MASTER'S THESIS OF NATURAL SCIENCE

Estimation of Potential for Mineral Water Using Tree Ensemble Machine Learnings on National Scale

트리 앙상블 기계학습을 이용한 전국 규모의 광천수 부존위치 추정

February 2021

Graduate School of Seoul National University
College of Natural Sciences
School of Earth and Environmental Sciences

Hye-Lim Lee

Estimation of Potential for Mineral Water Using Tree Ensemble Machine Learnings on National Scale

Submitting a master's thesis of Natural Science

January 2021

Graduate School of Seoul National University
College of Natural Science
School of Earth and Environmental Sciences

Hye-Lim Lee

Confirming the master's thesis written by Hye-Lim Lee

January 2021

Chair	김덕진	**(Seal)**
Vice Chair	이강근	**(Seal)**
Examiner	이상묵	**(Seal)**

ABSTRACT

Identifying groundwater potential zone in consideration of water quality is important to fulfill the increasing demand for drinking groundwater. Especially, reasonable and accurate detection of the potential zone is essential for efficient management of the groundwater resources and successful development of the groundwater for its drinking usage. This study covered mineral water among various drinks and estimated the spatial distribution of groundwater potential in the South Korea, using tree ensemble methods such as Boosted Regression Trees, Random Forest, and Extremely Randomized Trees. Total 6,135 groundwater quality data were collected on a nationwide scale to determine response variable. Environmental factors such as altitude, slope, drainage grade, effective soil depth, soil texture, land use, and hydrogeology were served as predictor variables. In precision recall curve analysis, all curves of the three methods were clearly distinguished from a baseline, which confirmed the applicability of the three classifiers to potential mapping. In addition, relative influence and partial dependence plot identified that factors related to contamination highly affected the modeling and potential mapping. Finally, the three validated models generated the spatial distribution of mineral water potential. The maps well reflected the distribution of imbalanced data and the results of variable influence for some predictors. Moreover, proper location for mineral water development could be determined by comparing the three maps. Consequently, the tree ensemble methods are promising in delineation of the groundwater potential zone for mineral water in national scale, and the analysis of the variable influence enable the extraction of the environmental conditions affecting the mineral water production.

Key words: groundwater potential map, mineral water, boosted regression trees, random forest, extremely randomized trees, variable influence

Student Number: 2019-20501

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Research Background

Groundwater is a valuable natural resource of freshwater for drinking, agricultural, and domestic uses, because it has widespread and continuous availability, excellent natural quality, and drought reliability (Todd and Mays, 2005). In particular, as the demand for high quality of drinking water increases, the extensive utility of groundwater for drinking purposes is expected. However, serious concerns have been raised over the contaminating groundwater quality due to the release and migration of pollutants from surface sources, seawater intrusion, radioactive materials and so on (Ikem et al., 2002; Kouzana et al., 2009; Khedr, 2013). Therefore, systematic groundwater management is necessary to retain groundwater quality, and targeting, monitoring and conserving groundwater potential zone are crucial for the management (Chowdhury et al., 2009).

As a part of the groundwater quality management, several studies have been conducted on the suitability assessment of groundwater for drinking purpose in some regions (Gowd, 2005; Peiyue et al., 2011; Al-Tabbal and Al-Zboon, 2012). Their groundwater quality monitoring and evaluation are very reliable and standard methods for drinking purpose, but it is only effective and applicable after the groundwater developed and pumped out. Such an approach for groundwater exploration is very time-consuming, costly, and labor intensive in terms of test drilling and water sampling (Sander et al., 1996; Chowdhury et al., 2009). Therefore, for the proper utilization and management of the groundwater resource,

it is important to make scientific decisions for the groundwater development suitable for drinking purposes.

Most of research related to groundwater quality mapping for drink use employed interpolation method (Anbazhagan and Nair, 2004; Kord and Moghaddam, 2014; Saha et al., 2018). Using observed groundwater quality data, it generates the ion contours where there is no quality data. Produced water quality maps were used to classify areas as undesirable or desirable for drinking water in accordance with WHO guideline. However, those methods cannot take account of any other factors that are related to target of the interpolation, but just distance of sample points. They encounter a problem of the uncertainty in the ion concentration distribution as the distance between groundwater wells increases.

In recent years, as another method for groundwater potential mapping, machine learning has been successfully applied (Naghibi et al., 2016; Rahmati et al., 2016; Lee et al., 2018). Because it trains the relationship between predictor variables and response, the estimated result can be independent of the uncertainty problem if the predictor variables are given. However, most of these studies have focused on the quantitative aspects of groundwater such as the spring location and groundwater productivity. Since the groundwater development for potable use must consider the water quality, it can be a relatively new subject to find groundwater potential zone for a particular drink purpose using machine learning. Among different kinds of beverage or drinking water, this study targeted mineral water, which is usually called as bottled water containing high total dissolved solid (TDS) concentration relative to tap water.

Groundwater quality is significantly influenced by environments conditions such as geological formation and anthropogenic activities (Zulu et al., 1996). Because the quality is determined by interaction of the environmental conditions along the groundwater evolution paths, it is indispensable to consider interactions between these factors when estimating groundwater potential for mineral water. Therefore, tree-based machine learnings which have ability to accommodate different types of predictor variables and to facilitate for “fitting” interactions between predictors can be a powerful tool in this research (Freidman and Meulman, 2003).

Moreover, the tree-based methods have advantage in calculating variable importance using well established methods such as mean decrease impurity importance and mean decrease accuracy (Louppe et al., 2013). Many studies have analyzed which predictor variables significantly influence on their model response, and demonstrated the practical utility of these importance measures in various groundwater studies (Naghibi et al., 2016; Knoll et al., 2019; Knierim et al., 2020). In this study, investigation of variable influence would help understanding which environmental variables influentially explain the estimation model for mineral water potential.

1.2 Objectives and Scope

The main purpose of this study is to estimate potential indicating suitability of the groundwater for mineral water using three tree ensemble machine learnings, and to assess the influence of environmental factors to response. To accomplish the main purpose, following four phases would progress systematically.

At first, the tree ensemble methods were applied to train the relationship between the environmental factors and groundwater quality which is converted to suitability for mineral water. This step should take precedence, prior to finding proper locations for mineral water without the water quality data. Several parameters in each method were combined and then tuned to derive three optimal ensemble models.

Secondly, the evaluation of the constructed models and the comparison of the results between the three models were conducted, considering the data imbalance and the objective of this research. Applied evaluation methods could help to verify the availability of the models and to assess model performance more reasonably than simply calculating the accuracy of the models.

Thirdly, variable influence of predictors representing the environmental factors was analyzed in the three ensemble models. This process was performed to examine the influence of each environmental factor on the model performance, response variations, and determination of the response. Interpretation of variable influence would help increase the reliability of the models and understand which environmental conditions are necessary to find location of groundwater for mineral

water, and moreover, which conditions have highly positive effects on the groundwater quality.

At last, three potential maps showing the spatial distribution of potential for mineral water were suggested by each method. In location without water quality data, prediction value indicating the suitability for mineral water could be estimated using trained models and the environmental factors. Determination of the location for suitable mineral water would be possible by plotting and comparing the potential maps.

2 METHODOLOGY

2.1 Tree-Based Ensemble Machine Learning

This study applied tree-based ensemble machine learnings to estimate the spatial distribution of groundwater potential for mineral water. Tree-based ensemble methods usually combine Classification and Regression Tree (CART) and ensemble techniques such as boosting and bagging. The tree ensemble methods applied in this study were Boosted Regression Trees (BRT), Random Forests (RF), and Extremely Randomized Trees (ERT).

2.1.1 Classification and Regression Tree (CART)

CART method (Breiman et al., 1984) segments the predictive space into smaller regions showing the most consistent responses to predictor variables (Fig. 2-1 (a)). For classification problems, the most major class in a region represents the district, and the average response of observations in a region is assigned to that sector when it is regression problems (Elith et al., 2008). In this study, it could be a classification problem because there are only two responses of suitable (1) or unsuitable (0) for mineral water, but it was performed as a regression problem for the following reasons. First, because classification problems only result in 0 or 1 response to data, it is difficult to assess the model performance in detail with graphical evaluation methods. Second, for the same reason, there is a limit to describe the change in response as the variable values change.

In CART, a region can be partitioned into smaller regions (R_1, R_2) by selecting predictor variables (X_j) and split-points (s) (Eq. 2-1) minimizing residual sum of

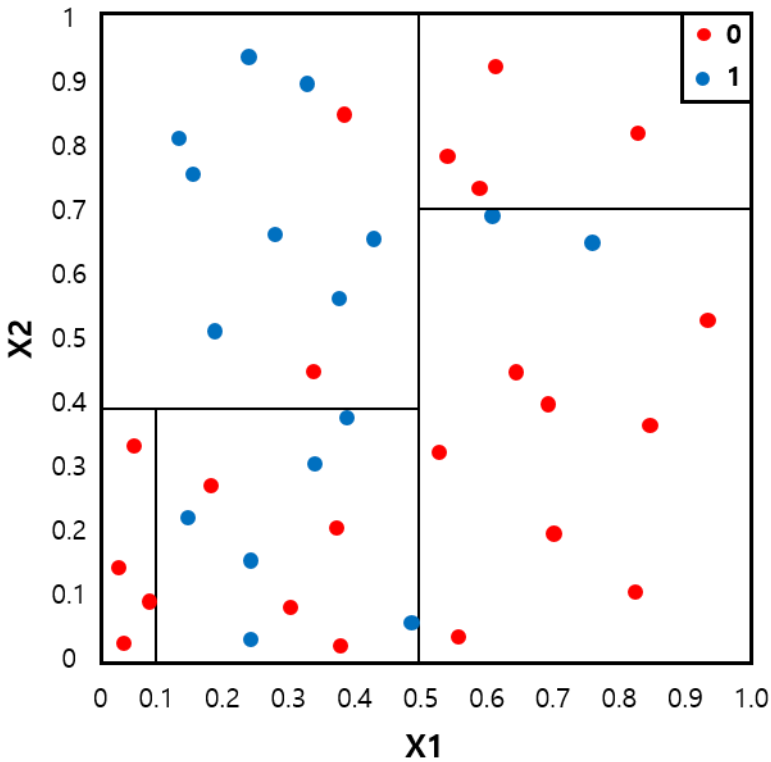
square (RSS) for regression problems (Eq. 2-2). Therefore, a predictor variable and a split-point can determine one split line as shown in Fig. 2-1.

$$R_1(j,s) = \{X|X_j < s\} \; and \; R_2(j,s) = \{X|X_j \geq s\} \qquad \text{(Equation 2-1)}$$

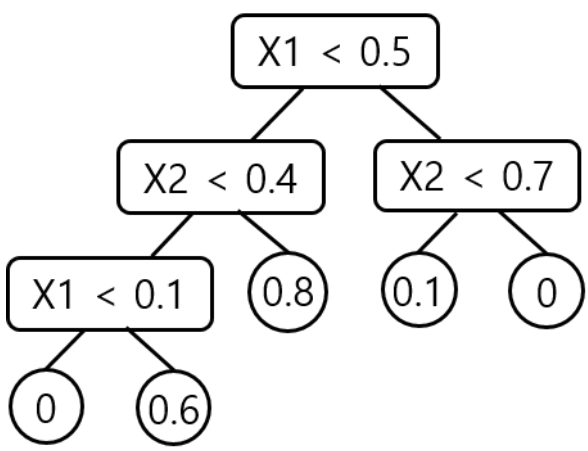
$$\sum_{i:\, x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:\, x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \qquad \text{(Equation 2-2)}$$

where x_i is the given observation, y_i is the response of the given observation, $\hat{y}_{R_1}$ is the averaged response for the training observations in $R_1(j,s)$ and $\hat{y}_{R_2}$ is the averaged response for the training observations in $R_2(j,s)$.

CART method is popular because it is intuitive, easy to visualize, and useful for interpretation. It has merits that various type of data (e.g. numeric, categorical, binary) can be used as input variables and responses, and complex interactions between inputs and responses can be modeled simply. Moreover, CART is insensitive to outliers or difference in scale between predictor variables (Breiman et al., 1984; Steinberg and Colla, 1995). However, its performance is poor compared to other machine learning methods, having difficulty in making smooth models, and overall, it has a possibility of overfitting when making a large tree for higher accuracy (De'Ath, 2007). To overcome these weaknesses, tree-based ensemble methods generating more powerful prediction models were applied in this study.



(a)



(b)

Figure 2-1. Schematic diagram of a CART method. (a) Predictor space composed of input variables (X1 and X2) and response variable (0 or 1) is partitioned with split lines. (b) Tree diagram of (a) described with decision and terminal nodes.

2.1.2 Boosted Regression Trees (BRT)

BRT (Friedman, 2001) is a type of additive model utilizing boosting as an ensemble technique. It sequentially generates multiple 'weak classifier', which is a single regression tree in the BRT method, to attain better predictive performance and to avoid overfitting. For BRT model, the first regression tree is modeled in a way of reducing the loss function maximally up to the selected tree size. The next following regression trees are fitted to the residuals of the previous trees (Elith et al., 2008). By reducing the residuals of the previous trees, the BRT model continuously tries to improve its performance as presented in Fig. 2-2.

BRT has two important parameters of tree complexity (*tc*) and learning rate (*lr*). The *tc* determines the number of splits in each tree so makes it possible to control interactions between predictor variables (Fig. 2-3). A *tc* of 1 generates trees composed of 1 split, which indicates the model does not consider interactions between input variables. The *lr* determines the contribution of each tree to the growing model, and higher *lr* can fit complete BRT model faster but cause instability (Elith et al., 2008). These two parameters determine the number of tree (*nt*) that are required for optimal outcome. By combining *tc* (1~5) and *lr* (0.01, 0.005, 0.001) and by comparing their results, parameters for optimal BRT model were induced in this study.

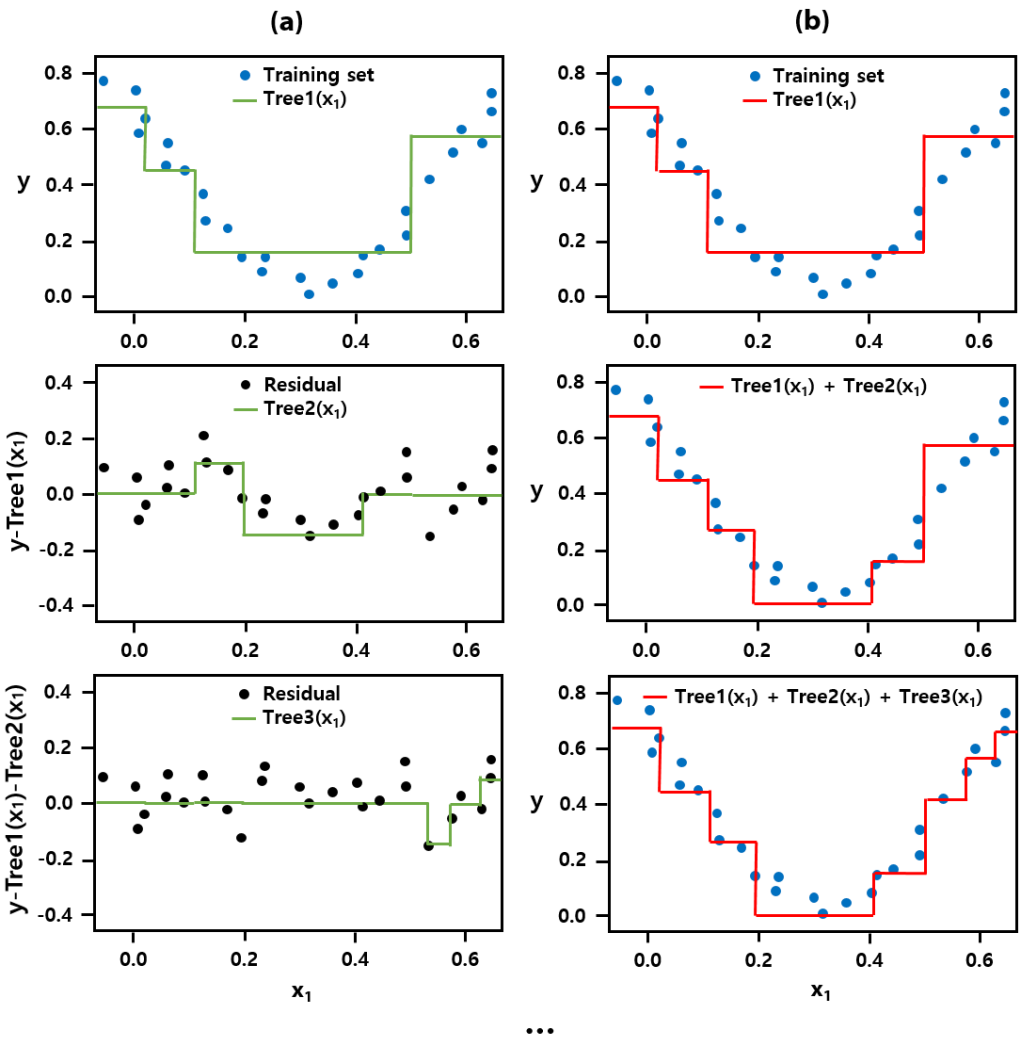


Figure 2-2. Process of improving model performance by sequentially generating trees in the BRT method. (a) The following tree is generated by focusing on fitting the residuals of the previous trees. (b) Final BRT model is an ensemble of the produced trees.

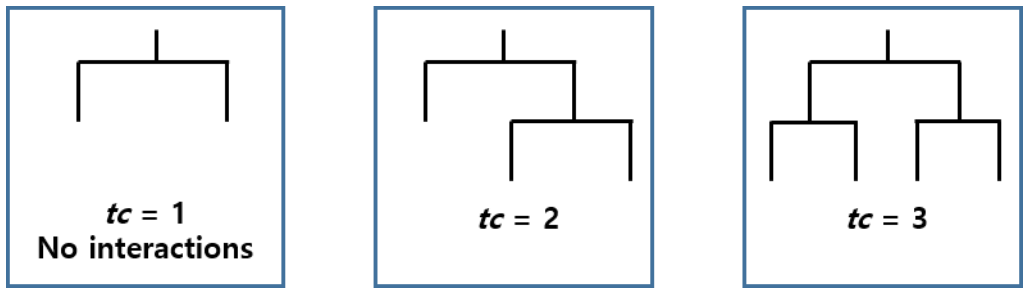


Figure 2-3. Tree complexity (*tc*) determining the number of splits in each tree controls interactions between predictor variables.

