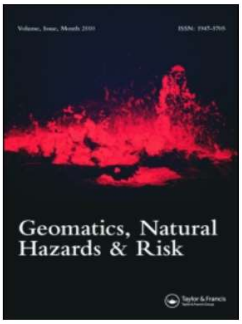

Geomatics, Natural Hazards and Risk

ISSN: 1947-5705 (Print) 1947-5713 (Online) Journal homepage: https://www.tandfonline.com/loi/tgnh20

Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea

Sunmin Lee, Jeong-Cheol Kim, Hyung-Sup Jung, Moung Jin Lee & Saro Lee

To cite this article: Sunmin Lee, Jeong-Cheol Kim, Hyung-Sup Jung, Moung Jin Lee & Saro Lee (2017) Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea, Geomatics, Natural Hazards and Risk, 8:2, 1185-1203, DOI: 10.1080/19475705.2017.1308971

To link to this article: https://doi.org/10.1080/19475705.2017.1308971

Published online: 10 Apr 2017.

GEOMATICS, NATURAL HAZARDS AND RISK, 2017
VOL. 8, NO. 2, 1185–1203
https://doi.org/10.1080/19475705.2017.1308971

Taylor & Francis
Taylor & Francis Group

Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea

Sunmin Lee[a,c], Jeong-Cheol Kim[a,b], Hyung-Sup Jung [a], Moung Jin Lee[c] and Saro Lee [d,e]

[a]Department of Geoinformatics, University of Seoul, Seoul, Korea; [b]National Institute of Ecology, Seocheon, Korea; [c]Center for Environmental Assessment Monitoring, Environmental Assessment Group, Korea Environment Institute (KEI), Sejong, Korea; [d]Geological Research Division, Korea Institute of Geoscience and Mineral Resources (KIGAM), Daejeon, Korea; [e]Korea University of Science and Technology, Daejeon, Korea

ABSTRACT

Since flood frequency increases with the impact of climate change, the damage that is emphasized on flood-risk maps is based on actual flooded area data; therefore, flood-susceptibility maps for the Seoul metropolitan area, for which random-forest and boosted-tree models are used in a geographic information system (GIS) environment, are created for this study. For the flood-susceptibility mapping, flooded-area, topography, geology, soil and land-use datasets were collected and entered into spatial datasets. From the spatial datasets, 12 factors were calculated and extracted as the input data for the models. The flooded area of 2010 was used to train the model, and the flooded area of 2011 was used for the validation. The importance of the factors of the flood-susceptibility maps was calculated and lastly, the maps were validated. As a result, the distance from the river, geology and digital elevation model showed a high importance among the factors. The random-forest model showed validation accuracies of 78.78% and 79.18% for the regression and classification algorithms, respectively, and boosted-tree model showed validation accuracies of 77.55% and 77.26% for the regression and classification algorithms, respectively. The flood-susceptibility maps provide meaningful information for decision-makers regarding the identification of priority areas for flood-mitigation management.

ARTICLE HISTORY
Received 11 October 2016
Accepted 8 March 2017

KEYWORDS
Flood susceptibility; random forest, boosted tree; GIS; Seoul metropolitan city

1. Introduction

Recently, as global warming continues, abnormal weather phenomena include precipitation and high temperatures occur frequently, leading to various environmental problems, including localized downpours, flooding, drought and heat waves, that could cause many casualties (Kim et al. 2016; Novelo-Casanova & Rodríguez-Vangort 2016). Floods can incur an enormous economic cost, particularly in city centres that comprise high levels of urban infrastructure that includes buildings and transportation facilities (Lee et al. 2015; Klaus et al. 2016). Because of the irreducible and huge damages to road, rail, bridges, electricity, water, phone services, the functionality of entire cities can become paralysed; therefore, the provision of adequate flood-protection and flood-prevention measurements is urgent.

To rapidly prevent the damage from a flood, an advance-warning system and prompt action by governments and institutions are needed. As mentioned, a flood could cause the loss of human life or injury, so a flood policy including the improvement of pumping stations and water facilities

CONTACT Saro Lee leesaro@kigam.re.kr

should not be an afterthought (Vojtek & Vojteková 2016). The prevention measures should include the establishment of countermeasures using the information on flood-prone areas; therefore, to minimize the damages from floods, a vulnerability assessment is needed to determine the priority areas according to the flood vulnerability, and it could be derived from previous flood damages.

In studies on natural disasters, various data are usually needed (Regmi et al. 2013), but it is not easy to obtain suitable data (Cao et al. 2016). For a flood study, the factors that significantly affect a flood are vegetation type, soil type, and geological, geomorphological and hydrological characteristics. The acceleration of a flood could be caused by the reckless destruction of nature by human beings (Chang & Chen 2016); indiscriminate development, over-harvesting and deforestation are examples. By using these factors with a geographic information system (GIS), the studies of flood areas calculate and evaluate the effects of floods.

The capital city of South Korea, Seoul metropolitan city, is vulnerable to floods during the summer season due to torrential rain. Especially during 2010 and 2011, extreme rainfall affected the southern region of the city, causing a lot of flood damage. A determination of the priority sectors is important for the city government to manage the floods beforehand to prevent damages. In September 21, 2010, 259.5 mm per a day of a torrential downpour caused considerable damage to the city centre, especially around the Gwangwhamun, Cheonggyecheon area (Figure 1(b)). Heavy rainfall in the central part of Korea during the morning of 27 July 2011 (Figure 1(c–d)), resulted in 164 mm of rainfall, and this exceeds the frequency value of over 100 years. The purpose of this study is the

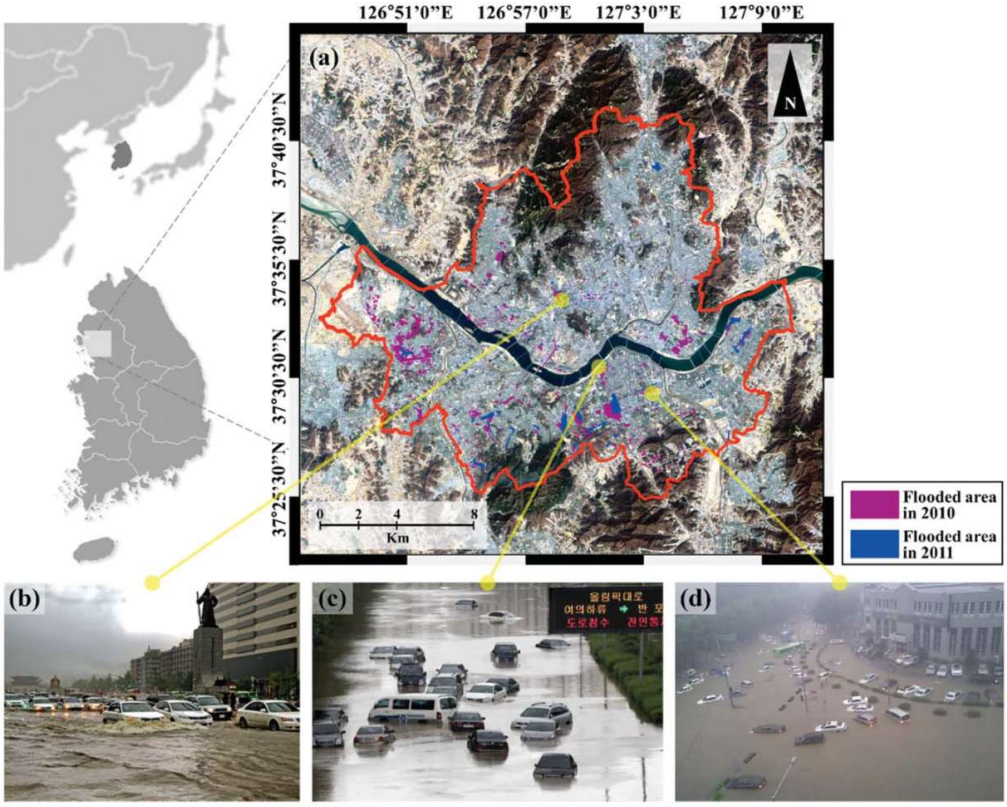

Figure 1. Study area with flooded area in 2010 and 2011 with pictures from flooded area: (b) Gwangwhamun in September 2010 (Yu, et al. 2011), (c) Olympic-daero, Seocho-gu (Choi 2010) and (d) Daechi-dong, Gangnam-gu in July 2011 (Jo 2011).

creation of flood-susceptibility maps for the Seoul metropolitan city using a random-forest model and a boosted-tree model in a GIS environment.

