Article

Terrain Proxy-Based Site Classification for Seismic Zonation in North Korea within a Geospatial Data-Driven Workflow

Han-Saem Kim [ID], Chang-Guk Sun *[ID], Moon-Gyo Lee and Hyung-Ik Cho [ID]

Earthquake Research Center, Korea Institute of Geoscience and Mineral Resources, Daejeon 34132, Korea; adoogen@kigam.re.kr (H.-S.K.); mglee@kigam.re.kr (M.-G.L.); hicho@kigam.re.kr (H.-I.C.)
* Correspondence: pungsun@kigam.re.kr; Tel.: +82-42-868-3265

Abstract: Numerous seismic activities occur in North Korea. However, it is difficult to perform seismic hazard assessment and obtain zonal data in the Korean Peninsula, including North Korea, when applying parametric or nonparametric methods. Remote sensing can be implemented for soil characterization or spatial zonation studies on irregular, surficial, and subsurface systems of inaccessible areas. Herein, a data-driven workflow for extracting the principal features using a digital terrain model (DTM) is proposed. In addition, geospatial grid information containing terrain features and the average shear wave velocity in the top 30 m of the subsurface (V_{S30}) are employed using geostatistical interpolation methods; machine learning (ML)-based regression models were optimized and V_{S30}-based seismic zonation in the test areas in North Korea were forecasted. The interrelationships between V_{S30} and terrain proxy (elevation, slope, and landform class) in the training area in South Korea were verified to define the input layer in regression models. The landform class represents a new proxy of V_{S30} and was subgrouped according to the correlation with grid-based V_{S30}. The geospatial grid information was generated via the optimum geostatistical interpolation method (i.e., sequential Gaussian simulation (SGS)). The best-fitting model among four ML methods was determined by evaluating cost function-based prediction performance, performing uncertainty analysis for the empirical correlations of V_{S30}, and studying spatial correspondence with the borehole-based V_{S30} map. Subsequently, the best-fitting regression models were designed by training the geospatial grid in South Korea. Then, DTM and its terrain features were constructed along with V_{S30} maps for three major cities (Pyongyang, Kaesong, and Nampo) in North Korea. A similar distribution of the V_{S30} grid obtained using SGS was shown in the multilayer perceptron-based V_{S30} map.

Keywords: digital terrain model; proxy; site classification; seismic site effect; V_{S30}; regression model; geographic information system; North Korea

1. Introduction

Numerous seismic activities such as earthquakes, collapses, and explosions have been recorded in North Korea. However, assessing the seismic hazards is challenging because of the lack of and/or discrepancies in the data from public seismic stations, damage inventories of earthquakes, and site investigation results. Historically, the largest earthquake (M_W 6.2; 19 March 1952) in the Korean Peninsula occurred near Pyongyang, North Korea [1]. Moreover, nuclear tests between 2006 and 2017 triggered events such as mine collapses [2]. In North Korea, the Phyongnam and Taebaeksan Basins are two major Paleozoic basins in the Rangnim massif range and the Gyeonggi and Yeongnam massifs, where seismic amplification and ground shaking tend to occur [3]. Therefore, seismic site effects due to the amplification of seismic waves have been suspected to cause aggravated earthquake hazards in the basin area.

The damage caused by earthquakes is enhanced by the reflection and dispersion of seismic waves near the surface or shallow subsurface [4]. In city-scale areas with high spatial variability of terrain or geology, certain wave propagation phenomena related to

local amplification have been investigated via numerical simulations and seismological observations [5]. In nationwide areas comprising sedimentary basins, valleys, and mountain ridges or tops, empirical cross-correlations with terrain, geology, and geophysical characteristics have been estimated, while site response indexes have been developed by dividing these areas via macro- or micro-seismic zonation.

Many counties have seismic codes to compute the seismic site effects at local sites that are based on the observed average shear wave velocity (V_S) in the top 30 m of the subsurface (V_{S30}) [6]. However, in the absence of site surface data or seismological observations, the validity of V_{S30} in tectonically active complex territories, particularly in geological contexts (such as the existence of inversions in the V_S profiles), has been challenged by certain researchers [7–9]. To indirectly predict V_{S30} distribution, Ahdi et al. [10] developed V_{S30} maps that included local proxy types, namely topographical slope and geomorphological landscape containing high spatial variability. The terrain features (curvature and alluvial fan) influence the subsurface strata and their stiffness. Accordingly, the geomorphological, landform, and land cover features should be incorporated into the subsoil investigation results for geotechnical or geology engineering to visualize the V_{S30} map.

The digital elevation model (DEM) and digital terrain model (DTM) are useful for developing V_{S30} maps based on geographic information system (GIS)-based regression analysis [11]. DEMs of various resolutions for several regions across the United States were used to examine the use of optimum resolution DEMs in developing V_{S30} maps [12,13]. A method of generating the automatic topography classification using a 50 m DEM was developed using multiple linear regression analysis of the logarithm correlations between the observed V_{S30} and the topographic attributes reported by Iwahashi and Pike [14]. However, seismic amplification associated with proxies is too nuanced for determining a single synthetic parameter such as V_{S30} [15,16].

Therefore, to integrate the proxies for determining V_{S30} and other site response parameters, a data-driven approach including a random effect procedure, based on artificial intelligence (AI), is essential. This data-driven approach is suitable for the identification of geomechanical risks and facilitates the avoidance of pitfalls (e.g., low correlation coefficient); hence, it can be standardized into a widely consistent workflow [17]. AI technology uses spatial, temporal, and anomalous features of systems for classification, regression, and clustering. In addition, it has been used extensively owing to its efficiency in producing useful information and finding hidden variables in various datasets [18]. Machine learning (ML) models have become increasingly prevalent in the simulation of the spatial distribution of soil properties [19–21]. The ML-based subsurface or soil texture mapping includes the use of artificial neural networks, boosted regression trees, random forests, and artificial neuro-fuzzy inference systems [22,23]. For early disaster response to floods, landslides, and earthquake risks, a combination of the ML workflow and geospatial information such as DEM [24], synthetic aperture radar (SAR) imagery [25], and area intensity [26] was proposed. Moreover, following the incorporation of stochastic methods and geological hazard assessment with nonlinear data handling by using varying scales/types of sources, the application of data mining and ML algorithms increased [27,28]. The best-fitting model using only surficial proxy presents a spatial distribution of V_{S30} based on the borehole measurements.

In this study, the V_{S30} of major cities in North Korea (Pyongyang, Kaesong, and Nampo) were mapped. In the training areas in South Korea, the borehole-based V_{S30} values and DTM-based terrain characteristics were identified as labels and features, respectively; these parameters were constructed as geospatial grid information for input into the regression models. The elevation, slope, and topographic position index (TPI)-based landform classes were extracted using the DTM. The V_{S30} grids from the borehole dataset were interpolated using geostatistical methods with the lowest residuals. The best-fitting model for V_{S30} regression, with the grids incorporating V_{S30} and terrain-based features, visualized the V_{S30} map of the test areas in North Korea. The novelty of the proposed

methodology for developing V_{S30} is the ML-based multivariate regression that integrates a large number of grids as interpreted from the DTM and borehole datasets.

2. Method and Data

2.1. Study Area

Geographically, the Korean peninsula is located on the eastern borders (33–43° N and 124–132° E) of the Eurasian continent. According to Sun et al. [29], different geologic strata from the Pre-Cambrian period to the Cenozoic era have been distributed throughout the peninsula. Most regions consist of plutonic and metamorphic rocks, mainly comprising Rangnim, Gyeonggi, and Yeongnam massifs [30] as well as a partial sedimentary rock distribution. The Choogaryong rift valley clearly divides geology into two regions in the north and south, which is mostly similar with the Military Demarcation Line in the Korean Peninsula. In South Korea, various geological rock types (e.g., granite and Mesozoic sedimentary rock) are distributed. However, North Korea is mostly formed by Pre-Cambrian metamorphic rocks, Paleozoic sedimentary rocks, and Quaternary volcanic rocks. From the geomorphological perspective, the Korean Peninsula is defined as an old landform due to its continuous erosion. Steep and high mountains are commonly distributed in North Korea, whereas low mountains and alluvial plains are common in South Korea. Nationally, geological and geomorphological differences are observed between South and North Korea; however, local similarities are observed in certain areas at a city level scale as the shallow soil layers above the engineering bedrock mostly influence the effects of the seismic site. Geographical, geomorphological, and geological similarities during the formation of the layers should be determined prior to building datasets.

Therefore, the study area comprised one training area in South Korea and three test areas in North Korea for the development of the best-fitting model for site classification by considering the similarities shown in Figure 1. The training area in South Korea included Seoul city, part of Incheon city, and Gyeonggi Province (Figure 1a). Seoul spans 605.25 km^2 and is approximately 15 km long; it is bisected by the Han River into the northern and southern halves. The city is bordered by eight mountains and comprises more lowlands in the Han River plains and in the west. Incheon, bordering Seoul and Gyeonggi Province to the east, is located in northwestern South Korea. Gyeonggi Province, which encompasses Seoul, covers 10% of the territory (10,171 km^2) in South Korea. Hence, the training area between the east longitude of 126°39′N and 127°0′N and north latitude of 37°27′N and 37°31′12″N is defined by considering the spatial proximity and connectivity of the largest metropolitan area in the northwestern region of South Korea. In the training area, the geology characteristic is largely granite, wherein the bedrock is widely intruded into the Jurassic Period of Mesozoic era [31]. In the northwest direction, granite is present in Seoul.

For V_{S30} mapping in North Korea, sections of Pyongyang, Kaesong, and Nampo were selected as the test areas based on the archived DTM. The minimum distances between the training and test areas are 176.3, 42, and 172.5 km for Pyongyang, Kaesong, and Nampo, respectively. Pyongyang is located in the west-central region of North Korea in a flat plain, which is one of the two largest plains (50 km east of Korea Bay) on the western coast of the Korean Peninsula (Figure 1b) [32]. Kaesong is the southernmost city of North Korea and is located close to the border of South Korea. The city center encompasses the significantly smaller mountainous northeastern region (average elevation of 103 m), and most areas consist of low hills with a height of less than 100 m [33]. Nampo is a seaport and lies on the northern shore of the Taedong River at a distance of 15 km east from the river's mouth [26,34]. The Imjingang Belt lies between the Rangnim and Gyeonggi massifs; therefore, were grouped the training area (Seoul, Incheon Gyeonggi Province) and Kaesong [30].

With respect to geography, the Pyongyang area located on the Taedong River has a plain area that is similar to that of Seoul. The mountainous region in the Kaesong area resembles the southern and northern parts of Seoul. The Nampo area is a seaport and can be regarded as the western coast of Incheon. The training and test areas are located

in the same Paleozoic basins (i.e., Phyongnam basin) and exhibit large distributions of Mesozoic igneous rocks. The shallow soil layer was mostly formed as Quaternary alluvial soil and weathered residual soil by fluvial or weathering processes. Despite the uncertainty in literature-based geological investigations in North Korea owing to the huge knowledge gap, the formation similarities between the shallow subsurface in the training and test areas were identified.

Figure 1. Study areas in the Korean Peninsula: (**a**) training and test areas in Korean Peninsula; (**b**) central-western region involving Seoul and partial Incheon for the training area of the site classification model; (**c**) test area in Pyongyang; (**d**) test area in Kaesong; (**e**) test area in Nampo. The base map has been obtained from the OpenStreetMap project.

2.2. Digital Terrain Model

DTMs in the training area of South Korea were archived from NASA's Shuttle Radar Topography Mission (SRTM, http://www2.jpl.nasa.gov/srtm/instr.htm, accessed date: 3 March 2020) and other local radar mission projects to render surficial terrain information for the target areas. Despite the archiving of the terrain information in the test areas, North Korea is inaccessible due to military security. The global DTM was acquired by the Advanced Land Observing Satellite from the Japan Aerospace Exploration Agency.