2.1.3 Random Forests (RF)

RF (Breiman, 2001) uses bootstrap aggregating (bagging) as an ensemble technique. Bagging repeatedly selects several random samples with replacement in training data and generates trees that fit to selected random samples (Fig. 2-4). This process improves predictive performance because it reduces the variance of the model by adopting different training data and by de-correlating each tree. RF method takes a random subset of whole predictor variables as candidates when adding a split line, while the general tree-based model takes whole predictor variables as candidates for a split line (Fig. 2-5). This process used in RF method has de-correlating effect between the trees (Liaw and Wiener, 2002; James et al., 2013). If one or few input variables highly influence to response variable, these predictors will be frequently selected for nodes, causing trees to become correlated. Application of a random subset of predictors can prevent lopsided selection, so it is able to solve the tree correlation problem. Final prediction value of RF is the average of all predictions in each tree.

RF has two important parameters of the number of trees in the forest (*ntree*) and the number of random predictor variables in each tree (*mtry*). Optimal value for *mtry* among 1 ~ 15 was chosen to minimize the root mean square error (RMSE) first. After generating 10,000 trees for RF model with the optimal *mtry*, *ntree* satisfying the minimal error was determined.

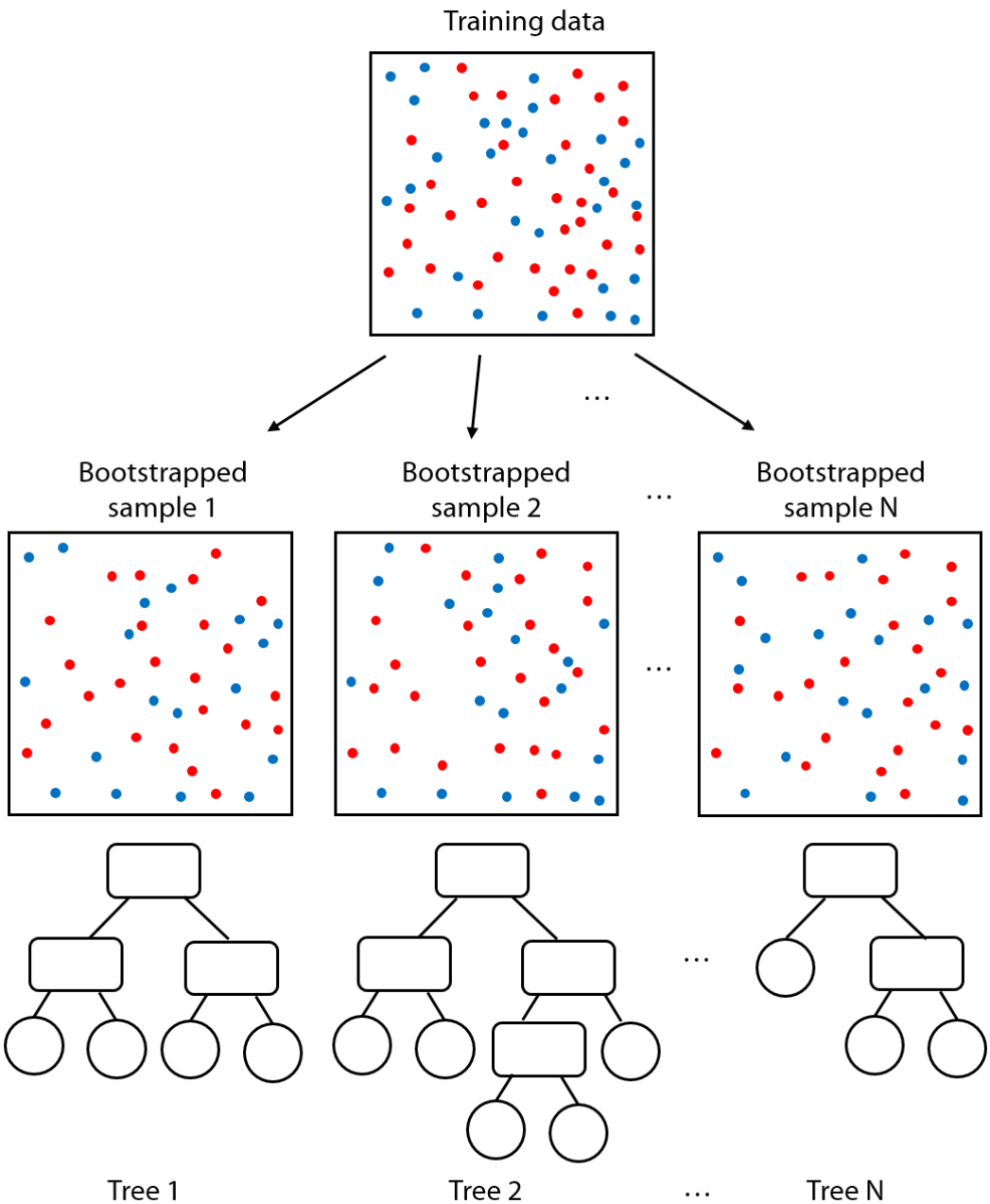


Figure 2-4. Schematic diagram of bagging. Bagging repeatedly selects several random samples with replacement in training data and generates trees that fit to selected random samples.

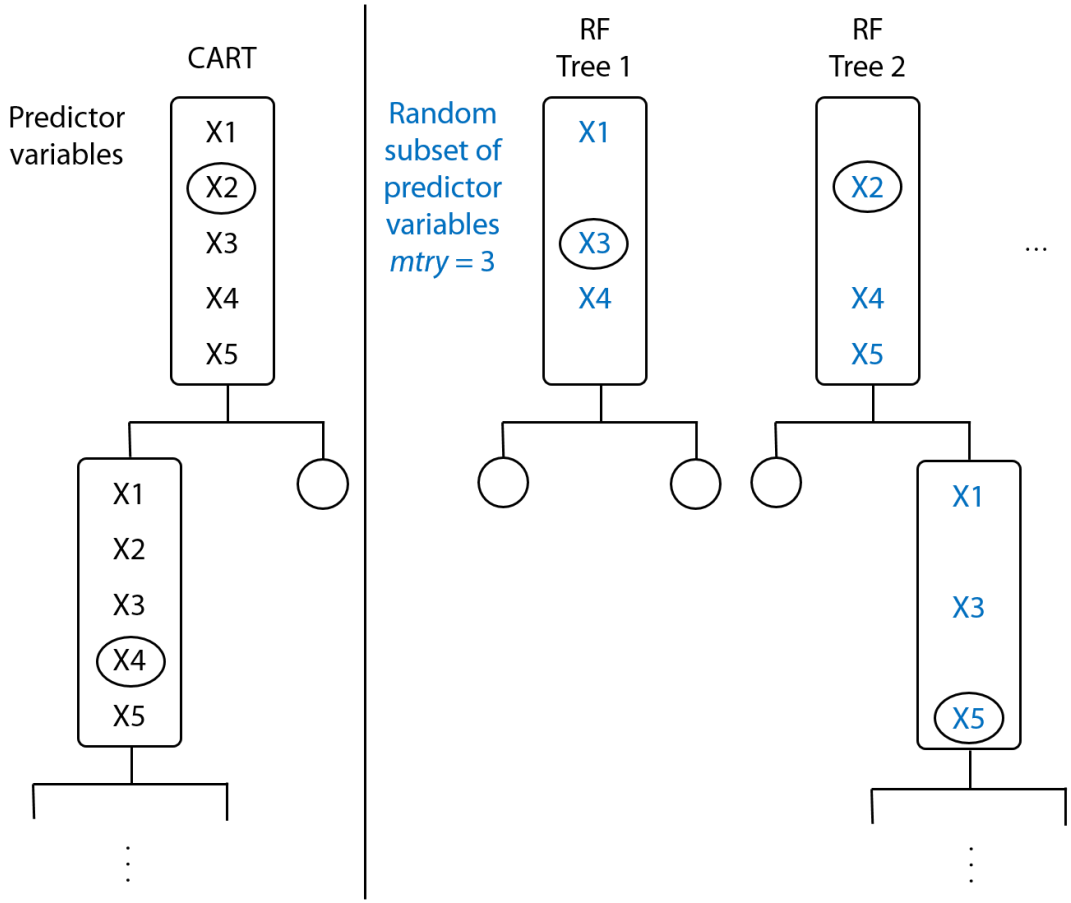


Figure 2-5. Comparative diagram of CART (left) and RF (right) methods. A random subset of whole predictor variables is taken as candidate when partitioning the predictor space in RF method, while CART method takes whole predictor variables as candidates for a split line.

2.1.4 Extremely Randomized Trees (ERT)

ERT (Geurts et al., 2006) is similar to RF method, in that it employs a random subset of predictors when adding split lines and averages outputs of all individual trees for the final model result. However, there are two differences: first, individual tree is modeled using whole learning samples rather than using the bootstrapped samples, and second, splitting values for split lines are additionally randomized in ERT. In ERT, randomness continues to accumulate by adding random split-points to a random subset of predictors. A fixed number (*numRandomcut*) of split-points are generated randomly for each random predictor and the best split line satisfying minimal RSS is determined among them (Fig. 2-6). This extreme randomness allows the model to reduce the variance of the results.

ERT has three important parameters of the number of random predictor variables in each tree (*mtry*), the number of random split-points for each predictor variable (*numRandomcut*), and the number of trees (*ntree*). Default value of *ntree* (500) was used, and combinations of *mtry* (1~15) and *numRandomcut* (1~5) were tried to find optimal ERT model, and parameters were determined by trial and error method finally.

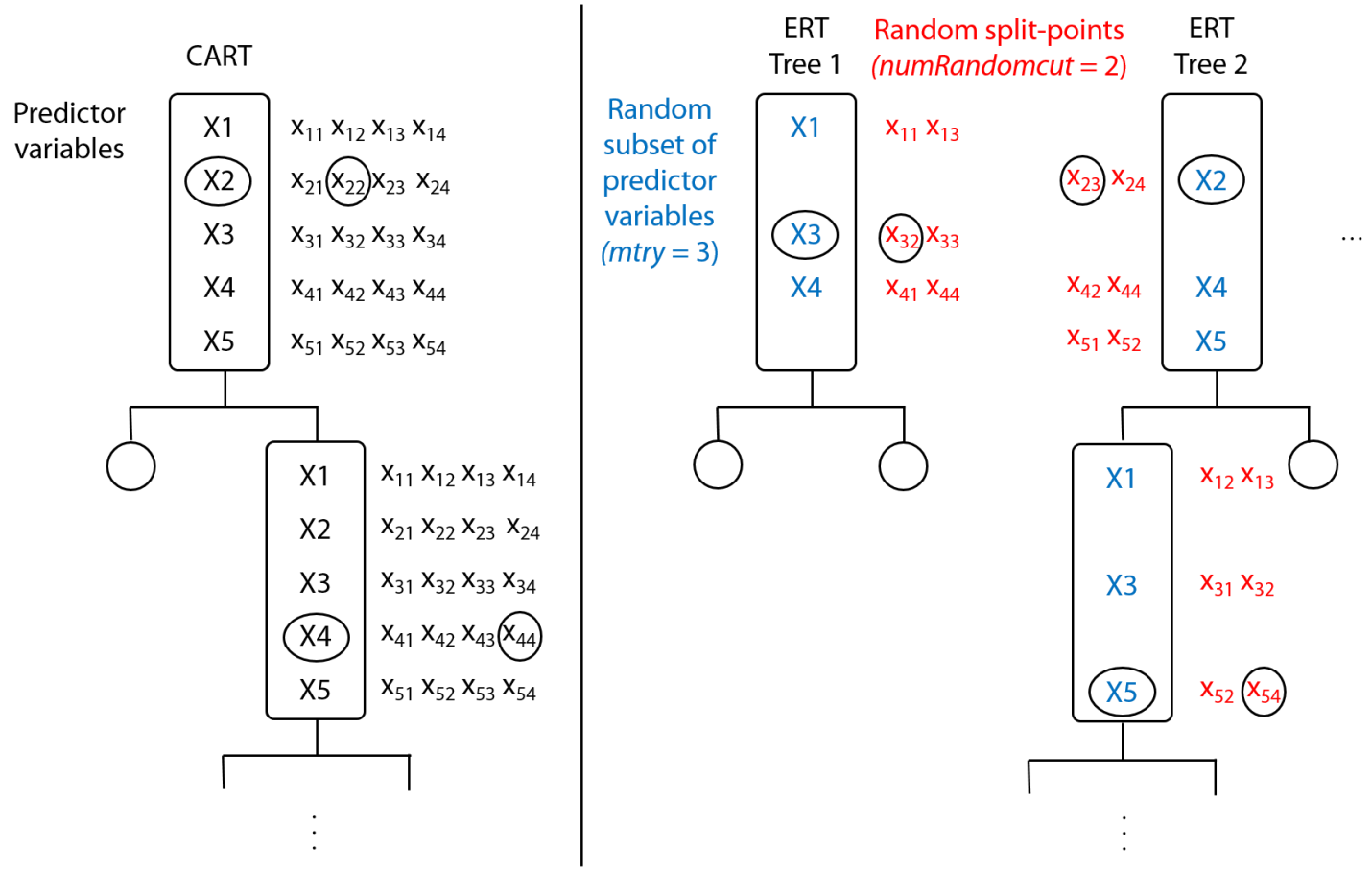


Figure 2-6. Comparative diagram of CART (left) and ERT (right) methods. Random split-points as well as a random subset of whole predictor variables are introduced in ERT method.

2.2 Data Processing

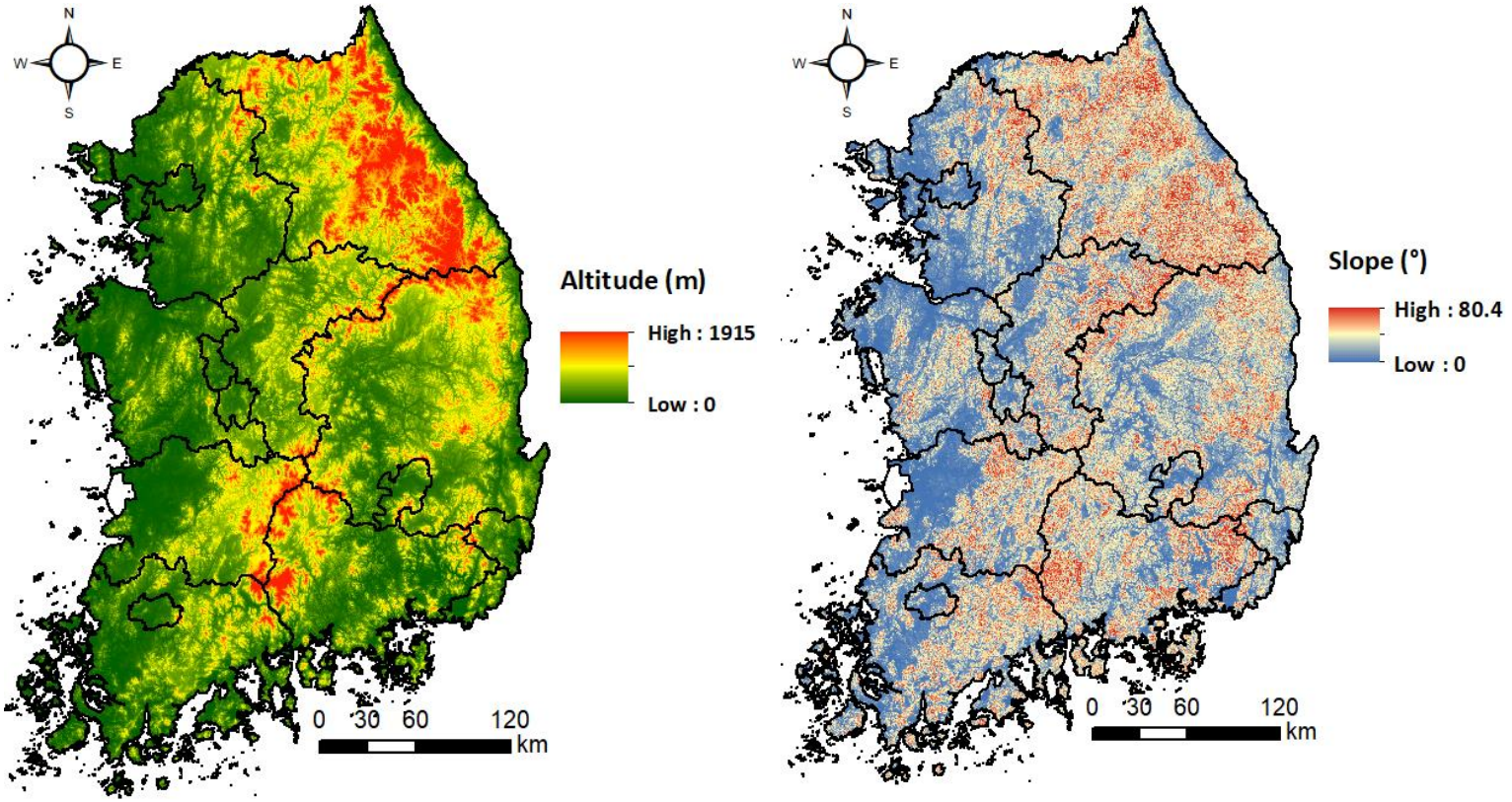
2.2.1 Data Collection

National-scale of groundwater quality data used for determining response variables were provided by the National Groundwater Information Management & Service Center (GIMS), the Korea Rural Community Corporation (KRCC), and the National Institute of Environmental Research (NIER). The water quality information included well ID, XY coordinates, concentration of eight major ions (Na^+, K^+, Ca^{2+}, Mg^{2+}, SO_4^{2-}, NO_3^-, Cl^-, and HCO_3^-), concentration of total dissolved solid (TDS), and aquifer type (i.e., bed rock or alluvial).

Groundwater quality is affected by natural and anthropogenic influence, and therefore, the composition of groundwater is dependent on geological, topographical, hydrological and biological factors (Khatri and Tyagi, 2015). Based on reference review, the following factors were identified as important layers influencing the spatial distribution of groundwater quality: altitude, slope, drainage grade, effective soil depth, soil texture, land use, and hydrogeology (Grootjans et al., 1988; Lerner and Harris, 2009; Igboekwe and Akankpo, 2011; Akankpo and Igboekwe, 2012; Thivya et al., 2013; Giri et al., 2017). Geographic Information System (GIS) based data of them were collected to be used as predictor variables in tree models. Data for altitude and slope are given as numerical value, and type of the others is categorical data. Detail information of the variables are described in Table 2-1 and maps for each GIS data are given in Fig. 2-7.

