrently.

Fig. 10(a) illustrates the initial partitioning of the area under analysis into 250 m × 250 m grids. Although most of the grids encompassed boreholes, nine grids did not. The decision to employ three-dimensional (3D) kriging was contingent on the borehole density within each grid, as evaluated in Fig. 10(b). Although 3D spatial interpolation is generally viable with a minimum of three data points, this study restricts its application to grids with at least four data points, as shown in Fig. 10(c). This threshold was to facilitate a more precise evaluation of spatial interpolation prediction errors, incorporating an extra data point for leave-one-out cross-validation (LOOCV), which reserves one data point as the test set while modeling the remaining data points.

Cross-validation techniques play a pivotal role in generating error estimates for the interpolated predictions, thereby guiding the selection of the most appropriate spatial interpolation model. These techniques include LOOCV, shuffle splitting, stratified shuffle splitting, and spatial k-folding. Therefore, researchers are empowered to select spatial interpolation models based on these errors. The methodology examined here employed LOOCV, a technique universally recognized for its efficacy in error estimation, as evidenced in the works of Dirks *et al*. (1998), Luo *et al.* (2008), Mair and Fares (2011), and Risk *et al.* (2022). This approach was instrumental in validating the precision of the spatial interpolation predictions, as illustrated in Fig. 10(d).

Fig. 10(e) shows the process of identifying the grids with the lowest prediction errors, thereby extracting reliable spatial interpolation results. In instances where a grid is deficient in boreholes (fewer than four), as shown in Fig. 10(f), the grid size is expanded to facilitate interpolation. Post-interpolation within this enlarged grid, the interpolated results are then resized to the original grid dimensions (250 m × 250m), as demonstrated in Fig. 10(g). Upon successfully completing spatial interpolation for each grid, the data were meticulously integrated into a cohesive mosaic grid to refine optimal spatial interpolation, specifically for geo-layer thickness. The final step involved transforming the interpolation results into a format suitable for practical applications. This was accomplished by converting the predicted geolayer thickness within the target area into a measure of the vertical distance from the mean sea level to specific locations based on DEM surface elevations. This measure, clearly depicted in Fig. 10(i), is essential for accurately representing geo-layer boundary elevations. The transformed data are then ready for use in various applications, including sophisticated 3D visualization and detailed design projects, thereby facilitating a deeper understanding and utilization of geospatial data in the context of the target area.

3.3 Optimal subsurface stratification in Seoul

The adaptive grid approach was leveraged to optimize the spatial interpolation for Seoul, maximizing the use of

