Application of GIS-based neural network models for subsurface stratification

Mingi Kim (Primary author)
Seoul National University, Seoul, South Korea, doulos.mk@gmail.com

Choong-Ki Chung (Corresponding author)
Seoul National University, Seoul, South Korea, geolabs@snu.ac.kr

ABSTRACT: Mathematical and geostatistical approaches such as the inverse distance weighted and kriging interpolation methods have been widely used to integrate the information of subsurface strata. These approaches have limitations, however, in that the spatial variability of the soil strata is not considered, or a statistical assumption is essential for interpolation. In this study, a method of subsurface stratification based on the application of GIS-based artificial neural networks was developed. This method is real-data-driven and does not require any geostatistical assumption. More than 30,000 site investigation data of Seoul from the integrated DB center for national geotechnical information in South Korea were adopted for training and validation of neural network models. Outlier analysis based on cross-validation using 4×4km grids was carried out for detecting and eliminating the outliers. The GIS-based topographic features, such as gradient and geological maps, were also utilized for modeling artificial neural networks. The spatial interpolation results of Seoul from the GIS-based neural network model and ordinary kriging were presented. Finally, cross-validation was implemented, and the root mean square errors of the methods were employed for quantitative performance evaluation.

Keywords: Subsurface stratification, neural network, GIS, database, outlier analysis

1. Introduction

The reliable prediction of subsurface geolayer information is essential for estimating the earth volume of infrastructure construction sites. The number of borehole data required for the prediction of geolayer information is limited, however, for economic and spatiotemporal reasons. In addition, it is impossible to secure reliable three-dimensional geolayer information from the one-dimensional borehole data points. The earth volume of construction sites has been estimated based on the engineer's experience, but as an empirical method, numerous errors occur at the actual site conditions.

The spatial integration methods for interpolating discontinuous spatial variables include mathematical techniques like the simple averaging, triangular, and inverse weighted distance methods as well as the geostatistical technique, which is represented by kriging methods. [1-3] In particular, geostatistical techniques have been widely used as a tool for solving spatial problems. They are known to be more reliable than mathematical techniques because the spatial variability in the distance between two points is considered for spatial interpolation.

Geostatistical techniques, however, are difficult to apply in construction areas with high uncertainty and variability. It also has the disadvantage of requiring various statistical assumptions. Therefore, there is a need to improve the prediction reliability and accuracy by introducing a data-driven approach instead of the conventional empirical and statistical approaches.

In this study, geographic information system (GIS)-based artificial neural network models were constructed for the reliable prediction of geolayer information, and their reliability was evaluated compared to the conventional method, ordinary kriging. First of all, in the whole area of Seoul, the capital of South Korea, a geo-database was built by collecting and standardizing borehole information from the integrated database (DB) center for national geotechnical information in South Korea. Using the constructed database, the artificial neural networks (ANNs) were learned, and models for the prediction of geolayer information were constructed. The models included a single-layer neural network model and a multiple-layer neural network model (multi-layer perceptron, MLP). The subsurface elevations of each geolayer were estimated from the coordinate information of the boreholes through two ANN models. The reliability of the prediction results of the ANN model and the geostatistical technique, ordinary kriging, was quantitatively evaluated using independent cross-validation.

2. Methodology

2.1 Ordinary kriging

Geostatistics is a field of statistics that solves spatiotemporal problems by analyzing the distribution and correlations of regionalized variables. [4] Ordinary kriging is one of the most representative methods of geostatistics, and it uses variograms to correlate spatial data and provide predictive values. Specifically, the properties of unsampled points are predicted through the linear weighted combination of the surrounding values already known. The ordinary kriging equation is expressed as shown below.

$$z^* = \sum_{i=1}^{n} \lambda_i z_i \ \ and \ \ \sum_{i=1}^{n} \lambda_i = 1$$

where z^* is the predicted value at an unknown point, z_i

is the value whose location and data are already known, λ_i is the weight of each point, and is the number of data used for kriging prediction.