Many research studies have been conducted using probabilistic models and the multi-criteria decision analysis for flood mapping (Wang et al. 2011; Masood & Takeuchi 2012). In addition, using GIS platform, the support vector machine with different kernel types were applied in Kuala Terengganu basin, Malaysia mainly using topographical input data (Tehrany et al. 2015). Frequency ratio, weights-of-evidence models were also applied to Golastan Province, Iran (Rahmati et al. 2016a), and in GIS environment, statistical models were integrated for flood-susceptibility mapping (Blanco-Vogt & Schanze 2014; Morelli et al. 2014). For flood mapping, many researchers exerted modelling techniques such as hydrologic modelling, hydraulic modelling and global hydrodynamic modelling (Grimaldi et al. 2013; Papaioannou et al. 2013; Chini et al. 2014; Curebal et al. 2016). Random forests were applied to various fields in environmental areas to characterize stand-level forest-canopy cover (Ahmed et al. 2015) and to model plant-species richness (Lopatin et al. 2016). A number of studies have applied the random-forest classifier for a flood mapping for which other methods are integrated, as follows: spectral-mixture analysis (Feng, Gong, et al. 2015) or unmanned aerial-vehicle remote sensing (Feng, Liu, et al. 2015). Modelling techniques have also been applied to random forests for analysis flooding (Albers et al. 2016) and to assess the flood-hazard risk (Wang et al. 2015). The boosted-tree algorithm, especially for the regression model, has been proven as an influential approach for ecological-modelling study areas including hydrological areas over the last few years (Elith et al. 2008; Leathwick et al. 2008; De'ath & Fabricius 2010; Nylén et al. 2015). A study that used the boosted-tree algorithm for flood mapping was also conducted recently (Coltin et al. 2016). The results of the analysis of the flood vulnerability in the Seoul area using artificial neural network are shown as 79.05% (Lee 2014), which shows no significant difference from the validated values of this study.

The difference of this study is the use of the random-forest and boosted-tree models for the application and validation processes in Seoul metropolitan city, Korea. Especially, the flooded areas of different periods were used for the training and validation data. Also, the predictor importance of each variable was calculated, and the random-forest and boosted-tree models were applied and compared.

The flood-susceptibility mapping was performed as follows: (1) data were collected and the related factors were extracted and calculated, (2) spatial datasets were established with a grid format, (3) the flood-susceptibility assessment was conducted using the random-forest model and the boosted-tree model, and the predictor-importance values of each factor were calculated in this process, and lastly (4) the validation of the susceptibility map was performed using the flooded areas of 2011.

2. Study area

Seoul metropolitan city is the capital and the biggest city of South Korea, and is located in the northwest part of South Korea. Seoul metropolitan area is less than 1% of the total area of the country, approximately 605.25 km^2. Seoul metropolitan city is composed of 25 administrative districts called Gu with nine regular subway lines and two special lines. The population of Seoul metropolitan city was over 10 million in 2011, which is more than a quarter of the whole population of South Korea, and it is also the 10th largest city in the world. As there is a large floating population, the transportation facilities are complicated, and there are continuous construction activities such as subway works and road constructions.

Located at 126°59'40" E and 37°33'59" N, the Han River bisects the city area into two parts, northern and southern (Figure 1(a)). Han River comes from areas such as Tancheon, Jungnangcheon and Anyangcheon, and the stream ratio is approximately 10%. Seoul metropolitan city is located in a basin, surrounded by Bukhansan (Seoul's highest peak at 836 m), Inwangsan (338 m),

Dobongsan (740 m) and Gwanaksan (629 m), as well as other substantial mountains. The average height becomes lower toward the city centre near Han River.

In terms of the climatological feature, Seoul metropolitan city has East Asian monsoon characteristics. Particularly from June to September, it is commonly hot and humid. The average temperature of August is from 22.4° C to 29.6° C, with the possibility of higher temperatures. Recently, a serious heat wave hit Seoul metropolitan city, and some have lasted for as long as 12 days. Otherwise, in winter, it is generally drier than summer, and the average temperature of January is from –5.9° C to 1.5° C. The annual precipitation rate in this area is 1370 mm. More than 70% of the city's rainfall is recorded in summer. In the monsoon season, heavy rains and strong winds come together, causing a rain front and flooding at the same time. Due to the seasonal difference of the frequency of heavy rain, high-strength damage control is required. Especially, due to the high population density, the flood of a big city such as Seoul metropolitan city could cause more serious damage than those of non-urban areas; therefore, the prevention of the damages to roads and railway facilities from flooding by a mapping of the flood susceptibility is required to reduce the great cost of disaster recovery.

3. Theory

3.1. Random-forest model

A random-forest model is an ensemble-learning technique for which a multitude of decision trees is constructed to explain the spatial relationships between the occurrence of floods and the related factors for classification and regression. It operates by constructing a multitude of decision trees at the training time, and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random-decision forests correct for the decision-tree habit of over-fitting to a training set.

The random-forest algorithm is an ensemble-learning method that comprises an ensemble of simple tree predictors for classification and regression (Breiman 2001). The first algorithm for random-decision forests was created using the random-subspace method (Ho 1995). From Ho's formulation, the implementation is achieved according to the stochastic-discrimination approach to classification (Kleinberg 2000). An extension of the algorithm was developed with a bagging idea and the random selection of the features to construct a collection of decision trees with a controlled variance (Breiman 2001), and this algorithm is used for this study. A result with a group of predictor values is produced for both classification and regression algorithms. When the final result has discrete values, it is called a classification (Ellis et al. 2014). In this study, flood vulnerability is classified based on given factor values. In general, the classification problem aims at obtaining the decision boundary. Conversely, regression algorithm carries a continuous value for the output or target value. Therefore, this study shows the degree of vulnerability as a continuous value with regression algorithm (Ellis et al. 2014). As mentioned, first, for the classification algorithm, a group of independent predictor values is classified with one of the groups that is present in the dependent variable. As an alternative, for the regression algorithm, the tree responds to the predictors with an estimate of the dependent variable by outputting the class that is the mode of the mean prediction.

A random number of simple trees determine the final predictors of a random-forest algorithm. In the classification algorithm, it constructs decision trees and output the classes. The response, the most-popular class, is decided by the ensemble of simple trees. For a regression algorithm, the average of the results is used to obtain an estimate of the dependent variable. A considerable improvement of the prediction accuracy could be derived from the use of tree ensembles. A random forest does not require any assumptions about the relationships between the explanatory and response variables; and it is a proper method for the analysis of hierarchical and non-linear interactions in large datasets (Olden et al. 2008). When a random-forest algorithm is applied to predict new data cases, a superior capability can be shown.

A set of predictor values is chosen independently, and it affects the response of each tree; also, it shows the same distribution for all of the trees in the forest, and it is a subset of the predictor values of the initial dataset. The optimal condition of the subset of the predictor variables is given by log2 (M + 1), where M is the number of inputs in the algorithm, and the mean-square error for a random forest is given by the following equation:

$$\varepsilon = \left(v_{\text{observed}} - v_{\text{response}}\right)^2, \tag{1}$$

where ε is the mean-square error from the algorithm, v_{observed} is the variables from the observed data, and v_{response} is the variables from the result.

The average of the predictions from the trees is calculated as follows:

$$\text{S} = \frac{1}{\text{K}} \sum K^{\text{th}} v_{\text{response}}, \tag{2}$$

where *S* is the random-forest prediction, and K applies to the individual trees in the forest.

For the classification algorithm, the definition of a margin function that counts the average number of excess votes for the correct class compared to other classes' average vote in the dependent variable is conducted after a determination of a set of simple trees and random predictor variables. The classification algorithm provides a practical method for the making of predictions, and it could also be a method for associating a confidence measure with the predictions.

For the regression algorithm, simple trees that are capable of producing a numerical response value are used to build random forests. Like the classification algorithm, a predictor set is randomly selected for all of the trees from the equal distribution.

Commonly, the missing data from the predictor variables in the random forest could be flexibly integrated. When the algorithm builds a model, the prediction for a specific case is created based on the last preceding (non-terminal) node in the respective tree when the missing data are included; therefore, the overall mean at the root node is used to derive the prediction variables when there is no valid data at a particular point in the sequence of trees. It is therefore unnecessary to remove the no-data cases for some of the predictors from the random-forest process.

3.2. Boosted-tree model

The boosted-tree algorithm is from one of the general computational approaches of stochastic-gradient boosting, which are also known as TreeNet (TM Salford Systems, Inc.) and MART (TM Jerill, Inc.). This technique has appeared as one of the most-influential methods for data mining and predictive modelling in the past few years. These potent algorithms could effectively be used for regression as well as classification with continuous and categorical predictors (Friedman, Getoor, et al. 1999; Friedman, Goldszmidt, et al. 1999; Hastie 2001).