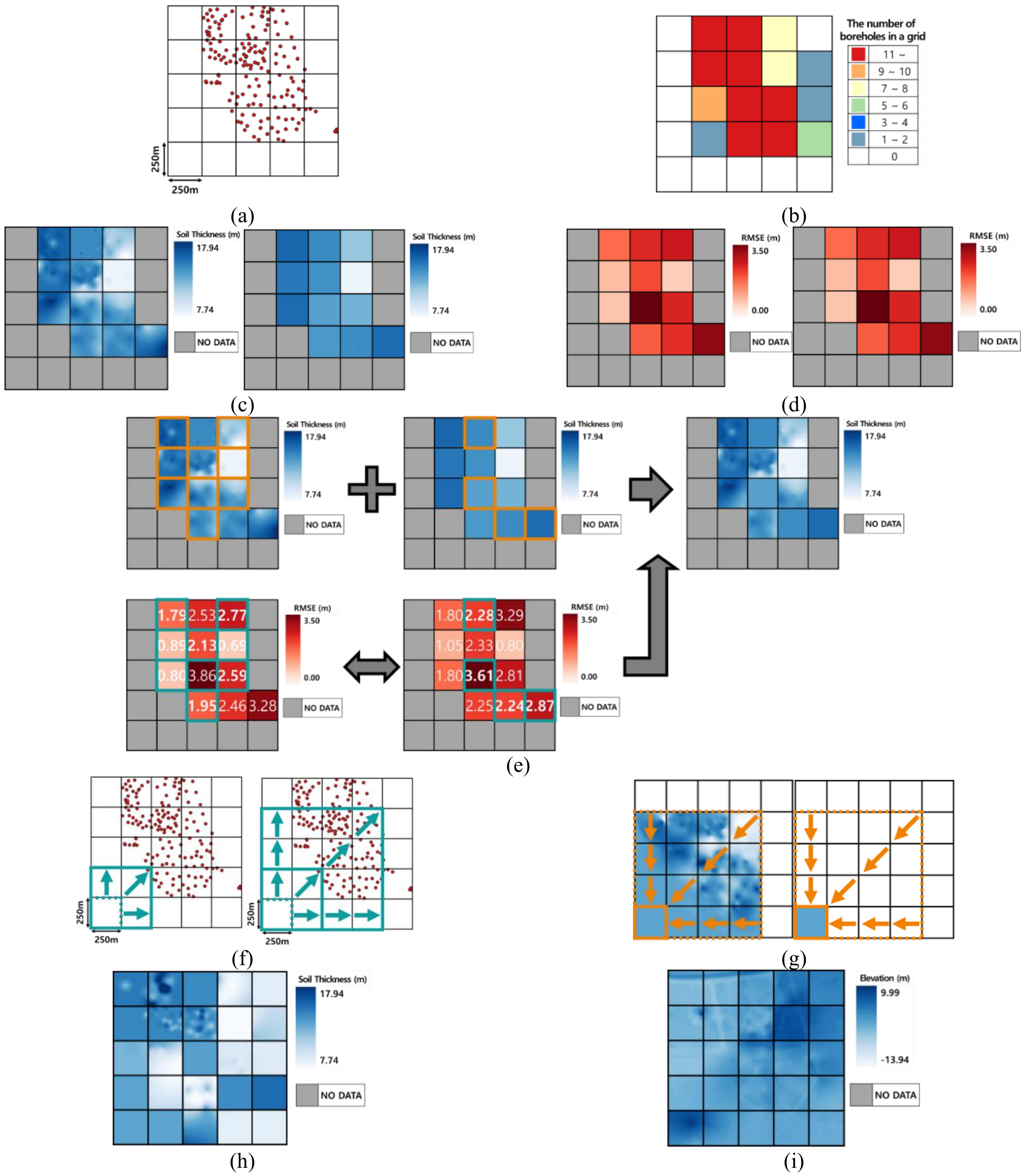


Fig. 10 Illustration of the optimal interpolation model: (a) Target area selection and grid division, (b) Borehole density check per grid, (c) Interpolation comparison (left: kriging, right: averaging), (d) Cross-validation comparison (left: kriging, right: averaging), (e) Optimal RMSE grid extraction, (f) Grid size increase for analysis areas with low borehole density (<4), (g) Results trimming to initial grid size (250 m× 250 m) [Clip grids], (h) Interpolation results merging [Mosaic grids] and (i) Geo-layers boundary elevation calculation

the 250 m×250 m grid-based spatial interpolation outcomes, as validated in Section 3.1 (see Fig. 11). A spatial interpolation was accomplished for 99.55% of Seoul by employing a spectrum of grid sizes, including 250 m × 250 m, 500 m×500 m, 1 km×1 km, and 2 km×2 km. The remaining 0.45%, predominantly along the city's administrative boundaries, exhibited deficiencies in borehole data for effective interpolation. The usage ratios for each grid size were as follows: 16.16% for 250 m×250 m, 30.83% for 500 m×500 m, 39.06% for 1 km ×1 km, and 13.95% for 2 km×2 km. In this adaptive grid interpolation, kriging emerged as the predominant method, accounting for approximately 82% of the usage, with averaging

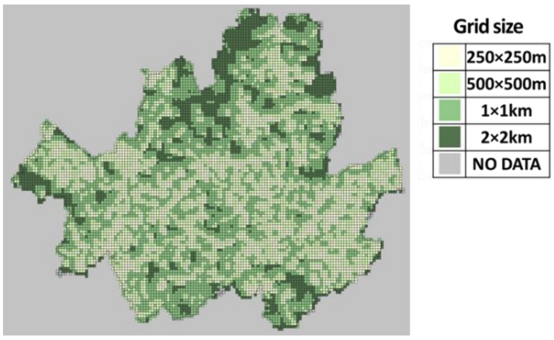


Fig. 11 Ratio of grids utilized in Seoul

(a) (b)