To integrate the boreholes and DEM precisely, the transverse Mercator spatial coordinate system in the DEM was simultaneously transformed based on the geographic coordinate system (latitude and longitude). The DTM with a 5 m × 5 m resolution (0.15 arcsec) was available; subsequently, interferometry was used to gather topographic (elevation) data. The horizontal and vertical accuracies of the root mean squared error (RMSE) were 5 m, and global site condition maps for deriving the topographic slope data as a proxy for the site amplification were collected from the SRTM 30 and 3 arcsec [12,13]. Stewart et al. [35] developed geologic and terrain-based site condition maps for California by using a geological map and 1, 3, 9, and 30 arcsec DEMs. Conversely, high-resolution topographic data for developing local site condition maps are essential, particularly for North Korea, where direct land surveys are impossible. The slope was extracted via remotely sensed applications using DTMs based on DTM-derived longitudinal profile analysis [36] implemented in ESRI GIS products. Figure 2 shows the DTM-based elevation and slope distribution maps.
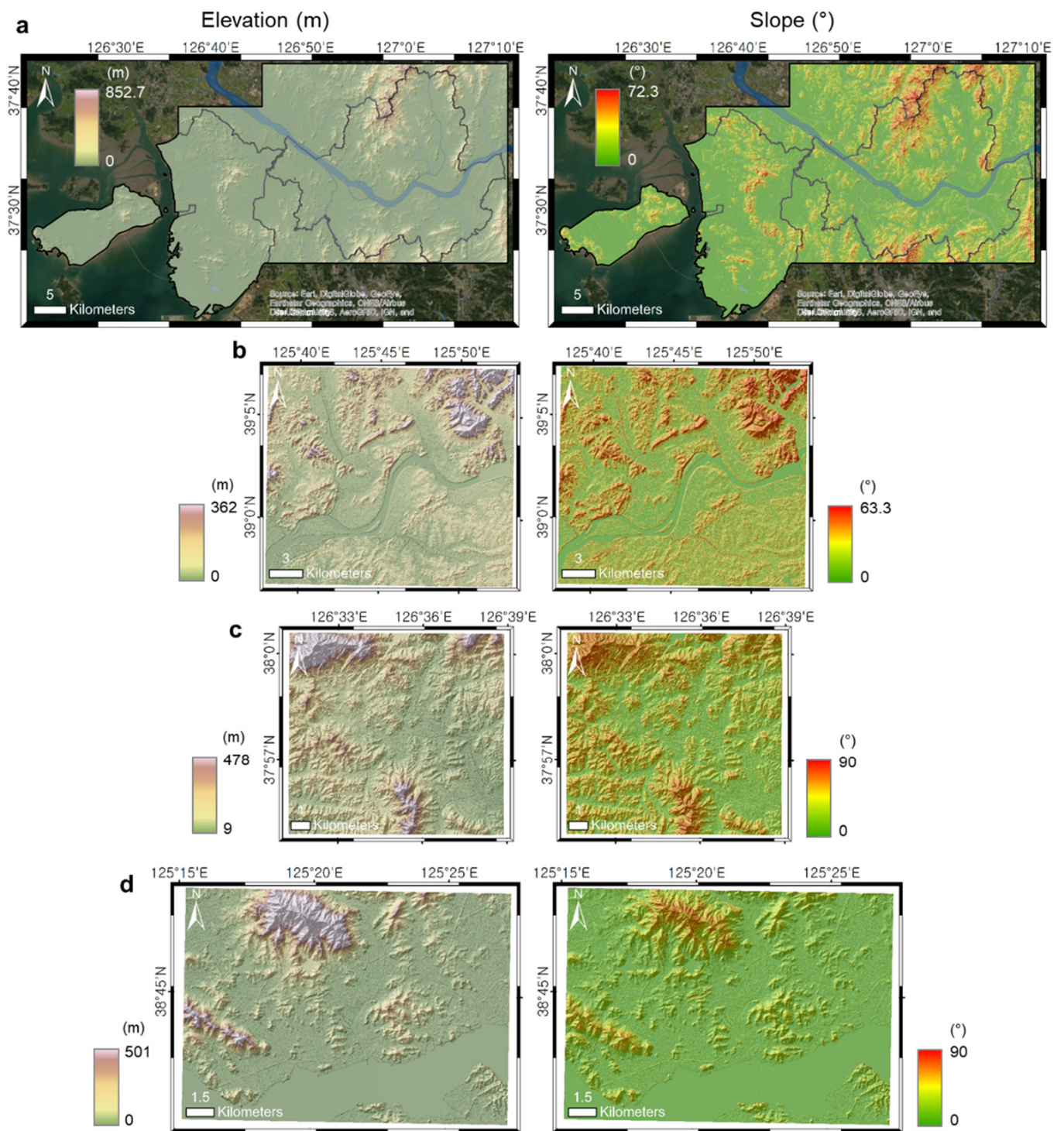
Figure 2. DTM-based elevation and slope: (**a**) training area in Seoul and Incheon; (**b**) test area in Pyongyang; (**c**) test area in Kaesong; and (**d**) test area in Nampo.

The maximum elevation is approximately 852 m in the northern and southern mountainous regions of the training areas, and the average standard deviation is 6.2 m (Figure 2a). The highest slope is approximately 72°. The maximum elevations in Pyongyang, Kaesong, and Nampo (the test areas) are approximately 363, 478, and 501 m, respectively, and their maximum slopes are approximately 63.3, 90, and 90°, respectively. In the Pyongyang area, the elevation and slope along the watershed were approximately 0–25 m and 0–10°, respectively. In the Kaesong area, the surrounding mountains, spatially occupying approximately 65% of the area, have an elevation of 110–478 m and a slope of 30–70°, excluding the central basin. In the Nampo area, most of the plain (70% area) exhibits an elevation of 0 to 100 m and a slope of 0 to 8°, except for certain mountainous areas in the northern region.

2.3. Geotechnical Datasets

In the test area, the borehole log datasets (31,033 points), including the engineering strata and standard penetration test (SPT) results, were collected from the Korea Civil Engineering and Building Technology Institute's Geo-Info system. Additionally, site visits for 350 locations were carried out to gather surface geo-knowledge information, mainly in regions that lacked borehole data. The site survey was conducted at locations where borehole data were not obtained using cone penetration, a global positioning system, and other techniques to validate the rock outcrop or cut slope and cross-check with the engineering strata from the adjacent borehole data [37]. The engineering strata were classified into five groups: fill layer, alluvial soil, weathered soil, weathered rock, and bedrock (soft rock).

In situ geophysical measurement of V_S such as downhole, cross-hole, and spectral analysis of surface waves is the preferred approach for calculating the small-strain shear properties [25]. It was also used in the site classification systems and ground motion equations. When the direct measurement of the V_S profile is not determined, the use of multiple indirect methods (i.e., penetration-based V_S correlations) is recommended; these methods require selecting a V_{S30} design according to the seismic design codes in many countries [38]. The site-specific correlations between the V_S and penetration tests, such as the cone penetration test, SPT, and Becker penetration test, have been obtained previously [39–43]. However, the correlations exhibit certain limitations: (1) low coefficient of determination values, (2) local uncertainty based on low quality data, and (3) disregard of the characteristic deformation behaviors of soils.

In this study, the direct V_S profiles were not stored in the Geo-Info system, and large-scale SPT results were yielded. SPT, detailed in ASTM D 1586 [44], provides penetration resistance related to the blow count (*n* value), which is widely used to determine engineering properties and soil design parameters. For uncertainty analysis with regard to the empirical relation of SPT-*N* versus V_S, the V_{S30} from each borehole location was computed using the representative transformation formulas for soil types (Table 1). If the *n* value was not registered above the engineering bedrock, the representative V_S values were calculated to be 190 m/s for the fill layer, 280 m/s for alluvial soil, 350 m/s for weathered soil, 650 m/s for weathered rock, and 1300 m/s for bedrock [45]. Correlation #1 [46], which was reasonably consistent with correlations from previous studies in other countries, was adopted in the domestic seismic design [4,47] and applied as the preferred approach. Correlation #2 [38] was recommended for N_{60} (normalized to 60% energy ratio), effective stress (σ'_v), and the geological era, and applied in the Pacific Earthquake Engineering Center (PEER), USA. Correlation #3 was derived by using N_{60} that included previous correlations [42]. The performance of ML models was evaluated depending on the *N* versus V_S correlations.

Table 1. N-V_S correlation equations (modified after related studies by Sun et al. [46], Wair et al. [38], and Hasancebi and Ulsay [42]).

Soil Type		Correlation #1 [46] V_S (m/s)	Correlation #2 [38] V_S (m/s)	Age Scaling Factors		Correlation #3 [42] V_S (m/s)
				Holocene	Pleistocene	
All soils		$65.64N^{0.407}$	$30N_{60}^{0.215}\sigma_v'^{0.275}$	0.87	1.13	$104.79N_{60}^{0.26}$
Alluvial soil	Clay & silt	$65.64N^{0.407}$	$26N_{60}^{0.17}\sigma_v'^{0.32}$	0.88	1.12	$107.63N_{60}^{0.237}$
	Gravel—Holocene	$78.63N^{0.361}$	$53N_{60}^{0.19}\sigma_v'^{0.18}$	-	-	$104.79N_{60}^{0.26}$
	Gravel—Pleistocene	$78.63N^{0.361}$	$115N_{60}^{0.17}\sigma_v'^{0.12}$	-	-	$104.79N_{60}^{0.26}$
	Sand & silt	$82.01N^{0.319}$	$30N_{60}^{0.215}\sigma_v'^{0.275}$	0.87	1.13	$104.79N_{60}^{0.26}$
	Sand	$65.64N^{0.407}$	$30N_{60}^{0.23}\sigma_v'^{0.23}$	0.90	1.17	$131N_{60}^{0.205}$
Weathered soil		$75.76N^{0.371}$	$30N_{60}^{0.215}\sigma_v'^{0.275}$	0.87	1.13	$104.79N_{60}^{0.26}$
Weathered rock		$107.94N^{0.418}$	$30N_{60}^{0.215}\sigma_v'^{0.275}$	0.87	1.13	$104.79N_{60}^{0.26}$

2.4. Methodology

2.4.1. Topographic Position Index and Land Formation

A strong correlation is observed between the topographic position and physical and biological processes that influence landscape, such as soil erosion and deposition, hydrological balance, wind exposure, and cold air drainage [48]. The TPI is used to compare the elevation of each cell with the average elevation of the neighborhood surrounding each cell to classify the landforms using the DTM or DEM [48–50]. In the center of the local window, the local mean elevation is subtracted and calculated via algorithm analysis using a geographic information system (GIS) [51]. The algorithm consists of a local neighborhood and a central point, representing the vertical position in length above or below the mean elevation of the neighborhood as follows:

$$TPI = Z_0 - \frac{\sum_{n-1} Z_n}{n} \tag{1}$$

where Z_0 is the elevation of the model point under elevation, and Z_n is the elevation of the cell within the local window.

Neighborhood sizes of 25 and 500 m were used for small and large elements of landform classes [52]. The TPI contributes to the classification of the DEM into 10 landform categories with a grid threshold TPI (25 or 500) (Table 2). TPI values are categorized into pitch positions depending on the pitch and how intense they are at each point. TPI values above a certain level may represent mountain tops or hilltops, while those below the threshold could signify bottoms of the valleys or depressions. TPI values of approximately 0 can be defined as plains (when the path is above 0) or midslope ridges (if the slope is above a certain threshold). For a better categorization of the landform class appropriate for the terrain proxy, the grouped classes are presented in Table 2.

2.4.2. Geospatial Interpolation for Surficial and Subsurface Grid Building

The geotechnical datasets, including the borehole data and results gathered from on-site studies, mainly comprised building lots, roads, and railroads. Grid generation, in conjunction with engineering strata and the SPT-N value based on geostatistical interpolation, is essential for comparing the geotechnical properties with the DEM-based terrain features under the same grid (or resolution). Kriging is a representative interpolation method for smoothing out the inherent distribution of the original dataset. However, in practice, the variance of the estimates is lower and, therefore, on average, high values are underestimated and low values are overestimated. Conditional simulation is an efficient method of resolving such limitations of the kriging estimates.

Table 2. Definition of Topographic Position Index (TPI)-based landform classes and grouped landform classes (modified after Weiss [48]).

Class	Landform	Proposed Grouped Class	Neighborhood TPI	
			Small (TPI$_{25}$)	Large (TPI$_{500}$)
1	Incised streams	LF-I	≤ -1	≤ -1
2	Midslope drainages	LF-II	≤ -1	> -1 and < 1
3	Upland drainages		≤ -1	≥ -1
4	U-Shaped valleys	LF-III	> -1 and < 1	≤ -1
5	Plain (slope $\leq 2°$)	LF-IV	> -1 and < 1	> -1 and < 1
6	Open slopes (slope $> 2°$)	LF-III	> -1 and < 1	> -1 and < 1
7	Upper slopes		> -1 and < 1	≥ -1
8	Local ridges		≤ -1	≤ -1
9	Midslope ridges	LF-II	≤ -1	> -1 and < 1
10	Mountain tops		≤ -1	≥ -1

In this study, we applied conventional geostatistical interpolation approaches: inverse distance weight (IDW), simple kriging (SK), ordinary kriging (OK), universal kriging (UK), empirical Bayesian kriging (EBK), and sequential Gaussian simulation (SGS). Instead of calculating the mean using the kriging algorithm, SGS produces many feasible realizations of a property. As a few attempts to assess the various reliabilities (5th, 50th, 100th) of different outcomes have been made, the final prediction (E-type) was identified by combining all results, that is, the mean simulation. Details on the variogram, kriging, and simulation methods are reported by Chiles and Delfiner [53] and Zuo et al. [54].

The 5 m × 5 m resolution of V_S was established by interpolating the borehole-based V_S value for unit depth (i.e., 1 m) and visualizing along a two-and-a-half-dimensional (2.5D) geospatial grid. The selection procedure for the optimized interpolation method was proposed for borehole data using a cross-validation technique [55,56]. Therefore, the optimal interpolation method with the lowest prediction error (residuals) was applied.

2.4.3. Multivariate Site Response Parameters and Classification System

Seismic zonation based on the multivariate site response parameters was assessed with a spatially assigned grid including elevation, slope, TPI-based landform class, and V_S. V_S profiles can be obtained using downhole, cross-hole, and surface wave methods from different sources for general estimates. Conventional parameters are primarily used to quantify the seismic site effect and define the criteria for site classification. The site classification scheme was applied to V_{S30} in the US seismic code after 1994 as the main categorization parameter, and the average V_S up to 30 m below ground level [57] can be calculated as follows:

$$V_{S30} = 30 / \sum_{i=1}^{n} \frac{D_i}{V_{Si}} \tag{2}$$

where D_i and V_{Si} represent the thickness and the V_S of the i_{th} layer, respectively. The engineering bedrock (H), also known as the analytical brief index, only represents the geometrical features without considering the stiffness of the V_S profile, if only the engineering strata are given. The natural site period results from the sharp contrast between the soil and engineering bedrock and is measured using strata thickness and V_S.