As the random-forest algorithm, boosted-tree algorithm is an ensemble-learning method that comprises an ensemble of simple tree predictors considering the performance of the previous classifier for classification and regression. The boosted-tree algorithm evolved from boosting-method applications into regression trees. The initial idea is the adding up of a sequence of simple trees on the preceding tree from the prediction residuals. The algorithm builds binary trees as the general classification, and the regression-tree models partition the data into two samples. At each split node of a boosted-tree algorithm, the best data-partitioning point is decided, and the deviations of the observation from the means, the residuals for each partition, are computed. Determining the stopping criteria is very important since simple tree has limitations in continuous variables and sensitivity to the size of the sample data; on the contrary, if the tree becomes complex, it shows overfitting problem which means algorithm creates overly set model to input data. Therefore, as for tree

models, such as random-forest and boosted-tree models have inherent stopping criteria in algorithm and various heuristics method that helps avoid overfitting problem. In this way, trees could ultimately produce a more-effective fit of the prediction values to the observation values, even though its relationships with the predictor and dependent variables are complex, such as those that are nonlinear; therefore, the boosted-tree algorithm can serve as a reliable machine-learning algorithm by fitting a weighted additive expansion of simple trees.

However, as with other machine-learning methods, it is difficult to determine when the training should be stopped for the boosted-tree algorithm. This problem generally occurs when it is used in predictive data mining, and it could also lead to the problem of overfitting. A common solution for this problem is the use of a test sample to predict the observation data that were unused for the previous estimation of the respective models. The evaluation of the quality of the fitted model is performed according to the test data. In this regard, the predictive accuracy could be tested, and it could be known whenever an overfitting problem has occurred. The regression problem, such as that regarding the prediction of a continuous dependent variable, has been strongly highlighted with respect to the boosted-tree algorithm. The algorithm could be applied to the management of classification problems (Friedman, Getoor, Koller & Pfeffer 1999).

First, the creation of a coded variable of the values for each class with the values 1 or 0 is proposed to show if the observation belongs to the respective class. When in sequence, boosting trees are, respectively, built for each class of the categorical dependent variable. The following step uses the logistic transformation to calculate the residuals from the following boosting step. Repeatedly, the logistic transformation is applied to the predictions for each 0 or 1 class to calculate the ultimate classification probabilities (Friedman, Getoor, Koller & Pfeffer 1999; Hastie 2001).

When the boosted-tree algorithm is applied to classification problems, a separation of the sequences of the boosted trees that are built for their class is required; therefore, the calculative attempt commonly becomes large by a multiple of what it takes to solve the simple regression-prediction problem for a single continuous dependent variable. When the number of classes is more than approximately 100, the analysis of the categorical dependent variables is not highly reliable, as the process might contain amounts of illogical effort and time.

3.3. Summary

To summarize, the random-forest and the boosted-tree algorithms used in this study are both based on the decision tree applying an ensemble-learning technique. Decision tree is a tree-shaped model which predicts the value of the target variable based on input variables; the single decision tree has a large variation of the performance (Quinlan 1986). Since the decision tree is determined according to the learning data which has small bias and large variance, deeply growing carry overfitting problem; it is difficult to be used generally (Lee & Lee 2015). Random-forest and boosted-tree algorithms construct the model by applying the ensemble-learning method to the decision tree and combining multiple decision trees.

The difference is that the random-forest algorithm extracts data from the input data randomly and determines the single decision tree. The process forming the single decision tree repeats for the number of times and the final random-forest model is determined based on the weights of the well-classified decision trees generated from the repeated number of times (Breiman 2001). Since the random-forest algorithm adds randomness to the sample variables, it maximizes the advantages of the ensemble-learning technique, resulting in high prediction and classification accuracy (Belgiu & Dragut 2016). However, there are remaining instabilities due to data samples which are resides in decision trees (Muchlinski et al. 2016).

The boosted-tree algorithm extracts arbitrary data from the input data, such as random forests and determines the decision tree with the extracted data. The process of creating the decision tree is repeated a certain number of times. In the process of extracting samples, data which is not properly classified in the previous step are selected preferentially. Therefore, the difference between the two

models depends on whether it considers the classification performance of the model in the previous step when extracting the sample and creating the model. The boosted-tree algorithm also remains stability in prediction based on a large number of trees (Naghibi et al. 2016, Tan et al. 2016). In particular, it is possible to lower the bias by sampling data within a high misclassification rate (Youssef et al. 2016a). However, there are also disadvantages of losing explanatory power, which is one of the advantages of the existing tree.

4. Spatial datasets

It is widely known that many hydrological factors are the influencing factors of flood susceptibility. Degree of slope, stream evolution, soil, vegetation, morphology, land use, lithology, geological structure and human activity are the known factors, but the relationships between the factors and flood susceptibility have not been examined with real factors; therefore, 12 input factors including the distance from the river, which is considered as a main indicator, and that are expected to have a relationship with flood susceptibility were adopted. The factors were then applied to the random-forest and boosted-tree models (Table 1, Figure 2).

In order to evaluate flood vulnerability, setting variables affecting floods should be preceded (Liu & De Smedt 2005). Among the impact variables on flood, slope is one of the most important factors (Tehrany et al. 2013). Thus, slope was most frequently used as a factor related to flood among the topographical factors (Pradhan 2010; Tehrany et al. 2014; Bui et al. 2016; Khosravi et al. 2016, Marconi et al. 2016, Rahmati et al. 2016b; Seekao & Pharino 2016b; Youssef et al. 2016b). Additionally, DEM (Pradhan 2010; Tehrany et al. 2014; Sowmya et al. 2015; Bui et al. 2016; Marconi et al. 2016; Rahmati et al. 2016b; Youssef et al. 2016b), distance to river (Bui et al. 2016; Khosravi et al. 2016; Marconi et al. 2016; Rahmati et al. 2016b; Youssef et al. 2016b), TWI and SPI(Bui et al. 2016; Khosravi et al. 2016) and plan curvature(Bui et al. 2016; Khosravi et al. 2016; Youssef et al. 2016b) are used to evaluate floods from topographical factors. SLF is also included to spatial datasets since it is related to water drainage (Borrelli et al. 2014) with soil-drainage factor (Tehrany et al. 2014; Seekao & Pharino 2016a; Youssef et al. 2016b). In addition, factors concerned with impermeability are used in flood vulnerability analysis such as green infra farmland (Tehrany et al. 2014; Khosravi et al. 2016), retarding basin (Seekao & Pharino 2016a, 2016b) and geology (Lawal et al. 2014; Tehrany et al. 2014).

The flood-area data of this study are in an inventory map that was generated through an area-based surveying that was under government control; the flood-area data includes 4,338 flooding zones. The size of the samples depends on many variables (Ohlmacher & Davis 2003). The 12 input

Table 1. Data layer related to flood of study area.

Category	Factors	Data type	Scale
Hazard	Flooded area	Polygon	1:1,000
Topographical map[a]	Ground elevation (m)	GRID	1:1,000
	Gradient – slope (°)		
	Distance from the river (m)		
	Slope Length Factor (SLF)		
	Topographic wetness index (TWI)		
	Stream power index (SPI)		
	Plan curvature		
Land-use & land registration map[b]	Impermeability layer	Polygon	1:25,000
	Green Infra Farmland		
	Retarding basin		
Soil map[c]	Soil drainage	Polygon	1:25,000
Geological map[d]	Geology	Polygon	1:25,000

[a]Topographical factors were extracted from digital topographic map by National Geographic Information Institute (NGII; http://www.ngii.go.kr).
[b]The land-use map is published by Korea Ministry of Environment (http://eng.me.go.kr/).
[c]The detailed soil map produced by Rural Development Administration (RDA; http://www.rda.go.kr).
[d]The geological map produced by the Korea Institute of Geoscience & Mineral Resource (KIGAM; http://www.kigam.re.kr).