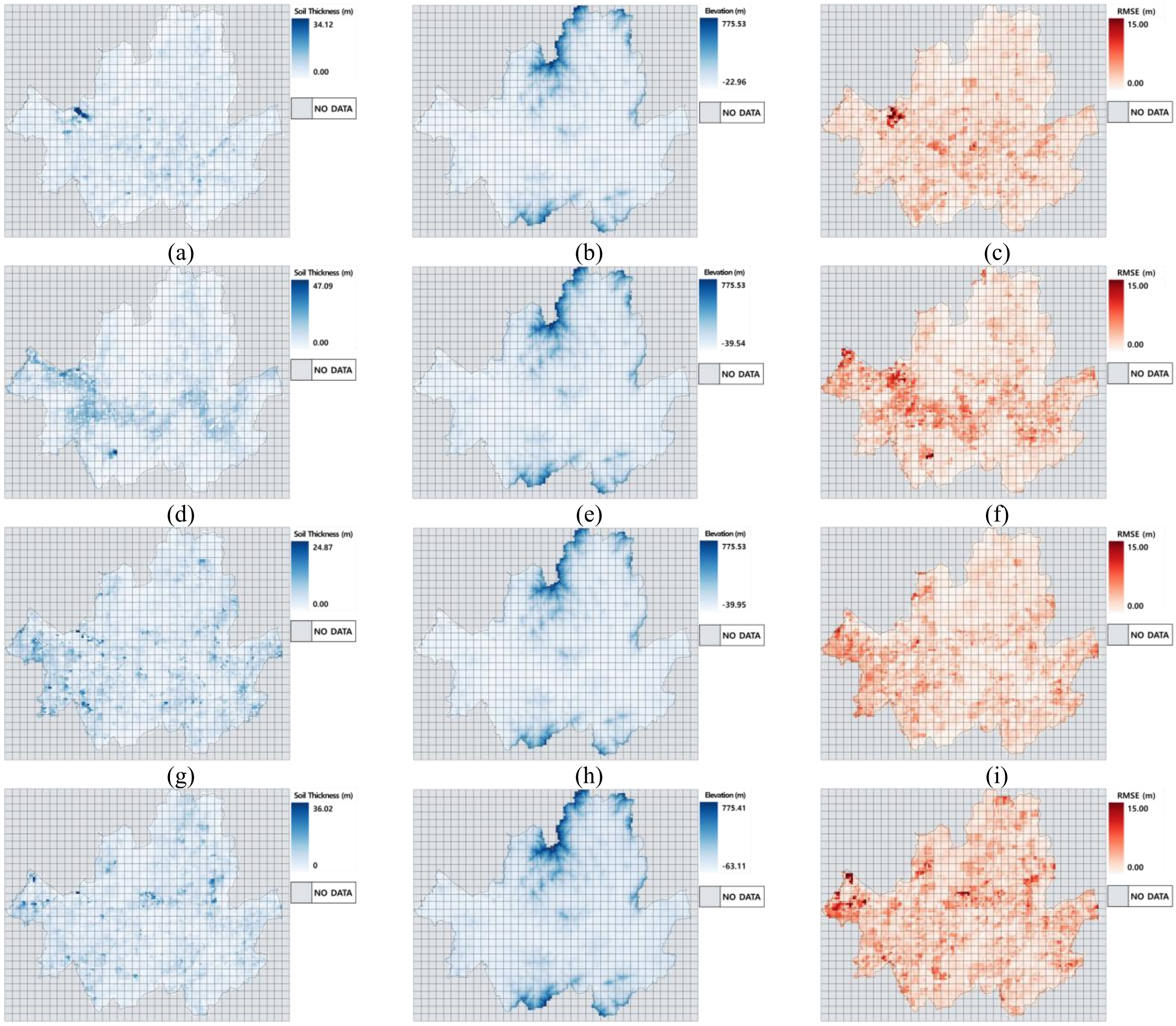
(c) (d)

Fig. 12 Ratio of spatial interpolation methods utilized in Seoul: (a) Fill soil, (b) Sedimentary soil, (c) Residual soil and (d) Weathered rock

constituting the remaining 18%.