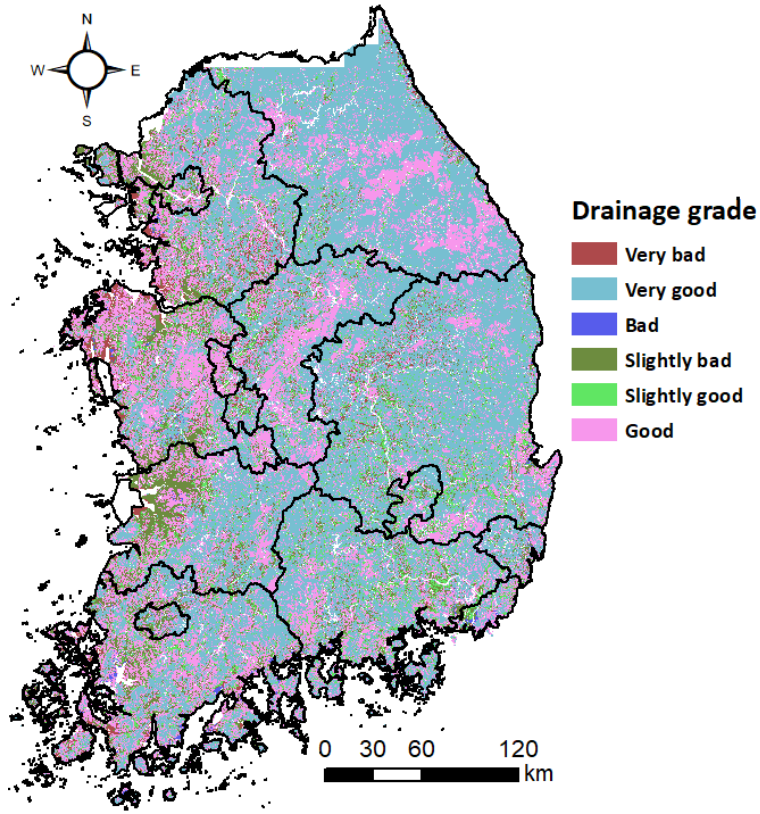
Table 2-1. Detail information of the predictor variables.

Predictor variable	Sub-variables	Variable type	Data source
Altitude		Numerical	National Geographic Information Institute
Slope		Numerical	National Geographic Information Institute
Drainage grade	Very good, Good, Slightly good, Slightly bad, Bad, Very bad	Categorical	Rural Development Administration
Effective soil depth	< 20, 20 - 50, 50 - 100, > 100 cm	Categorical	Rural Development Administration
Soil texture	Loamy very fine sand, Loamy sand, very fine sandy loam, Sandy loam, Loam, Clay loam, Silty loam, Silty clay loam	Categorical	Rural Development Administration
Land use	Residential area, Industrial area, Commercial area, Amusement area, Transportation facility, Other bare land, Public facility, Golf course, Rice paddy, Field, Orchard, House cultivation, Other cultivation, Coniferous forest, Broadleaf forest, Mixed forest, Natural grassland, Other grassland, Inland water, Inland wetland	Categorical	Ministry of Environment
Hydrogeology	Intrusive igneous rock, Unconsolidated clastic sediment, Metamorphic rock, Clastic sedimentary rock, Carbonate rock	Categorical	Korea Institute of Geoscience and Mineral Resources

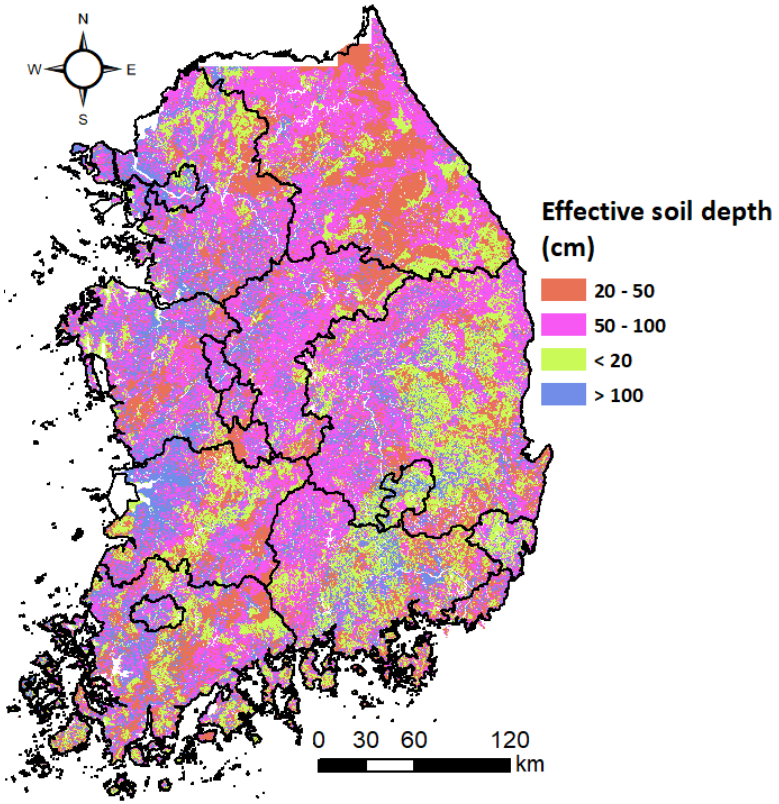


(a) (b)

Figure 2-7. Maps for each predictor variable. (a) altitude. (b) slope. (c) drainage grade. (d) effective soil depth. (e) soil texture. (f) land use. (g) hydrogeology.

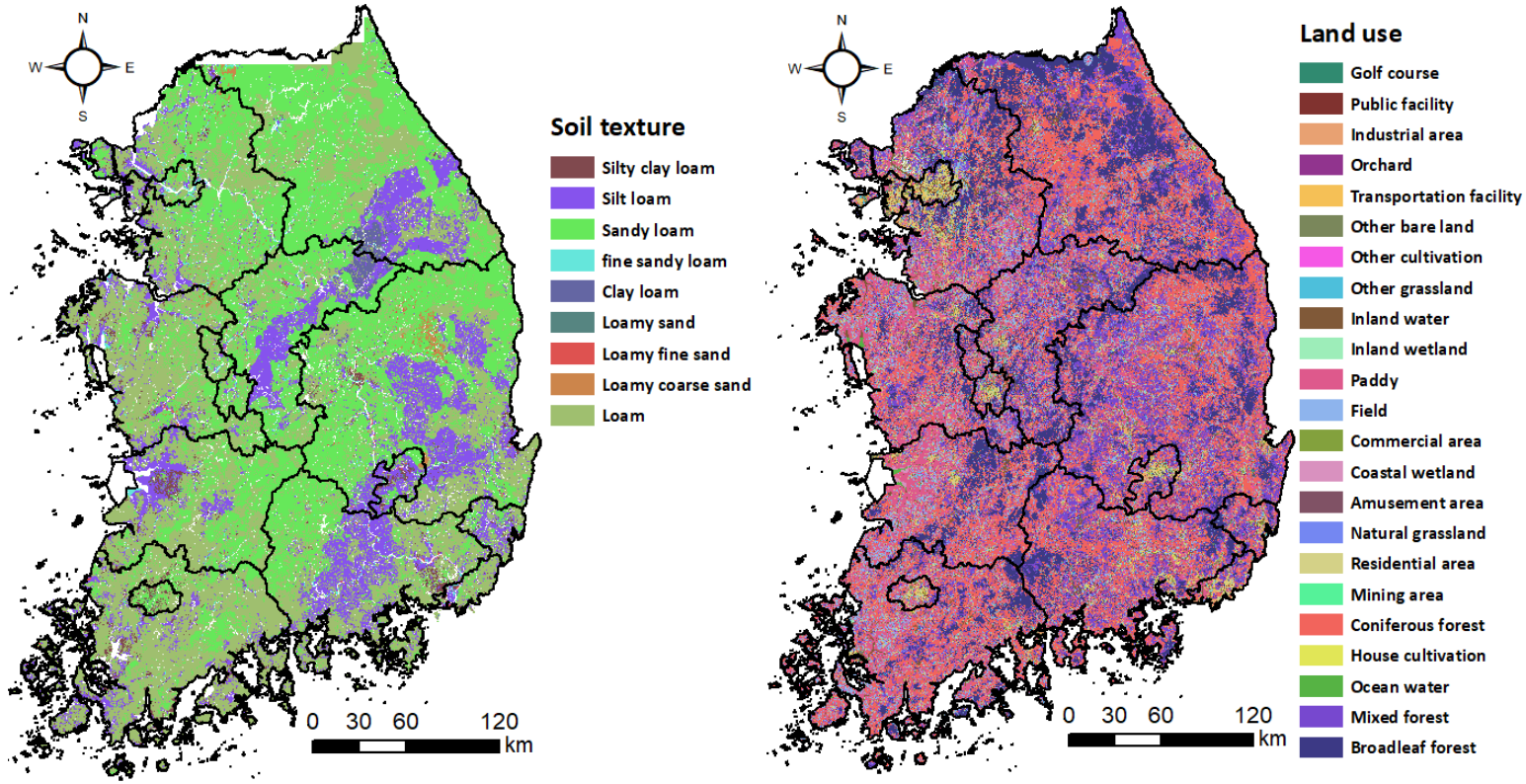


(c)



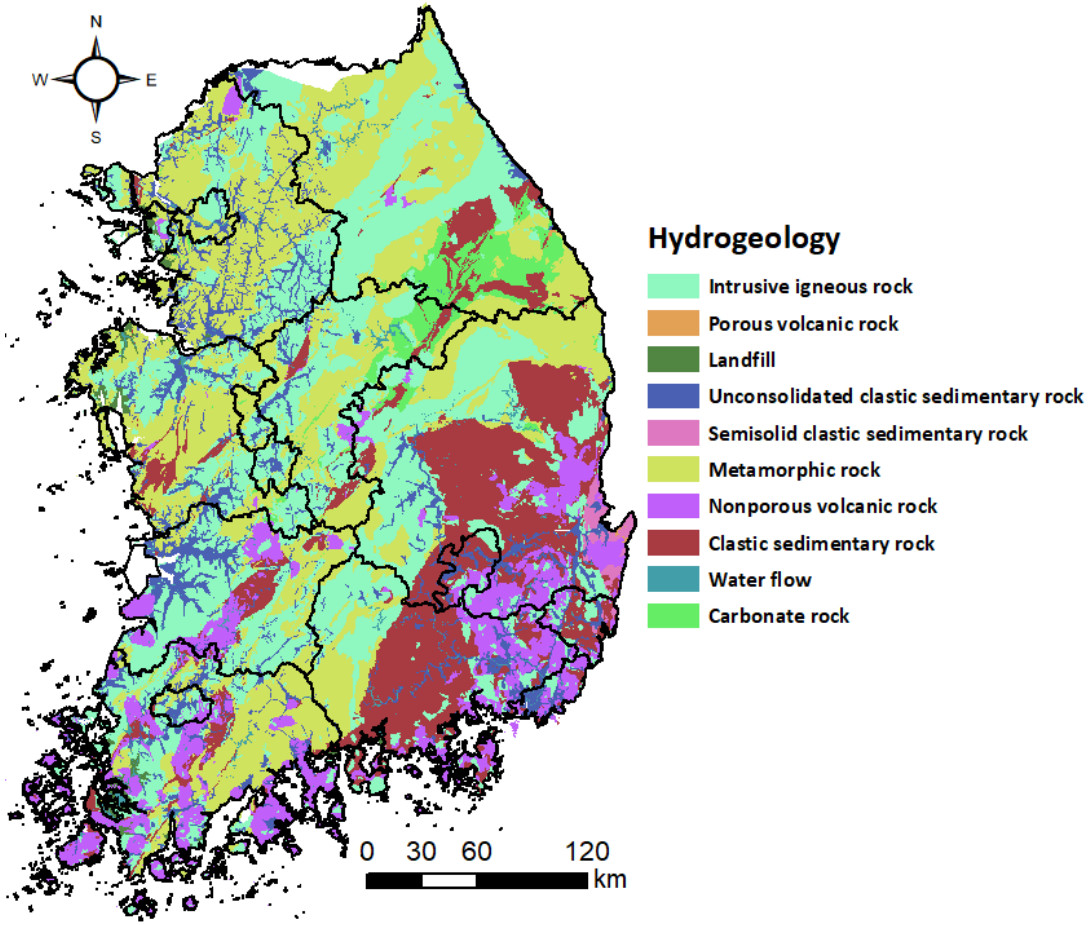
(d)

Figure 2-7. Continued.



(e) **(f)**

Figure 2-7. Continued.



(g)

Figure 2-7. Continued.

2.1.2 Data Preprocessing

Response variable indicating suitable or unsuitable for mineral water was determined by establishing several criteria using well information and water quality data (Fig. 2-8). At first, alluvial type of wells were removed from the collected datasets, because groundwater development for drinking purpose is hardly conducted at the alluvial aquifer. Afterwards, the remaining data were classified into suitable (1) or unsuitable (0) groundwater for mineral water using nitrate-N concentration, presence of potential contaminant source, and TDS concentration. Nitrate-N value more than 3 mg/L indicates anthropogenic influence on groundwater (Madison and Brunett, 1985), and the anthropogenic influence was regarded as harmful impact on groundwater quality. TDS concentration is closely related to the taste of mineral water and well scaling. Less than 600 mg/L of TDS is preferable for the taste of mineral water (Table 2-2), but also water with low TDS concentrations may not be acceptable because it has a flat taste. In addition, more than 500 mg/L of TDS can cause the scaling of groundwater well and shorten the service life of the well (Fawell et al., 2003). For the taste of mineral water and the sustainable groundwater well, acceptable TDS range for mineral water was determined to be 100 ~ 500 mg/L in this study. Therefore, groundwater samples with nitrate-N concentration less than 3 mg/L, with TDS concentration range of 100 ~ 500 mg/L, and absence of potential contamination sources within 100 m radius buffer were given response of suitable (1) (Fig. 2-8).

Categorical environmental factors at a point were transformed into numerical data by setting a 100 m radius buffer zone and by calculating a ratio of area

corresponding to each sub-variable as shown in Fig. 2-9. This process was carried out in order to consider the influence of the surrounding environment, not just the point where the groundwater located. For numerical input variables like altitude and slope, their value at a well coordinate was extracted.

The total number of the national dataset was 6,135, and the number of data that have ‘suitable (1)’ and ‘unsuitable (0)’ response was 254 and 5,881, respectively. Fig. 2-10 showed the location of groundwater wells finally applied to the tree-based ensemble methods by responses. About 70 percent of total data (4,298) were used for training and about 30 percent of them (1,837) were served as test, maintaining the proportion of the response variable. There is a huge difference between the number of ‘1’ and ‘0’ responses, meaning imbalance of the dataset, and therefore careful interpretation of model results is needed.

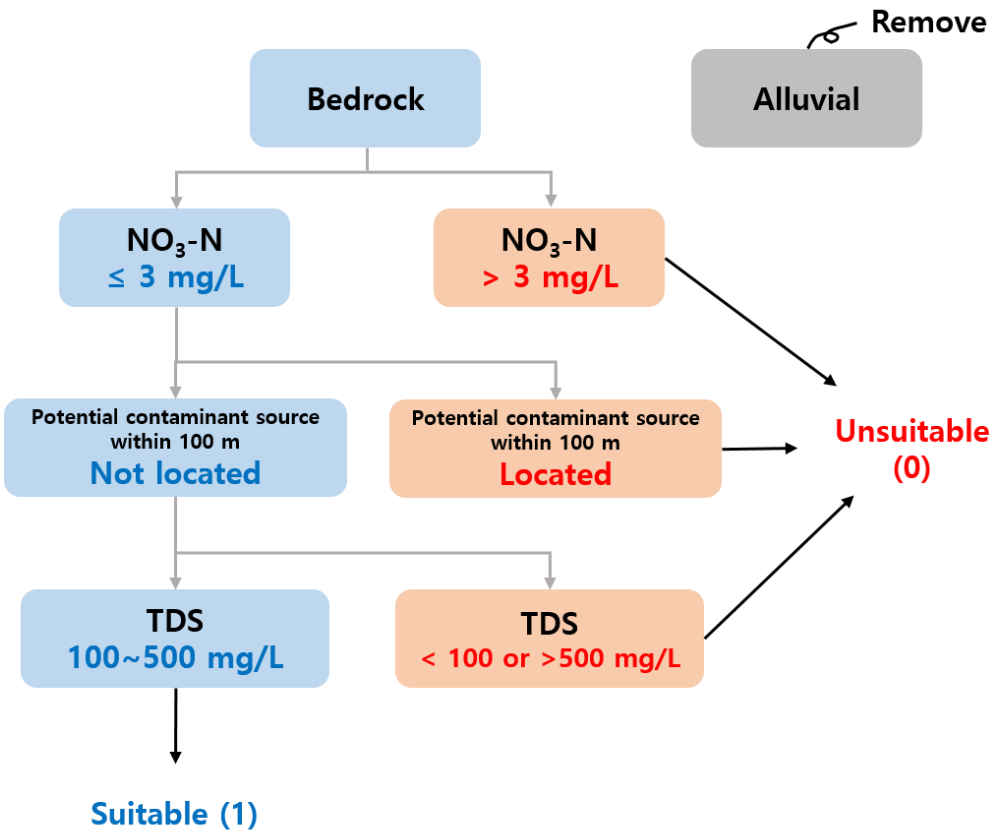
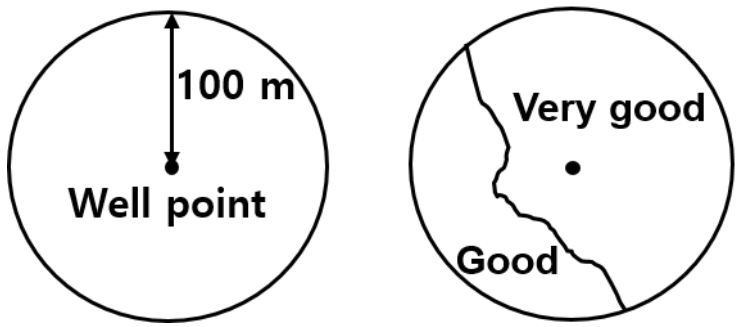


Figure 2-8. Flowchart for response variable preprocessing. Nitrate-N concentration, presence of potential contaminant source, and TDS concentration were used for response classification.

Table 2-2. Rated taste of water with different TDS concentrations. Less than 600 mg/L of TDS is preferable (excellent and good) for the taste of water.

Level of TDS (mg/L)	Rating
Less than 300	Excellent
300 ~ 600	Good
600 ~ 900	Fair
900 ~ 1,200	Poor
Above 1,200	Unacceptable



Drainage grade					
Very good	Good	Slightly good	Slightly bad	Bad	Very Bad
0.7	0.3	0	0	0	0

Figure 2-9. An example of setting the 100 m radius buffer to transform categorical variable (drainage grade) into numeric format.