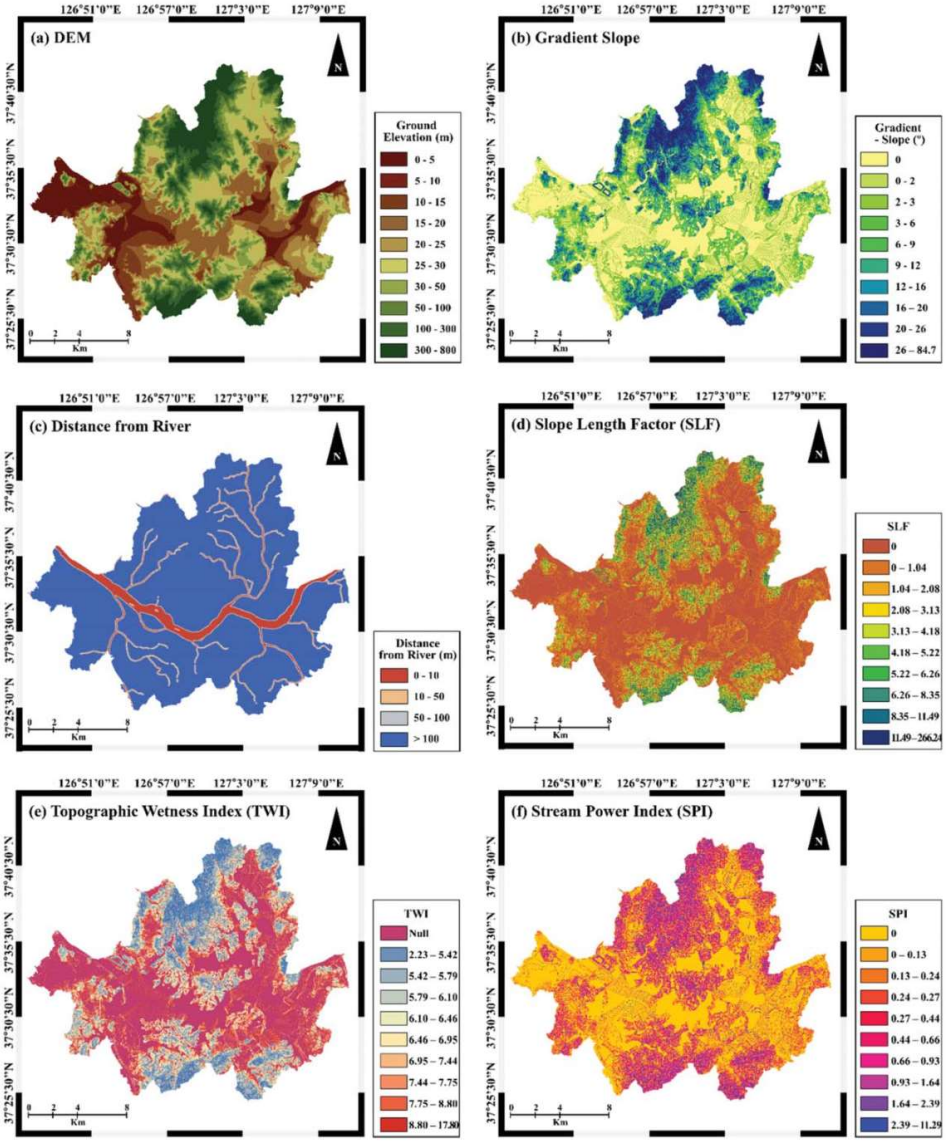

Figure 2. Constructed spatial database for flood productivity potential index analysis.

factors were prepared from other maps from various national administrations (Table 1). According to Table 1, the flood-related hydrological factors were collected from topographical, land-use, soil and geological maps. A digital elevation model (DEM) was prepared through a topographic map by digitizing the contours with a 5 m interval. The DEM could then be calculated into the gradient (slope), the distance from the river, the slope length factor (SLF), the topographic wetness index (TWI), the stream power index (SPI) and the plan curvature.

Using the FILL tool in ArcGIS 10.1, a removal of the internal drainage from the elevation grid was performed. The DEM was made into a 30 m × 30 m resolution through the making of a triangulated irregular network (TIN) for which the elevation value was used. Through the DEM input factor, the following factors could be calculated: slope, SLF, TWI, SPI and plan curvature.

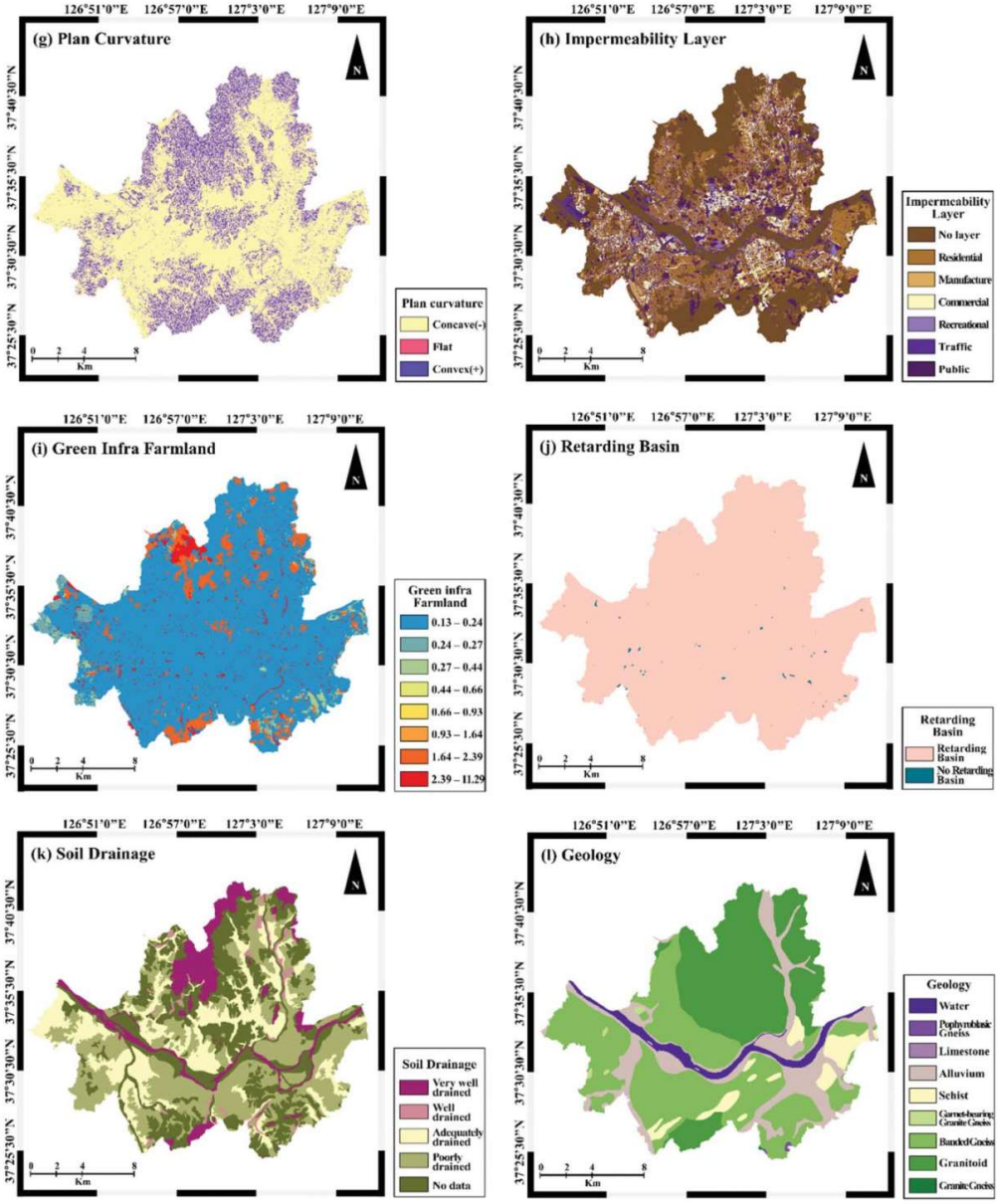

Figure 2. Continued.

The geomorphological environment of the study area could be represented using the topographic factors that influence the flood possibilities. The topographic factors such as slope and curvature are closely related to hydrological conditions such as groundwater flow, slope stability and soil moisture. Particularly, soil-moisture characteristics can be represented with topographic factors (Moore et al. 1991). From DEM data, the topographic-index slope map was generated using the SLOPE tool from ArcGIS 10.1, and the distance from the river was also calculated from the topographic map using the DISTANCE tool. The plan curvature was derived using the same method, but with the CURVATURE tool. After the calculation of the factor slope, the SLF for the average erosion was calculated

using the Revised Universal Soil Loss Equation (Wischmeier & Smith 1978). The SLF given a slope length λ defined as

$$\mathrm{SLF} = \left(\frac{\lambda}{72.6}\right)^{m}, \tag{3}$$

where the constant 72.6 is the unit of Revised Universal Soil Loss Equation in feet. The variable m is slope-length exponent (Wischmeier & Smith 1978). The TWI index was also calculated within the run-off model (Beven & Kirkby 1979), and it was calculated for the downslope edge of each value using the parameter value and the distribution of the accumulated area among the downslope cells. The TWI commonly signifies the tendency of water, and it is useful for the quantifying of the topographic control on hydrological processes. The TWI predicts the water accumulation at any point and shows the tendency of gravitational forces to move water in a downward direction. The water infiltration affects the soil strength, and is contrastively affected by material characteristics such as pore-water pressure and permeability. The SPI (Moore et al. 1991) indicates the erosion power of the stream and is regarded as a contributor to the provision of stability in the study area; therefore, it could be used to determine the places where soil-conservation measures can decrease the erosive effects of concentrated surface runoff.

The topographic map is provided in a grid form by the National Geographic Information Institute (NGII). The scale of this map is 1:1000. The land-use and land-registration map is published by the Korea Ministry of Environment with a form of polygon coverage and a scale of 1:25000. The land-use and the land-registration were derived from SPOT-5 images of November 2007. The soil map and the geological map are also vector maps with the polygon-coverage form; the scales of the maps are both 1:25000, and they are produced by the National Academy of Agricultural Science (RDA) and the Korea Institute of Geoscience and Mineral Resources (KIGAM), respectively. Impermeability layer, green infra farmland and retarding basin data are derived from land-use and land-registration map. The soil-drainage values were acquired from the soil map, and the geological units were obtained from geological map. After on-screen digitizing to the data, the attribute values were added. Finally, coordinate system is converted into Korean coordinate system as other geospatial data in the datasets. The geological characteristics of the study area indicate that it comprises 24 geological units of bedrock that can be classified into geological groups such as alluvium, granitoid and Precambrian metamorphic rocks (gneiss, schist and limestone).