V_S topographical calculations also positively correlate with DEMs, geological maps, and other geomorphic indicators [12,13,58]. The scientific concept of a terrain proxy criterion was established in a metropolitan area in South Korea, including mountains, valleys, plains, and sites with diverse geological features [37,45,59]. Empirical correlations with site response parameters were suggested in South Korea with respect to site-specific geospatial conditions to regional zoning of seismic site effects (Table 3). As V_{S30} decreases, the stiffness of the strata weakens, and the elevation and slope decrease.

Table 3. Multivariate site classification system for seismic site effects (modified after Kim et al. [60]).

Generic Description	Site Class		Geotechnical Criteria						Geo-Proxy-Based Criteria	
			H (m)	V_{S30} (m/s)	T_G (s)	f_0 (Hz)	Slope (%)	Elevation (m)	Lithology	
									Geological Era	Stratigraphy
Rock	B		<6	>760	<0.06	>16.67	>5.6	>80	Paleozoic	Plutonic/metamorphic rocks
Weathered Rock and		C1	<10	>620	<0.10	>10.00	>3.5	>60		Cretaceous fine-grained sediments
Very Stiff Soil	C	C2	<14	>520	<0.14	>7.14	>2.0	>45	Mesozoic	Sheared/weathered crystalline rocks
		C3	<20	>440	<0.20	>5.00	>1.1	>31		Oligocene–Cretaceous sedimentary rocks
Intermediate Stiff Soil		C4	<29	>360	<0.29	>3.45	>0.62	>22		Oligocene coarse-grained younger material
		D1	<38	>320	<0.38	>2.63	>0.23	>18		Miocene fine-grained sediments
		D2	<46	>280	<0.46	>2.17	>0.08	>14	Cenozoic	Coarse younger alluvium
Deep Stiff Soil	D	D3	<54	>240	<0.54	>1.85	>0.023	>11		Holocene alluvium
		D4	<62	>180	<0.62	>1.61	>0.006	>9		Fine-grained alluvial/estuarine deposits
Deep Soft Soil	E		≥62	≤180	≥0.62	≤1.61	≤0.006	≤9	-	Inter-tidal mud

2.4.4. Multivariate Regression

The best-fitting regression algorithms were compared with approaches focused on cost functions to predict the V_{S30} value in this analysis. The V_{S30} value as the label attribute and other original features are included in the grid properties: elevation, slope, and landform class. The elevation and slope are subgrouped with the landform class for predicting the V_{S30} and validation of the best-fitting models. It is imperative to preprocess the data before inputting them into regression models. There are five basic stages of data preprocessing: (1) handling missing values, (2) scaling, (3) one-hot encoding, (4) feature selection, and (5) splitting the dataset into training and test datasets [21]. In this study, the input attributes of the training area in Seoul and Incheon were defined as the training datasets. After preprocessing, the grouped attributes with landform classes in the training datasets were input into the fitted regression models. V_{S30} was predicted using the test datasets and terrain information in the three test areas in North Korea.

Given the different numbers of attributes inputted into the training datasets, a resample method, such as K-fold cross-validation, was required to determine the best-fitting model. K-fold cross-validation removes the convergence of the K-fold data partition and yields K. Each K-fold can be used for end-of-line testing, and all other folds are used simultaneously for training. In this study, 10 folds are generated for the subsurface grid in the training area. Prior to splitting the grids, shuffles and a robust pseudorandom number generator, denoted as Mersenne Twister [61] in Python, are applied. We designed four representative regression algorithms of AI: logistic regression (LR), neighbors K-nearest (K-NN), support vector regression (SVR), and multilayer perceptron (MLP). The models were trained using each algorithm, while hyperparameters were incorporated and optimized model outputs on the K-fold validation dataset were contrasted. In this study, the top-performing model pipeline was determined based on the Auto-Sklearn [62], which is an open-source library for data transformation and machine learning algorithms based on the Bayesian optimization search procedures [63] in Python.

The classical statistical approach used for binary classification is LR and has been adopted as a simple ML model. The LR model uses a logistic feature to compress the output of a linear equation between 0 and 1, rather than fitting a straight line or hyperplane. As linear regression assumes that the data are linear, LR uses a sigmoid to model the data.

$$g(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

For the K-NN procedure, the predicted values for the neighboring measurements of the variables were obtained as weighted averages. One of the main challenges of this approach is the need to choose an ad hoc similarity metric, particularly for heterogeneous datasets of different types and sizes that exhibit varying interrelationships between the extracted features [64]. The proximity to a certain spot in a high-dimensional space is scarce and leads to significant variety. K-NN regression is a nonparametric method that intuitively approximates the relations between the independent and continuous effects by integrating observations in the same neighborhood. K-NN regression uses distance functions and an important and less concerned approach for producing K-NN rankings (e.g., Euclidean distance, Manhattan distance, and Minkowski distance). Manhattan distance (denoted as block distance or taxicab geometry), which is the distance (d) between two points (x_i and y_i) at right angles to the axes, has been calculated as follows:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{4}$$

SVR is a linear or nonlinear regression method based on the concept of support vector machines (SVMs). When formulating SVMs as convex optimization problems [65–67], SVM resolves binary classification problems. The SVR methodology uses linear quadratic programming methods to handle data in high-dimensional space [68]. We also sought to

reduce the LR error rate. During SVR, we attempted to match the error under any threshold. The former condition generates the objective function in equation ($\text{Min}\frac{1}{2}\|w\|^2$) where the size of the normal vector to the surface is approximated.

MLP is a static neural structure consisting of layers that transmit and transfer information through synaptic links represented by weight adjustment. The MLP has also been used as a regression approximation tool. A conversion function is used to transfer the weighted number of inputs and the terms of bias to the activating step [69,70]. A feed-forward neural network is an artificial and non-cyclic neural network. Every perception is completely linked to all the nodes in a single layer. The input, hidden, and output layers are usually composed of MLP modeling. Hyperparameter tuning was performed for the best-fitting model of the MLP classifier by adjusting numerous parameters such as the number of hidden layers, number of nodes of the hidden layer, activation function, optimizer, number of epochs, batch size, and learning rate. In particular, the process of finding the optimal network was applied, namely grid search; meanwhile, various cases in the hyperparameter were altered. In this analysis, two hidden layers in 80 nodes were linked to the link layers. The input values were weighted and generated according to the activation function of the layer above [71]. As an activation function, rectified linear units (ReLU) were used for the hidden layers and the Adam optimizer was set up. If the input was less than 0 and the raw output of ReLU was different, the output was considered to be 0. The random state and solver are eight, and the stochastic gradient descent, respectively [70].

3. Results

3.1. Topographic Position Index-Based Classification of Landforms

The TPI was computed, and its grid-based landform class was determined based on the relative height difference and slope inclination (Figure 3). The zonal proportion based on the occupied grid was determined against 10 TPI classes for each training and test area (Figure 4). TPI-based landform maps support the site-specific characterization of topographic site effects on seismic waves to induce important gradients during the ground acceleration in grids while elucidating their relationship with V_{S30}.

In the training area, the plain (class 5) that accounted for approximately 33% of the total area was mainly distributed along the riverside and reclaimed land. The widest landform was formed by the U-shaped valleys (class 4) that accounted for approximately 35% of the test area; these valleys were mainly observed on the northern hillside and lower slopes. The upper slopes (class 7) and open slopes (class 6) were regionalized in the northern and southern mountainous areas. The U-shaped valleys and upper slopes accounted for approximately 32% and 24%, respectively, of the test area in Pyongyang. The partial river basin (plain, class 5) occupying 17% of the area was classified as a flat relief or lowland along the Taedong River. The upper slope (class 6), which accounted for approximately 25% of the test area in Pyongyang, was also distributed and blended with the plain area of an urban district. The test area of Kaesong exhibited a major landform that was classified as U-shaped valleys (class 4), which accounted for approximately 35% of the total test area; these valleys comprised most of the urban and industrial zones; a plain area that accounted for only 7% of the test area in Pyongyang was also identified. For the test area in Nampo, the U-shaped valleys (class 4) accounted for approximately 28% of the northern mountainous region. The plain (class 5), excluding the sea area, accounted for approximately 28% of the test area and was distributed in the blended area with upper slopes (class 7) that accounted for 18% of the test area.

The test area in Pyongyang was relatively similar to the training area when compared to the U-shaped valleys and plain area. For the U-shaped valleys, the test area in Kaesong had a terrain characteristic resembling that of the northern mountainous area in the training area. Kaesong exhibited more terrain features (classes 1, 2, 3, 8, 9, and 10) of highland regions compared to the others. The test area in Nampo was similar to the northwestern area in the test area with regard to only the plain area.
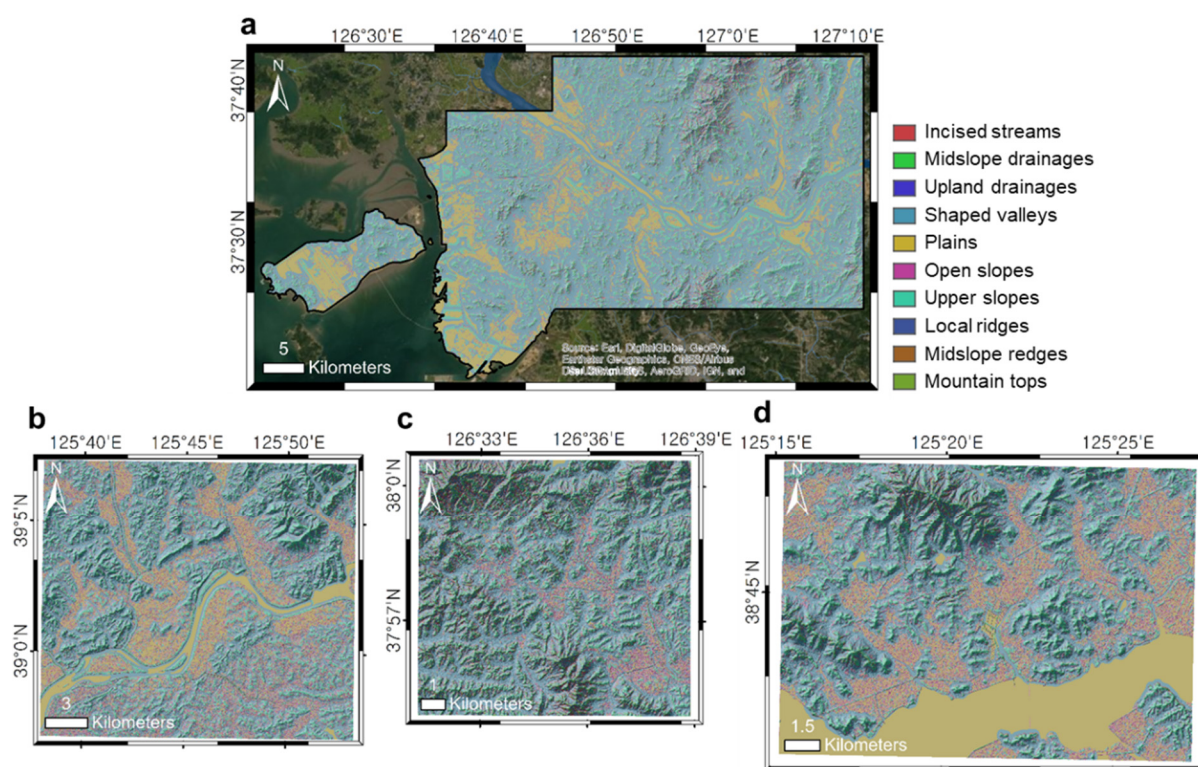
Figure 3. Topographic position index (TPI)-based classification of landforms: (**a**) the training area in Seoul and Incheon; (**b**) test area in Pyongyang; (**c**) test area in Kaesong; (**d**) test area in Nampo.
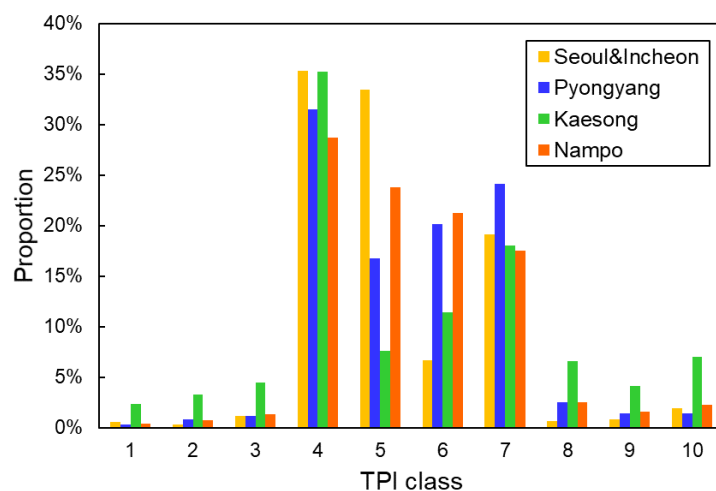

Figure 4. Histogram of the TPI classes in the training and test areas.