An assessment of the predictive accuracy of adaptive grid interpolation using LOOCV is presented in Table 3. The overall mean error (RMSE) was found to be 1.97 meters, representing a significant improvement of approximately 19% in reliability over the kriging results using the entire borehole dataset (error of 2.77 meters, as shown in Table 1). This error reduction was consistent across the varying grid sizes, although larger grid sizes generally correlated with increased errors. Notably, the 1 km×1 km grid size demonstrated higher reliability than the 500 m×500 m grid in residual and weathered rock layers, likely owing to topographic influences, underscoring the need for further terrain-based analyses in future studies.

The RMSE progressively increased with increasing depth, suggesting greater uncertainty in deeper geological layers. Urban development and construction activities can potentially disrupt and alter the arrangement and composition of the existing geological layers. Human activities, such as excavation, filling, and terracing, can lead to the deposition of new materials and reconfiguration of existing geological layers. Urbanization often involves the construction of infrastructure, such as roads and buildings, altering natural drainage patterns and sediment transport processes, resulting in sediment deposition or erosion and forming new geological layers. Industrial activities and pollution introduce contaminants into the environment,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 13 2D results and confidence maps for geo-layers with optimal spatial interpolation: (a-c) Fill soil (thickness, boundary elevation between fill and sedimentary soil, RMSE), (d-f) Sedimentary soil (thickness, boundary elevation between sedimentary and residual soil, RMSE), (g-i) Residual soil (thickness, boundary elevation between residual soil and weathered rock, RMSE), (j-l) Weathered rock (thickness, boundary elevation between weathered rock and soft/moderate rock, RMSE)

Table 3 Results of spatial interpolation with adaptive grids of geo-layer thickness over Seoul

Grid size	Spatial interpolation method	RMSE (m)				
		Fill soil	Sedimentary soil	Residual soil	Weathered rock	Average by grid size
250 m×250 m	Kriging	1.33	1.34	1.74	2.46	1.60
	Averaging	0.73	0.49	0.92	1.98	
500 m×500 m	Kriging	1.53	1.87	1.90	2.62	1.94
	Averaging	1.53	1.51	1.43	2.39	
1 km×1 km	Kriging	1.69	1.95	1.87	2.51	2.02
	Averaging	1.86	2.21	1.71	2.64	
2 km×2 km	Kriging	1.77	2.20	1.98	2.60	2.14
	Averaging	1.90	2.07	1.78	2.84	
Average RMSE by geo-layer		1.60	1.83	1.86	2.56	1.97

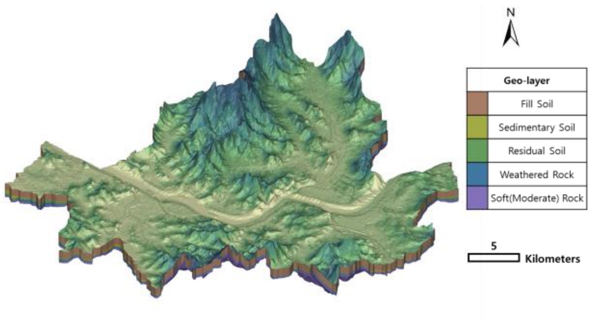


Fig. 14 3D visualization result of optimal spatial interpolation method

which accumulate and alter geological layers (Birkeland 1984). Although human-induced factors can create exceptions, the inherent uncertainties in geological structures become increasingly pronounced with depth.

The two-dimensional (2D) final results (i.e., layer thickness, boundary elevation by layer, and confidence maps) are depicted in Fig. 13. A 3D visualization of these results using CTech's EVS software is presented in Fig. 14. For clarity in visualizing the geological boundaries, the scale of the z-axis was exaggerated ten-fold.

4. Conclusions

This study aimed to determine the most efficacious spatial interpolation technique using borehole data that had undergone meticulous preprocessing. This includes the precise removal of outliers to enhance the robustness and reliability of the outcomes substantially. The process entailed a comprehensive and detailed comparison of the borehole data against DEMs, which is a critical step in achieving standardization that ensures both the accuracy and dependability of the dataset. We derived several key conclusions through this sophisticated approach, which combines the principles of kriging with systematic averaging of ground conditions. This methodological amalgamation not only augments the precision of our spatial analyses but also fortifies the integrity and applicability of our findings in the broader field of geospatial research.