5. Methodology

The flowchart of Figure 3 shows the data-processing steps that were used in this study; the random-forest and boosted-tree models were mainly applied to compare the two different data-mining models. The flooded areas from 2010 and 2011 were extracted so that they could be applied to these data-mining models as training and validation data. The statistical program STATISTICA and a GIS program ArcGIS were used for the application of the models.

To analyse the correlation between the hydrological factors and the flood susceptibility, the relevant factors were collected or calculated from the topographic, land-use, soil and geological maps. Then, to apply the random-forest and boosted-tree models, the related factors were converted into ASCII data in ArcGIS. Also, the flood-occurrence areas were obtained from the 2010 and 2011 data after the study area was determined. The flood-occurrence areas were set as dependent variables for the analysis, and the other factors that influence flood occurrence were set as independent variables. The topography, land-use, soil and geology data were also determined as independent variables. A total of 12 susceptibility factors were selected from the spatial datasets. The flooded area of 2010 and the flooded area of 2011 were used for the training and validation data, respectively.

A comprehensive analysis of the flood-susceptibility-related maps was conducted in a vector format. The factors (Table 1) from the maps were resampled into a 30 m $\times$ 30 m grid format with RESAMPLE tool in ArcGIS. Specifically, a bilinear interpolation which causes smoothing of the data

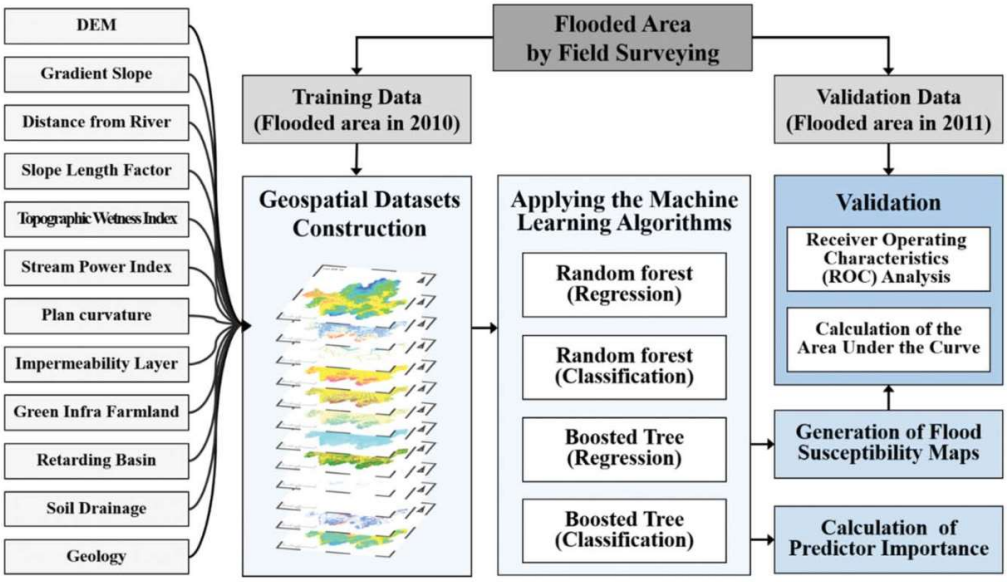

Figure 3. Flowchart for spatial prediction of flood susceptibility.

was used. It determines the resampled value from the four nearest input cell centre according to a weighted distance average of input cell centres. As a result, the total number of cells is 1,247,350, and the dimensions of the study-area grid are 1,010 rows by 1,235 columns. The numbers of cells of the flooded areas of 2010 and 2011 are 19,855 and 7,253, respectively. The training and validation data from the flooded areas were also adjusted using the same spatial resolution.

When the data from ASCII file finished converting into the STATISTICA format, the random-forest and boosted-tree models were applied in the programme. The flooded area of 2010 was used for the training data with the factors set as independent variables. In this step, all of the variables were divided into continuous data and categorical data before they were set up. The continuous variables contained within the DEM are distance from the river, TWI, SPI, slope gradient, SLF and plan curvature. The remaining categorical variables are geology, land use, soil drainage, retarding basin and green farmland.

With the predictor-importance values, the significance of each response variable could be computed. The predictor importance is calculated by summing up either the drop of the node impurity for the classification or the resubstitution estimate for the regression over all of the nodes. The predictor importance in this study is different from the variable importance that was used in a previous study (Breiman 2001), whereby the node-impurity values and the resubstitution-estimate values were summed in the classification and regression, respectively, for the actual split variables in each respective split; in this study, the predictor-importance values are built by summing up all of the predictors over all of the nodes.

After the application of the models on the training data from the 2010 flooded area, the validation was performed. The predicted flood-susceptible-area maps should represent the future flood-susceptibility areas effectually; therefore, the data from later flood events could be used for the certification rather than the training data. A validation set from the 2011 flooded area was used to perform the validation analysis. Eventually, the flood-susceptibility maps from the applied models were validated by comparing the known flooded areas of 2010 and 2011. In succession, a receiver operation curve (ROC) was organized with a calculation of the area under the curve (AUC); the calculated values from the result in the study area were sorted in a descending order to obtain the relative ranks for the patterns of each prediction result. Subsequently, 100 classes consisting of the accumulated 1% intervals were composed with all of the values (Lee 2006). The AUC was used for the evaluation of the prediction ability of the model, whereby a higher AUC represents a more-

effective prediction and the flood-possibility prediction accuracy of the applied model and the factors (Lee et al. 2012).

6. Result

The random-forest and boosted-tree models were used for the mapping of the flood susceptibility to observe the correlation between the flooded area and each factor that was extracted from the provided maps. To make a visual interpretation, the index is composed of five classes that are based on the area. For a visual and easy interpretation of the areas, the index areas were classified with the following five groups: very high, high, moderate, low and very low. Respectively, 10%, 10%, 20%, 20% and 40% are included in each group of area criterion that distinguishes susceptibility steps within the boundary of the study area. The predicted maps for flood susceptibility are delineated concretely by five categories. The data of the flooded areas of this study, which were acquired from the flood-survey data, covers a large extent. The predicted maps show aspects that are similar to those seen previously in the flooded areas (Figure 4).

Table 2 and Figure 6 show the predictor-importance values of each predictor variable in the analysis. The predictor importance comprises the importance rating on a 0 to 1 scale. The predictor variables can be distinguished to determine which of the variables that are used in the analysis could

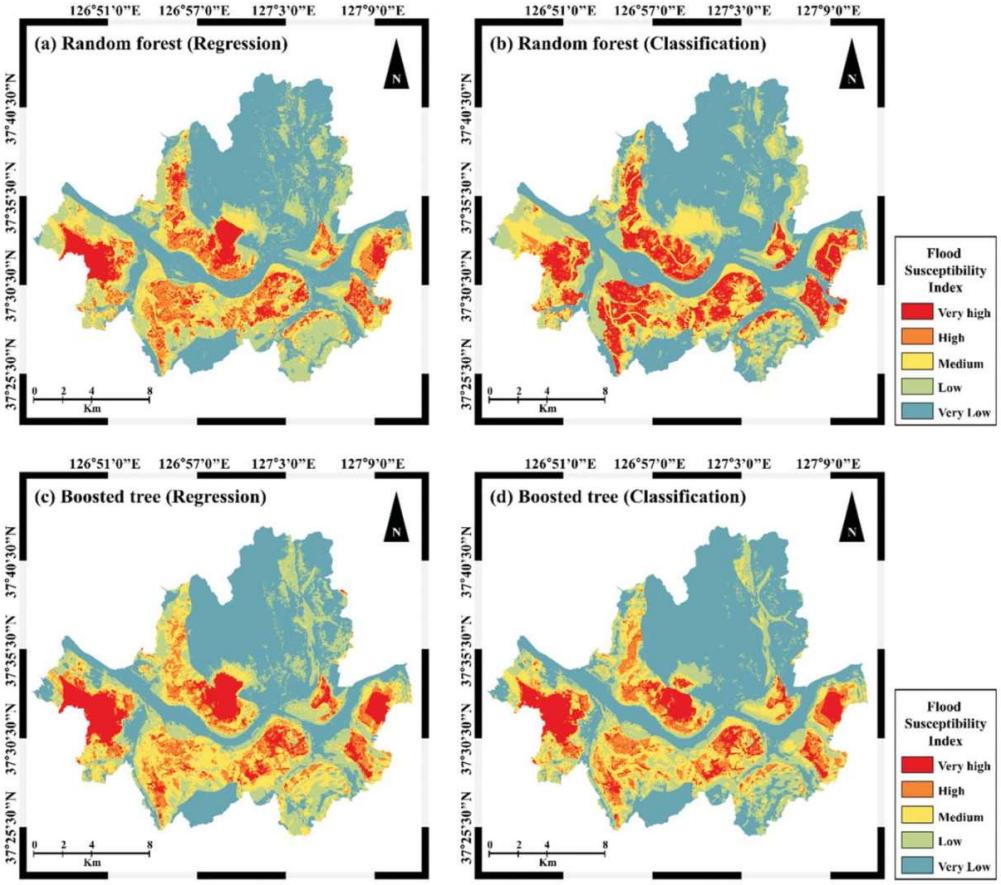

Figure 4. Flood-susceptibility map using (a) random-forest regression model, (b) random-forest classification model, (c) boosted-tree regression model and (d) boosted-tree classification model.