3.2. Geospatial Grids Assigned with V_{S30} and in Test Area

Geospatial grids (5 m × 5 m resolution) assigned with V_{S30} (Figure 5b–j) were developed using the borehole datasets (Figure 5a), based on the proposed geostatistical method (IDW, SK, OK, UK, EBK, and SGS). V_S was transformed from the SPT-N value using three correlations (Table 1), and N-V_S correlation #1 was used for selecting the optimized interpolation method. The predicted V_{S30} was generally high when focusing on the outside mountainous region of the administrative divisions of Seoul. A deep stiff soil with a V_{S30} value lower than 360 m/s was found to be distributed along the Han River and the southern part of Incheon, including the Incheon International Airport. On the V_{S30} maps in the test area, there were minor spatial variations according to the interpolation methods.
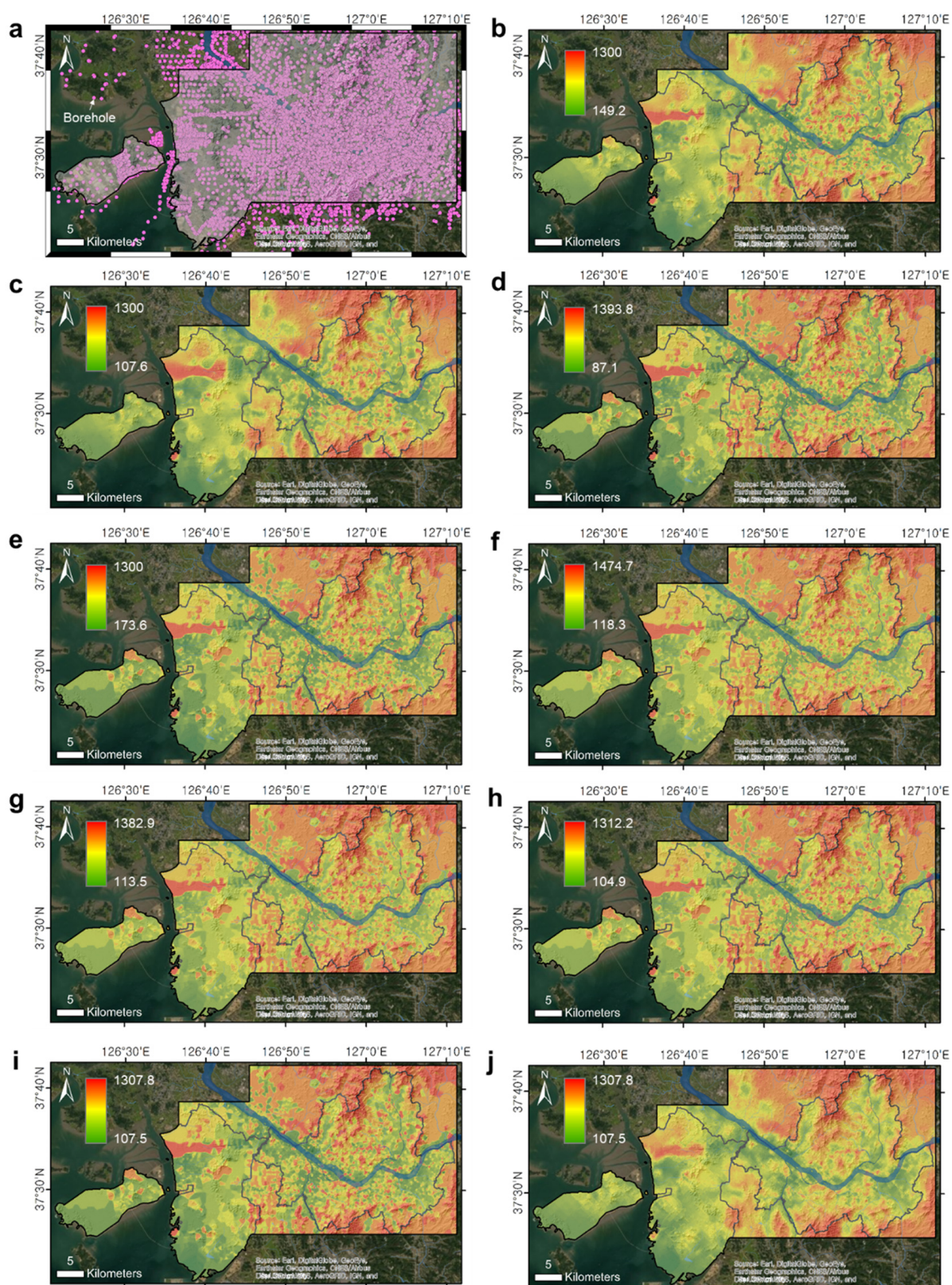
Figure 5. V_{S30} mapping using (**a**) borehole (SPT-*N*) datasets in the training area based on the interpolation algorithms: (**b**) inverse distance weight; (**c**) simple kriging; (**d**) ordinary kriging; (**e**) universal kriging; (**f**) empirical Bayesian kriging; (**g**) sequential Gaussian simulation—5th realization; (**h**) sequential Gaussian simulation—50th realization; (**i**) sequential Gaussian simulation—100th realization; (**j**) sequential Gaussian simulation—E-type. V_{S30} was calculated using V_S that was transformed from the SPT-*N* value by using correlation #1 (Table 1). The unit of V_{S30} is m/s.

The proposed optimization procedure was also conducted by considering the cost function of cross-validation. The residuals and scatter of the plots between the measured and predicted V_{S30} values were analyzed (Figure 6). When the actual V_{S30} values were below 760 m/s, the threshold of the engineering bedrock depth [72] implied that V_{S30} values were overestimated. When the actual V_{S30} values exceeded 760 m/s, V_{S30} was underestimated. As the 70% boring log was mostly investigated prior to the engineering bedrock, soil strata were largely scattered, leading to over-fitting and poor validation due to the smoothing effects of kriging [73]. Although most outcrop bedrocks in mountainous regions were studied and their V_S values were assumed to be 1300 m/s as per the site visiting survey, the measured V_S for bedrock was lower than 1300 m/s. Moreover, the high stochastic/spatial variation of V_S for bedrock in local areas induced low spatial correlation coefficients and under-fitting of the V_{S30} predictions. The residuals of UK are larger than those of the other methods because UK was developed precisely to recognize and quantify the linear spatial trends in datasets [74].
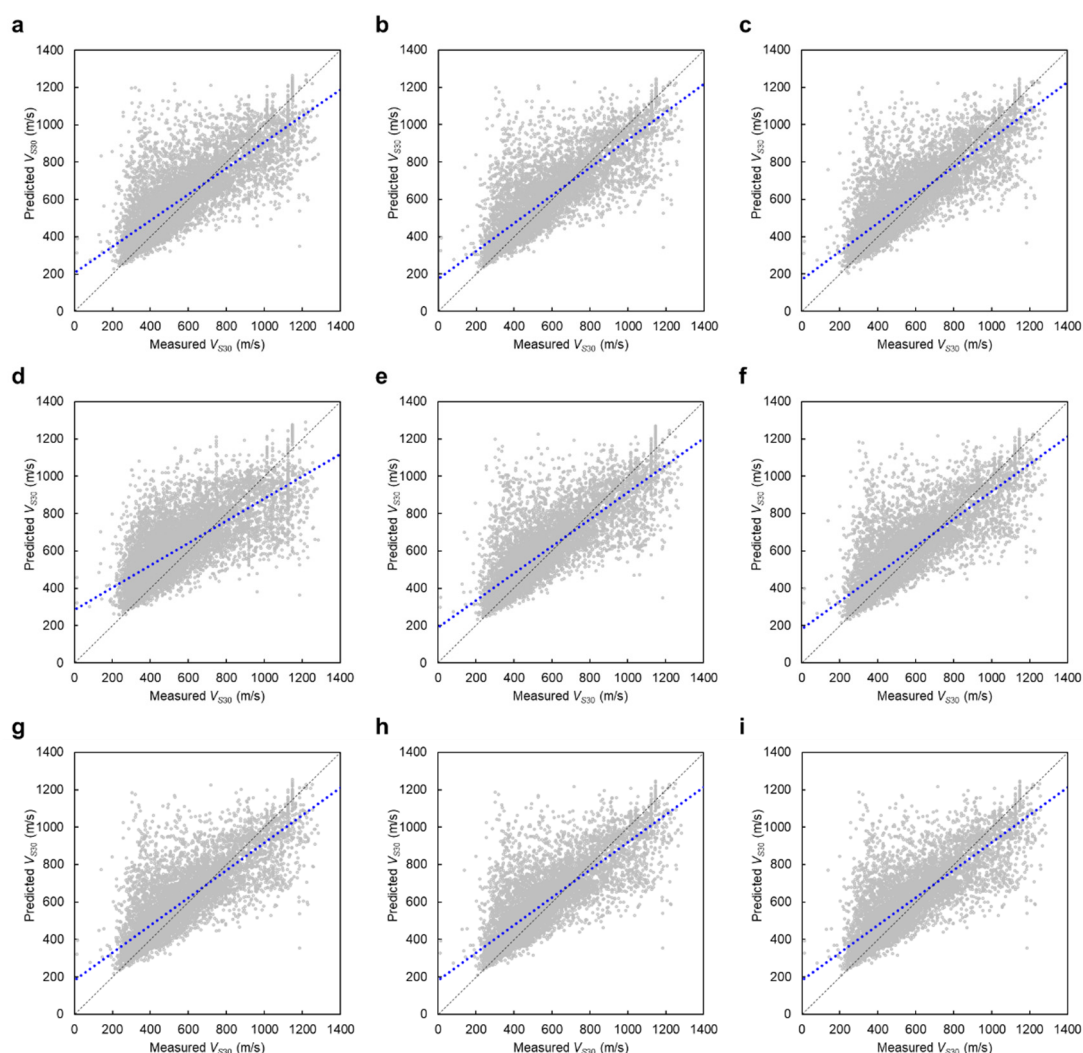

Figure 6. Cross-validation results on borehole location. The measured versus predicted V_{S30} for: (**a**) inverse distance weight; (**b**) simple kriging; (**c**) ordinary kriging; (**d**) universal kriging; (**e**) empirical Bayesian kriging; (**f**) sequential Gaussian simulation—5th realization; (**g**) sequential Gaussian simulation—50th realization; (**h**) sequential Gaussian simulation—100th realization; (**i**) sequential Gaussian simulation—E-type. The black dotted line represents the 1:1 ratio. The blue-dotted line indicates the linear regression line.

To better compare the residuals in regression analysis, this study evaluated four error indices: mean absolute error (MAE), RMSE, root relative squared error (RRSE),

and coefficient of determination (R^2) (Table 4). The metrics of the indices are calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$ (5)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$ (6)

$$RRSE = \sqrt{\frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2}}$$ (7)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2}$$ (8)

where $\hat{y}$ is the predicted value of y, and $\overline{y}$ is the mean value of y. The indices indicate the difference between the measured and predicted V_{S30}; the higher the accuracy, the closer the difference value is to 0. R^2 shows how closely the values complement the original value; the higher the value, the better the model. The SGS-E-type method showed the optimal prediction results when comparing the RMSE and RRSE values; these were the lowest for the SGS-E-type method. The MAE for the SK method had the lowest values. The R^2 values of the SK, OK, and SGS-E-types were closer to 1, indicating a higher precision. When the error distribution is assumed to be Gaussian, RMSE and RRSE are more suitable for reflecting the model results than MAE. In particular, RMSE is more useful if significant errors are avoided, and it is considered that RRSE reduces the error to the same dimensions as the quantity being predicted. Accordingly, the SGS-E-type as the best-fitting method was used to construct the geospatial grid assigned with V_{S30} in the test area. The proportions of grids, which were built by interpolation methods and allocated for the site class (Table 3), have been summarized in Table 4. In general, a similar distribution for proportion is presented; the ratio of the B, C, D, and E classes is 42%, 31%, 13.3%, and 13.7% in the training area, respectively.

Table 4. Performance and statistical summary of geospatial interpolation model in regression analysis.