The kriging method assumes that the error between the predicted and true values should be minimized for the determination of the weight, and that the estimated value should not be biased. Due to this characteristic, kriging shows excellent prediction ability when the raw data distribution is gentle and follows the normal distribution. It is disadvantageous, however, in that the variance of kriging predictions tends to decrease sharply when raw data with large variance are used, and in that the characteristic distributions containing extreme values are difficult to predict. [5]

2.1. Artificial neural network models

ANNs are machine learning algorithms created by mathematically simulating the process of human neural networks. Researches on ANN have been conducted for a long time, but since the development of a coefficient estimation method called "back-propagation," ANN has been used to solve various problems. [6-8] Especially in the current age of big data, it is widely used to process various characters, graphics, and sound information. [9]

ANNs in machine learning are classified into supervised and unsupervised learning according to the purpose of analysis. The supervised neural network models for prediction include the multi-layer perceptron, recurrent neural network, convolutional neural network, etc. In this study, the MLP models were used for the prediction of geolayer information. Fig. 1 shows the framework for the development of a subsurface geolayer prediction model using ANN.

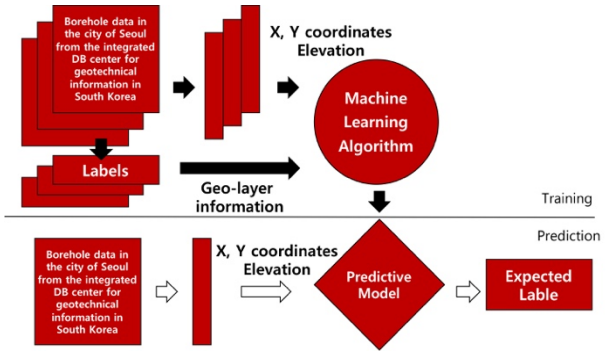

Figure 1. Framework for the development of a subsurface geolayer prediction model using ANN: training and prediction phases.

The neural network is composed of an input layer, a hidden layer (or hidden layers), and an output layer. In the input layer, there are neurons in which each input variable matches 1:1. In the hidden layer, there are neurons generated by combining the neurons of the input layer and the weight. The complexity of the model is determined according to the number of layers in the hidden layer, and when the number of hidden layers becomes 2 or more, the model is called "deep neural network" or "deep learning." In the output layer, there are neurons generated by combining the neurons in the hidden layer with the weights, and the number of output layers is determined according to the type of dependent variable (numeric, binary, or multinomial) to be predicted. The neurons in the hidden and output layers perform the function of calculating the sum of the input values and the weights of the previous layer. In addition, an activation function for outputting a signal as a weighted sum of neurons is performed.

In this study, the independent variables of the input layer used the x, y coordinates and the elevation of the boreholes, and the dependent variables were set as geolayer information in elevation form. Two neural network models (single- and multiple-layer) were created, as shown in Fig. 2 and 3, respectively. In the single-layer neural network model, there is a hidden layer between the independent and dependent variables, and the activation function creates a nonlinear effect, like the generalized linear model.

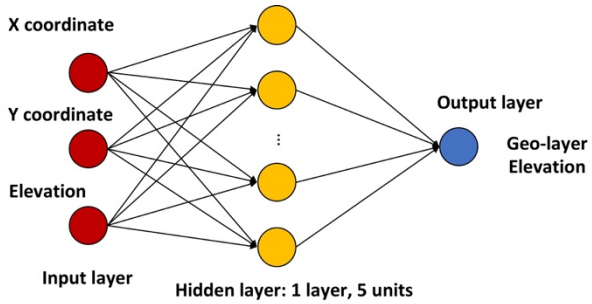

Figure 2. A single-layer neural network model for subsurface geolayer stratification, with one hidden layer.

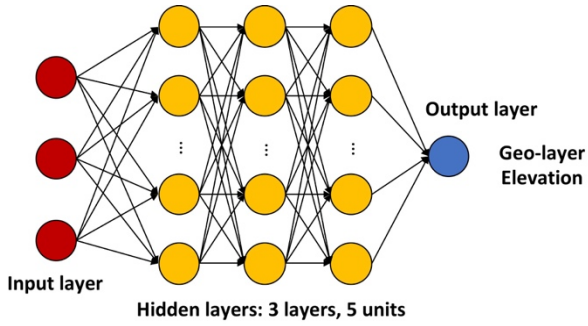

Figure 3. A multiple-layer neural network model for subsurface geolayer stratification, with three hidden layers.