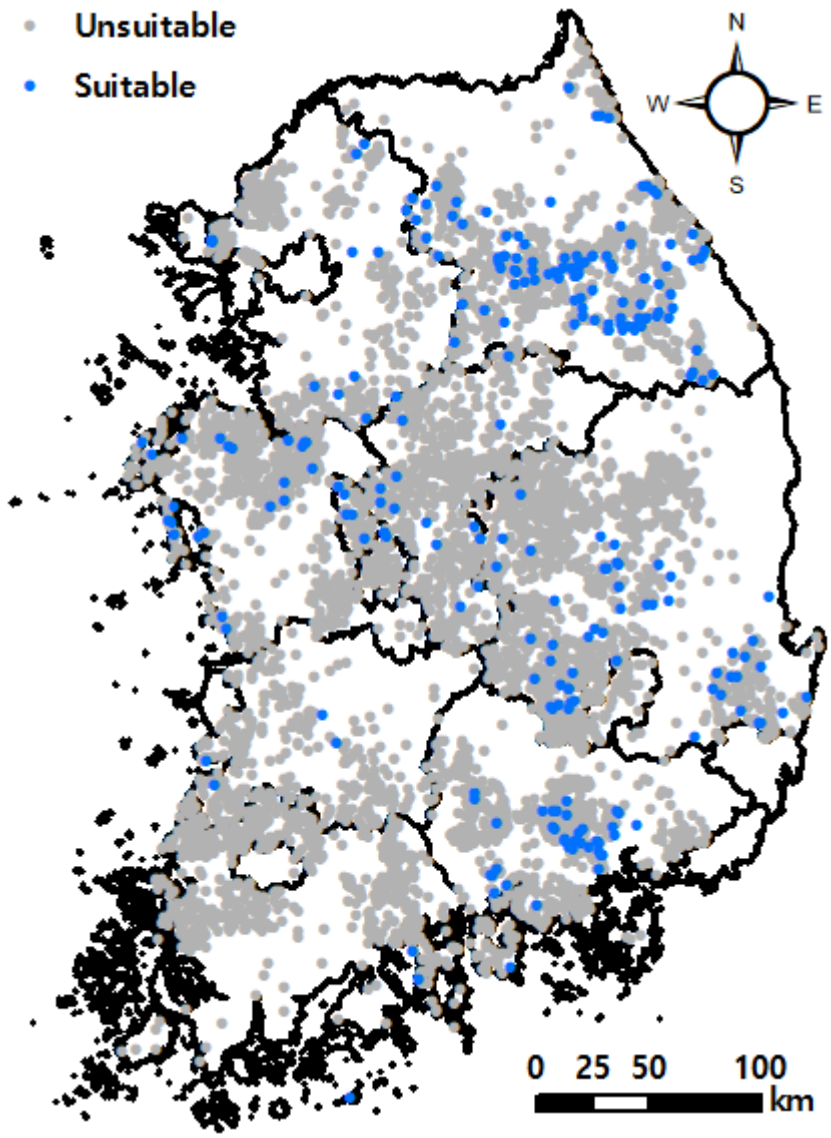


Figure 2-10. Location of groundwater wells finally applied to the tree-based ensemble methods by responses.

2.3 Evaluation Methods

After fitting the tree-based ensemble models with training data, test data were applied to the trained models for their validation. Prediction results were obtained from the models and they were compared with real output (observation), so made it possible to evaluate the model. Area Under a Curve (AUC) of Receiver Operating Characteristics (ROC) curve, AUC of Precision Recall (PR) curve, F1 measure, F0.5 measure, and G measure were used as the main evaluation measures, and basic metrics for them are calculated as follow:

$$\textbf{False positive rate } (= \textbf{Sensitivity}) = \frac{\boldsymbol{FP}}{\boldsymbol{TN} + \boldsymbol{FP}} \quad \text{(Equation 2-3)}$$

$$\textbf{True positive rate } (= \textbf{Recall}) = \frac{\boldsymbol{TP}}{\boldsymbol{FN} + \boldsymbol{TP}} \quad \text{(Equation 2-4)}$$

$$\textbf{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{FP} + \boldsymbol{TP}} \quad \text{(Equation 2-5)}$$

Table 2-3 shows the confusion matrix after classifying the prediction value into response of 1 or 0 with threshold, which describes the terms in the eq. 2-3 ~ 2-5.

ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis, and PR curve plots the precision on the x-axis and the recall on the y-axis. Both curves are drawn along the change in threshold, which means the trade-off between the two rates of axes (Negnevitsky, 2005). In the ROC curve, model satisfying low false positive rate and high true positive rate at the same time shows good performance in prediction. Therefore, the closer the ROC curve to the top-left corner, the better the model performance is, so performance could be quantified as the AUC of the ROC curve (ROC AUC). In the PR curve, model

satisfying high precision and high recall value at the same time shows good performance in prediction, and AUC value in PR curve (PR AUC) also can represent the model performance. Fig. 2-11 illustrates the ROC and PR curves that can be obtained from an ideal model. In addition, the model is evaluated to be better than random classifier by confirming whether the PR curve exists above a baseline (Saito and Rehmsmeier, 2015) which has a y value of baseline in Eq. 2-6 and is parallel to x-axis.

$$\text{Baseline} = \frac{the\ number\ of\ positive\ data}{the\ total\ number\ of\ training\ data} \qquad \text{(Equation 2-6)}$$

F1 measure, F0.5 measure and G measure provide a single measurement combining precision and recall, and they summarize the model performance. Generalized formular of F measure is in eq. 2-7, and F1 measure and F0.5 measure are computed by giving 1 and 0.5 for beta, respectively. F1 measure is the harmonic mean of precision and recall, which gives equal weight to precision and recall, and F0.5 measure doubles the weight of precision over recall to emphasize the effect of precision. G measure, determined by the geometric mean of precision and recall, is also used as the evaluation parameter (eq. 2-8). All three measures can get a high value only when the values of both precision and recall are sufficiently large. Therefore, they are suitable for evaluating model performance in imbalanced classifications and widely applied in those problems (Sun et al., 2013).

$$\textbf{F measure} = (\mathbf{1} + \boldsymbol{\beta^2})\frac{\boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{\beta^2} * \boldsymbol{precision} + \boldsymbol{recall}} \qquad \text{(Equation 2-7)}$$

$$\boldsymbol{G\ measure} = \sqrt{\boldsymbol{precision} * \boldsymbol{recall}} \qquad \text{(Equation 2-8)}$$

Threshold for calculating the precision and the recall was determined by reflecting the purpose of this study. The main objective of this study is to estimate location of mineral water in terms of well exploration, and therefore to find the true suitable water (true positive) and to avoid the false suitable water (false positive) are simultaneously important. Considering these characteristics, the threshold was set as follow: first, thresholds that induce a recall of less than 0.2 were excluded from consideration. Among the threshold candidates, a threshold showing the highest precision was selected finally.

Table 2-3. Confusion matrix when the threshold is determined to be *a*.

Tree-based ensemble (Threshold = *a*)		Observation	
		0 (unsuitable)	1 (suitable)
Prediction	0 (unsuitable)	True Negative (TN)	False Negative (FN)
	1 (suitable)	False Positive (FP)	True Positive (TP)

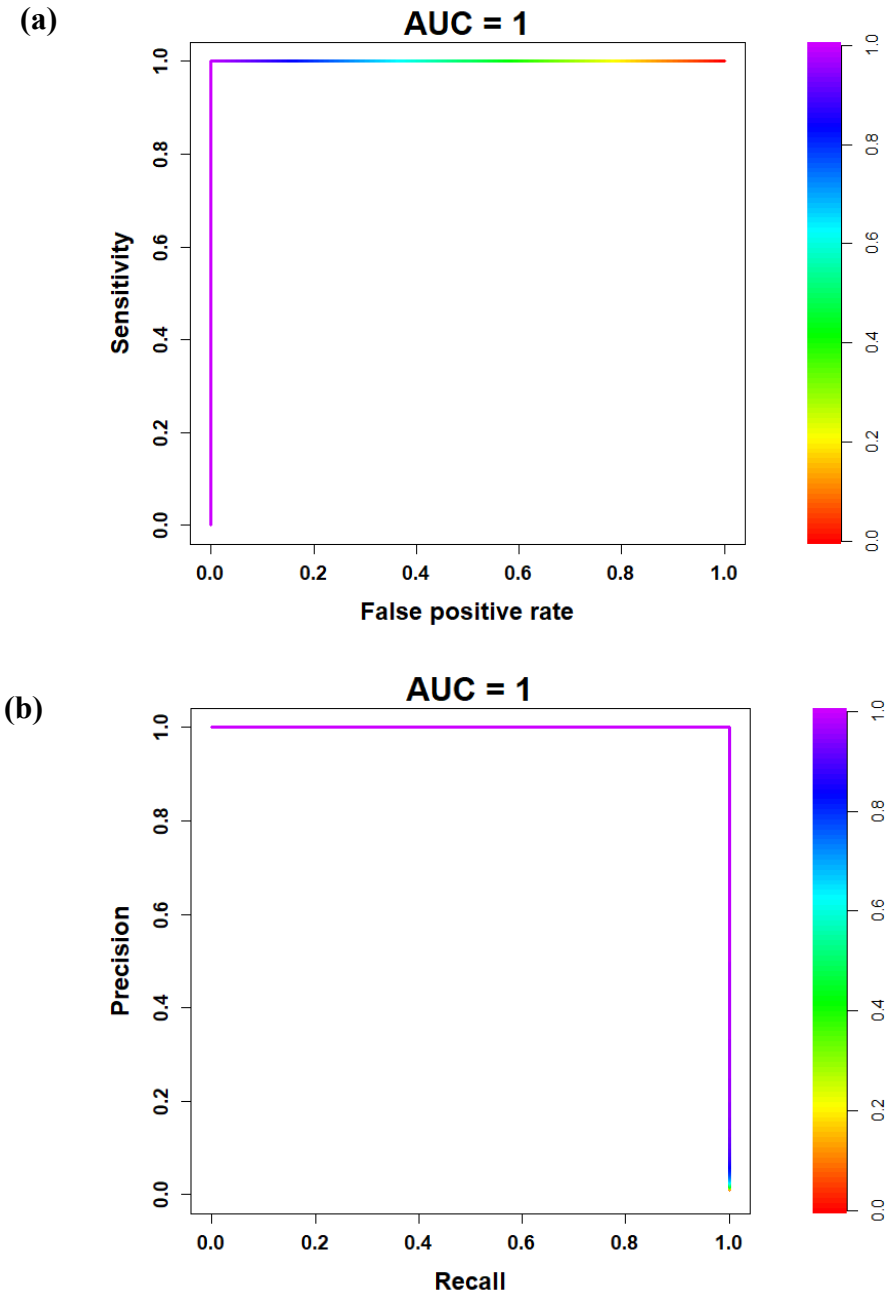


Figure 2-11. Graphical model evaluation methods. (a) ROC curve of an ideal model. (b) PR curve of an ideal model. AUC value of 1 in both of curves reflects a perfectly accurate test.

2.4 Variable Influence

Applying ensemble techniques to a single tree model of CART can achieve better performance but decrease the model interpretability (Hastie et al., 2009; Molnar, 2020). To compensate the weakness of the predictive ensemble models, variable influence of predictors was additionally examined.

2.4.1. Variable Importance

For the BRT model, variable importance was calculated by averaging over all trees of the number of times a predictor selected for splitting, weighted by the residual improvement to the model at each split (Friedman and Meulman, 2003). Then relative influence of predictors was obtained so that the sum of the variable importance equals one hundred. For the RF and ERT models, variable importance was measured as an increased mean square error after permuting the values of a specific input while all others are left unchanged (Liaw and Wiener, 2002). Variable importance of the two models were also transformed to relative influence for comparison with the BRT model. The results of the variable importance indicate which variables have a strong influence on the response, with higher number implying stronger impact on the model performance.

2.4.2. Partial Dependence Plot

Partial dependence plots of each predictors were produced for the three models to depict the dependence between the input variables and the target response, which showed whether the relationship between them is linear, monotonic or more complex (Molnar, 2020). The x-axis of the plot has a range of minimum to maximum values of each variable in training data. The y value in the plot is calculated as the average for all expected response of data point when substituting the data value of a specific predictor to fixed x value. Therefore, the y-axis of the plot represents the marginal effect of the predictor to the response. In this study, it was able to determine which predictor variables contributed significantly to the existence of suitable (or unsuitable) groundwater for mineral water from the results of the partial dependence plots.

2.5 Potential Mapping

Validated models can be applied to delineate the potential of mineral water even in locations without groundwater quality data. For the three models, prediction was carried out for 500 * 500 m grids across the country and potential maps were completed using interpolation method. First, potential maps were presented as continuous values to examine the trends and characteristics of the spatial distribution of the predicted responses. To demonstrate the variable influence results of some predictors, GIS-based data of them was intersected on the continuous potential maps, and they were compared. Next, binary potential maps classified by the optimal thresholds in each model were also plotted to figure out the spatial distribution of suitable groundwater for mineral water. Using the binary maps, the area ratio of the suitable response to the total area of a district was calculated (Eq. 2-9), and the regions where mineral water existed at a high rate was identified.

$$\boldsymbol{Ratio}\ (\%) = \frac{\boldsymbol{Area\ occupied\ by\ suitable\ response}}{\boldsymbol{Total\ area\ of\ a\ district}} \times \mathbf{100} \qquad \text{(Equation 2-9)}$$

3 RESULTS AND DISCUSSION

Parameters in each tree-based machine learning algorithm (BRT, RF and ERT) were determined so that the models can optimally solve the problem. Candidates of multiple parameters were set first, and by selecting parameters that showed the least error, optimal models of the three algorithms were induced. Table 3-1 (a) contains the parameters and their candidate values used in the combination to find hyperparameters in each method, and determined values of the parameters are presented in Table 3-1 (b).

3.1 Evaluation of the Model Performance

Fig. 3-1 displays ROC curves and their AUC values of the three optimal models (BRT, RF and ERT). The three ROC curves were all close to the upper left corner, and the AUC values in the ROC curves of the BRT, RF and ERT models were 0.968, 0.966, and 0.958, respectively, indicating that all three methods had excellent model performance. However, because of the large denominator in false positive rates caused by the imbalanced dataset, the value of the false positive rates did not change significantly with the threshold variation. Therefore, the evaluation of models using the ROC curve can mislead the interpretation of model performance in the unbalanced dataset (Saito and Rehmsmeier, 2015; Movahedi et al., 2020), so PR curve was mainly used for assessing the model performance.

Result of PR curves and their AUC values of the three optimal models are shown in Fig. 3-2. In the PR curve, the AUC values for the BRT, RF and ERT models were 0.486, 0.500, and 0.470, respectively, and a baseline was calculated

as 0.0416 according to the Eq. 2-7. It appeared that the PR curves of all three methods were not close to the upper right corner, but the PR curves existed above the baseline and their AUC values greatly exceeded the value of baseline. Therefore, all three models are highly applicable to estimating groundwater suitable for mineral water.

Optimal threshold was determined to classify the continuous prediction values into suitable or unsuitable responses for mineral water. The thresholds for the BRT, RF, and ERT models were 0.401, 0.432, and 0.430, respectively, and confusion matrixes constructed by them are given in Table 3-2. Several metrics value such as accuracy, precision and recall calculated from the confusion matrixes are listed in Table 3-3. The accuracies of all three models were very high and there was not much difference between the three models in accuracy. However, because high accuracy values can be easily reached in this study due to the imbalanced dataset, precision and recall were primarily considered in the model evaluation. BRT model showed the highest values for all metrics, which means the BRT model has the highest performance for the selected threshold reflecting the purpose of this study. Precision was higher in the RF model (0.640) than in the ERT model (0.548), but recall was higher in the ERT model (0.227) than the RF model (0.213). To summarize the model performance with a single score combining precision and recall, F1 measure, F0.5 measure and G measure were also calculated for the three models (Fig. 3-3). As a result, BRT model outperformed the other two models, which indicates the BRT model detects more true suitable water (true positive) with higher precision than the RF and ERT models. F1 measures of the RF and ERT

models were comparable, but the RF model had better result than the ERT model with respect to F0.5 measure and G measure. Although the AUC results in PR curve confirmed the applicability of the three models, the BRT model was the best in terms of groundwater development, followed by the RF and ERT models in order.

Table 3-1. (a) Parameters and their values used to find the best parameters in each method. (b) Parameter values determined for the optimal model in each method.

(a)

Method	Parameter	Value
BRT	*learning rate (lr)*	0.01, 0.005, 0.001
	tree.complexity (tc)	1~5
RF	*mtry*	1~15
ERT	*mtry*	1~15
	numRandomCut	1~5

(b)

Method	Parameter	Value
BRT	*learning rate (lr)*	0.001
	tree.complexity (tc)	4
RF	*mtry*	14
ERT	*mtry*	14
	numRandomCut	2

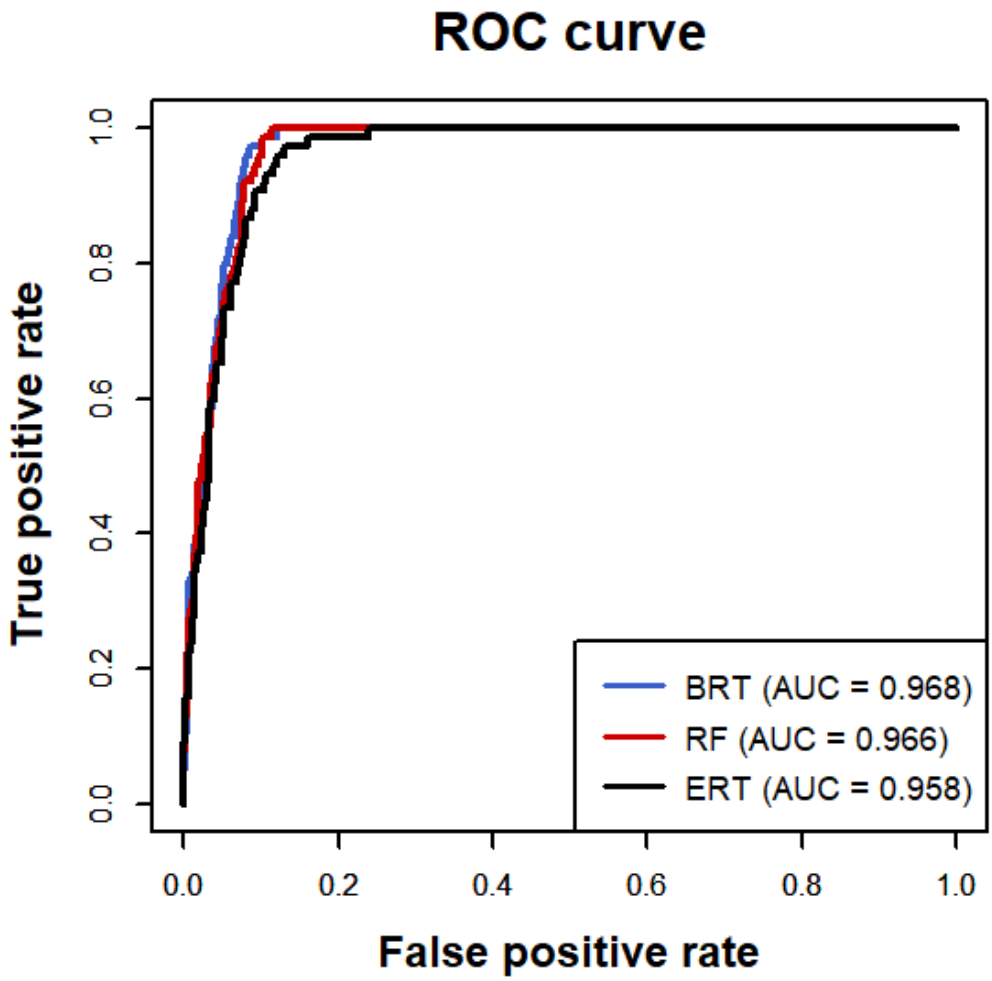


Figure 3-1. ROC curves for BRT (blue), RF (red), and ERT (black) models and their AUC values are 0.968, 0.966, and 0.958, respectively.

