Groundwater Productivity Potential Mapping Using Logistic Regression and Boosted Tree Models: The Case of Okcheon City in Korea

Saro Lee, Chang-Wook Lee, and Jeong-Cheon Kim

Abstract

This study analyzed Groundwater Productivity Potential (GPP) using different models in a geographic information system (GIS) in Okcheon area, Korea. These models used the relationship between groundwater-productivity data, including specific capacity (SPC) and transmissivity (T), and its related hydrogeological factors. Data about related factors, including topography, lineament, geology, forest and soil were constructed to a spatial database. Additionally, T and SPC data were collected from 86 well locations. Then, GPP were mapped using the Logistic Regression (LR) and Boosted Tree Regression (BT) models. The resulting GPP maps were validated using Area Under Curve (AUC) analysis with the well data. The GPP maps using the LR and BT models had accuracies of 85.04 and 81.66% with T value, respectively. And the GPP maps using the LR and BT models had accuracies of 82.22 and 81.53% with SPC value, respectively. These results indicate that LR and BT models can be useful for GPP mapping.

Keywords

Groundwater potential • GIS • Logistic regression • Boosted tree • Korea

1 Introduction

Groundwater is one of the important natural resources used in agriculture, industry and public water supply. In Korea, the use of groundwater increased by more than 225% between 1994 and 2014, and the current national supply of groundwater no longer meets the needs of society. Therefore, reliable analytical models predicting locations of groundwater are needed for efficient management use of groundwater. So, the purpose of the study was to develop and apply the GIS based Groundwater Productivity Potential (GPP) model using Logistic Regression (LR) and Boosted Tree (BT) models in the Okcheon country of Korea. The GPP is defined as the probability of finding out groundwater in an area. Especially, the study mainly used topographical factors among various others, because groundwater is most affected by such factors. Recently, many GPP mapping studies that have been published used new models such as Frequency Ratio (FR) [1], Artificial Neural Network (ANN) [2], Random Forest (RF) [3], Logistic Regression (LR) [4], Boosted Regression Tree (BTR) [5] and Support Vector Machine (SVM) [2].

For the GPP mapping, T (Transmissivity) and SPC (specific capacity) point data were obtained and randomly classified as either training data (50%) or validation data (50%). Geology, topography, soil texture, and land cover data were combined into a spatial database. Hydrogeological factors, including slope, aspect, slope gradient, relative slope position, hydraulic slope, valley depth, topographic wetness index (TWI), slope length (LS) factor, convergence index, depth from groundwater, distance from lineament, distance from channel network, and so forth, were extracted from spatial databases. Then T and SPC data were selected (T values ≥ 2.6, SPC values ≥ 4.875) as training data for the three models. Finally, the GPP maps were assessed using AUC techniques.

S. Lee (✉)
Korea Institute of Geoscience and Mineral Resources (KIGAM), 124 Gwahang-no, Yuseong-gu, Daejeon, 305-350, South Korea
e-mail: leesaro@kigam.re.kr

C.-W. Lee
Kangwon National University, 1 Kangwondaehak-Gil, Chuncheon-si, Gangwon-Do 200-701, South Korea

J.-C. Kim
National Institute of Ecology (NIE), 1210 Geumgang-Ro, Maseo-Myeon, Seocheon-gun, Chungcheongnam-Do 33657, South Korea

H. M. El-Askary et al. (eds.), *Advances in Remote Sensing and Geo Informatics Applications*, Advances in Science, Technology & Innovation, https://doi.org/10.1007/978-3-030-01440-7_69

2 Data and Method

The study area is the Okcheon country of South Korea. This area lies between 36°10′N and 36°26′N latitude and 127°29′E and 127°53′E longitude and covers 537.06 km^2. Since groundwater is associated with drinking and irrigation water supplied to communities, it is very meaningful to estimate GPP.