Table 2. Predictor importance.

	Random forests		Boosted trees	
	Regression	Classification	Regression	Classification
Distance from the river	1.000000[a]	0.887842[a]	1.000000[a]	1.000000[a]
Geology	0.613653[a]	1.000000[a]	0.618868[a]	0.938907[a]
DEM	0.383702[a]	0.889842[a]	0.787912[a]	0.800607[a]
Soil drainage	0.281214	0.544684	0.750739	0.886870
Land use	0.262450	0.519128	0.718148	0.792216
Slope gradient	0.193123	0.346925	0.518220	0.524749
SLF	0.146831	0.288270	0.331230	0.336169
SPI	0.116328	0.260325[b]	0.142089[b]	0.138259[b]
TWI	0.114760	0.301313	0.202899	0.204070
Green infra farmland	0.108932[b]	0.250822[b]	0.522558	0.449083
Plan curvature	0.108574[b]	0.275975	0.109680[b]	0.106436[b]
Retarding basin	0.009553[b]	0.026744[b]	0.040141[b]	0.026901[b]

[a]Top three variables of the predictor importances of related variables.
[b]Three bottom variables of the predictor importances of related variables.

make a major or a minor contribution to the prediction of the dependent variable of interest. In Table 2, the top three variables for both of the random-forest models are distance from the river, geology and DEM. For the boosted trees, soil drainage ranked instead of geology in the regression model and instead of DEM in the classification model; therefore, the area occupied by these variables

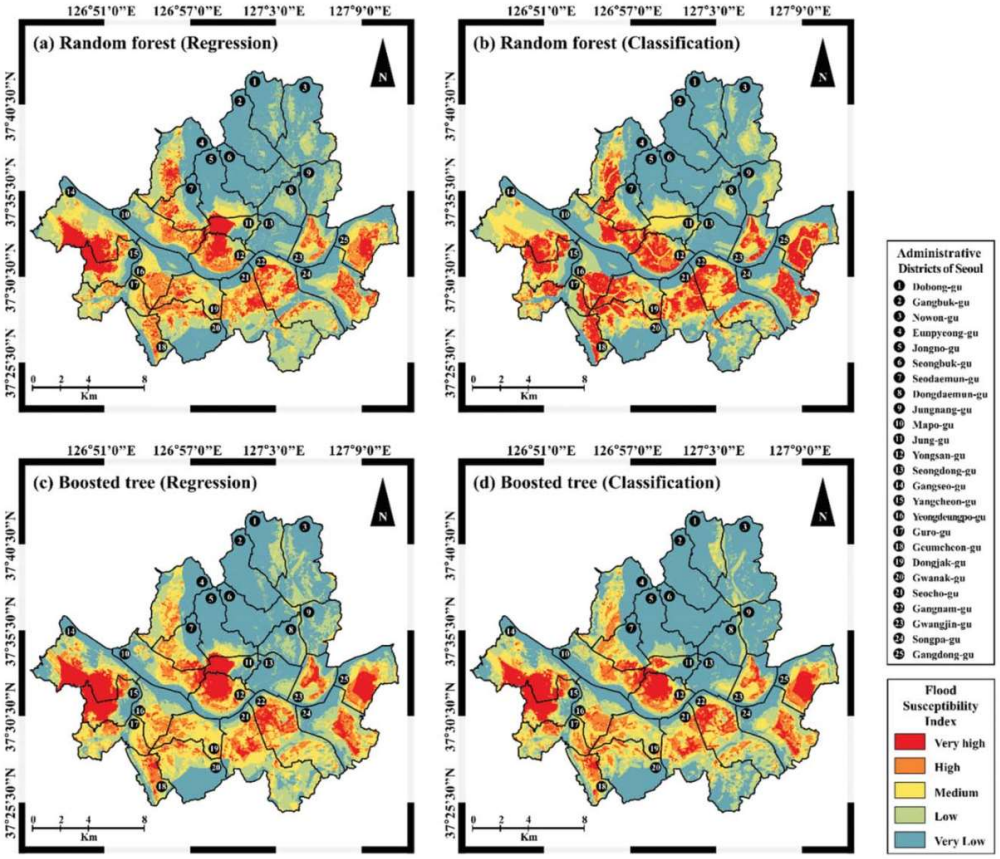

Figure 5. Flood-susceptibility map using (a) random-forest regression model, (b) random-forest classification model, (c) boosted-tree regression model and (d) boosted-tree classification model with administrative district of Seoul metropolitan city.

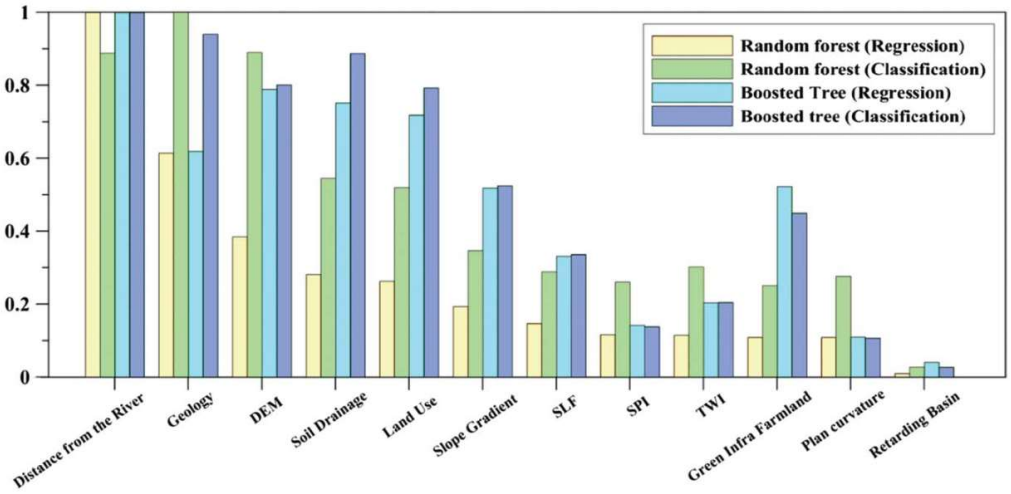

Figure 6. Bargraph of predictor importances.

shows a maximum susceptibility to flooding in the study area. Identically, the bottom variable is retarding basin in both models. SPI, green infra farmland and plan curvature also occupied the third-bottom group on three occasions. Figure 6 shows the relative importance of the predictor variables.

As a result, for the validation of the flood maps, the AUC is simply used. Figure 7 shows the validation rate of the predicted flood-susceptible areas, and the areas of the random-forest model showed a 78.78% (0.7878) accuracy in the regression model and a 79.18% (0.7918) accuracy in the classification model. As a consequence of the boosted-tree model, the regression model showed a 77.55% (0.7755) accuracy, while the classification model indicated a 77.26% (0.7726) accuracy.

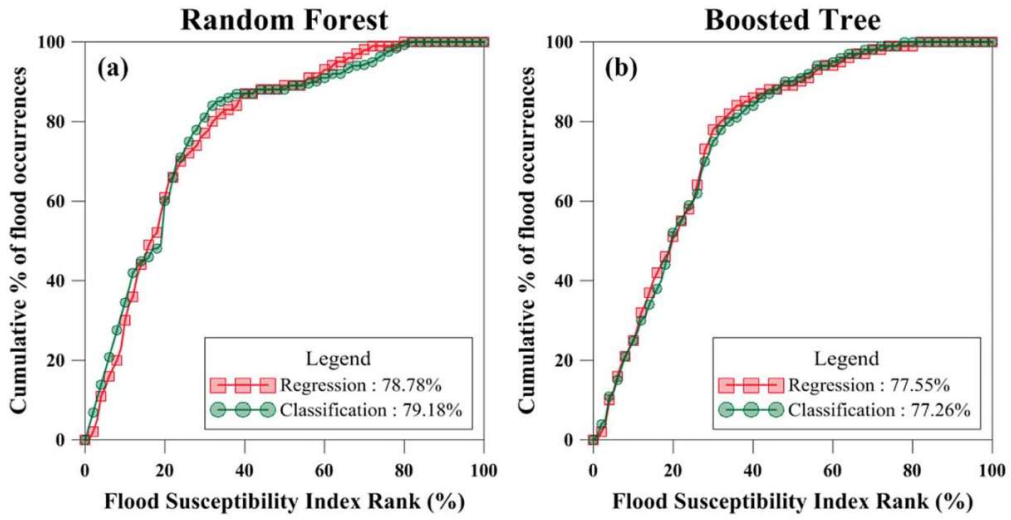

Figure 7. Cumulative frequency diagram showing flood susceptibility from (a) random-forest results and (b) boosted-tree results (*x*-axis) occurring in cumulative percent of flood occurrence (*y*-axis).