Rigorous verification of the accuracy of the borehole data extracted from various portals demonstrated a notable enhancement in reliability. The initial MAE significantly decreased from 5.20 meters to 2.39 meters post-processing. Similarly, the initial RMSE decreased substantially from 8.78 meters to 5.75 meters, marking an overall improvement of approximately 44%. This remarkable progress underscores the essential role of advanced statistical pre-processing techniques in improving the reliability of spatial interpolation data. Such methods are instrumental not only in refining data representation but also in bolstering the integrity and applicability of the data across a spectrum of geotechnical and civil engineering analyses.

The analysis of spatial interpolation results, delineated by varying grid sizes, revealed a discernible pattern of escalating errors correlated with the expansion of the grid size. Consequently, this study strategically optimized the use of smaller grid sizes, which consistently demonstrated higher reliability. Notably, the RMSE for each geolayer indicated an increasing trend with depth, suggesting that as exploration delves deeper into the subsurface layers, the uncertainty surrounding the geological structure becomes increasingly pronounced.

The proposed optimal spatial interpolation method with adaptive grids introduces a more refined representation of spatial variations within metropolitan areas like Seoul. This is achieved through meticulous grid division, which facilitates localized value estimation. Additionally, the averaging approach applied to borehole data effectively addresses the challenges posed by insufficient or unevenly distributed data, thereby reducing prediction errors, particularly in terrains with less undulation. However, most of these findings are based on data from 2016. Consequently, caution should be exercised when utilizing these interpolations in scenarios involving post-borehole investigation developments or alterations because they may not provide a reliable reference for future construction plans and designs. The execution of additional borehole investigations, particularly in mountainous regions, is vital for enhancing the precision of geotechnical assessments. Such efforts have demonstrated the value of highly reliable subsurface maps, which are indispensable in civil engineering, architecture, and urban development projects, especially during the design phase. Moreover, these investigations provide pragmatic solutions to the challenges posed by the scarcity of geotechnical data due to spatial and temporal constraints. By leveraging known geotechnical characteristics to predict conditions at unsampled locations, these studies play a critical role in proactively identifying ground-related risks such as subsidence, liquefaction, seismic activity, and slope instability, thereby serving as key instruments for disaster prevention and risk management.

From a geotechnical engineering perspective, subsurface maps and geotechnical databases can be effectively utilized in urban planning and infrastructure development by enabling data-driven land suitability analysis, optimizing foundation design, and supporting the strategic placement of critical facilities. When integrated with GIS platforms, these datasets allow urban planners to visualize underground conditions in three dimensions, evaluate seismic site classifications, and avoid geohazard-prone zones during early planning stages. Furthermore, continuous updating and digitization of these databases can facilitate smart city initiatives and resilient infrastructure development by providing real-time access to spatially referenced subsurface information.

Acknowledgments

This study was supported by the project titled "Local Assessment on Earthquake Induced Multi-Geohazard in Urban Area based on Site-specific Geotechnical Characterization" funded by the Daelim Suam Scholarship and Cultural Foundation of Korea and the Research Institute of Construction and Environmental Engineering at Seoul National University. And research for this paper was carried out under the KICT Research Program (project no. 20250270-001, Intelligent Geotechnical Data Analysis and Underground Infrastructure Stability Assessment Based on Data Fusion) funded by the Ministry of Science and ICT.

References

Aziz, M., Khan, T.A. and Ahmed, T. (2017), "Spatial interpolation of geotechnical data: A case study for Multan City, Pakistan", *Geomech. Eng.*, **13**(3), 475-488. https://doi.org/10.12989/gae.2017.13.3.475.

Baecher, G.B. and Christian, J.T. (2005), Reliability and Statistics in Geotechnical Engineering, *John Wiley & Sons, Hoboken*, NJ, USA.

Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data*, (3rd Ed.), Wiley, New York, NY, USA.