Methods	Regression Analysis				Proportion of Site Class (%)			
	MAE (m/s)	RMSE (m/s)	RRSE (%)	R^2	B	C	D	E
IDW	98.59	146.17	60.00	0.64	41.5	30.8	13.9	13.9
SK	84.59	134.30	55.68	0.69	41.6	30.7	13.8	13.8
OK	85.58	133.88	55.68	0.69	45.5	28.7	12.9	12.9
UK	125.69	168.35	67.82	0.54	41.5	30.8	13.9	13.8
EBK	92.21	140.20	58.31	0.66	48.5	27.1	12.2	12.2
SGS-5th	89.05	137.65	57.45	0.67	45.1	28.9	13.0	13.0
SGS-50th	88.66	137.48	57.45	0.67	42.1	30.5	13.7	13.7
SGS-100th	88.38	135.39	56.57	0.68	41.9	30.6	13.8	13.8
SGS-E-type	88.38	133.39	55.68	0.69	41.9	30.6	13.8	13.8

3.3. Adaptation of Terrain Proxy-Based Site Class

To investigate the interrelationship between V_{S30} and landform, the combination of landform classes, which are clearly defined as the principal components for the regression model, have been examined using the box and whisker plots. The grids assigned with V_{S30} in the training area were built using SGS and grouped by the TPI-based landform class, which has been denoted as LF (Figure 7a). Class 1 (LF1) had the broadest range (400–1100 m/s between the first and third quartiles) of V_{S30} as well as the highest V_{S30} (approximately 1300 m/s). LF2, LF3, LF8, LF9, and LF10 represented the drainages or ridges in the mountainous area and exhibited a similar range (approximately 400–800 m/s

between the first and third quartiles) of V_{S30}. LF4, LF6, and LF7, which were classified as valleys and slopes, ranged from approximately 400 to 600 m/s between the first and third quartiles of V_{S30}. LF5 (plain area) was verified as having the lowest range (approximately 300–420 m/s between the first and third quartiles) of V_{S30}. Accordingly, the LFs were grouped into four classes (LF-I, -II, -III, and -IV) with similar ranges of V_{S30}. (Figure 7b, Table 2). LF-I represented deeply incised streams characterized as metamorphic rocks [75] and exhibited the highest V_{S30}. LF-II was defined to include drainage and ridge sites (TPI$_{25}$ ≤ −1). LF-III represented valleys and slopes, with TPI$_{25}$ > −1 and < 1, and a slope > 2°. LF-IV denoted a plain where the mean V_{S30} was 400 m/s.
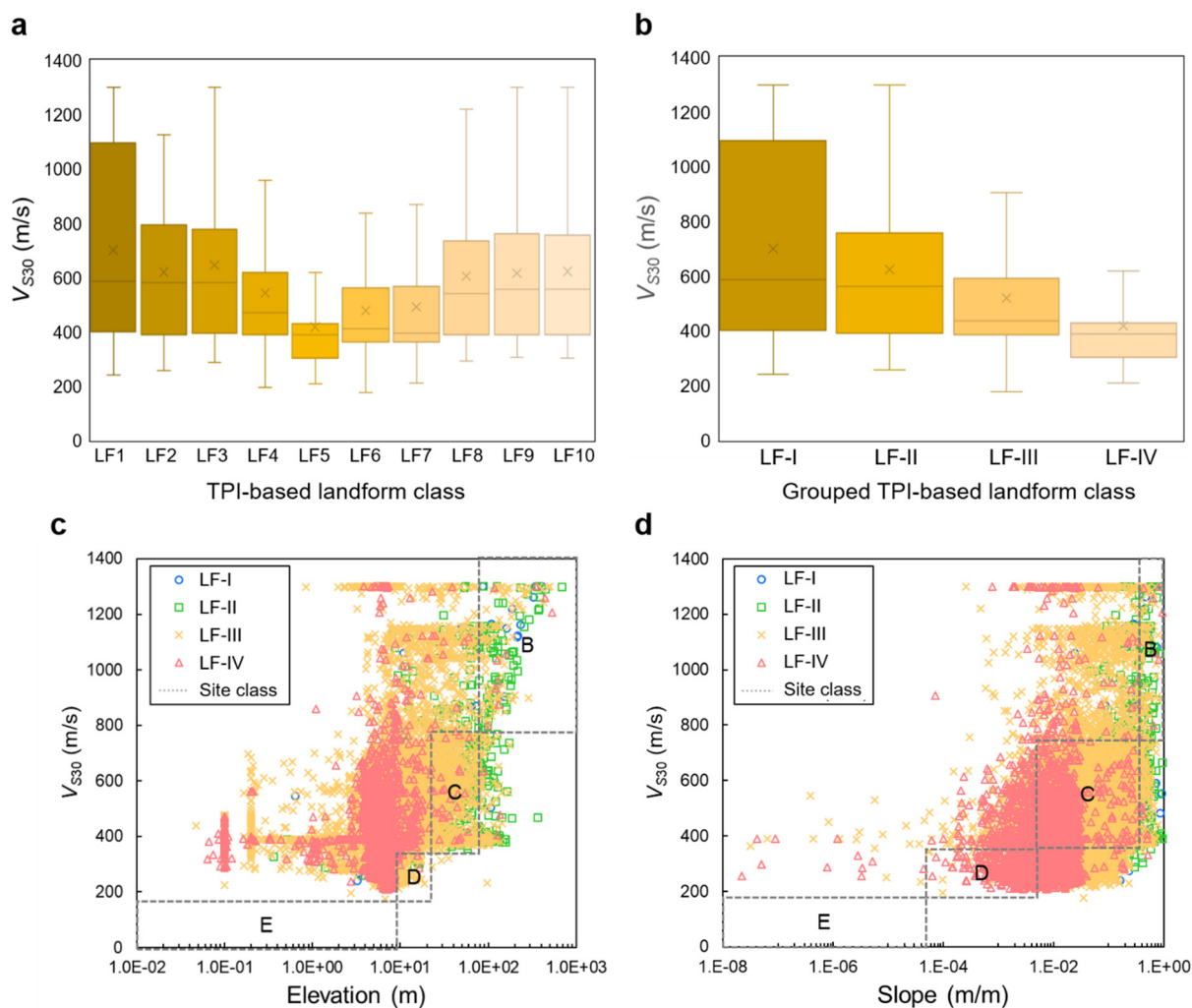
Figure 7. Relations of SGS-E-type based V_{S30} and terrain proxy values classified by the TPI-based landform class in the training area: (**a**) relations between TPI-based landform class and V_{S30}; (**b**) relations between grouped TPI-based landform class and V_{S30}; (**c**) relations between elevation and V_{S30} classified by the grouped TPI-based landform class; (**d**) relations between slope and V_{S30} classified by the grouped TPI-based landform class. The criteria of terrain proxy-based site classification are also presented.

The coinciding of grids of different sites classified by V_{S30} (Table 3) was evaluated along with the four groups of the LFs by comparing the elevation and slope as a proxy of V_{S30}. The minimum first quartile of the V_{S30} value for each class was mostly 400 m/s because the grids with V_{S30} values of 400 m/s had a wide range of surface elevations and slopes (Figure 7c,d). No grid was determined as a class E. The grid area associated with the elevation and slope of LF-IV accounted for approximately 32% and 89% of the total area for the D class, respectively. Among class C, the grid area associated with the elevation and

slope of LF-III accounted for approximately 83% and 65% of the total area, respectively. The grid area associated with the elevation and slope of LF-II accounted for approximately 68% and 59% of the total area for B class, respectively. The grids classified as LF-I accounted for only 2% of the entire test site, although they exhibited scattered relations with V_{S30}. The grid area associated with the elevation and slope of LF-I accounted for 95% and 89% of the total area for B class, respectively. Therefore, LF-II (including LF-I), LF-III, and LF-IV were strongly related to the site classes B, C, and D, respectively. The statistical characteristics of terrain proxies (elevation and slope) and V_{S30} values are summarized using the grouped TPI-based landform classes (Table 5). From LF-I to LF-IV, V_{S30} increases; otherwise, the elevation and slope generally increase.

Table 5. Statistical summary of terrain proxies (elevation and slope) and V_{S30} values as per the grouped TPI-based landform class.

Landform Class	Grid-Statistics	V_{S30} (m/s)	Elevation (m)	Slope (m/m)
LF-I	Min.	240.72	0.00	0.02
	Mean	700.56	700.56	0.27
	Max.	1299.98	1299.98	0.92
	Std.	320.70	320.70	0.20
	Count		980	
LF-II	Min.	257.76	0.00	0.00
	Mean	623.83	59.88	0.30
	Max.	1299.99	695.24	1.55
	Std.	258.73	63.96	0.19
	Count		8347	
LF-III	Min.	175.13	0.00	0.00
	Mean	519.61	15.73	0.07
	Max.	1299.99	504.24	1.47
	Std.	227.23	24.69	0.11
	Count		103,704	
LF-IV	Min.	208.00	0.00	0.00
	Mean	417.05	5.73	0.01
	Max.	1299.99	533.28	0.96
	Std.	195.54	13.90	0.04
	Count		57,970	

3.4. Multivariate Regression Model for Terrain Proxy-Based V_{S30} Classification

Using the grids (171,000 cells) in the training area constructed by SGS, the elevation and slope were subgrouped using the proposed four LFs and defined as input feature datasets for the training procedure in the proposed four regression models. The labeled data were V_{S30}. Using the four modeled regression algorithms for each LF group, the four input combinations (elevation and slope) were trained and verified using K (i.e., 10-fold cross-validation). The averaged residuals between the predicted and actual V_{S30} values (Figure 8) were calculated for each cross-validation using the metrics of the indices: MAE, RMSE, RRSE, R^2. The validation data for the model were then considered to be a single subsample (i.e., 17,100 cells), and the remaining K-1 samplings (i.e., 153,900 cells) were used as the training data. The prediction performance of N-V_S correlations (Table 3) for deriving the V_{S30} grid were also compared with the cost functions.
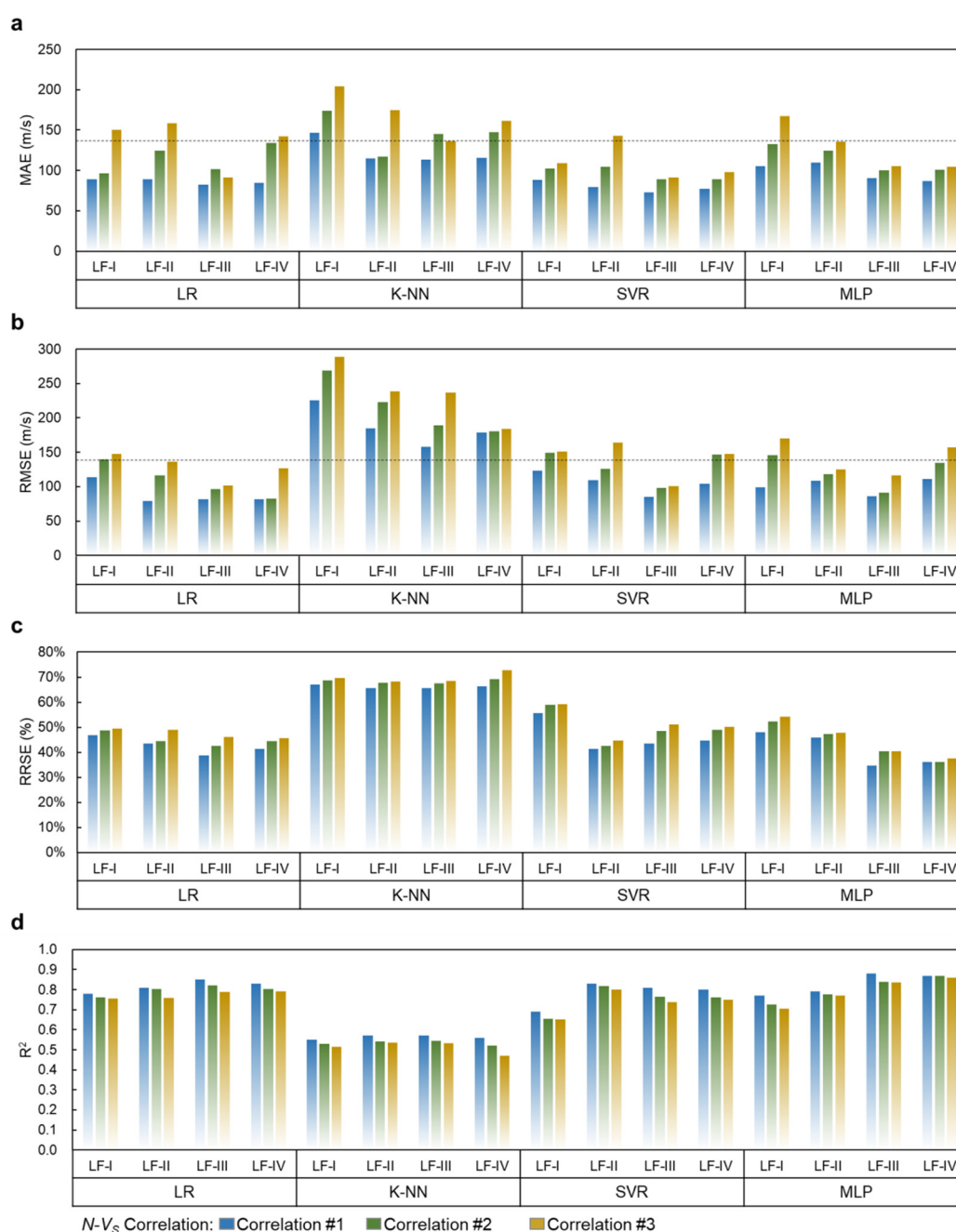
Figure 8. Bar charts of four cost function-based prediction performance of the predicted V_{S30} by the best-fitting model within four regression methods and three N-V_S correlations. The metrics of the indices are (**a**) MAE, (**b**) RMSE, (**c**) RRSE, and (**d**) R^2. The black dotted line in MAE and RMSE indicates the minimum deviation (140 m/s) of V_{S30} threshold in the site classification system (Table 3).