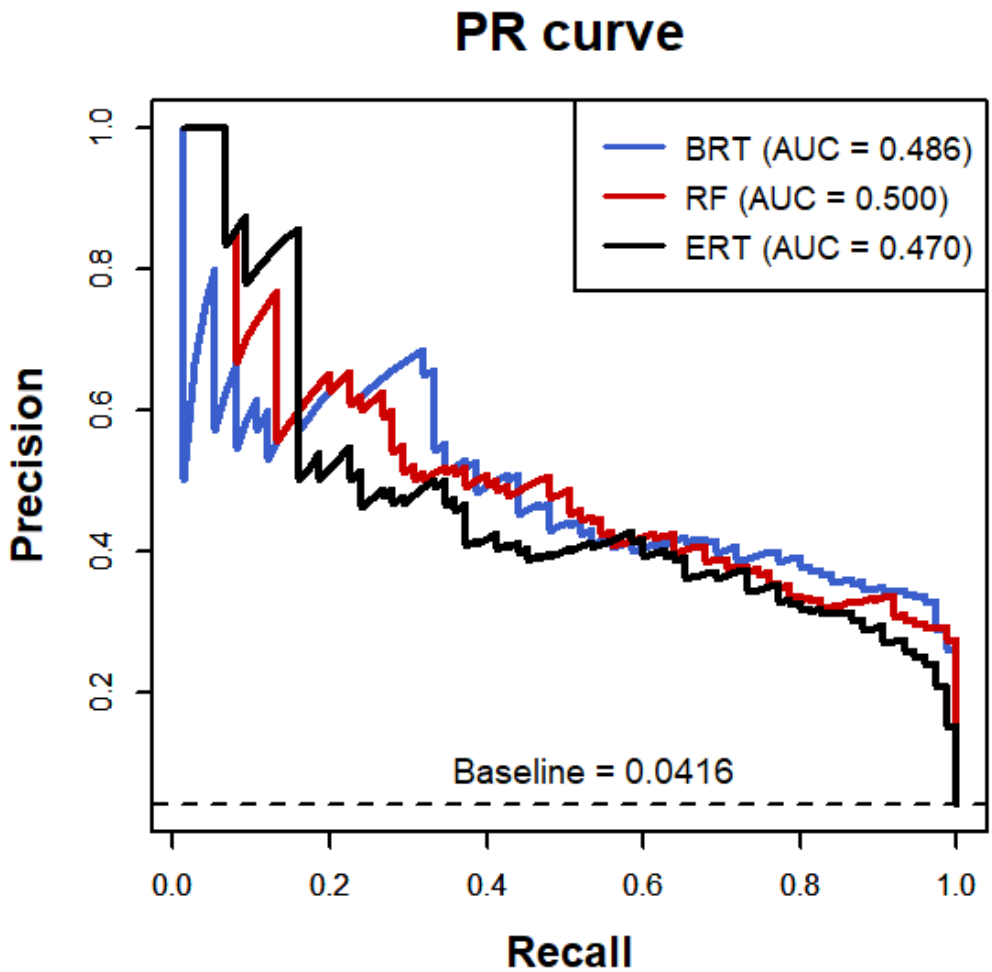


Figure 3-2. PR curves for BRT (blue), RF (red), and ERT (black) models and their AUC values are 0.486, 0.500, and 0.470, respectively.

Table 3-2. Confusion matrixes for BRT, RF, and ERT methods generated with each optimal threshold considering groundwater exploration.

BRT method (Threshold = 0.401)		Observation	
		0 (unsuitable)	1 (suitable)
Prediction	0 (unsuitable)	1751	51
	1 (suitable)	11	24

RF method (Threshold = 0.432)		Observation	
		0 (unsuitable)	1 (suitable)
Prediction	0 (unsuitable)	1753	59
	1 (suitable)	9	16

ERT method (Threshold = 0.430)		Observation	
		0 (unsuitable)	1 (suitable)
Prediction	0 (unsuitable)	1748	58
	1 (suitable)	14	17

Table 3-3. Accuracy, precision, and recall calculated from the confusion matrixes of the BRT, RF, and ERT models.

	BRT	RF	ERT
Accuracy	0.966	0.963	0.961
Precision	0.686	0.640	0.548
Recall	0.320	0.213	0.227

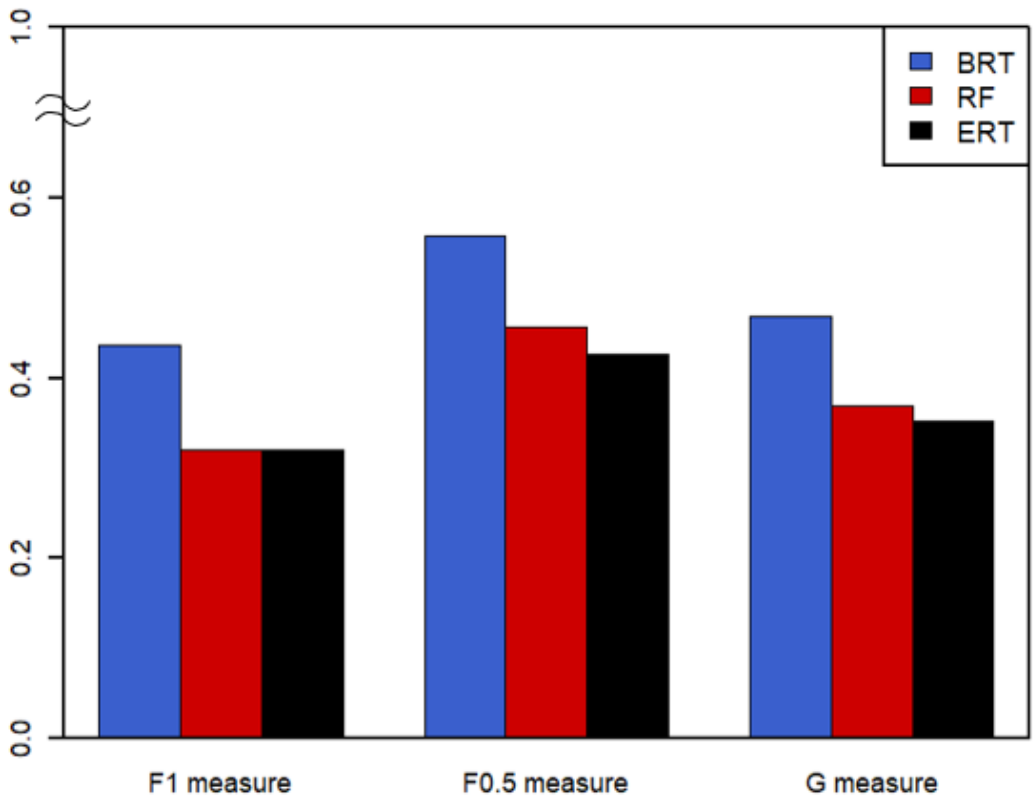


Figure 3-3. F1 measures, F0.5 measures, and G measures for BRT (blue), RF (red), and ERT (black) models. BRT model is superior to the other two RF and ERT models in all measures.

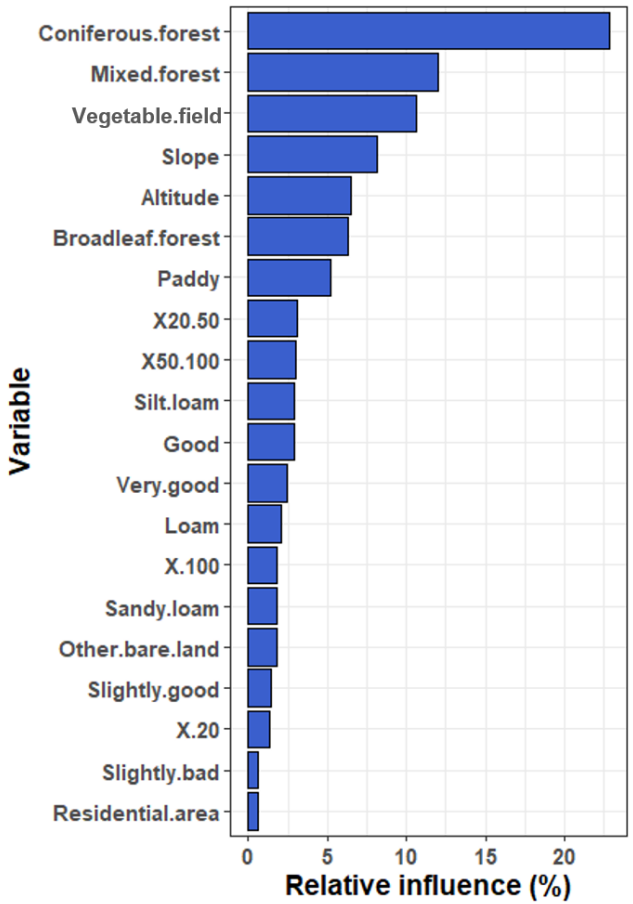
3.2 Results of Variable Influence

3.2.1. Variable Importance

Relative influence of every variables was computed by normalizing the variable importance so that the sum of them to be 100 %. Top 20 variables of the relative influence in the BRT, RF, and ERT models are summarized in Fig. 3-4, 3-5, and 3-6, respectively, in order of decreasing contribution. The ranking and the relative influence of predictor variables differed between the three models, but there were some similarities. In all three models, most dominating predictor was coniferous forest with relative influences of 22.8 %, 18.0 % and 21.5 % for BRT, RF, and ERT models, respectively. Other forest-related variables such as mixed forest and broadleaf forest and cultivation-related variables such as vegetable field and paddy were also important variables in all three models. Overall, variables of land use category ranked high, which implies that land use is the primary environmental factor representing the spatial groundwater suitability for mineral water in South Korea. In addition, three models were also slightly explained by slope and altitude variables, and especially, slope accounted for more than 5 % of the response variability in three models.

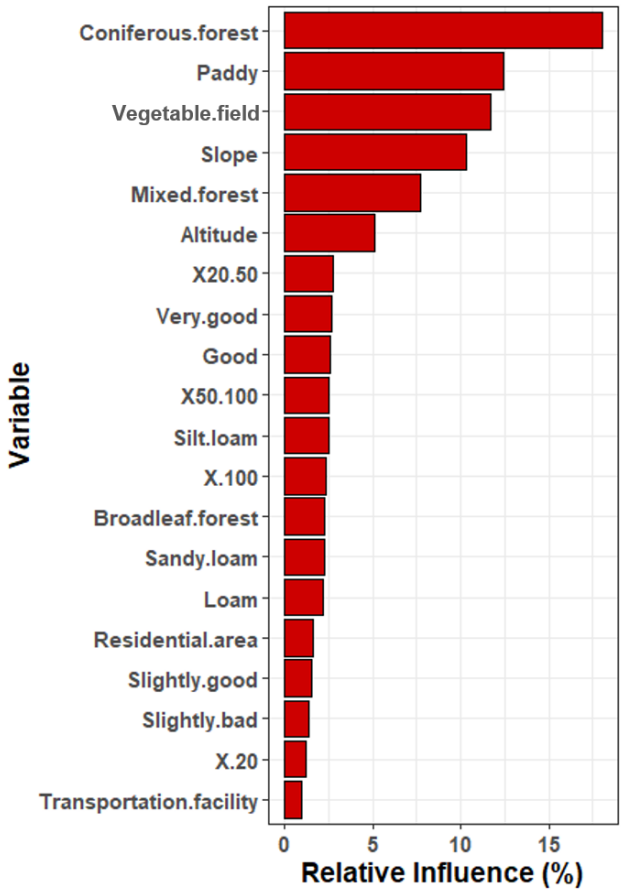
Seventeen out of the top 20 variables are matched in the three models, which indicates there was consistency in the variables to make accurate model for estimating suitable groundwater for mineral water. Overlapping variables in the three influence results and their averaged values are listed in descending order in

Table 3-4, and they affect not only the accuracy of the models but also the variation of the response.



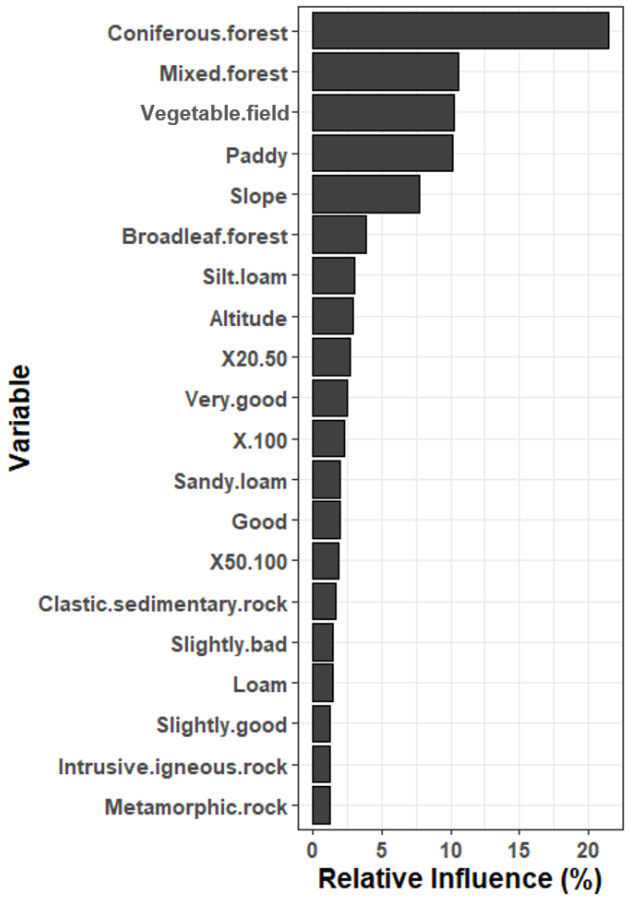
Predictor variable	Variable category	Relative influence (%)
Coniferous forest	Land use	22.8
Mixed forest	Land use	12.0
Vegetable field	Land use	10.6
Slope	-	8.1
Altitude	-	6.5
Broadleaf forest	Land use	6.3
Paddy	Land use	5.2
20 ~ 50 cm	Effective soil depth	3.1
50 ~ 100 cm	Effective soil depth	3.0
Silt loam	Soil texture	2.9
Good	Drainage grade	2.9
Very good	Drainage grade	2.4
Loam	Soil texture	2.1
More than 100 cm	Effective soil depth	1.8
Sandy loam	Soil texture	1.8
Other bare land	Land use	1.8
Slightly good	Drainage grade	1.4
Less than 20 cm	Effective soil depth	1.4
Slightly bad	Drainage grade	0.6
Residential area	Land use	0.6

Figure 3-4. Top 20 variables of relative influence and their summary in the BRT model in order of decreasing contribution.



Predictor variable	Variable category	Relative influence (%)
Coniferous forest	Land use	18.0
Paddy	Land use	12.4
Vegetable field	Land use	11.7
Slope	-	10.3
Mixed forest	Land use	7.7
Altitude	-	5.1
20 ~ 50 cm	Effective soil depth	2.8
Very good	Drainage grade	2.7
Good	Drainage grade	2.6
50 ~ 100 cm	Effective soil depth	2.5
Silt loam	Soil texture	2.5
More than 100 cm	Effective soil depth	2.4
Broadleaf forest	Land use	2.3
Sandy loam	Soil texture	2.2
Loam	Soil texture	2.2
Residential area	Land use	1.6
Slightly good	Drainage grade	1.5
Slightly bad	Drainage grade	1.4
Less than 20 cm	Effective soil depth	1.2
Transportation facility	Land use	0.9

Figure 3-5. Top 20 variables of relative influence and their summary in the RF model in order of decreasing contribution.



Predictor variable	Variable category	Relative influence (%)
Coniferous forest	Land use	21.5
Mixed forest	Land use	10.5
Vegetable field	Land use	10.2
Paddy	Land use	10.1
Slope	-	7.8
Broadleaf forest	Land use	3.9
Silt loam	Soil texture	3.0
Altitude	-	2.9
20 ~ 50 cm	Effective soil depth	2.7
Very good	Drainage grade	2.5
More than 100 cm	Effective soil depth	2.3
Sandy loam	Soil texture	1.9
Good	Drainage grade	1.9
50 ~ 100 cm	Effective soil depth	1.8
Clastic sedimentary rock	Hydrogeology	1.6
Slightly bad	Drainage grade	1.5
Loam	Soil texture	1.5
Slightly good	Drainage grade	1.3
Intrusive igneous rock	Hydrogeology	1.3
Metamorphic rock	Hydrogeology	1.2

Figure 3-6. Top 20 variables of relative influence and their summary in the ERT model in order of decreasing contribution.

Table 3-4. Seventeen overlapping variables and their averaged values in the results of relative influence from the BRT, RF, and ERT models.

Predictor variable	Variable category	Average of relative influence (%)
Coniferous forest	Land use	20.8
Vegetable field	Land use	10.8
Mixed forest	Land use	10.1
Paddy	Land use	9.2
Slope	-	8.7
Altitude	-	4.8
Broadleaf forest	Land use	4.2
20 ~50 cm	Effective soil depth	2.8
Silt loam	Soil texture	2.8
Very good	Drainage grade	2.6
Good	Drainage grade	2.5
50 ~ 100 cm	Effective soil depth	2.4
More than 100 cm	Effective soil depth	2.2
Sandy loam	Soil texture	2.0
Loam	Soil texture	1.9
Slightly good	Drainage grade	1.4
Slightly bad	Drainage grade	1.1

3.2.2. Partial Dependence Plot (PDP)

Partial dependence, which indirectly describes the expected response as a function of a particular variable, was plotted for the overlapping variables in the three influence results suggested in Table 3-4 (Fig. 3-7). The RF and ERT models showed almost similar trends in the partial dependences of the selected predictor variables. The BRT model showed similar plots to the other two models, but slightly differed in several variables such as slope, altitude, very good, and 50 ~ 100 cm, as their values of x increase. The partial dependence plots of the ERT model were smoother comparing to those of the BRT and RF models. These smooth features were also shown in a study by Rhee et al. (2020), which estimates drought severity with multiple linear regression, decision trees, adaptive boosting, RF, and ERT methods. This finding suggests that the variation in output values from ERT model is relatively insensitive to change in predictor values regardless of research topic, and the smooth features in partial dependence plots are own characteristic of the ERT model.