In the multiple-layer neural network model, the number of hidden layers is added to the single-layer neural network model. The number of hidden layers was 3, and each hidden layer was composed of 5 unit neurons. The neural network finds the weight that minimizes the prediction error through the gradient descent method. In the neural network models for the prediction of geolayer information, an Adam optimizer is used to finely adjust the learning speed so as to minimize the prediction error. The dropout technique is also applied to prevent the overfitting problem that may be caused by the presence of many hidden layers. The dropout method simplifies the learning model by randomly selecting a constant percenter of input or hidden layer neurons during each learning process. In addition, the Xavier initial value setting technique was applied. The neural network models were created by tensorflow, an open-source software developed by Google and released in 2015.

3. Geo-Database Construction

3.1. Borehole information of Seoul

For the prediction of the subsurface geolayer information of Seoul, site investigation data were collected from the integrated DB center for national geotechnical information in South Korea (https://www.geoinfo.or.kr). Seoul is the capital of South Korea and has a 605.25km^2 area. The collected data were 33,867 boring data as of 2017, and the distribution of the data is shown in Fig. 4 below.

Geolayer information is distributed in a specific order, depending on the depth of each stratum according to the geological formation process. In the case of the subsurface of Seoul, bedrock, soft rock, weathered rock, weathered residual soil, sedimentary soil, and fill soil are generally distributed in order from the bottom. For the prediction of geolayer information, the thickness of each stratum or the elevation value between the strata can be used. In this study, the elevation values between the strata were used as a dependent variable.

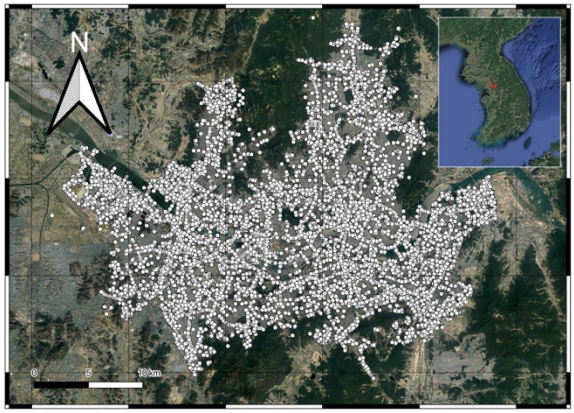

Figure 4. Distribution of the boreholes in Seoul for the development of ANN models (total of 33,867 boreholes as of 2017).

3.2 Outlier analysis

Unfortunately, 15,301 of the boring data of Seoul were found not to have subsurface geolayer information. In addition, even if the subsurface geolayer information was included, there were some cases in which the ground elevation of the boreholes was significantly different from that in the digital elevation model (DEM). DEM contains data that show the shape of the terrain by storing the elevation value of the terrain as a numerical value. There were 2,710 boreholes with a more-than-100-meter ground elevation deviation from DEM. This deviation was due to the fact that the notation of ground elevation was different for each construction project, and the ground elevation values of 2,439 boreholes were corrected with the ground elevation from DEM. [10]

It is important to use reliable data for the learning of neural network models. The site investigation data, however, may include outliers due to human factors like measurement errors as well as the inherent homogeneity and spatial variability of the soil. Outlier analysis was performed to remove the outliers included in the boring data of Seoul.

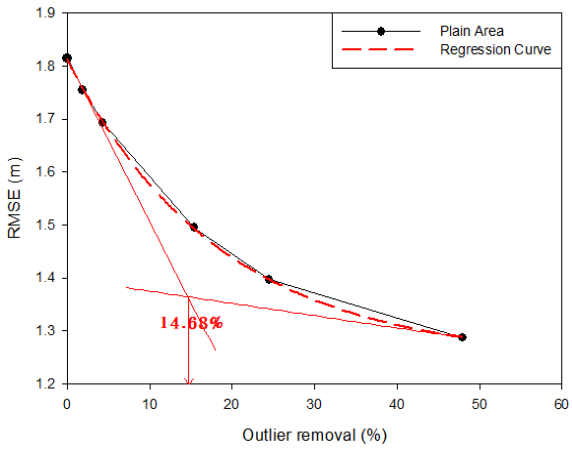

Figure 5. Outlier analysis result of the borehole dataset of Seoul using cross-validation: one of the 4×4km plain areas.