This study using LR and BT models are based on the relationship between groundwater productivity data (SPC and T) and hydrogeological factors (Table 1). To calculate groundwater productivity, SPC and T are set as dependent variables and various hydrogeological factors are set as independent variables. SPC is the amount of water that can be produced per unit drawdown. Also, T is the rate of flow under a unit hydraulic gradient through a unit width of aquifer of given saturated thickness. The groundwater productivities respond to a total of 86 cells (each 43 cells (including the T data of ≥ 2.6 m^2/d, SPC ≥ 4.875 m^3/d/m) for training and 43 cells for validation.

The LR model is to help find the best expression to describe the relationship between dependent variables and various independent variables. The BT model is a general calculation method of stochastic gradient amplification. Ultimately, this approach allows fitting the best estimate of the observed values to yield better results. In summary, the GPP mapping was performed as follows: (1) geospatial data were constructed and the related factors were extracted or calculated, (2) a geospatial database was founded with a grid, (3) the GPP assessment was conducted using the LR and BT models, and (4) the validation of the potential map was achieved using AUC.

3 Results

The GPP maps using the LR and BT models results are shown in Fig. 1. The AUC was recalculated since the total area used the well data that had not been used for the training the models. From the validation, the LR and BT models produced AUC values of 0.8113 and 0.8372 by T value, respectively. Also, the validation of the GPP maps, the LR and BT models produced AUC values of 0.8024 and 0.8080 by SPC value, respectively.

4 Discussion and Conclusion

This study applied and assessed the LR (statistics) and BT (data mining) models for groundwater potential. As a result, the accuracies were computed as 85.04 and 81.66% for LR and BT models with T value, 82.22 and 81.53% for LR and BT models with SPC value, respectively. Therefore, it can be concluded that LR with T value had the best performance. In addition, other models using T or SPC values in this study also showed a good accuracy of over 80% when predicting spatially groundwater potential.

From the result of calculated LR models table or predictor importance of BT model, in order of influence, the relationships between well data and the examined factors were

Table 1 Data layers of the study area

Original data	Factors	Data type	Scale
Yield	T[m^2/d/m], SPC [m^3/d/m]	Point	
Topographical map[a]	Slope [°], Aspect, Relative slope position Plan curvature, Topographic Wetness Index (TWI), Slope Length factor (LS-factor), Convergence index, Lineament density, Drainage basin, Hydraulic slope [m], Valley depth [m], Depth to groundwater[m]	Grid	1:5000
Geological map[b]	Hydrogeology	Polygon	1:50,000
Soil map[c]	Soil texture	Polygon	1:25,000
Land cover map[d]	Land cover, Distance from fault [m], Distance from lineament [m], Distance from channel network [m]	Polygon	1:5000

[a]Topographical factors were extracted by the National Geographic Information Institute (NGII)
[b]The geology map offered by the Ministry of Land, Transport and Maritime Affairs (MLTM)
[c]The soil map was offered by the National Institute of Agricultural Science and Technology
[d]The land cover map was offered by the Korean Ministry of Environment