7. Conclusion and discussion

As a result of recent global warming, the world has suffered great losses due to unforeseen weather phenomena. The frequency and intensity of major floods increased into the twenty-first century. Flood risk should be managed effectively, especially for the extreme events that are set to become more frequent with the impact of climate change. The damages from floods led to the necessity of a flood-risk map, and flood-risk maps have actually been produced by governments; however, the previous generation of flood-risk maps is limited to the concept of risk only. A flood-susceptibility-map production that is based on sophisticated numerical results in conjunction with the vulnerability to flood disasters is absent.

The random-forest model and the boosted-tree model were therefore employed in this research for flood-susceptibility mapping. After selecting 12 variables that are related to flooding, the flood-susceptible areas were calculated according to the two models, whereby the relationships between particular flooded-area data and the related variables were demonstrated. Spatial datasets was set up with the factors that are relevant to flood occurrences, and the correlation is indicated. The maps were validated using the flooding data of 2011 after the flooding data of 2010 was used for training.

According to the results, in the southern area under the river, Gangdong-gu, Songpa-gu, Dongjak-gu, Yangcheon-gu, the southern parts of Gangnam-gu, Seocho-gu and Gwanak-gu are susceptible. Gangseo-gu and Guro-gu are also highly susceptible even though they are relatively lower than the other regions of the district that is south of the Han River (Figure 5). Contrastively, the northern region including Dobong-gu, Gangbuk-gu, Seongbuk-gu, Jongno-gu, Dongdaemun-gu and Seongdong-gu showed a low susceptibility to flooding. The area north of the Han River is less vulnerable to flooding due to its high elevation; however, the neighboring regions such as Mapo-gu, Yongsan-gu and Gwangjin-gu showed a high susceptibility.

Table 2 shows that the distance from the river, geology and DEM stand out as the most-important predictors. It is also distinct that the flood occurrence decreases with the increasing of the distance from the river. For geology, it is related to the alluvium area where the flood formed; accordingly, in Figure 2(a), the southeastern and western regions are covered by lowlands. This finding agrees with the flood data since most of the floods occurred on low-elevation regions. Alternatively, retarding basin is less important among the predictor variables, while the lesser importance of SLF, SPI, TWI and plan curvature are also confirmed. According to the outcome, it was assumed that these variables are from the same data, the DEM.

The flood-susceptibility maps that were generated using the random-forest model and the boosted-tree model were validated. Particularly, the random-forest curve shows that 60% of the floods are captured within 20% of the result maps, while 55% of the floods are captured within 20% of the result maps in the boosted-tree model. All of the models reached 100% with 80% of the result maps. Consequently, every result was considered as satisfactory when the accuracies (77% to 79%) are over 75% for every result, since a steeper slope at the beginning part of the curve shows a more-effective predictive reliability (Conforti et al. 2012).

The damages inflicted by floods continuously undergo a series of changes over time, and this imposes a limitation on a spatial analysis of flood inundation. Above all, the tidal structure of the inundation area could not be observed in real time effectively. The accurate value of the data that were used for the analysis is difficult to establish because it is spatially constructed through a survey that was conducted within the administrative-district system. Incorrect location information could cause substantial problems in a spatial analysis. Similarly, the water supply, drainage facilities and water-supply system might have a significant effect on flooding, but the facility data with the correct information is very difficult to obtain; therefore, additional data acquisition and further studies are required together with more-accurate data.

With GIS and remote sensing technology, other data could be used to make a decision, whereby the data are manipulated; for example, the classified remote sensing data from satellite images could be used (Adinarayana & Krishna 1996). Also, the large scales of the spatial and temporal data could

be evaluated using GIS techniques. Several of the factors that represent the surface characteristics are important in water management and they easily influence flood susceptibility; therefore, for the effective management of water, all of the information should be comprehensively built according to spatial datasets, whereby a multilateral effort is put into the solving of the problems related to water, in consideration of a combination of various studies.

The variables and methodology that are used in this study are applicable to flood-susceptibility mapping for other study areas. The flood-susceptibility maps could be used to identify the current flood susceptibility in each of the administrative districts. A rapid acquisition of the information of the potential target areas could provide evacuation information, and losses of life and property could be preventable. This information and the maps that are generated from it could therefore be applied to flood prevention and management. In addition, regarding quantitative data such as those used for the establishment of river master plans, land-use plans and flood-protection plans, the results of this study could be used to formulate a standard for value judgments. Also, continued research on a variety of factors that affect flooding and the new models is required.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM) funded by the Minister of Science, ICT and Future Planning of Korea. This research (NRF-2015R1A2A2A01005018) was also supported by the Mid-Career Researcher Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (MEST). This research was conducted by Korea Environment Institute (KEI) with support of a grant (16CTAP-C114629-01) from Technology Advancement Research Program (TARP) funded by Ministry of Land, Infrastructure and Transport of Korean government.

ORCID

Hyung-Sup Jung http://orcid.org/0000-0003-2335-8438
Saro Lee http://orcid.org/0000-0003-0409-8263

References

Adinarayana J, Krishna NR. 1996. Integration of multi-seasonal remotely-sensed images for improved landuse classification of a hilly watershed using geographical information systems. Int J Remote Sens. 17:1679–1688.

Ahmed OS, Franklin SE, Wulder MA, White JC. 2015. Characterizing stand-level forest canopy cover and height using landsat time series, samples of airborne LiDAR, and the random forest algorithm. ISPRS J Photogramm Remote Sens. 101:89–101.

Albers SJ, Déry SJ, Petticrew EL. 2016. Flooding in the Nechako River Basin of Canada: A random forest modeling approach to flood analysis in a regulated reservoir system. Can Water Res J/Revue canadienne des ressources hydriques. 41:250–260.

Belgiu M, Drăgut L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens. 114:24–31.

Beven K, Kirkby MJ. 1979. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrolog Sci J. 24:43–69.

Blanco-Vogt A, Schanze J. 2014. Assessment of the physical flood susceptibility of buildings on a large scale–conceptual and methodological frameworks. Nat Hazards Earth Syst Scis. 14:2105–2117.

Borrelli P, Märker M, Panagos P, Schütt B. 2014. Modeling soil erosion and river sediment yield for an intermountain drainage basin of the Central Apennines, Italy. Catena. 114:45–58.

Breiman L. 2001. Random forests. Mach learn. 45:5–32.