The partial dependence plots of coniferous forest, which accounted for the largest proportion in the relative influences, showed a positive correlation between the variable and prediction value for all three models. Response of the plot for coniferous forest recorded increase of about 0.3 as value of x increases. It is the largest positive variation among all the variables, indicating positively dominant contribution of coniferous forest to estimate suitable groundwater for mineral water. In addition to coniferous forest, mixed forest, slope, broadleaf forest, 20 ~ 50 cm (effective soil depth), and silt loam showed the positive correlations with the

response. It implies that these environmental factors contribute to the existence of suitable groundwater for mineral water. On the other hand, cultivation-related variables such as paddy and vegetable field showed negative correlation with the response. Variables principally contributing to the existence of unsuitable groundwater for mineral water were paddy and vegetable field.

The results of variable influence can be interpreted indirectly, considering the criteria such as nitrate-N concentration used to determine the response of each groundwater samples. Groundwater samples with nitrate-N concentration less than 3 mg / L have been given suitable response in this study. The positive influence of forest-related variables on the prediction can be slightly explained by the nitrate concentration. The percentage of forest within a buffer zone is one of the dependable predictors of nitrate concentration in groundwater. In addition, the probability of detecting nitrate concentration greater than 2 mg/L decreases with the percentage of forest increases (Gardner and Vogel, 2005). It can partly explain the positive influence of forest-related variables, which are less contaminated by nitrate. Similar to the positive influence, the negative influence of cultivation variables on the output can also be interpreted in relation to the nitrate concentration. In the studies of predicting nitrate concentration in groundwater, Knoll et al. (2019) reported that arable land shows the largest impact on nitrate concentration in multiple linear regression and BRT methods. Also, Wick et al. (2012) observed the statistically significant positive effect of cropland on nitrate contamination of groundwater because of nitrogen fertilizers. These studies suggest that the agricultural activities mainly affect nitrate leaching to groundwater, and

thus the cultivation variables would primarily lead groundwater samples to unsuitable response for mineral water in this study.

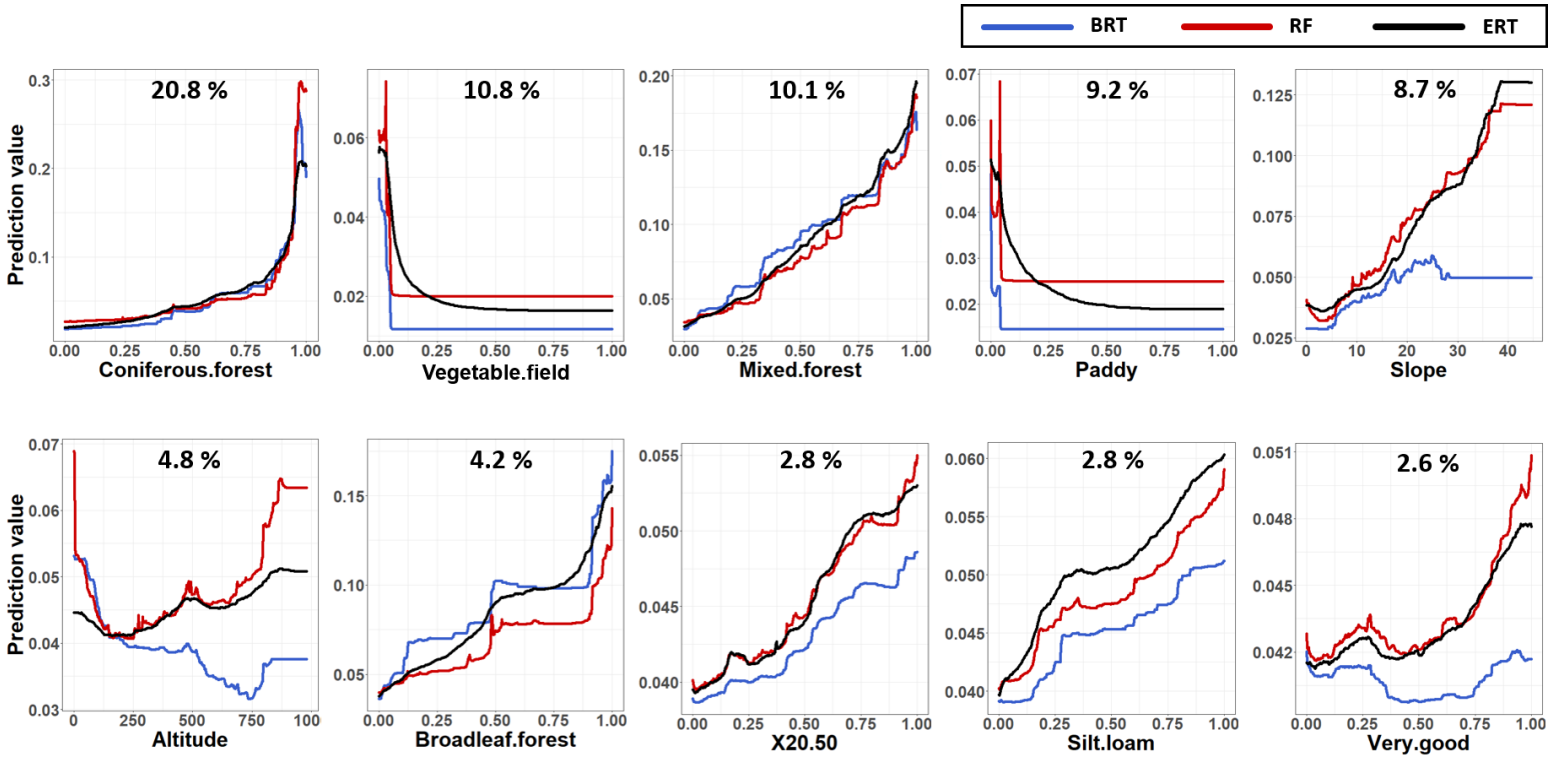


Figure 3-7. Partial dependence plots of the overlapping variables in the three relative influence results. Note that the range of y-axis corresponding to the prediction value varies with the predictor variables.

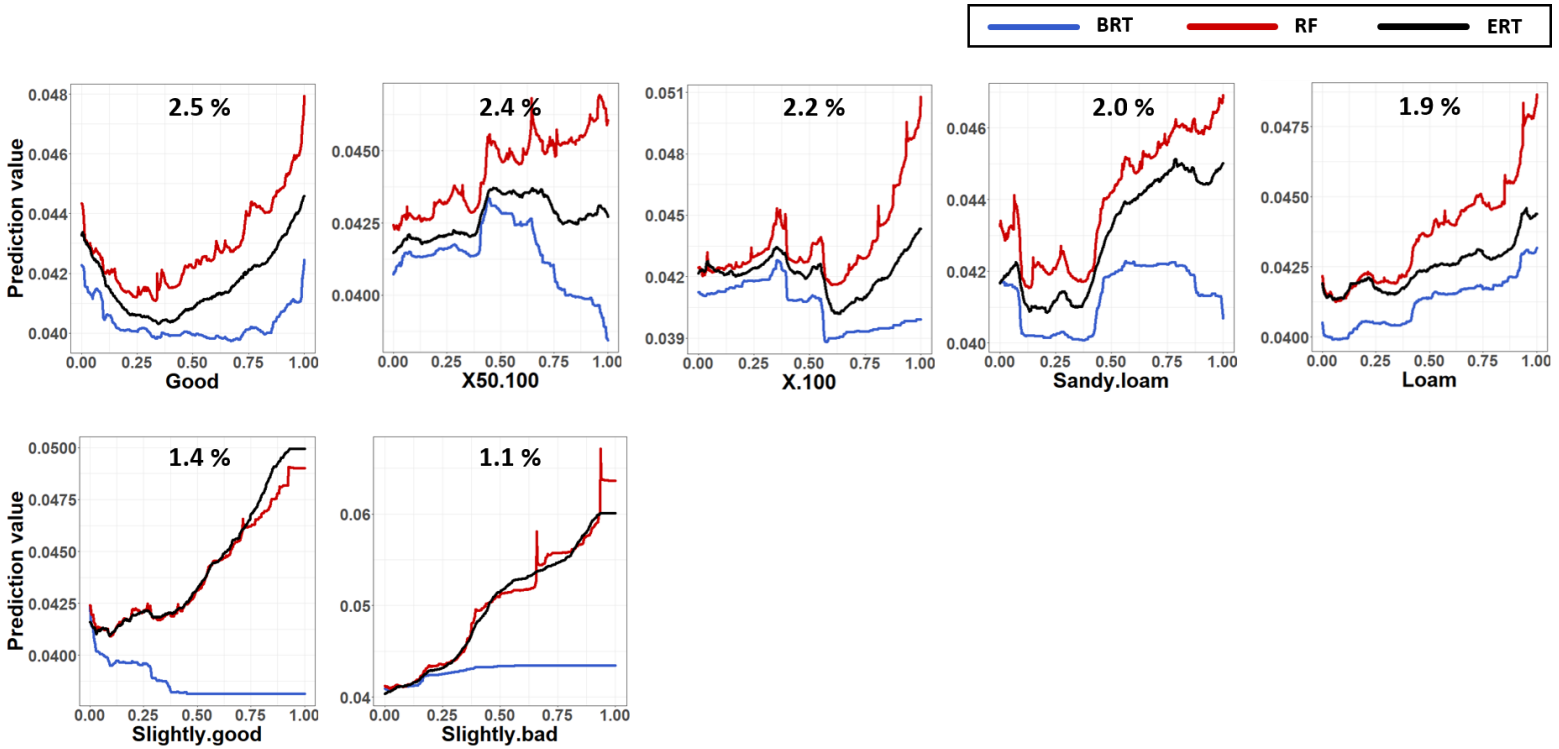


Figure 3-7. Continued.

3.3 Potential Map of Mineral Water

The applicability and validity of the tree ensemble models for estimating groundwater suitable for mineral water was confirmed in section 3.1. Therefore, gridded potential maps for mineral water could be produced for national scale by applying the three models.

3.3.1. Potential maps described with continuous values

Continuous potential maps delineating the spatial distribution of potential for mineral water and histograms of prediction values in the maps are shown in Fig. 3-8, 3-9, and 3-10 for the BRT, RF, and ERT methods, respectively. The darker the blue color on the maps, the closer the prediction value is to 1, the more likely groundwater is suitable for mineral water. In the histograms, a large number of prediction values were close to zero and the distribution was highly skewed to the right in all three models. Originally unbalanced dataset, which means the number of unsuitable (0) response overwhelmed that of suitable response (1) in the applied data, affects these distribution results and the final three potential maps. Reflecting the right-skewed distribution of the prediction values, the potential maps were mainly colored by bright blue, especially in the western side of South Korea. It seems because most of the cultivation regions such as paddies and vegetable fields are concentrated in the western part as depicted in Fig. 3-11. The result very well revealed the negative effects of the cultivation-related variables on the responses mentioned in section 3.2.2. Correlation coefficients between the three potential

maps were very high, above 0.89 (Table 3-5). Similar patterns between the three applicable models enhanced reliability of the illustrated potential maps.

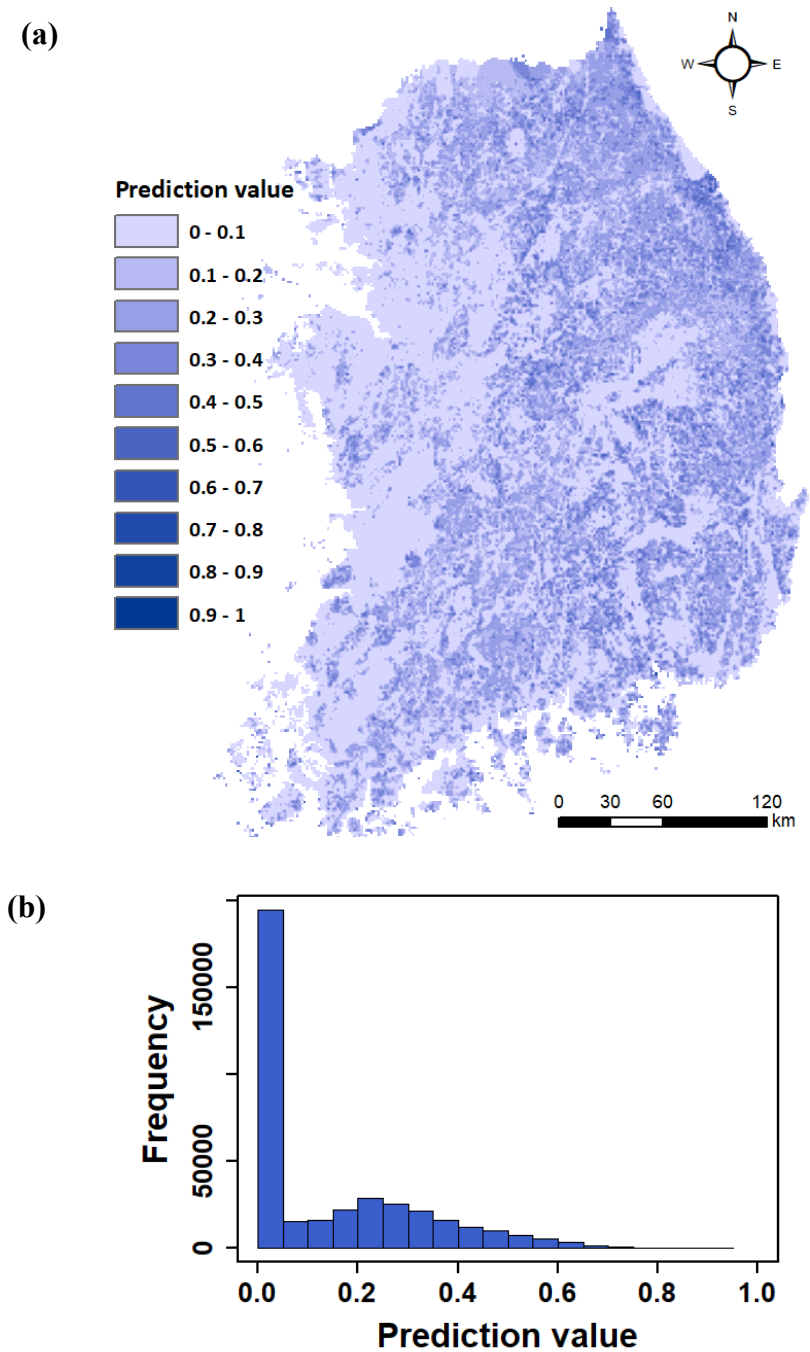


Figure 3-8. (a) Potential map describing the spatial distribution of predicted value for mineral water potential estimated by the BRT model. (b) Histogram of prediction values at grid points with 500 m equal interval for the BRT model.

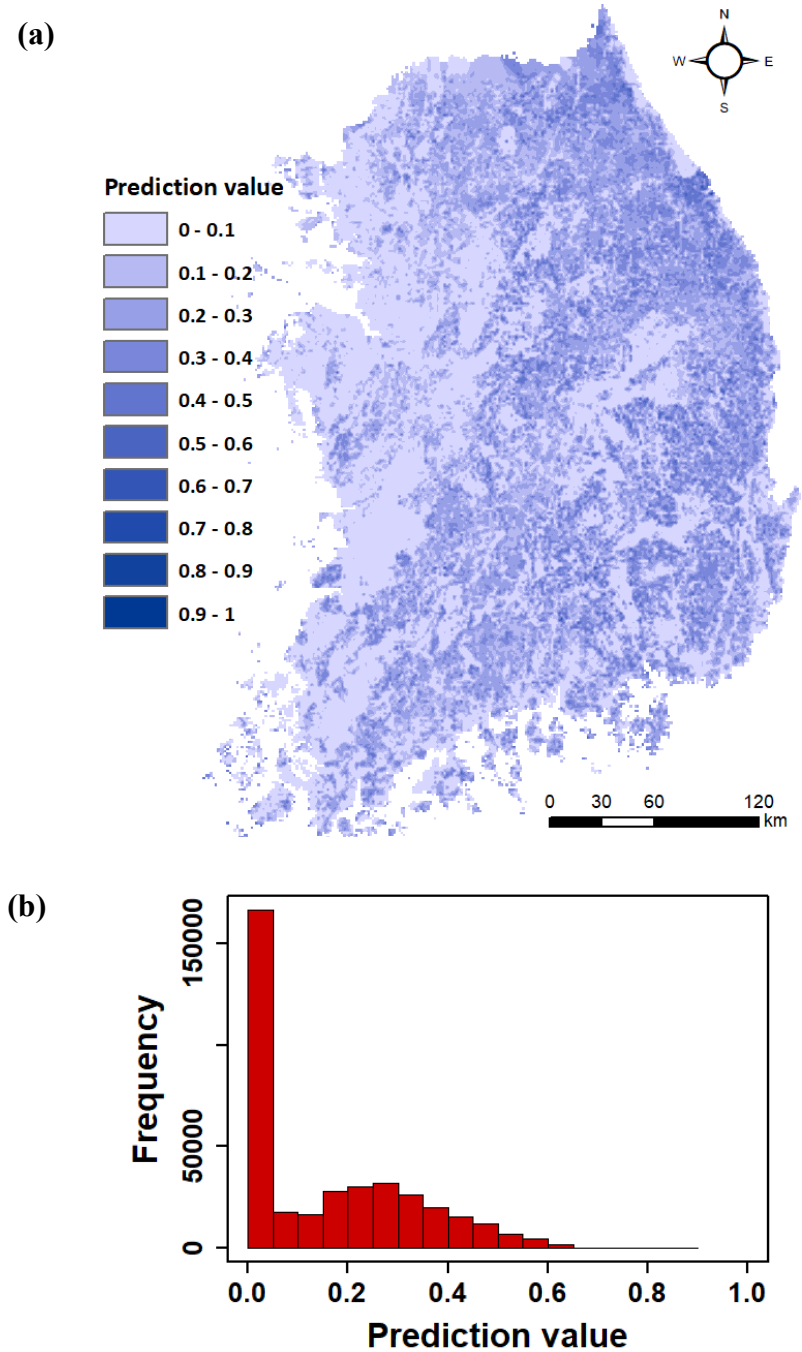


Figure 3-9. (a) Potential map describing the spatial distribution of predicted value for mineral water potential estimated by the RF model. (b) Histogram of prediction values at grid points with 500 m equal interval for the RF model.