The outlier detection method based on cross-validation was applied to the boring data of Seoul. [11][12] The entire area of Seoul was divided into 4x4km subdivisions, and outlier analysis was performed for each divided area. Fig. 5 shows the result of one of the outlier analyses of the borehole dataset of Seoul. As outliers increased, the root mean square error (RMSE) decreased. As the percentage of boreholes removed increased, however, the number of boring data available for neural network learning decreased. Therefore, the method to determine the optimal outlier ratio was adopted, such as the threshold point of the RMSE and outlier ratio to total number of data.

Table 1. Numbers of boreholes from the geo-database of Seoul used to construct ANN models

Total boreholes	33,867 boreholes
Boreholes with ground elevation data	18,566 boreholes
Boreholes with a more-than-100-meter ground elevation deviation from DEM	2,710 boreholes
Boreholes with corrected ground elevation	2,439 boreholes
Boreholes with outliers	2,506 boreholes
Boreholes without outliers	15,789 boreholes
Boreholes for learning	11,052 boreholes
Boreholes for validation	4,737 boreholes

Table 1 shows the numbers of geo-database borehole information used for neural network traning and validation. The number of boreholes exclued by outlier analysis among the boring data with geolayer information was 2,506. Of the data excluding the outliers, 11,502 boring data (70%) were used for learning, and 4,737 boring data (30%) were used for validation.

4. Spatial Interpolation Results of the Geolayer Information of Seoul

The subsurface elevation values of each geolayer were interpolated using ordinary kriging, the single-layer neural network model, and the multiple-layer neural network model. A part of Seoul was selected for the target area for spatial interpolation. Fig. 6 shows the air view of the target area. Its size was 766x651 m. The 74 boring information from the geo-database was used as the spatial interpolation of the target area.

As mentioned before, three independent variables were used as input values: the x, y coordinates and the ground elevation, and for the output values, the dependent variable of geolayer information was used. Geolayer information refers to the four elevations of strata boundary: the elevation between the fill and sedimentary soil layer, the elevation between the sedimentary and residual soil layer, the elevation between the weathered residual soil and weathered rock layer, and the elevation between the weathered and soft rock layer. [3]

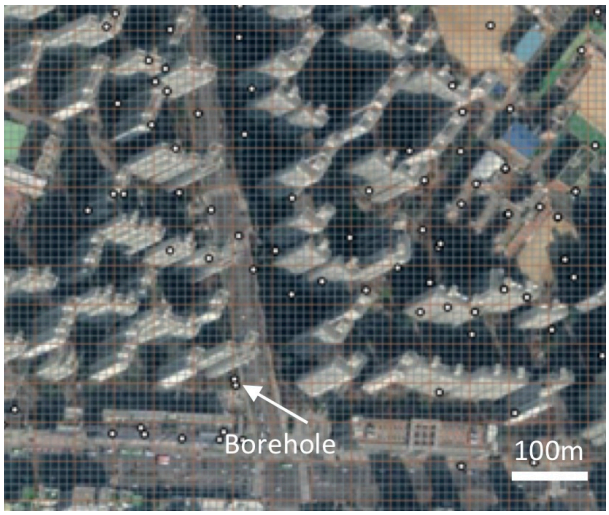

Figure 6. Application area of subsurface geolayer stratification.

Root mean square error was used as an indicator for the quantitative analysis of the reliability of each spatial interpolation result. The RMSE values of the neural network models were obtained during the traning and validation phases, and the RMSE values of ordinary kriging were calculated through cross-validation. RMSE refers to the difference between the actual field test values and the estimated values, and is expressed by the equation below.

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

where y_i is the actual experimental (observed) value, $\hat{y}_i$ is the estimated value, and N is the total number of boreholes.

The cross-validation method used to obtain the RMSE values from ordinary kriging is classified as leave-one-out, k-fold, or independent cross-validation according to the number of excluded points used in the analysis process. In this study, the independent cross-validation method was used. Table 2 shows the RMSE values through the reliablity analysis results.