The best-fitting model for each LF group was determined using the cost functions. The MAE for SVR had the lowest value and was considered to be a more optimized method. When RMSE and RRSE exhibited their lowest values, LR and MLP showed optimal prediction performance. In particular, MAE and RMSE were compared with the minimum deviation (140 m/s) of V_{S30} threshold in the site classification system. The indices were lower than 140 m/s except for K-NN, indicating the low possibility of encountering errors during site classification when using the regression algorithms. The R^2 values of LR and MLP were also closer to 1, indicating relatively high precision. In particular, K-NN was the least reliable in predicting V_{S30}. The optimal predictable LF was determined

by comparing the cost functions. Similar to the optimization of geospatial interpolation models, the LR presented performed optimally in the prediction of V_{S30} because using the RMSE is more appropriate for realizing a Gaussian error distribution. To select the eclectic regression model that almost completely satisfies the cost functions, the priorities of the optimal performing regression algorithm are as follows: LR, MLP, SVR, and K-NN.

The predicted grids classified as LF-III and comprising maximum data (103,704 cells) yielded the lowest metrics of residual indices for all the models. The input feature of the landform class that represents valleys and slopes may be consistent with the slope and elevation. Otherwise, LP-I, which is characterized by its deeply incised streams and identified as class B accruing to site classification, shows the highest level of uncertainty in predicting the V_{S30} in the training area. Characterizing the representative V_{S30} in bedrock is complicated when using the sparse distribution of investigated rock outcrops in the training area. After comparing the N-V_S correlations in the cost functions, correlation #1 was selected and used for the transformation of N to V_S to calculate the V_{S30} as it yielded the lowest MAE, RMSE, and RRSE values and the highest R^2. Nevertheless, the reasons these values when compared to other stochastic relationships are as follows: (1) low number of geotechnical investigations were performed in mountain and rock outcrop areas; (2) uncertainty in the geospatial interpolation of sparsely distributed and low correlated features; and (3) possible underfitting in the auto fitting procedures of hyperparameters.

In this study, the predicted residuals were elucidated and the difference between the predicted V_{S30} values with the measured V_{S30} values (Figure 9) was evaluated based on the N-V_S correlation #1. The best-fitting regression model out of the four algorithms was applied to the training areas to compare the spatial variability in the predicted V_{S30} map (Figure 10). The scatter patterns in LR, SVR, and MLP were very similar, whereas the residual application of the K-NN was significantly scattered. For all grouped LFs, the V_{S30} value was under approximately 400 m/s; at this value, the site classes D or E were overestimated. The grids assigned with low V_{S30} were mainly located at LF-IV, where elevation and slope were the smallest (5.73 m and 0.01°, respectively) (Table 5). However, these grids in LF-I or LF-II may be a source of noise parameters in the regression models. Conversely, when the V_{S30} value exceeded approximately 400 m/s, site classes B and C were underestimated.

The grids allocated in the high V_{S30} map (Figure 5) mostly exhibited top or ridge characteristics: LF-I or LF-II, high elevation and slope, and outcrop bedrock or stiff soil. The grid with a high V_{S30} in the plain would be a source of noise parameters. When the V_{S30} value was either higher or lower than 400 m/s, the number of residuals increased. V_{S30} was mapped in the training area (Figure 10) and test areas (Figure 11) based on the best-fitting model of the four regression methods. The maps from LR and SVR presented similar spatial trends, particularly in the northern and southern mountainous regions where V_{S30} exceeded 800 m/s. The grids were defined as class B using LR, K-NN, SVR, and MLP, which occupied 32%, 39%, 31%, 38% of the training area, respectively. The grid areas assigned with class C for the application of LR, K-NN, SVR, and MLP occupied 19%, 25%, 20%, and 30% of the training area, respectively. Class D was distributed at 49%, 46%, 49%, and 42%, respectively, in the four V_{S30} maps. The MLP-based V_{S30} map had a similar proportion of site class with the V_{S30} grid, according to the geospatial interpolation models.
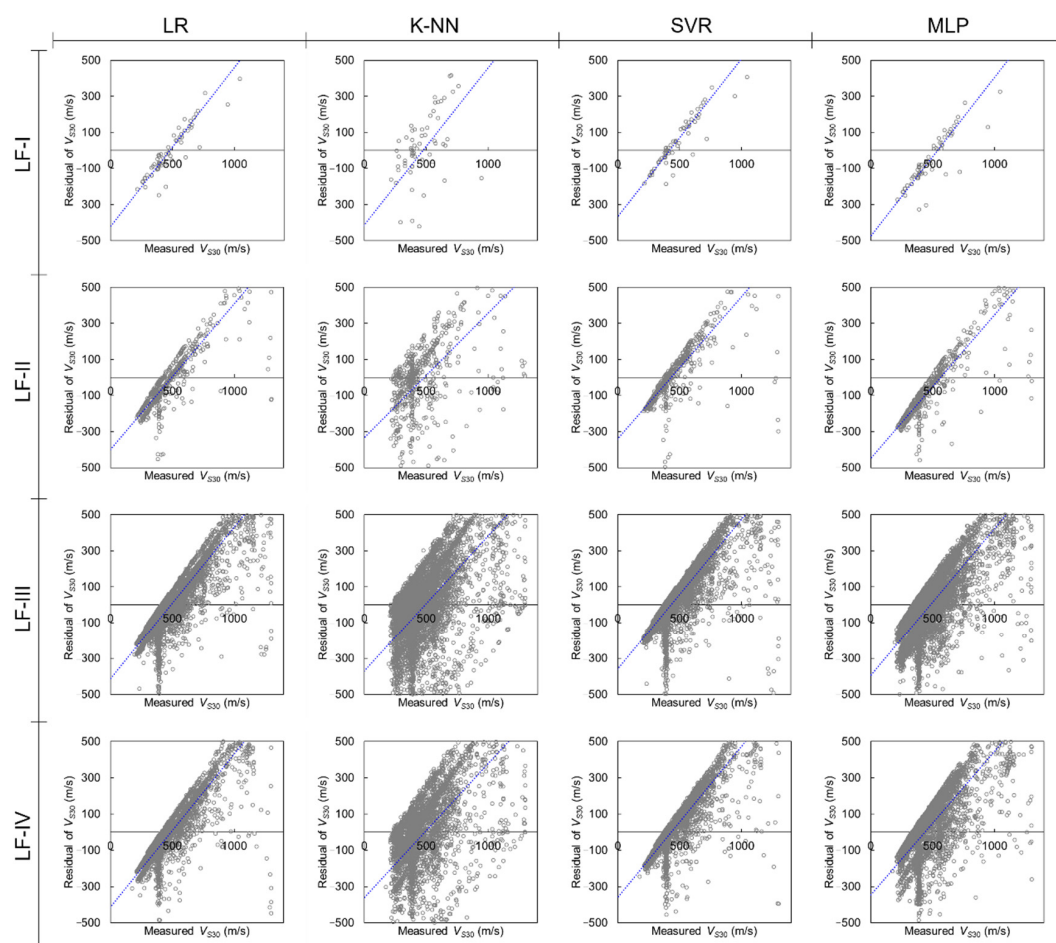
Figure 9. Correlations between estimated V_{S30} and prediction residuals using four regression models for the grouped TPI-based landform class based on N-V_S correlation #1. The blue dotted line indicates linear relations between the measured and residual values of V_{S30}.
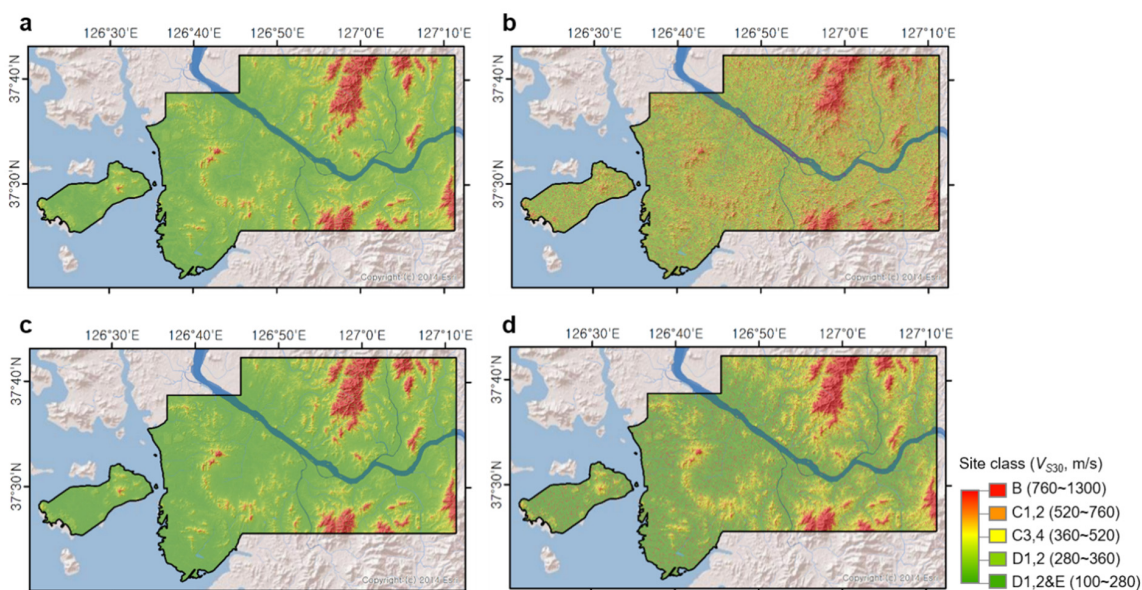

Figure 10. V_{S30} mapping for the training area in Seoul and Incheon using regression models: (**a**) logistic regression; (**b**) K-nearest neighbors; (**c**) support vector regression; and (**d**) multilayer perceptron.
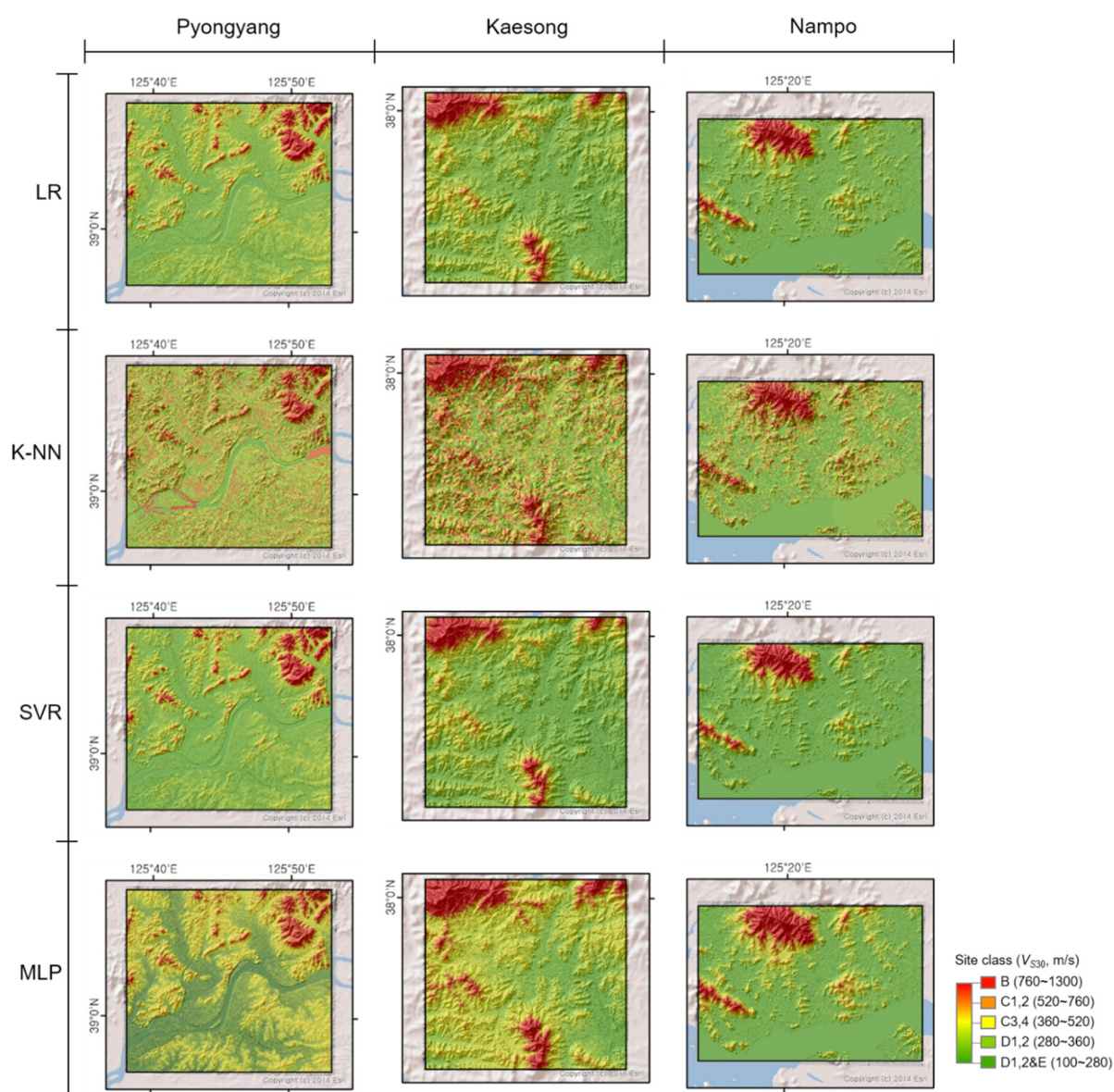
Figure 11. V_{S30} mapping for the test areas in Pyongyang, Kaesong, and Nampo, using regression models: logistic regression; K-nearest neighbors; support vector regression; multilayer perceptron.