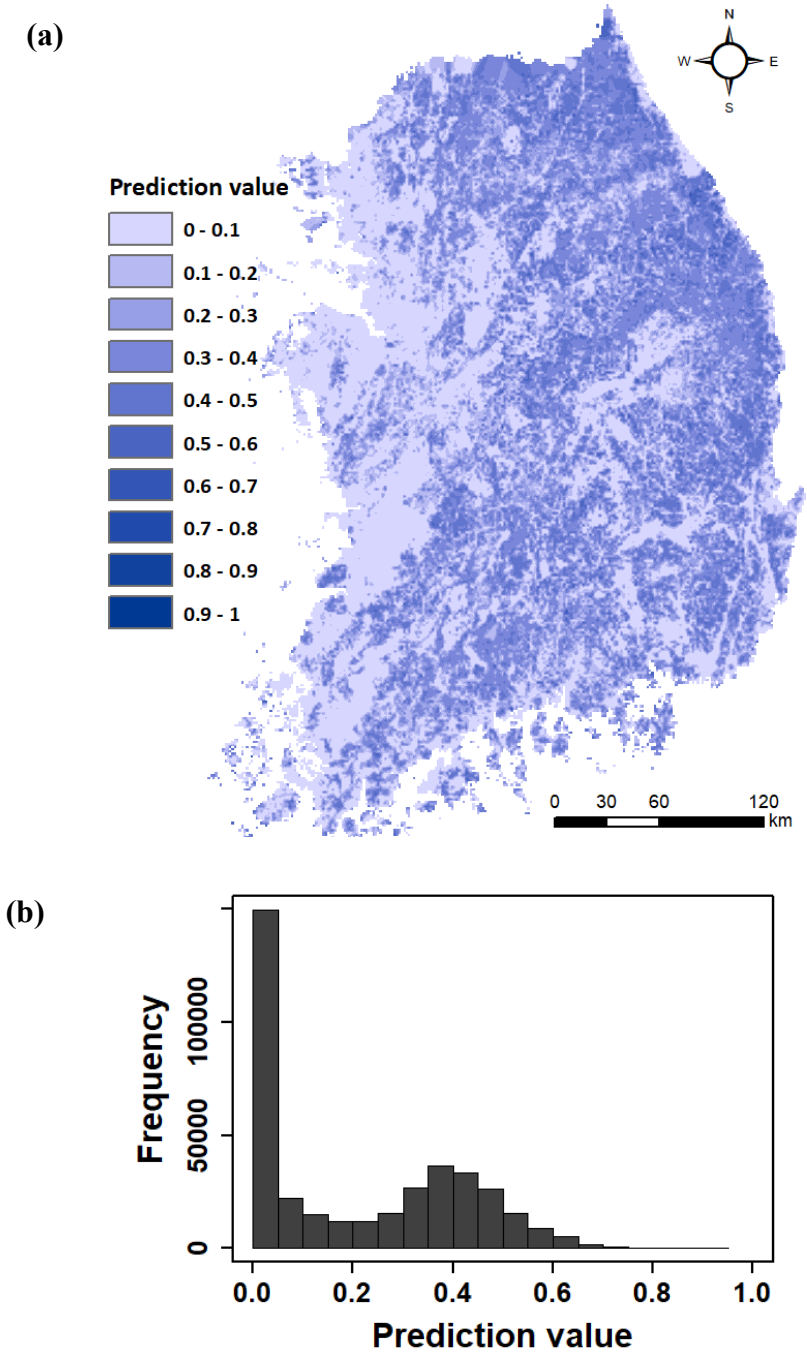


Figure 3-10. (a) Potential map describing the spatial distribution of predicted value for mineral water potential estimated by the ERT model. (b) Histogram of prediction values at grid points with 500 m equal interval for the ERT model.

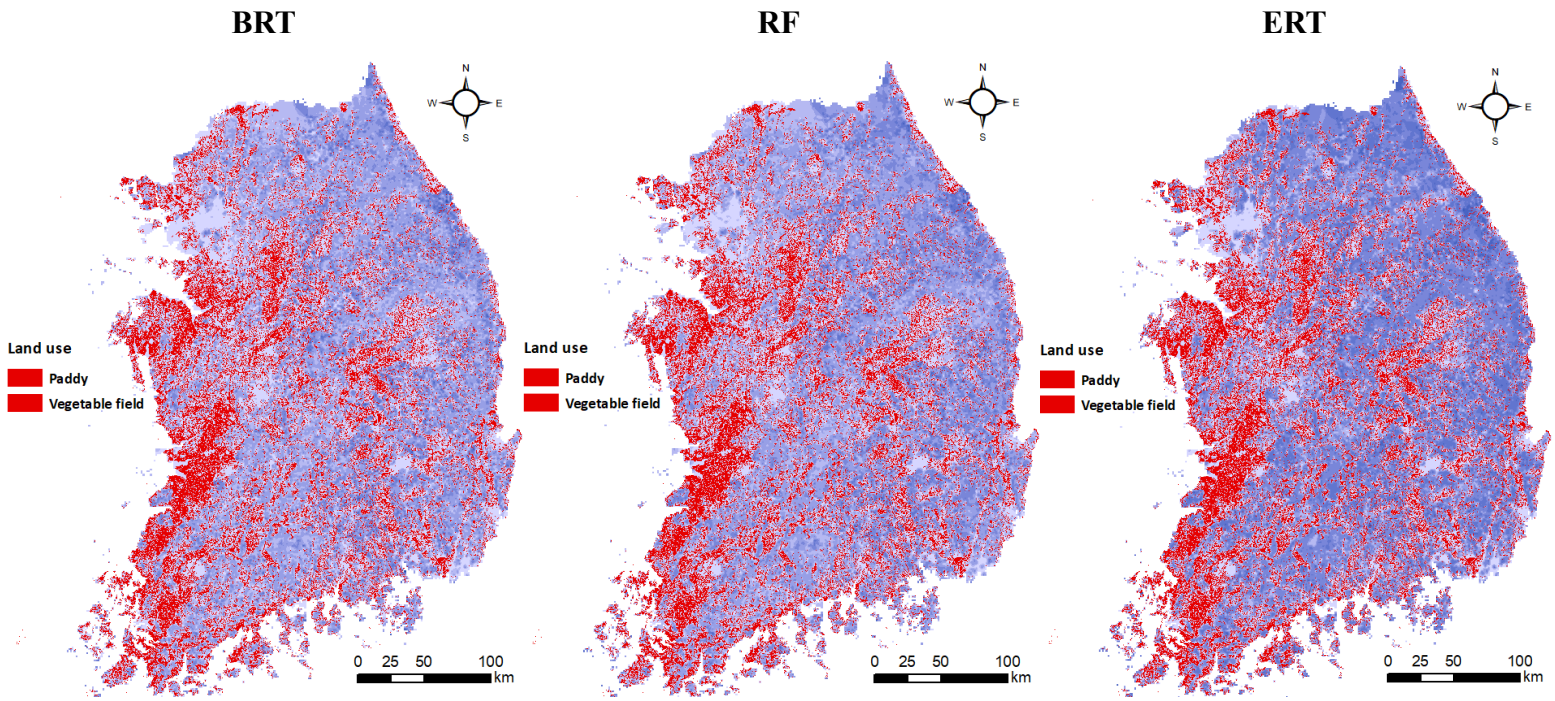


Figure 3-11. Location of cultivation-related variables is plotted on the BRT (left), RF (center), and ERT (right) potential maps. Areas exhibiting unsuitable response (bright blue color in potential map) are mostly covered by paddies and vegetable fields.

Table 3-5. A correlation matrix between the predictoin values in the BRT, RF, and ERT potential maps.

	BRT	RF	ERT
BRT	1	0.930	0.897
RF	0.930	1	0.953
ERT	0.897	0.953	1

3.3.2. Binary potential maps

Optimal thresholds for distinguishing the suitable or unsuitable responses in each model were determined in section 3.1. Applying these thresholds, three binary potential maps indicating the location of suitable or unsuitable groundwater for mineral water were illustrated (Fig. 3-12 ~ 3-14). Below the binary maps, the optimal threshold in each model was marked on the histogram showing the distribution of prediction values. In the ERT model, there were more prediction values above the optimal threshold than in the BRT and RF models, resulting in more suitable responses on the ERT binary map. Reflecting the characteristic of the originally imbalanced dataset, unsuitable response (0) widely occupied all three binary maps.

Ratios of the suitable response area to the total area for each model were calculated by regions, and their results are displayed in Fig. 3-15. The higher the proportion, the darker the blue on the ratio map. Range of the proportions varies with model because of the different threshold and distribution of prediction values between the three models. ERT ratio map had the widest range, with a range of approximately 0 to 40 %. The top five regions with the highest ratio for suitable response in each model are presented in Fig. 3-16 with their locations. Gangneung, Donghae, and Sokcho regions matched in the three results, implying the overlapped districts had high percentage of suitable area for mineral water in good agreement.

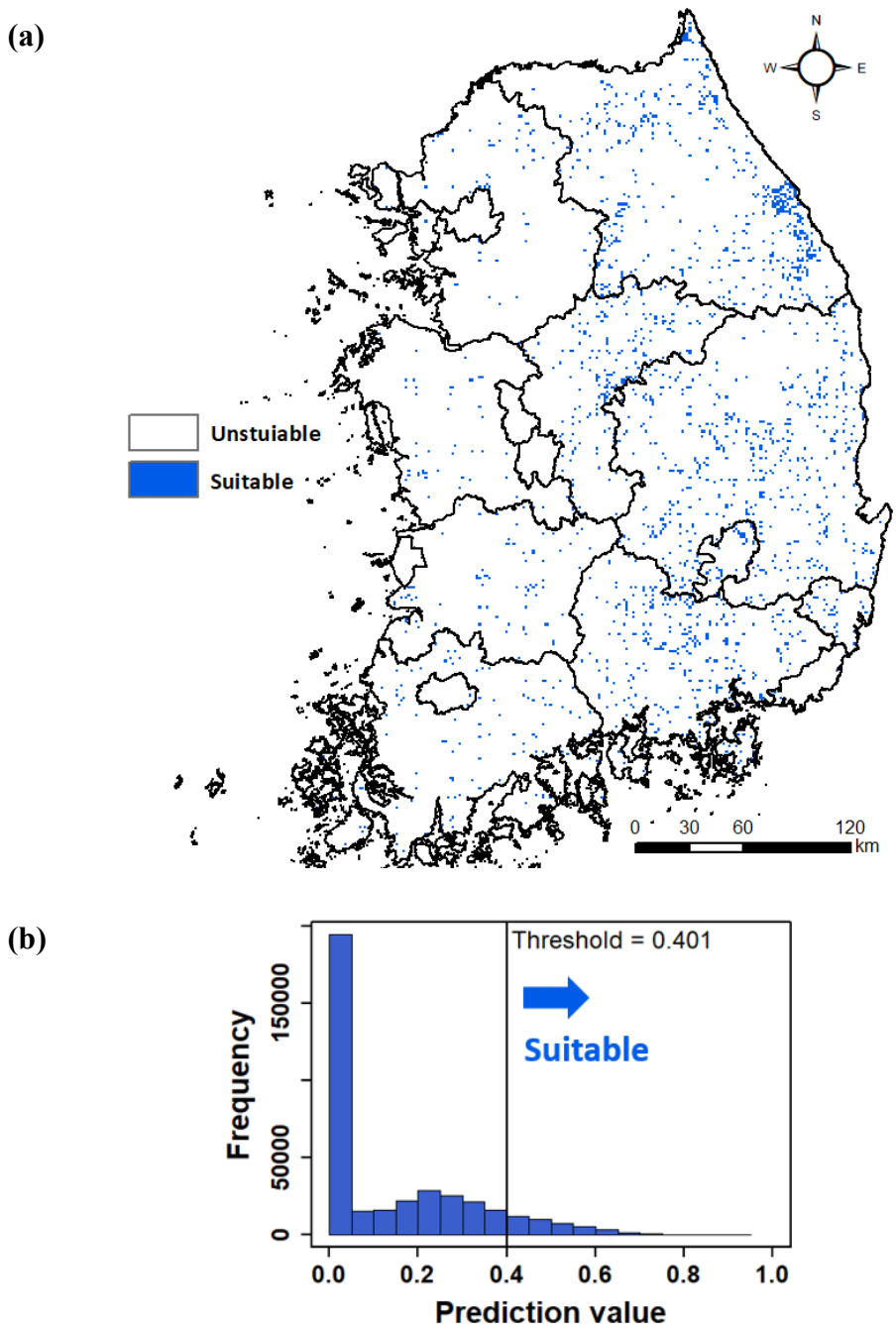


Figure 3-12. (a) Binary potential map delineating the spatial distribution of suitable groundwater for mineral water estimated by the BRT model. (b) Optimal threshold in the BRT model is marked on the histogram of the prediction values.

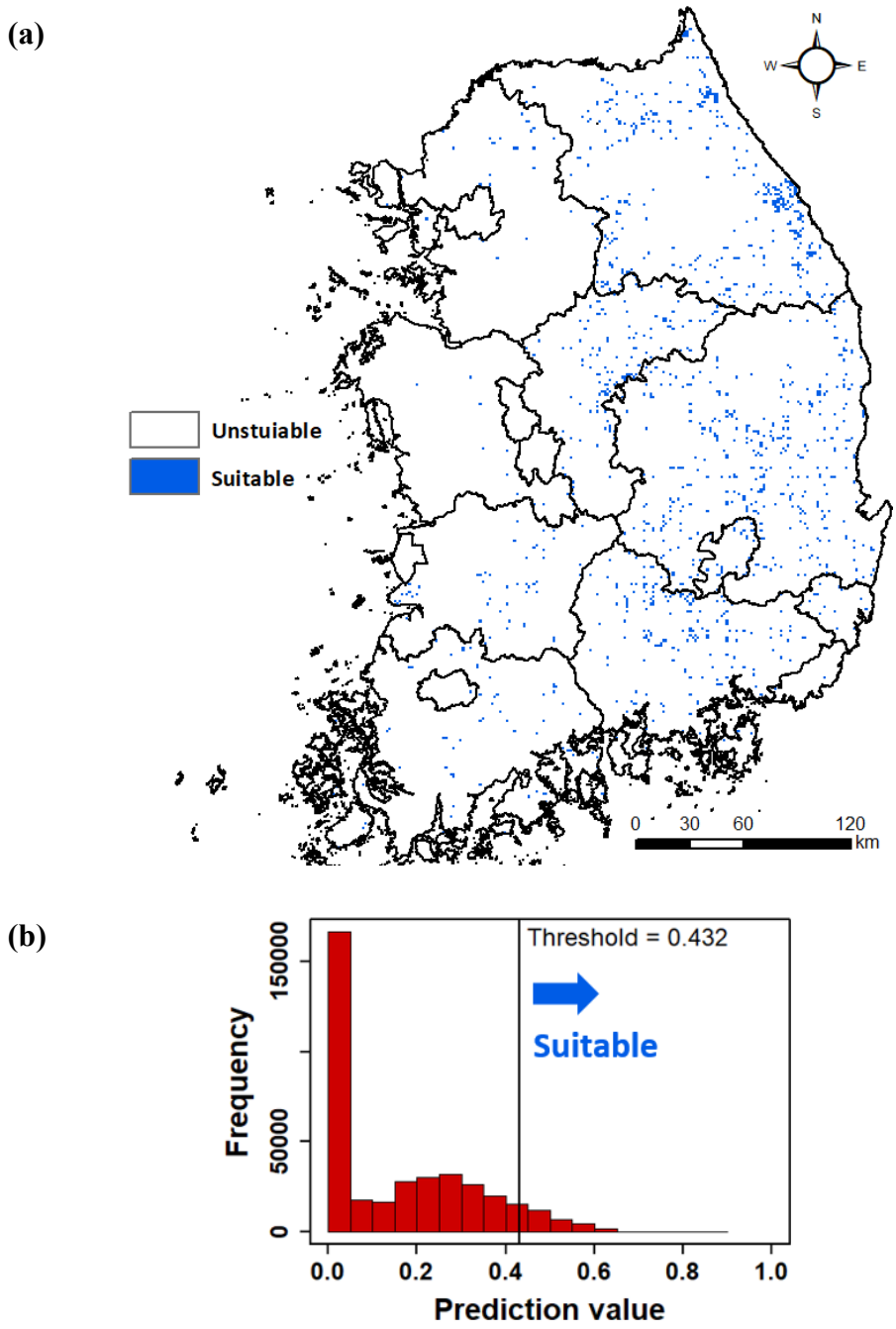


Figure 3-13. (a) Binary potential map delineating the spatial distribution of suitable groundwater for mineral water estimated by the RF model. (b) Optimal threshold in the RF model is marked on the histogram of the prediction values.

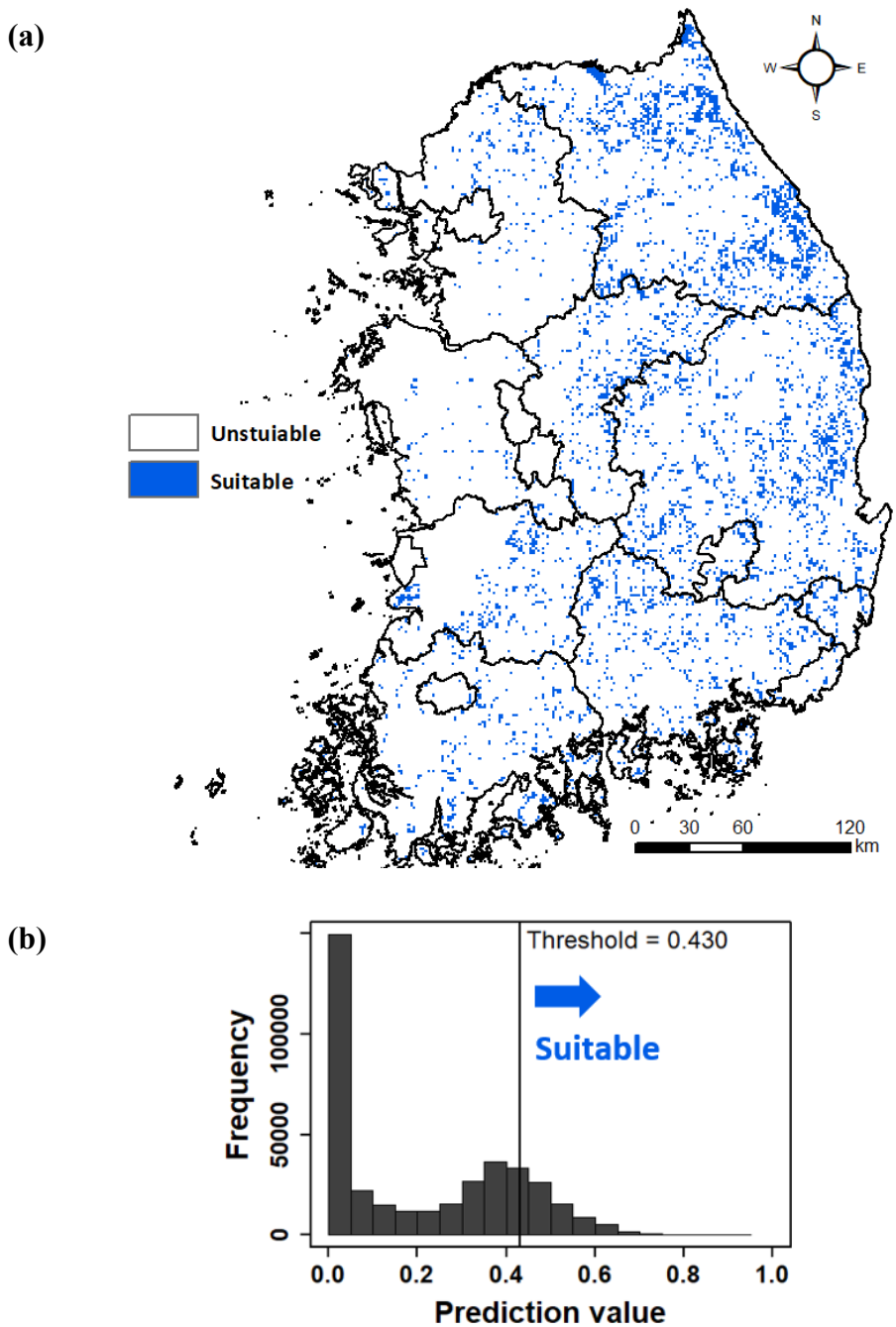


Figure 3-14. (a) Binary potential map delineating the spatial distribution of suitable groundwater for mineral water estimated by the ERT model. (b) Optimal threshold in the ERT model is marked on the histogram of the prediction values.