Birkeland, P.W. (1984), *Soils and Geomorphology*, Oxford University Press, New York, NY, USA.

Carr, J.R. (1995), *Numerical Analysis for the Geological Sciences*, Prentice Hall, Englewood Cliffs, NJ, USA.

Chen, P.P.S. (1976), "The entity-relationship model—toward a unified view of data", *ACM Trans. Database Syst.*, **1**(1), 9-36. https://doi.org/10.1145/320434.320440.

Chiles, J.P. and Delfiner, P. (2012), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, Hoboken, NJ, USA.

Choe, J. (2013), *Geostatistics*, Sigmapress, Seoul, Korea.

Chung, T.K., Kim, H.S., Sun, C.G. and Chung, C.K. (2025), "Mosaic-based spatial interpolation method considering topography for subsurface mapping using borehole data in Seoul", *KSCE J. Civ. Eng.*, **29**(6), 100091. https://doi.org/10.1007/s12205-025-100091.

Chung, T.K., Kim, M.G. and Chung, C.K. (2018), "Development of integrated geotechnical information management system based on opensource framework with standardization of geotechnical information", *Proceedings of the 2018 Spring Geotechnical Engineering Conference*, Seoul, Korea.

Date, C.J. (2004), *An Introduction to Database Systems*, Pearson Education India, New Delhi, India.

Daw, R.H., Watson, R. and Pearson, E.S. (1972), "Miscellanea studies in the history of probability and statistics. XXX. Abraham De Moivre's 1733 derivation of the normal curve: A bibliographical note", *Philosophy*, **59**(3).

Denkovski, D., Atanasovski, V., Gavrilovska, L., Riihijärvi, J. and Mähönen, P. (2012), "Reliability of a radio environment map: Case of spatial interpolation techniques", *Proceedings of the 7th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, Stockholm, Sweden, June.

Dirks, K.N., Hay, J.E., Stow, C.D. and Harris, D. (1998), "High-resolution studies of rainfall on Norfolk Island: Part II: Interpolation of rainfall data", *J. Hydrol.*, **208**(3-4), 187-193. https://doi.org/10.1016/S0022-1694(98)00155-3.

Gangrade, R., Mooney, M. and Trainor-Guitton, W. (2020), "Incorporating spatial uncertainty into site investigations for tunneling applications", *Proceedings of the Geo-Congress 2020*, Minneapolis, MN, USA. https://doi.org/10.1061/9780784482797.034.

Grasmick, J.G. (2019), "Modeling spatial geotechnical parameter uncertainty and quantitative tunneling risks", Ph.D. Dissertation, Colorado School of Mines, Golden, CO, USA.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, USA.

Hawkins, D.M. (1980), *Identification of Outliers*, (Vol. 11), Chapman and Hall, London, UK.

Hingorani, K., Gittens, D. and Edwards, N. (2017), "Reinforcing database concepts by using entity relationships diagrams (ERD) and normalization together for designing robust databases", *Issues Inf. Syst.*, **18**(1), 148-155. https://doi.org/10.48009/1_iis_2017_148-155.

Hong, E.K. (2012), *Database Course (Based MSSQL)*, Life Power Press, Seoul, Korea.

Hubert, M. and Vandervieren, E. (2008), "An adjusted boxplot for skewed distributions", *Comput. Stat. Data Anal.*, **52**(12), 5186-5201. https://doi.org/10.1016/j.csda.2007.11.008.

Institution of Civil Engineers (1991), *Inadequate Site Investigation*, Thomas Telford, London, UK.

Isaaks, E.H. and Srivastava, R.M. (1989), *Applied Geostatistics*, Oxford University Press, New York, NY, USA.

Jaksa, M., Kaggwa, W., Fenton, G. and Poulos, H. (2003), "A framework for quantifying the reliability of geotechnical investigations", *Proceedings of the 9th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP)*, San Francisco, CA, USA.

Kasugai, M., Oshima, A., Hamada, T. and Yamamoto, K. (2009), "Liquefaction potential in alluvial sand layers in Osaka area and effects of countermeasures by lowering groundwater levels", *Proceedings of the Symposium on Groundwater and Geo-Environment*, 2009.