3.5. V_{S30} Mapping in the Test Area, North Korea

Based on the best-fitting regression model, the V_{S30} values in four test areas in North Korea were mapped using elevation, slope, and the grouped LF from the DTM (Figure 11). In Pyongyang, the results of LF and K-NN were similar to those of the training area. K-NN predicted the discontinuous zone, which was classified as class B, along the Taedong River. In the MLP-based V_{S30} map, the C3 and C4 classes, representing the intermediate stiff soil, were distributed on the basin-edge area and occupied a wider area (28%) than those in the LR, K-NN, and SVR-based maps (10%, 19%, and 9%, respectively). The D and E classes were distributed in the basin and embankment areas, except in the K-NN-based map.

The LF-I zone in the test area, Kaesong, was mainly classified as class B. The LF-II zone was classified as class C based on the MLP-based map, whereas other maps mostly identified the zone as class D. As class C denotes the weathered rock or stiff soil and is generally distributed in the drainage and ridge areas (Sun and Kim, 2017), the MLP-based map provided reasonable distribution of the site class. In every map, classes D or E were determined by focusing on LF-IV. The V_{S30} maps of the test area in Nampo were obtained

from the four regression models that generally represented a similar pattern, with LF-I and LF-IV being classified as classes B and D, respectively. Class D was widely distributed, exclusively in the northern and northwestern mountains, where the elevation exceeded 350 m and the slope exceeded 50°.

The LR- and SVR-based maps showed a similar V_{S30} distribution to that observed in the training area. The K-NN-based V_{S30} maps showed discontinuous zoning despite having the same LF. The MLP-based map determined class C for a wider area than the maps based on other models. The normal distribution of V_{S30} obtained from the best-fitting model in the four regression methods was derived for the four test areas (Figure 12). The probability density of the SGS-E-type in the training area was also compared based on the distribution patterns (skewness and kurtosis). The probability density distributions of LF, SVR, and MLP-based V_{S30} maps were more peaked (leptokurtic) than that of the SGS-E-type. If V_{S30} was leptokurtic, most of the V_{S30} grids were distributed mainly in the V_{S30} range of 600–800 m/s as compared to that observed in a mesokurtic or platykurtic distribution (i.e., SGS-E-type and K-NN applications). The kurtosis of V_{S30} developed from the K-NN was similar to that of the SGS-E-type. In the training area, the skewness coefficients of LF and SVR were zero, while K-NN and MLP were typically negative. In the test areas of Pyongyang and Nampo, the distribution patterns were similar to those in the training area. Similar to the resemblance in geography and terrain of the training and test areas, the predicted V_{S30} grid conserved the surficial and subsurface characteristics. In contrast, the negative skewness of V_{S30} for the four models was estimated for the test area in Kaesong. As most areas of Kaesong are low hills, a higher V_{S30} value (approximately 700–900 m/s) than that for the training area was possible.
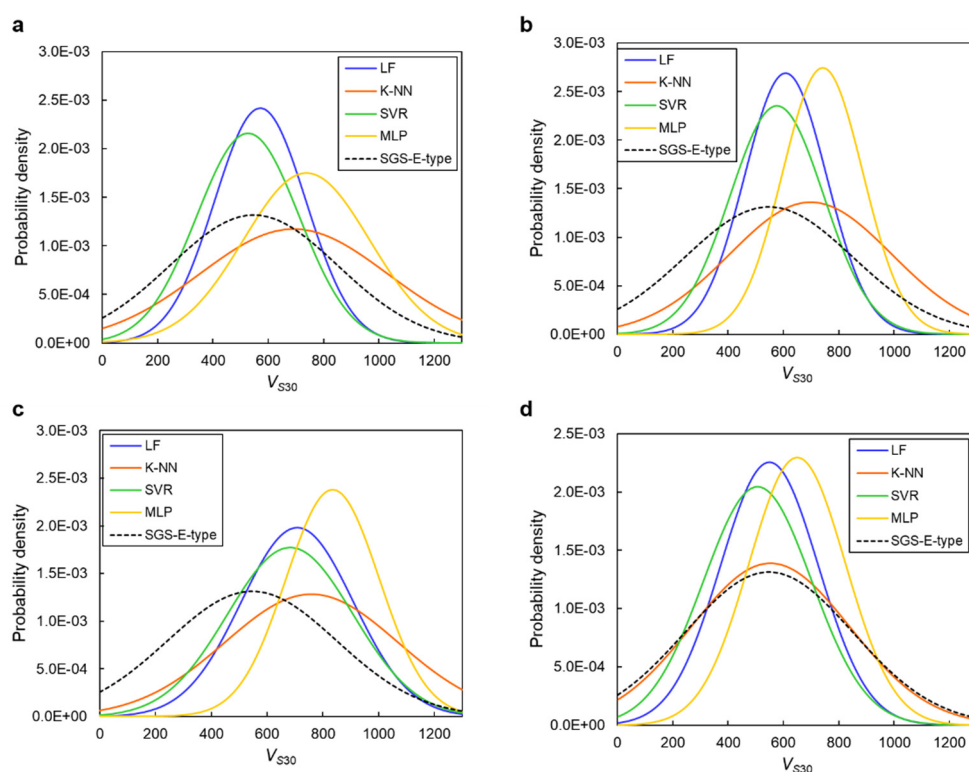

Figure 12. Normal distribution of V_{S30} map from the regression models for the (**a**) training area in Seoul and Incheon; (**b**) test area in Pyongyang; (**c**) test area in Kaesong; and (**d**) test area in Nampo.

4. Discussion

The large-scale V_{S30} map predicted using only the terrain features supports site classification against seismic site effects. In particular, the configuration of subsurface

characteristics was complicated by the given confined terrain information, despite the important global or local spots in the Korean Peninsula. The TPI-based landform characteristics were newly defined as the principal components of the learning model to reclassify the conventional criteria of various terrain proxies. The terrain-based proxy usually uses the slope gradient and elevation, and in combination with LF, adequately predicts V_{S30}. The training area in South Korea was representative of the geography and terrain characteristics of the Korean Peninsula. The subareas (basins, ridges, and valleys) in the training area resembled those in the three test areas in North Korea based on the LF. As the LFs were classified into 10 landform categories with a grid threshold TPI (25 or 500 m), the subgrouping of LFs was appropriate for clearly categorizing the relations between LF and V_{S30}.

The previous methodologies suggested the individual proxy dependence of V_{S30} related to geomorphological, terrain, and geological properties through a stochastic approach, such as a lognormal linear regression model [11–14,44]. The best-fitting models for the training area had the lowest residuals for various cost function metrics and presented a good performance by applying K-fold cross-validation. However, determining the optimal regression model by considering only the cost functions is challenging. Hence, geospatial interpolation was applied to verify the reliability of the models based on the site class proportion of V_{S30} grids. The V_{S30} maps in North Korea primally visualized site classes against seismic site effects using the best-fitting models of the four regression methods. Although verification is impossible without geotechnical or geophysical investigations, the coefficients of the site amplification can be preliminarily interpreted for seismic hazard assessment or resistance designs.

Prior to the site classification, regression, and mapping workflow, the archiving of the V_{S30} profile and its site response from the geophysical survey is essential. However, if the V_S profile is not distributed for a large-scale area, the empirical correlations with other physical value can be assumed as the best alternative approach for determining the V_S profile. Nevertheless, the applicability of N-V_S correlations for regression modeling were validated; the direct V_S profile should be archived or supplemented for the site-specific characterization of seismic amplification. Therefore, the proposed regression model is recommended only in n-value-based V_{S30} mapping.

Geological unit-based proxies for seismic site characterization maps in high seismicity regions have been effectively used to demonstrate the shallow shear wave profile and its index properties. Accordingly, the integrated proxy index is possible considering the local cross-correlation between terrain and geological features, given the geological map of North Korea. An ensemble learning approach is necessary to combine several regression models to produce an optimal predictive model for an extended follow-up research.

5. Conclusions

In this study, the DTM was used to map large-scale V_{S30} in the test site in North Korea according to the proposed procedures: (1) TPI-based landform classification; (2) geostatistical interpolation for building a geospatial grid with geotechnical and DTM features; (3) calculation of V_{S30} grids; (4) reclassification and subgrouping of the TPI-based landform class with respect to V_{S30} grids; (5) uncertainty analysis with regression modes, landform class, and N-V_S correlations; and (6) multivariate regression for V_{S30} grid prediction using terrain proxy-based principal components. The reclassified TPI-based landform was strongly correlated with V_{S30}, and its site response coefficients were similar to the correlations between V_{S30} and elevation or slope. In particular, the grouped LF divided the sedimentary site with a lower V_{S30} from the outcrop bedrock site in the training and test sites. The geospatial grid, which has been developed based on the SGS-E-type, also supported the characterization of shallow subsoil and land cover conditions in addition to V_{S30}. The ML techniques for predicting the V_{S30} grid, which were initially allocated with the DTM, determined the site-specific V_{S30} map at the test sites: Pyongyang, Kaesong, and Nampo. The best-fitting models of LR, K-NN, SVR, and MLP were determined based on

K-fold cross-validation. The MLP-based V_{S30} map represented a similar distribution of the geospatial grid by using the SGS-E-type. On the V_{S30} map in the test areas of North Korea, feasible site classification against the seismic site effect was visualized, which will probably be used as a reference map for preliminary seismic hazard assessments.

References

1.　Lee, S.J.; Rhie, J.; Kim, S.; Kang, T.S.; Cho, C.S. 1-D velocity model for the North Korean Peninsula from Rayleigh wave dispersion of ambient noise cross-correlations. *J. Seismol.* **2020**, *24*, 121–131. [CrossRef]
2.　Walter, W.R.; Dodge, D.A.; Ichinose, G.; Myers, S.C.; Pasyanos, M.E.; Ford, S.R. Body-wave methods of distinguishing between explosions, collapses, and earthquakes: Application to recent events in North Korea. *Seismol. Res. Lett.* **2018**, *89*, 2131–2138. [CrossRef]
3.　Zhai, M.; Zhang, X.H.; Zhang, Y.B.; Wu, F.Y.; Peng, P.; Li, Q.L.; Li, Z.; Guo, J.; Li, T.; Zhao, L.; et al. The geology of North Korea: An overview. *Earth Sci. Rev.* **2019**, *194*, 57–96. [CrossRef]
4.　Pedersen, H.A.; Campillo, M.; Sánchez-Sesma, F.J. Azimuth dependent wave amplification in alluvial valleys. *Soil Dyn. Earthq. Eng.* **1995**, *14*, 289–300. [CrossRef]
5.　Theodulidis, N.; Bard, P.-Y.; Archuleta, R.; Bouchon, M. Horizontal-to-vertical spectral ratio and geological conditions: The case of Garner Valley downhole array in southern California. *Bull. Seismol. Soc. Am.* **1996**, *86*, 306–319.
6.　Castellaro, S.; Mulargia, F.; Rossi, P.L. V_{S30}: Proxy for seismic amplification? *Seismol. Res. Lett.* **2008**, *79*, 540–543. [CrossRef]
7.　Wald, L.A.; Mori, J. Evaluation of methods for estimating linear site-response amplifications in the Los Angeles region. *Bull. Seismol. Soc. Am.* **2000**, *90*, S32–S42. [CrossRef]
8.　Mucciarelli, M.; Gallipoli, M.R. Comparison between V_{S30} and Other Estimates of Site Amplification in Italy. In Proceedings of the First European Conference on Earthquake Engineering and Seismology, Geneva, Switzerland, 3–8 September 2006.
9.　Di Giacomo, D.; Gallipoli, M.R.; Mucciarelli, M.; Parolai, S.; Richwalski, S.M. Analysis and modeling of HVSR in the presence of a velocity tnversion: The case of Venosa, Italy. *Bull. Seismol. Soc. Am.* **2005**, *95*, 2364–2372. [CrossRef]
10.　Ahdi, S.K.; Stewart, J.P.; Ancheta, T.D.; Kwak, D.Y.; Mitra, D. Development of V_S profile database and proxy-based models for V_{S30} prediction in the Pacific Northwest region of North America. *Bull. Seismol. Soc. Am.* **2017**, *107*, 1781–1801.
11.　Karimzadeh, S.; Feizizadeh, B.; Matsuoka, M. DEM-Based V_{S30} map and terrain surface classification in nationwide scale—A case study in Iran. *ISPRS. Int. J. Geo Inf.* **2019**, *8*, 537. [CrossRef]
12.　Wald, D.J.; Allen, T.I. Topographic slope as a oroxy for seismic site conditions and amplification. *Bull. Seismol. Soc. Am.* **2007**, *97*, 1379–1395. [CrossRef]
13.　Allen, T.I.; Wald, D.J. On the use of high-resolution topographic data as a proxy for seismic site conditions (V_{S30}). *Bull. Seismol. Soc. Am.* **2009**, *99*, 935–943. [CrossRef]
14.　Iwahashi, J.; Pike, R.J. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology* **2007**, *86*, 409–440. [CrossRef]
15.　Hartzell, S.; Carver, D.; Williams, R.A. Site response, shallow shear-wave velocity, and damage in Los Gatos, California, from the 1989 Loma Prieta earthquake. *Bull. Seismol. Soc. Am.* **2001**, *91*, 468–478. [CrossRef]
16.　Frankel, A.D.; Carver, D.L.; Williams, R.A. Nonlinear and linear site response and basin effects in Seattle for the M 6.8 Nisqually, Washington, earthquake. *Bull. Seismol. Soc. Am.* **2002**, *92*, 2090–2109. [CrossRef]

17. McGaughey, W.J. Data-Driven Geotechnical Hazard Assessment: Practice and Pitfalls. In Proceedings of the First International Conference on Mining Geomechanical Risk, Australian Centre for Geomechanics, Perth, Australian, 9–11 April 2019; pp. 219–232.