Bui DT, Pradhan B, Nampak H, Bui QT, Tran QA, Nguyen QP. 2016. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. J Hydrol. 540:317–330.
Cao C, Xu P, Wang Y, Chen J, Zheng L, Niu C. 2016. Flash flood hazard susceptibility mapping using frequency ratio and statistical index methods in coalmine subsidence areas. Sustainability. 8:948.
Chang HS, Chen TL. 2016. Spatial heterogeneity of local flood vulnerability indicators within flood-prone areas in Taiwan. Environ Earth Sci. 75:1484.
Chini M, Giustarini L, Matgen P, Hostache R, Pappenberger F, Bally P. 2014. Flood hazard mapping combining high resolution multi-temporal SAR data and coarse resolution global hydrodynamic modelling. Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium. New York (NY): IEEE.
Choi, S. 2010, Sep 24. 'Super Urban Flood' in the Heart of Seoul, swallowed in an instant like a guerrilla. The JoongAng Ilbo [Internet]. [cited 2016 Oct 10]. Available from: http://news.joins.com/article/4471618
Coltin B, McMichael S, Smith T, Fong T. 2016. Automatic boosted flood mapping from satellite data. Int J Remote Sens. 37:993–1015.
Conforti M, Robustelli G, Muto F, Critelli S. 2012. Application and validation of bivariate GIS-based landslide susceptibility assessment for the Vitravo river catchment (Calabria, south Italy). Nat Hazards. 61:127–141.
Curebal I, Efe R, Ozdemir H, Soykan A, Sönmez S. 2016. GIS-based approach for flood analysis: case study of Keçidere flash flood event (Turkey). Geocarto Int. 31:355–366.
De'ath G, Fabricius K. 2010. Water quality as a regional driver of coral biodiversity and macroalgae on the Great Barrier Reef. Ecol Appl. 20:840–850.
Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. J Anim Ecol. 77:802–813.
Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. 2014. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas. 35:2191.
Feng Q, Gong J, Liu J, Li Y. 2015. Flood mapping based on multiple endmember spectral mixture analysis and random forest classifier – The Case of Yuyao, China. Remote Sens. 7:12539–12562.
Feng Q, Liu J, Gong J. 2015. Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier – a case of Yuyao, China. Water. 7:1437–1455.
Friedman N, Getoor L, Koller D, Pfeffer A. 1999. Learning probabilistic relational models. Proceedings of the 1999 International Joint Conferences on Artificial Intelligence. San Francisco (CA): Morgan Kaufmann Publishers.
Friedman N, Goldszmidt M, Wyner A. 1999. Data analysis with Bayesian networks: A bootstrap approach. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. Burlington (MA): Morgan Kaufmann Publishers.
Grimaldi S, Petroselli A, Arcangeletti E, Nardi F. 2013. Flood mapping in ungauged basins using fully continuous hydrologic–hydraulic modeling. J Hydrol. 487:39–47.
Hastie R. 2001. Problems for judgment and decision making. Ann Rev Psychol. 52:653–683.
Ho TK. 1995. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. New York (NY): IEEE.
Jo, H. 2011, Aug 12. Is it a Disaster or Resources. The Hankyoreh [Internet]. [cited 2016 Oct 10]; Available from: http://h21.hani.co.kr/arti/special/special_general/30216.html
Khosravi K, Pourghasemi HR, Chapi K, Bahri M. 2016. Flash flood susceptibility analysis and its mapping using different bivariate models in Iran: a comparison between Shannon's entropy, statistical index, and weighting factor models. Environ Monitor Assess. 188:656.
Kim D, Jung HS, Baek W. 2016. Comparative analysis among radar image filters for flood mapping. J Korean Soc Surv Geodesy Photogrammetry and Cartography. 34:43–52.
Klaus S, Kreibich H, Merz B, Kuhlmann B, Schröter K. 2016. Large-scale, seasonal flood risk analysis for agricultural crops in Germany. Environ Earth Sci. 75:1289.
Kleinberg EM. 2000. On the algorithmic implementation of stochastic discrimination. IEEE Trans Pattern Anal Mach Int. 22:473–490.
Lawal DU, Matori AN, Yusuf KW, Hashim AM, Balogun AL. 2014. Analysis of the flood extent extraction model and the natural flood influencing factors: a GIS-based and remote sensing analysis. In: IOP Conference Series: Earth and Environmental Science (Vol. 18, No. 1). Bristol: IOP Publishing; p. 012059.
Leathwick J, Elith J, Chadderton W, Rowe D, Hastie T. 2008. Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non−diadromous fish species. JBiogeogr. 35:1481–1497.
Lee MJ, Kang JE, Kim G. 2015. Application of fuzzy combination operators to flood vulnerability assessments in Seoul, Korea. Geocarto Int. 30:1052–1075.
Lee S, Lee CW. 2015. Application of decision-tree model to groundwater productivity-potential mapping. Sustainability. 7:13416–13432.
Lee S. 2014. Geological Application of Geographic Information System. Daejeon: International School for Geoscience Resources of Korea Institute of Geoscience and Mineral Resources. p. 143.

Lee S, Song KY, Oh HJ. Choi, J. 2012. Detection of landslides using web-based aerial photographs and landslide susceptibility mapping using geospatial analysis. Int J Remote Sens. 33:4937–4966.
Lee S, Sambath T. 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ Geology. 50:847–855.
Liu YB, Smedt FD. 2005. Flood modeling for complex terrain using GIS and remote sensed information. Water Resour Manage. 19:605–624.
Lopatin J, Dolos K, Hernández H, Galleguillos M, Fassnacht F. 2016. Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. Remote Sens Environ. 173:200–210.
Marconi M, Gatto B, Magni M, Marincioni F. 2016. A rapid method for flood susceptibility mapping in two districts of Phatthalung Province (Thailand): present and projected conditions for 2050. Nat Hazards. 81:329–346.
Masood M, Takeuchi K. 2012. Assessment of flood hazard, vulnerability and risk of mid-eastern Dhaka using DEM and 1D hydrodynamic model. Nat Hazards. 61:757–770.
Moore ID, Grayson R, Ladson A. 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol Process. 5:3–30.
Morelli S, Battistini A, Catani F. 2014. Rapid assessment of flood susceptibility in urbanized rivers using digital terrain data: Application to the Arno river case study (Firenze, northern Italy). Appl Geog. 54:35–53.
Muchlinski D, Siroky D, He J, Kocher M. 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Anal. 24:87–103.
Naghibi SA, Pourghasemi HR, Dixon B. 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monitor Assess. 188:44.
Novelo-Casanova DA, Rodríguez-Vangort F. 2016. Flood risk assessment. Case of study: Motozintla de Mendoza, Chiapas, Mexico. Geomat Nat Haz Risk.7:1538–1556.
Nylén T, Hellemaa P, Luoto M. 2015. Determinants of sediment properties and organic matter in beach and dune environments based on boosted regression trees. Earth Surface Process Landforms. 40:1137–1145.
Ohlmacher GC, Davis JC. 2003. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Eng Geolog. 69:331–343.
Olden JD, Lawler JJ, Poff NL. 2008. Machine learning methods without tears: a primer for ecologists. The Quarterly Rev biol. 83:171–193.
Papaioannou G, Loukas A, Georgiadis C. 2013. The effect of riverine terrain spatial resolution on flood modeling and mapping. Proceedings of the First International Conference on Remote Sensing and Geoinformation of Environment. Bellingham: International Society for Optics and Photonics.
Pradhan B. 2010. Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. J Spatial Hydrol. 9:1–18.
Rahmati O, Pourghasemi HR, Zeinivand H. 2016a. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. Geocarto Int. 31:42–70.
Rahmati O, Zeinivand H, Besharat M. 2016b. Flood hazard zoning in Yasooj region, Iran, using GIS and multi-criteria decision analysis. Geomat Nat Haz Risk. 7:1000–1017.
Regmi NR, Giardino JR, Vitek JD. 2013. Hazardousness of a Place. In: Encyclopedia of Natural Hazards. Heidelberg: Springer. p. 435–447.
Seekao C, Pharino C. 2016a. Assessment of the flood vulnerability of shrimp farms using a multicriteria evaluation and GIS: a case study in the Bangpakong Sub-Basin, Thailand. Environ Earth Sci. 75:1–13.
Seekao C, Pharino C. 2016b. Key factors affecting the flood vulnerability and adaptation of the shrimp farming sector in Thailand. Int J Disaster Risk Reduct. 17:161–172.
Sowmya K, John CM, Shrivasthava NK. 2015. Urban flood vulnerability zoning of Cochin City, southwest coast of India, using remote sensing and GIS. Nat Haz. 75:1271–1286.
Tan L, Scarton C, Specia L, van Genabith J. 2016. Saarsheff at semeval-2016 task 1: semantic textual similarity with machine translation evaluation metrics and (extreme) boosted tree ensembles. Proceedings of SemEval. Stroudsburg: Association for Computational Linguistics. p. 628–633.
Tehrany MS, Pradhan B, Mansor S, Ahmad N. 2015. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena. 125:91–101.
Tehrany MS, Lee MJ, Pradhan B, Jebur MN, Lee S. 2014. Flood susceptibility mapping using integrated bivariate and multivariate statistical models. Environ Earth Sci. 72:4001–4015.
Tehrany MS, Pradhan B, Jebur MN. 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol. 504:69–79.
Vojtek M, Vojteková J. 2016. Flood hazard and flood risk assessment at the local spatial scale: a case study. Geomat, Nat Haz Risk. 7:1973–1992.
Wang Y, Li Z, Tang Z, Zeng G. 2011. A GIS-based spatial multi-criteria approach for flood risk assessment in the Dongting Lake Region, Hunan, Central China. Water Resour manage. 25:3465–3484.

Wang Z, Lai C, Chen X, Yang B, Zhao S, Bai X. 2015. Flood hazard risk assessment model based on random forest. J Hydrol. 527:1130–1141.

Wischmeier WH, Smith DD. 1978. Predicting rainfall erosion losses – a guide to conservation planning. Hyattsville (MD): USDA, Science and Education Administration.

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM. 2016a. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides. 13:839–856.

Youssef AM, Pradhan B, Sefry SA. 2016b. Flash flood susceptibility assessment in Jeddah city (Kingdom of Saudi Arabia) using bivariate and multivariate statistical models. Environ Earth Sci. 75:12.

Yu, S, Park H, Lim J. 2011, Jul 28. Disaster Prevention Requires a New Paradigm "Mandatory Detention". The Hankyoreh [Internet]. [cited 2016 Oct 10]. Available from: http://www.hani.co.kr/arti/society/society_general/489527.html

Quinlan JR. 1986. Induction of decision trees. Mach Learn. 1:81–106.