Kilkenny, M.F. and Robinson, K.M. (2018), "Data quality: 'Garbage in – garbage out'", *Health Inf. Manage. J.*, **47**(3), 103-105. https://doi.org/10.1177/1833358318774357.

Kim, H.S. and Kim, H.K. (2019a), "Optimizing site-specific geostatistics to improve geotechnical spatial information in Seoul, South Korea", *Arab. J. Geosci.*, **12**, 1-20. https://doi.org/10.1007/s12517-018-4171-5.

Kim, H.S., Kim, H.K., Shin, S.Y. and Chung, C.K. (2012), "Application of statistical geo-spatial information technology to soil stratification in the Seoul metropolitan area", *Georisk*, **6**(4), 221-228. https://doi.org/10.1080/17499518.2012.744248.

Kim, H.S., Sun, C.G. and Cho, H.I. (2019b), "Site-specific zonation of seismic site effects by optimization of the expert GIS-based geotechnical information system for western coastal urban areas in South Korea", *Int. J. Disaster Risk Sci.*, **10**(1), 117-133. https://doi.org/10.1007/s13753-018-0208-3.

Kim, M., Chung, C.K., Han, J.W. and Kim, H.S. (2023), "Three-dimensional geostatistical modeling of subsurface stratification and SPT-N value at dam site in South Korea", *Geomech. Eng.*, **34**(1), 29-41. https://doi.org/10.12989/gae.2023.34.1.029.

Korean National Geographic Information (2013), "General Survey Work Regulation".

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013), "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", *J. Exp. Soc. Psychol.*, **49**(4), 764-766. https://doi.org/10.1016/j.jesp.2013.03.013.

Li, J. and Heap, A.D. (2008), "A review of spatial interpolation methods for environmental scientists", *Ecol. Inform.*, **6**(3-4), 228-241. https://doi.org/10.1016/j.ecoinf.2010.12.003.

Li, J. and Heap, A.D. (2014), "Spatial interpolation methods applied in the environmental sciences: A review", *Environ. Model. Softw.*, **53**, 173-189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Lightfoot, E. and O'Connell, T.C. (2016), "On the use of biomineral oxygen isotope data to identify human migrants in the archaeological record: Intra-sample variation, statistical methods and geographical considerations", *PLoS One*, **11**(4), e0153850. https://doi.org/10.1371/journal.pone.0153850.

Littlejohn, G.S., Cole, K. and Mellors, T.W. (1994), "Without site investigation ground is a hazard", *Proc. Inst. Civil Eng.-Civil Eng.*, **102**(2), 72-78.

Luo, W., Taylor, M.C. and Parker, S.R. (2008), "A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales", *Int. J. Climatol.*, **28**(7), 947-959. https://doi.org/10.1002/joc.1583.

Mair, A. and Fares, A. (2011), "Comparison of rainfall interpolation methods in a mountainous region of a tropical island", *J. Hydrol. Eng.*, **16**(4), 371-383. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000330.

Man, Y., Wei, L., Gang, H. and Juntao, G. (2010), "A novel data quality controlling and assessing model based on rules", *Proceedings of the 2010 Third International Symposium on Electronic Commerce and Security*, Guangzhou, China, July. https://doi.org/10.1109/ISECS.2010.15.

Mans, R.S., van der Aalst, W.M.P. and Vanwersch, R.J.B. (2015), "Data quality issues", *Springer Briefs in Business Process Management*, 79-88. https://doi.org/10.1007/978-3-319-16071-9_6.

Ministry of Land Transport and Transportation (2015), "Railway Design Standard: Roadbed", Ministry of Land Transport and Transportation, Sejong, South Korea.

Ministry of Land, Infrastructure and Transport (2021), "Enforcement Decree of the Act on Promotion of Construction Technology", Ministry of Land, Infrastructure and Transport, Sejong, South Korea.

Moivre, A. (2013), *The Doctrine of Chances or, A Method of Calculating the Probability of Events in Play*, Cambridge University Press, Cambridge, UK.

National Research Council (1984), "Geotechnical Site Investigations for Underground Projects: Volume 1", National Academies Press, Washington, D.C., USA.