Table 2. Reliability analysis results obtained using independent cross-validation: RMSE (m)

Elevation between strata	Ordinary kriging	Single-layer neural network	Multiple-layer neural network
Fill soil and sedimentary soil	3.31	4.20	3.19
Sedimentary soil and weathered residual soil	3.10	6.79	2.52
Residual soil and weathered rock	4.69	10.01	5.68
Weathered rock and soft rock	6.91	9.55	6.18

Among the three spatial interpolation methods, the single-layer neural network model recorded the largest RMSE values from the reliability test. In particular, it had the largest RMSE in the boundary layer between residual soil and weathered rock. The RMSE values obtained from the reliability analysis of the multiple-layer neural network model and ordinary kriging had similar tendencies for all the boundary layers. The RMSE of the multiple-layer neural network model was smaller than that of ordinary kriging, except for the boundary layer between the residual soil and weathered rock.

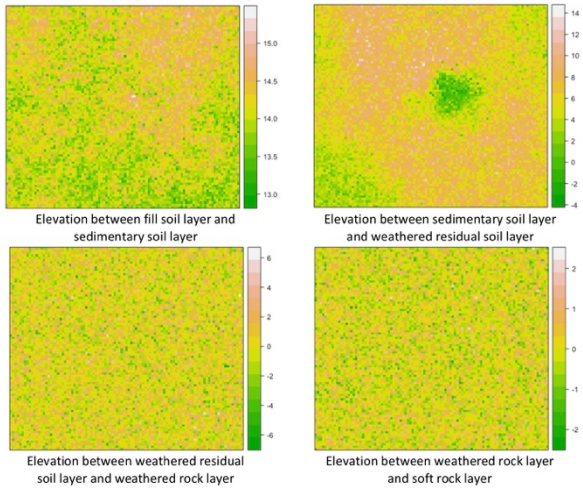

Figure 7. Subsurface stratification results of the geolayers obtained using the multiple-layer neural network model.

The results of the reliability test showed that there is slightly higher reliability of the multiple-layer neural network model over the geostatistical method that is conventionally used to predict the geolayer information. Thus, the applicability of the multiple-layer neural network model in the prediction of geolayer information was confirmed. As can be seen in Fig. 7, however, the multiple-layer neural network model cannot completely

reproduce the continuity of the geolayer information. In the shallow boundary layers, the prediction result showed some spatial variability, but in the deep boundary layers, random values were predicted without any spatial tendency. This might have been due to the insufficient data for learning and validation, incorrect neural network model construction, and lack of independent input variables.

5. Conclusions and Further Research

In this study, geographic information system (GIS)-based artificial neural network (ANN) models were constructed for the reliable spatial interpolation of subsurface geolayer information, and their reliability was evaluated compared to that of the conventional geostatistical method, ordinary kriging. The study results and the further researches needed are summarized below.

(1) A site investigation geo-database was constructed with the boring data obtained from the integrated DB center for national geotechnical information in South Korea to predict and interpolate the geolayer elevation.

(2) A single-layer neural network model and a multiple-layer neural network model were created by learning the boring data from the geo-database. The x, y coordinates and the ground elevation were set as input variables, and the elevation layers between the strata were set as the output variables.

(3) Spatial interpolation of the subsurface elevation of each geolayer was carried out using the two aforementioned neural network models and ordinary kriging. For the results of the reliability test, the single-layer neural network model had the lowest reliability. The predictive reliability of the multiple-layer neural network model was not significantly different from that of ordinary kriging.

(4) Apart from the reliability analysis results, the spatial variability and continuity of the geolayer information were not properly reproduced in some boundary layers between the subsurface strata.

(5) This research is part of an ongoing project with the purpose of assessing the applicability of the prediction and spatial interpolation of subsurface geolayer information. Further research will be conducted to compensate for the shortcomings of the above ANN models. In such research, for instance, the topographic elements will be considered input variables of the ANN model. For this purpose, a GIS-based database of various topographic elements was constructed, as shown in Fig. 8. The topographic factors under consideration are the slope angle, dip direction, topographic index, wetness index, and gradient curvature.