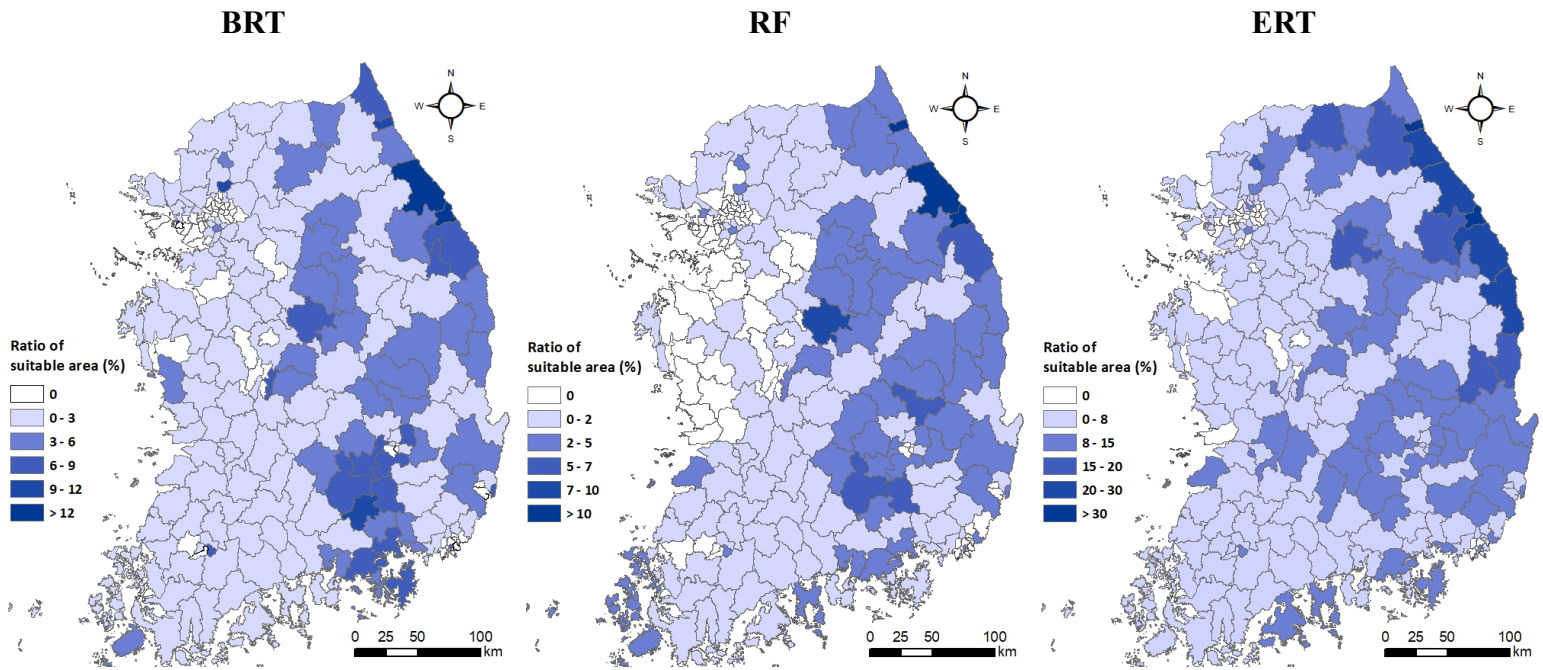


Figure 3-15. Three maps presenting the ratio of the suitable response area to the total area for BRT (left), RF (center), and ERT (right) models by region.

BRT

Rank	Province	Ratio (%)
1	Gangneung	15.2
2	Donghae	13.7
3	Sokcho	11.5
4	Uijeongbu	10
5	Uiryeong	9.3

RF

Rank	Province	Ratio (%)
1	Sokcho	22.4
2	Donghae	11.6
3	Gangneung	10.6
4	Goesan	7
5	Hapcheon	6.2

ERT

Rank	Province	Ratio (%)
1	Sokcho	40.3
2	Donghae	33.4
3	Yangyang	27.3
4	Gangneung	22.4
5	Samcheok	21.5

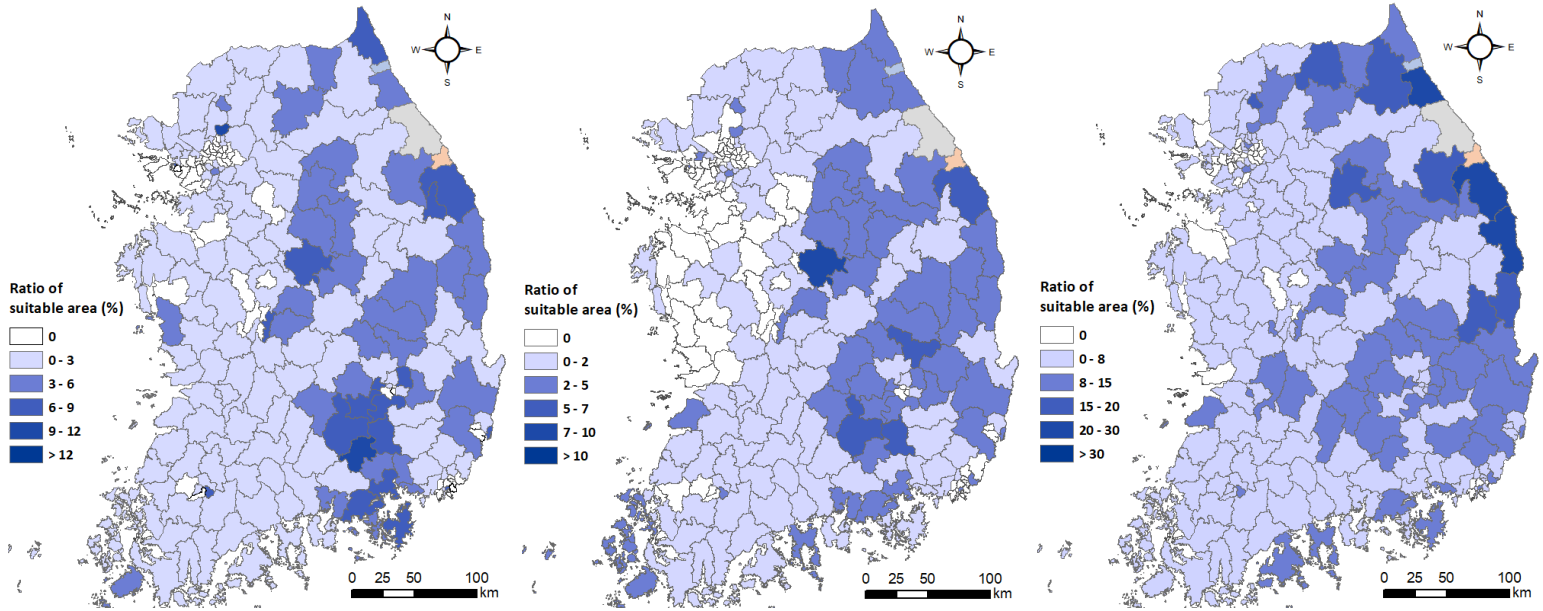


Figure 3-16. (a) The top five regions with the highest ratios for suitable response in the BRT (left), RF (center), and ERT (right) models. (b) Maps showed the location of the regions which the result of (a) overlapped between the three models.

4 CONCLUSION

This study estimated the spatial distribution of groundwater potential for mineral water in South Korea using tree ensemble approaches (BRT, RF, and ERT) and investigated which environmental indicators dominantly determine the groundwater quality for mineral water presence. Preferentially, the relationship between the groundwater quality transformed to the suitability for mineral water and the GIS-based environmental factors was trained to build models with the three methods. The AUC results of the PR curve demonstrated that the three constructed models are effective and applicable in describing the spatial suitability of groundwater for mineral water. Especially, among the three models, the BRT model estimated more suitable water with higher proportion, representing the best performance. Moreover, variable influence of each predictor was analyzed to understand the relationship between the variables and the response. As a result, some predictors of land use category mainly contribute to the model performance and the response variations in all three models, which indicates these factors are practical and important conditions for representing groundwater potential for mineral water. In the partial dependence plots, environmental conditions that dominantly determines the unsuitable response were cultivations, and the unsuitable distribution in the potential maps was well explained by the cultivation-related variables (Fig. 3-11). On the other hand, environmental conditions that primarily influence on the suitable response were forests, slope, and so on. The correlation (negative or positive) between some factors and the response appears to be related to nitrate concentration according to some references.

In both of the continuous potential maps and the binary potential maps, the three models generated remarkable similar distribution in the spatial patterns of the responses.

Similarity between the applicable models would enhance the utility of the potential maps and enable the determination of proper location for mineral water development with higher probability. Consequently, the tree ensemble machine learnings are proficient methods to relate the environmental factors and groundwater quality, when identifying groundwater potential zone for mineral water, and would be useful in efficient groundwater development and management.

5 REFERENCES

Akankpo, A. O., & Igboekwe, M. U. (2012). Application of geographic information system in mapping of groundwater quality for Michael Okpara University of Agriculture Umudike and its environs, Southeastern Nigeria. Appl. Sci. Res, 4, 1483-1493.

Al-Tabbal, J. A., & Al-Zboon, K. K. (2012). Suitability assessment of groundwater for irrigation and drinking purpose in the northern region of Jordan. Journal of Environmental Science and Technology, 5(5), 274-290.

Anbazhagan, S., & Nair, A. M. (2004). Geographic information system and groundwater quality mapping in Panvel Basin, Maharashtra, India. Environmental Geology, 45(6), 753-761.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Olshen. R., & Stone, C. (1984). Classification and Regression Trees, Wadsworth.

Chowdhury, A., Jha, M. K., Chowdary, V. M., & Mal, B. C. (2009). Integrated remote sensing and GIS-based approach for assessing groundwater potential in West Medinipur district, West Bengal, India. International Journal of Remote Sensing, 30(1), 231-250.

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. Ecology, 88(1), 243-251.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

Fawell, J. K., Lund, U., & Mintz, B. (2003). Total dissolved solids in drinking-water. Background document for development of WHO guidelines for drinking-water quality. World Health Organization, Geneva.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. Statistics in medicine, 22(9), 1365-1381.

Gardner, K. K., & Vogel, R. M. (2005). Predicting ground water nitrate concentration from land use. Groundwater, 43(3), 343-352.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

Giri, A., Bharti, V. K., Kalia, S., Kumar, K., Raj, T., & Kumar, B. (2017). Utility of multivariate statistical analysis to identify factors contributing groundwater quality in high altitude region of Leh-Ladakh, India. Asian Journal of Water, Environment and Pollution, 14(4), 61-75.

Gowd, S. S. (2005). Assessment of groundwater quality for drinking and irrigation purposes: a case study of Peddavanka watershed, Anantapur District, Andhra Pradesh, India. Environmental Geology, 48(6), 702-712.

Grootjans, A. P., Van Diggelen, R., Wassen, M. J., & Wiersinga, W. A. (1988). The effects of drainage on groundwater quality and plant species distribution in stream valley meadows. Vegetatio, 75(1-2), 37-48.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Igboekwe, M. U., & Akankpo, A. O. (2011). Application of Geographic Information System (GIS) in mapping groundwater quality in Uyo, Nigeria. International Journal of Geosciences, 2(4), 394.

Ikem, A., Osibanjo, O., Sridhar, M. K. C., & Sobande, A. (2002). Evaluation of groundwater quality characteristics near two waste sites in Ibadan and Lagos, Nigeria. Water, Air, and Soil Pollution, 140(1-4), 307-333.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, pp. 3-7). New York: springer.

Khatri, N., & Tyagi, S. (2015). Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. Frontiers in Life Science, 8(1), 23-39.

Khedr, M. G. (2013). Radioactive contamination of groundwater, special aspects and advantages of removal by reverse osmosis and nanofiltration. Desalination, 321, 47-54.

Knierim, K. J., Kingsbury, J. A., Haugh, C. J., & Ransom, K. M. (2020). Using Boosted Regression Tree Models to Predict Salinity in Mississippi Embayment Aquifers, Central United States. JAWRA Journal of the American Water Resources Association.

Knoll, L., Breuer, L., & Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the total environment, 668, 1317-1327.

Kord, M., & Moghaddam, A. A. (2014). Spatial analysis of Ardabil plain aquifer potable groundwater using fuzzy logic. Journal of King Saud University-Science, 26(2), 129-140.

Korea Rural Community Corporation (KRCC). (2020). Homepage. https://www.groundwater.or.kr.

Kouzana, L., Mammou, A. B., & Felfoul, M. S. (2009). Seawater intrusion and associated processes: case of the Korba aquifer (Cap-Bon, Tunisia). Comptes Rendus Geoscience, 341(1), 21-35.

Lee, S., Hong, S. M., & Jung, H. S. (2018). GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. Geocarto international, 33(8), 847-861.

Lerner, D. N., & Harris, B. (2009). The relationship between land use and groundwater resources and quality. Land use policy, 26, S265-S273.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. Advances in neural information processing systems, 26, 431-439.

Madison, R. J., & Brunett, J. O. (1985). Overview of the occurrence of nitrate in ground water of the United States: US Geological Survey Water-Supply Paper 2275.

Molnar, C. (2020). Interpretable Machine Learning. Lulu. com.

Movahedi, F., Padman, R., & Antaki, J. F. (2020). Limitations of ROC on Imbalanced Data: Evaluation of LVAD Mortality Risk Scores. arXiv preprint arXiv:2010.16253.

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environmental monitoring and assessment, 188(1), 44.

National Groundwater Information Management & Service Center (GIMS). (2020). Home page. http://www.gims.go.kr/gims_start.do.

Negnevitsky, M. (2005). Artificial intelligence: a guide to intelligent systems. Pearson education.

Peiyue, L., Qian, W., & Jianhua, W. (2011). Groundwater suitability for drinking and agricultural usage in Yinchuan Area, China. International journal of Environmental sciences, 1(6), 1241-1249.

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. Catena, 137, 360-372.

Rhee, J., Park, K., Lee, S., Jang, S., & Yoon, S. (2020). Detecting hydrological droughts in ungauged areas from remotely sensed hydro-meteorological variables using rule-based models. Natural Hazards, 103(3), 2961-2988.

Saha, R., Dey, N. C., Rahman, S., Galagedara, L., & Bhattacharya, P. (2018). Exploring suitable sites for installing safe drinking water wells in coastal Bangladesh. Groundwater for Sustainable Development, 7, 91-100.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.

Sander, P., Chesley, M. M., & Minor, T. B. (1996). Groundwater assessment using remote sensing and GIS in a rural groundwater project in Ghana: lessons learned. Hydrogeology Journal, 4(3), 40-49.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425, 76-91.

Steinberg, D., & Colla, P. (1995). CART: tree-structured non-parametric data analysis. San Diego, CA: Salford Systems.

Thivya, C., Chidambaram, S., Singaraja, C., Thilagavathi, R., Prasanna, M. V., Anandhan, P., & Jainab, I. (2013). A study on the significance of lithology in groundwater quality

of Madurai district, Tamil Nadu (India). Environment, development and sustainability, 15(5), 1365-1387.

Todd, D. K., & Mays, L. W. (2005). Groundwater hydrology edition. Welly Inte.

Wick, K., Heumesser, C., & Schmid, E. (2012). Groundwater nitrate contamination: factors and indicators. Journal of Environmental Management, 111, 178-186.

Zulu, G., Toyota, M., & Misawa, S. I. (1996). Characteristics of water reuse and its effects on paddy irrigation system water balance and the riceland ecosystem. Agricultural Water Management, 31(3), 269-283.

국문 초록

음용 지하수에 대해 점차 증가하고 있는 수요를 충족시키기 위해 수질을 고려하여 지하수 부존가능 지역을 파악하는 것이 중요하다. 특히, 지하수 자원의 효율적인 관리와 음용 지하수의 성공적인 개발을 위해서는 지하수 부존가능 지역에 대한 합리적이고 정확한 탐지가 필수적이다. 본 연구에서는 여러 음용도 중 광천수를 대상으로 하여, 부스팅 회귀 나무, 랜덤 포레스트, 익스트림 랜덤 트리를 사용하여 남한 전역의 광천수에 대한 지하수 부존가능성의 공간적 분포를 추정하였다. 반응 변수를 결정하기 위해 총 6,135 개의 지하수질 데이터를 전국규모로 수집하였으며, 고도, 경사, 배수등급, 유효토심, 표토토성, 토지이용, 수문지질의 환경적 요인들이 예측변수로 작용했다. 정밀 리콜 곡선 분석 결과, 적용된 세 가지 방법의 모든 곡선들이 기준선과는 명확하게 구분되었으며, 지하수 부존지역 매핑에 대한 세 분류기의 적용가능성을 확인해주었다. 또한, 상대 영향도와 부분 의존도 플롯으로 오염 관련 요인들이 모델링과 부존가능성 매핑에 영향을 크게 끼쳤음을 파악하였다. 마지막으로, 세 가지의 검증된 모델로 광천수 부존가능성에 대한 공간적 분포를 생성하였다. 생성된 지도들은 불균형한 데이터의 분포와 몇몇 예측 변수들의 영향도 결과를 잘 반영하였다. 또한, 세 지도를 비교함으로써 광천수 개발에 적절한 위치를 결정할 수 있었다. 결과적으로, 나무 앙상블 방법은 전국 규모의 광천수에 대한 지하수 부존가능성을 기술하는 데 있어 그 활용이 기대되며, 변수 영향도 분석을 통해서는 광천수 산출에 영향을 미치는 환경 조건을 파악할 수 있을 것이다.

주요어: 지하수 부존가능성 지도, 광천수, 부스팅 회귀 나무, 랜덤 포레스트, 익스트림 랜덤 트리, 변수 영향도