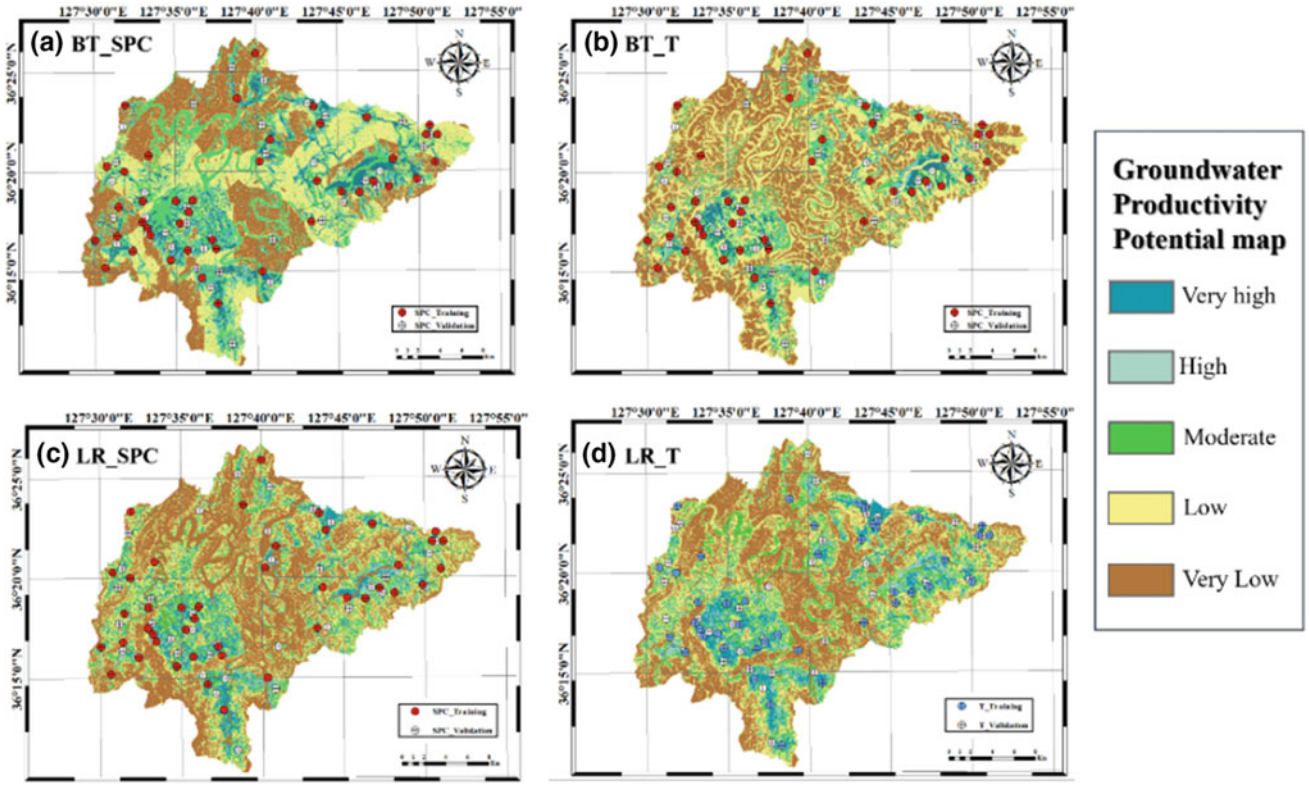

Fig. 1 GPP maps using logistic regression (LR) and boosted tree (BT) models

as follows. With gentle slope & hydraulic slope, lower relative slope position, and shorter slope length, GPP was estimated to be higher. However, with steeper slope & hydraulic slope, higher relative slope position, and longer slope length, GPP was estimated to be higher because rainfall running off in the upper region is accumulated in the lower region and influences the aquifer. On the other hand, the distance from the fault, distance from lineament, distance from channel network showed a negative correlation with GPP. The closer the channel is, the greater the GPP will be because the rivers have gotten water from the underground.

The proposed GPP mapping method can be applied to groundwater use planning and management, such as regional groundwater development planning, water system control based on systematic and objective planning. Finally, it can be deduced that new models of more recently developed statistics and data mining models could provide better results in future studies.

References

1. Jothibasu, A., Anbazhagan, S.: Spatial mapping of groundwater potential in Ponnaiyar River basin using probabilistic-based frequency ratio model. Model. Earth Syst. Environ. **3**(1), 33 (2017)
2. Lee, S., Hong, S.-M., & Jung, H.-S.: GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. Geocarto. Int., 1–15 (2017)
3. Rahmati, O., Pourghasemi, H.R., Melesse, A.M.: 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 (2016)
4. Park, S., Hamm, S.-Y., Jeon, H.-T., Kim, J.: Evaluation of logistic regression and multivariate adaptive regression spline models for groundwater potential mapping using R and GIS. Sustainability **9** (7), 1157 (2017)
5. Mousavi, S.M., Golkarian, A., Naghibi, S.A., Kalantar, B., Pradhan, B.: GIS-based groundwater spring potential mapping using data mining boosted regression tree and probabilistic frequency ratio models in Iran. AIMS Geosci. **3**(1), 91–115 (2017)