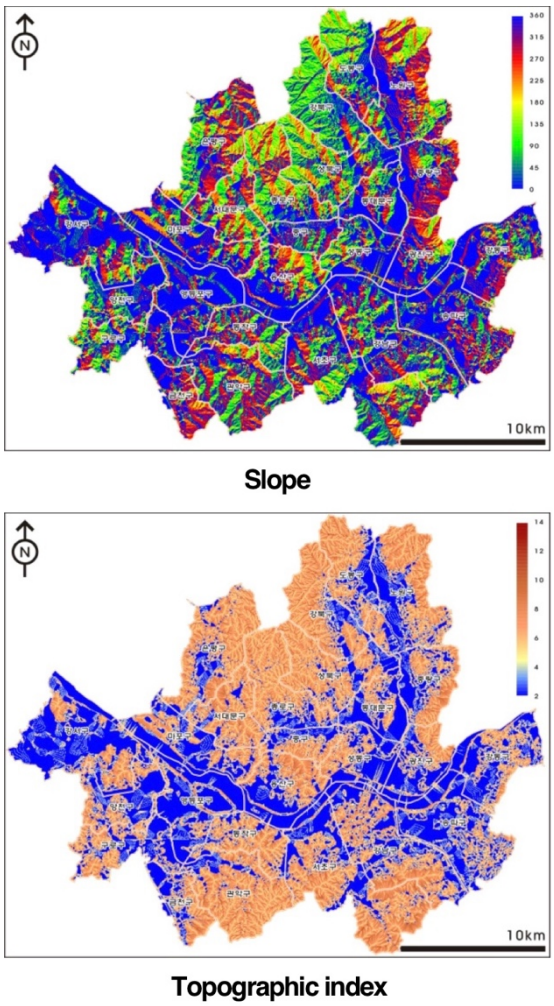

Figure 8. Example of a GIS-based topographic element database in Seoul: slope and topographic index.

Acknowledgement

This research was funded by the Korea Institute of Civil Engineering and Building Technology and supported by Institute of Engineering Research, Seoul National University.

References

[1] S. Altun, A. GÖKTEPE, and A. Sezer, "Geostatistical interpolation for modelling SPT data in northern Izmir," *Sadhana*, vol. 38, no. 6, pp. 1451–1468, Dec. 2013.

[2] E. Hosseini, R. Gholami, and F. Hajivand, "Geostatistical modeling and spatial distribution analysis of porosity and permeability in the Shurijeh-B reservoir of Khangiran gas field in Iran," *J Petrol Explor Prod Technol*, vol. 9, no. 2, pp. 1051–1073, Nov. 2018.

[3] S-H. Chun, C-G. Sun, and C-K. Chung, "Application of Geostatistical Method for Geo-Layer Information," *Journal of The Korean Society of Civil Engineers*, vol. 25, no. 2, pp. 103–115, 2005.

[4] J. P. Chiles and P. Delfiner, *Geostatistics: modeling spatial uncertainty*. 2009.

[5] J. K. Yamamoto, "Correcting the Smoothing Effect of Ordinary Kriging Estimates," *Mathematical Geology*, vol. 37, no. 1, pp. 69–94, Jan. 2005.

[6] J. Kim, "Franchise Business Sales Forecasting by comparison of Neural Network models," *Journal of the Korean Marketing Association,* vol. 33, no. 3, pp. 73–90, Aug. 2018.

[7] J. K. Kumar, M. Konno, and N. Yasuda, "Subsurface Soil-Geology Interpolation Using Fuzzy Neural Network," *Journal of*

Geotechnical and Geoenvironmental Engineering, vol. 126, no. 7, pp. 632–639, Jul. 2000.

[8] F. Farrokhzad, A. Barari, A. J. Choobbasti, and L. B. Ibsen, "Neural network-based model for landslide susceptibility and soil longitudinal profile analyses: Two case studies," Journal of African Earth Sciences, vol. 61, no. 5, pp. 349–357, Dec. 2011.

[9] P. Samui and T. G. Sitharam, "Site Characterization Model Using Artificial Neural Network and Kriging," Int. J. Geomech., vol. 10, no. 5, pp. 171–180, Oct. 2010.

[10] T-K. Chung, "Optimal spatial interpolation method considering topography and density of geotechnical information, Master thesis, 2018

[11] J-J. Kim, "3D geostatistical integration based on site-specific correlations with borehole data and geophysical data", Master thesis, 2015

[12] H-S. Kim, "Geo-spatial data integration for subsurface stratification of dam site with outlier analyses", Environ Earth sci., vol. 75:168, 2016.