18. Mignan, A.; Scolobig, A.; Sauron, A. Using reasoned imagination to learn about cascading hazards: A pilot study. *Disaster Prev. Manag.* **2016**, *25*, 329–344. [CrossRef]

19. Taghizadeh-Mehrjardi, R.; Emadi, M.; Cherati, A.; Heung, B.; Mosavi, A.; Scholten, T. Bio-inspired hybridization of artificial neural networks: An application for mapping the spatial distribution of soil texture fractions. *Remote Sens.* **2021**, *13*, 1025. [CrossRef]

20. Raiyani, K.; Gonçalves, T.; Rato, L.; Salgueiro, P.; Marques da Silva, J.R. Sentinel-2 image scene classification: A comparison between Sen2Cor and a machine learning approach. *Remote Sens.* **2021**, *13*, 300. [CrossRef]

21. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Ubiquitous GIS-based forest fire susceptibility mapping using artificial intelligence methods. *Remote Sens.* **2020**, *12*, 1689. [CrossRef]

22. Zhao, Z.; Chow, T.L.; Rees, H.W.; Yang, Q.; Xing, Z.; Meng, F.R. Predict soil texture distributions using an artificial neural network model. *Comput. Electron. Agric.* **2009**, *65*, 36–48. [CrossRef]

23. Song, X.D.; Liu, F.; Zhang, G.L.; Li, D.C.; Zhao, Y.G. Estimation of soil texture at a regional scale using local soil-landscape models. *Soil Sci.* **2016**, *181*, 435–445. [CrossRef]

24. Gholami, H.; Mohammadifar, A.; Bui, D.T.; Collins, A.L. Mapping wind erosion hazard with regression-based machine learning algorithms. *Sci. Rep.* **2000**, *10*, 1–16.

25. Okada, G.; Moya, L.; Mas, E.; Koshimura, S. The Potential Role of News Media to Construct a Machine Learning Based Damage Mapping Framework. *Remote Sens.* **2021**, *13*, 1401. [CrossRef]

26. Chousianitis, K.; Del Gaudio, V.; Sabatakakis, N.; Kavoura, K.; Drakatos, G.; Bathrellos, G.D.; Skilodimou, H.D. Assessment of Earthquake-Induced Landslide Hazard in Greece: From Arias Intensity to Spatial Distribution of Slope Resistance Demand Assessment of Earthquake-Induced Landslide Hazard in Greece. *Bull. Seismol. Soc. Am.* **2016**, *106*, 174–188. [CrossRef]

27. Schäfer, A.M.; Wenzel, F. Global megathrust earthquake hazard—maximum magnitude assessment using multi-variate machine learning. *Front. Earth Sci.* **2019**, *7*, 136. [CrossRef]

28. Gitis, V.G.; Derendyaev, A.B. Machine learning methods for seismic hazards forecast. *Geosciences* **2019**, *9*, 308. [CrossRef]

29. Sun, C.G.; Kim, D.S.; Chung, C.K. Geologic site conditions and site coefficients for estimating earthquake ground motions in the inland areas of Korea. *Eng. Geol.* **2005**, *81*, 446–469. [CrossRef]

30. Zhai, M.; Zhu, X.; Zhou, Y.; Zhao, L.; Zhou, L. Continental crustal evolution and synchronous metallogeny through time in the North China Craton. *J. Asian Earth Sci.* **2020**, *194*, 104169. [CrossRef]

31. Choi, B.Y.; Yun, S.T.; Yu, S.Y.; Lee, P.K.; Park, S.S.; Chae, G.T.; Mayer, B. Hydrochemistry of urban groundwater in Seoul, South Korea: Effects of land-use and pollutant recharge. *Environ. Geol.* **2005**, *48*, 979–990. [CrossRef]

32. Palka, E.J.; Galgano, F.A. *North Korea: A Geographical Analysis*; United States Military Academy: New York, NY, USA, 2003.

33. Doucette, J.; Lee, S.O. Experimental Territoriality: Assembling the Kaesong industrial complex in North Korea. *Pol. Geogr.* **2015**, *47*, 53–63. [CrossRef]

34. Jo, J.C.; Ducruet, C. Maritime Trade and Port Evolution in a Socialist Developing Country: Nampo, Gateway of North Korea. *Korea Spat. Plan. Rev.* **2006**, *51*, 3–24.

35. Stewart, J.P.; Klimis, N.; Savvaidis, A.; Theodoulidis, N.; Zargli, E.; Athanasopoulos, G.; Pelekis, P.; Mylonakis, G.; Margaris, B. Compilation of a local V_S profile database and its application for inference of V_{S30} from geologic-and terrain-based proxies. *Bull. Seismol. Soc. Am.* **2014**, *104*, 2827–2841. [CrossRef]

36. Vianello, A.; Cavalli, M.; Tarolli, P. LiDAR-derived slopes for headwater channel network analysis. *Catena* **2009**, *76*, 97–106. [CrossRef]

37. Sun, C.G.; Kim, H.S. GIS-based regional assessment of seismic site effects considering the spatial uncertainty of site-specific geotechnical characteristics in coastal and inland urban areas. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1592–1621. [CrossRef]

38. Wair, B.R.; DeJong, J.T.; Shantz, T. *Guidelines for Estimation of Shear Wave Velocity Profiles*; Pacific Earthquake Engineering Research Center: Berkeley, CA, USA, 2012.

39. Imai, T.; Tonouchi, K. Correlation of N-value with s-wave velocity and shear modulus. In Proceedings of the 2nd European Symposium on Penetration Testing, Amsterdam, The Netherlands, 24–27 May 1982; pp. 57–72.

40. Ohsaki, Y.; Iwasaki, R. On dynamic shear moduli and Poisson's ratio of soil deposits. *Soils Found.* **1973**, *13*, 61–73. [CrossRef]

41. Ohata, Y.; Goto, N. Empirical shear wave velocity equations in terms of characteristic soil indexes. *Earthq. Eng. Struct. Dyn.* **1978**, *6*, 167–187. [CrossRef]

42. Hasancebi, N.; Ulusay, R. Empirical correlations between shear wave velocity and penetration resistance for ground shaking assessments. *Eng. Geol. Environ.* **2007**, *66*, 203–213. [CrossRef]

43. Dikmen, U. Statistical correlations of shear wave velocity and penetration resistance for soils. *J. Geophys. Eng.* **2009**, *6*, 61–72. [CrossRef]

44. ASTM. Standard test method for penetration test and splitbarrel sampling of soils (D 1586–99). In *2002 Annual Book of ASTM Standards, sect. 4, vol. 04.08*; American Society of Testing and Materials: Philadelphia, PA, USA, 2002.

45. Sun, C.K.; Kim, H.S.; Chung, C.K.; Chi, H.C. Spatial zonations for regional assessment of seismic site effects in the Seoul metropolitan area. *Soil Dyn. Earthq. Eng.* **2014**, *56*, 44–56. [CrossRef]

46. Sun, C.G.; Cho, C.S.; Son, M.; Shin, J.S. Correlations between shear wave velocity and in-situ penetration test results for Korean soil deposits. *Pure Appl. Geophys.* **2013**, *170*, 271–281. [CrossRef]
47. MOLIT. *General Seismic Design. KDS 17 10 00*; Ministry of Land, Infrastructure and Transport: Sejong, Korea, 2018.
48. Weiss, A. Topographic Position and Landforms Analysis. In Proceedings of the Poster Presentation, ESRI User Conference, San Diego, CA, USA, 9–13 July 2001; Volume 200.
49. Guisan, A.; Weiss, S.B.; Weiss, A.D. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* **1999**, *143*, 107–122. [CrossRef]
50. Jenness, J. Topographic Position Index (tpi_jen.avx) Extension for ArcView. Volume 3 (x). Jenness Enterprises; p. v. 1.3a. Available online: http://www.jennessent.com/arcview/tpi.htm (accessed on 15 January 2021).
51. Seif, A. Using topography position index for landform classification (case study: Grain Mountain). *Bull. Environ. Pharmacol. Life Sci.* **2014**, *3*, 33–39.
52. De Reu, J.; Plets, G.; Verhoeven, G.; De Smedt, P.; Bats, M.; Cherretté, B.; De Maeyer, W.; Deconynck, J.; Herremans, D.; Laloo, P.; et al. Towards a three-dimensional cost-effective registration of the archaeological heritage. *J. Archaeol. Sci.* **2013**, *40*, 1108–1121. [CrossRef]
53. Chiles, J.P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 497.
54. Zuo, R.; Xiong, Y.; Wang, J.; Carranza, E.J.M. Deep learning and its application in geochemical mapping. *Earth Sci. Rev.* **2019**, *192*, 1–14. [CrossRef]
55. Kim, H.S.; Sun, C.G.; Cho, H.I. Geospatial big data-based geostatistical zonation of seismic site effects in Seoul metropolitan area. *ISPRS. Int. J. Geo Inf.* **2017**, *6*, 174. [CrossRef]
56. Kim, M.; Kim, H.S.; Chung, C.K. A three-dimensional geotechnical spatial modeling method for borehole dataset using optimization of geostatistical approaches. *KSCE J. Civ. Eng.* **2020**, *24*, 778–793. [CrossRef]
57. Dobry, R.; Borcherdt, R.D.; Crouse, C.B.; Idriss, I.M.; Joyner, W.B.; Martin, G.R.; Power, M.S.; Rinne, E.E.; Seed, R.B. New site coefficients and site classification system used in recent building seismic code provisions. *Earthq. Spectra* **2000**, *16*, 41–67. [CrossRef]
58. Michelini, A.; Faenza, L.; Lauciani, V.; Malagnini, L. ShakeMap implementation in Italy. *Seismol. Res. Lett.* **2008**, *79*, 688–697. [CrossRef]
59. Sun, C.G.; Kim, H.S. Geostatistical assessment for the regional zonation of seismic site effects in a coastal urban area using a GIS framework. *Bull. Earthq. Eng.* **2016**, *14*, 2161–2183. [CrossRef]
60. Kim, H.S.; Sun, C.G.; Cho, H.I. Geospatial assessment of the post-earthquake hazard of the 2017 Pohang earthquake considering seismic site effects. *ISPRS Int. J. Geo Inf.* **2018**, *7*, 375. [CrossRef]
61. Masood, F.; Boulila, W.; Ahmad, J.; Sankar, S.; Rubaiee, S.; Buchanan, W.J. A novel privacy approach of digital aerial images based on mersenne twister method with DNA genetic encoding and chaos. *Remote Sens.* **2020**, *12*, 1893. [CrossRef]
62. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.T.; Blum, M.; Hutter, F. Auto-sklearn: Efficient and robust automated machine learning. In *Automated Machine Learning*; Springer: New York, NY, USA, 2019; pp. 113–134.
63. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. *arXiv* **2012**, arXiv:1206.2944.
64. Yao, Z.; Ruzzo, W.L. A Regression-Based K Nearest Neighbor Algorithm for Gene Function Prediction from Heterogeneous Data. In *BMC Bioinformatics*; BioMed Central: London, UK, 2006; Volume 7, pp. 1–11.
65. Vapnik, V. The Support Vector Method of Function Estimation. In *Nonlinear Modeling*; Springer: Boston, MA, USA, 1998; pp. 55–85.
66. Dibike, Y.B.; Velickov, S.; Solomatine, D.; Abbott, M.B. Model induction with support vector machines: Introduction and applications. *J. Comput. Civ. Eng.* **2001**, *15*, 208–216. [CrossRef]
67. Liong, S.Y.; Sivapragasam, C. Flood stage forecasting with support vector machines 1. *J. Am. Water Resour. Assoc.* **2002**, *38*, 173–186. [CrossRef]
68. Lin, S.W.; Ying, K.C.; Chen, S.C.; Lee, Z.J. Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Syst. Appl.* **2008**, *35*, 1817–1824. [CrossRef]
69. Venkatesan, P.; Anitha, S. Application of a radial basis function neural network for diagnosis of diabetes mellitus. *Curr. Sci.* **2006**, *91*, 1195–1199.
70. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
71. Jodouin, J.F. *Les Réseaux de Neurones: Principes et Définitions*; Hermès: Paris, France, 1994.
72. Ministry of Public Safety and Security (MPSS). *Report on the 9.12 Earthquake and Countermeasures*; Ministry of Public Safety and Security (MPSS): Seoul, Korea, 2016.
73. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
74. Menafoglio, A.; Secchi, P.; Dalla Rosa, M. A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electron. J. Stat.* **2013**, *7*, 2209–2240. [CrossRef]
75. Skentos, A.; Ourania, A. Landform analysis using terrain attributes. A Gis application on the island of Ikaria (Aegean Sea, Greece). *Ann. Valahia Univ. Targoviste Geogr. Ser.* **2017**, *17*, 90–97. [CrossRef]