Phoon, K.K. and Kulhawy, F.H. (1999), "Evaluation of geotechnical property variability", *Can. Geotech. J.*, **36**(4), 625-639. https://doi.org/10.1139/t99-039.

Raheem, A.M. and Omar, N.Q. (2021), "Investigation of distinctive physico-chemical soil correlations for Kirkuk city using spatial analysis technique incorporated with statistical modeling", *Int. J. Geo-Eng.*, **12**(1), 18. https://doi.org/10.1186/s40703-021-00147-2.

Risk, C. and James, P.M.A. (2022), "Optimal cross-validation strategies for selection of spatial interpolation models for the Canadian forest fire weather index system", *Earth Sp. Sci.*, **9**(2). https://doi.org/10.1029/2021EA002019.

Rousseeuw, P.J. (1991), "Tutorial to robust statistics", *J. Chemometrics*, **5**(1), 1-20. https://doi.org/10.1002/cem.1180050103.

Santos, F. (2020), "Modern methods for old data: An overview of some robust methods for outliers detection with applications in osteology", *J. Archaeol. Sci. Rep.*, **32**, 102423. https://doi.org/10.1016/j.jasrep.2020.102423.

Sisman, R., Sahin, A. and Hori, M. (2017), "Automatic 3D soil model generation for southern part of the European side of Istanbul based on GIS database", *Geomech. Eng.*, **13**(6), 893-906. https://doi.org/10.12989/gae.2017.13.6.893.

Sun, C.G., Cho, H.I., Kim, H.S. and Lee, M.G. (2022), "Determining N value from SPT blows for 30 cm penetration in weathered strata", *Geomech. Eng.*, **28**(6), 625-636. https://doi.org/10.12989/gae.2022.28.6.625.

Sun, C.G., Chun, S.H., Ha, T.G., Chung, C.K. and Kim, D.S. (2008), "Development and application of a GIS-based tool for earthquake-induced hazard prediction", *Comput. Geotech.*, **35**(3), 436-449. https://doi.org/10.1016/j.compgeo.2007.08.001.

Todo, H., Yamamoto, K., Mimura, M. and Yasuda, S. (2013), "Japan's nation-wide electronic geotechnical database systems by Japanese geotechnical society", *Geotech. Geol. Eng.*, **31**, 941-963. https://doi.org/10.1007/s10706-012-9562-x.

Toews, M. (2015), "Welcome to the Users PostGIS Wiki!", Retrieved from http://trac.osgeo.org/postgis/wiki/UsersWikiMain.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, USA.

Uzielli, M., Lacasse, S., Nadim, F. and Phoon, K.K. (2007), "Soil variability analysis for geotechnical practice", *Characterization Eng. Nat. Soils*, **3-4**, 1653-1752.

Whyte, I.L. (1995), "The financial benefit from a site investigation strategy", *Ground Eng.*, **28**(8), 33-36.

Willmott, C.J. (1982), "Some comments on the evaluation of model performance", *Bull. Am. Meteorol. Soc.*, **63**(11), 1309-1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2.

Yasuda, S. and Watanabe, H. (2010), "A study on vibration of ground at Central Tokyo during earthquakes using the nation-wide electronic geotechnical database systems", *Proceedings of the 13th Japan Earthquake Engineering Symposium*, Ibaraki, Japan.

Yu, D., Sheikholeslami, G. and Zhang, A. (2002), "FindOut: Finding outliers in very large datasets", *Knowl. Inf. Syst.*, **4**, 387-412. https://doi.org/10.1007/s101150200013.

Yuen, K.V. and Mu, H.Q. (2012), "A novel probabilistic method for robust parametric identification and outlier detection", *Prob. Eng. Mech.*, **30**, 48-59. https://doi.org/10.1016/j.probengmech.2012.06.002.

Zheng, S., Zhu, Y.X., Li, D.Q., Cao, Z.J., Deng, Q.X. and Phoon, K.K. (2021), "Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning", *Geosci. Front.*, **12**(1), 425-439. https://doi.org/10.1016/j.gsf.2020.03.017.

Zumrawi, M. (2014), "Effects of inadequate geotechnical investigations on civil engineering projects", *Int. J. Sci. Res.*, **3**(6), 927-931.

IC