6th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Improving of groundwater level estimation using SMOTE technique

Emine Dilek Taylan*[1], Tahsin Baykal [1], Özlem Terzi [2]

[1]*Suleyman Demirel University, Civil Engineering Department, Isparta, Türkiye*
[2]*Isparta Applied Sciences University, Civil Engineering Department, Isparta, Türkiye*

Keywords
Groundwater
SMOTE
ET algorithm
Denizli

Abstract

Accurate estimation of the groundwater level is very important for sustainable water management and planning of water resources. Since the groundwater field studies are time consuming and costly, the use of machine learning techniques for the groundwater level estimation is proposed in this study. Also, synthetic minority oversampling technique (SMOTE) algorithm was used to increase the success of the extra tree (ET) algorithm, which is one of the machine learning techniques, in estimating the groundwater level. For this, precipitation, elevation, slope, and curvature data, which are effective on groundwater, were used. First, the groundwater level was classified as very low, low, medium high and very high. When the result of the model developed with ET algorithm was evaluated, the accuracy value was calculated as 0.53 and the Cohen's Kappa value as 0.36. Then, to increase the success of the extra tree model, data irregularities were removed with the SMOTE algorithm. It was observed that the accuracy of the SMOTE-ET model increased to 0.69 and the Cohen's Kappa value to 0.62. When the results are evaluated, it is thought that the SMOTE algorithm increases the success of the ET algorithm in estimating the groundwater level.

1. Introduction

Accurate estimation of groundwater level is one of the requirements for sustainable use and management of water resources. Modeling of groundwater level provides benefits to managers and engineers in decision making, especially due to increasing groundwater demands for agricultural use. In recent years, groundwater level modeling studies done by many researchers (Daliakopoulos 2005; Poursaeid et al. 2022; Sahoo and Jha 2013; Yadav et al. 2020) Afzaal et al. (2020) estimated groundwater levels for the Baltic River and Long Creek watersheds in Canada using deep learning and artificial neural networks with stream level, stream flow, precipitation, relative humidity, mean temperature, evapotranspiration, heat degree days, dew point temperature, and evapotranspiration variables. They said that these techniques are convenient and accurate. Dash et al. (2010) tried the ANN-GA hybrid model for groundwater level estimation and found that it was successful. Synthetic minority oversampling technology (SMOTE) is used to eliminate class imbalance and improve model performance with new samplings between minority classes and their neighbors in the input dataset. It has been widely preferred in the field of class imbalance in recent years. Wang et al. (2019) produced landslide susceptibility maps with higher accuracy using SMOTE with machine learning techniques such as support vector machine (SVM), logistic regression (LR), artificial neural network (ANN) and random forest (RF). Tang et al. (2022) stated that more appropriate estimations will be made with the XGBoosting method using SMOTE for the precipitation forecast in the Danjiangkou River basin. Razali et al. (2020) stated that the decision tree machine learning technique applied using SMOTE achieved 99.92% success in flood risk estimation for the Kuala Krai region in Kelantan, Malaysia.

In this study, the model was developed to estimate groundwater level with extra tree (ET) algorithm using precipitation, elevation, slope, and curvature data for Denizli province in Türkiye. In addition, improvement of ET model was made by SMOTE.

2. Method

2.1. Study region and data

Denizli province, located in the Aegean Region, is located between 28° 30' – 29° 30' east meridians and 37°

* **Corresponding Author**

*(dilektaylan@sdu.edu.tr) ORCID ID 0000 – 0003 – 0734 – 1900
(tahsinbaykal@gmail.com) ORCID ID 0000 – 0001 – 6218 – 0826
(ozlemterzi@isparta.edu.tr) ORCID ID 0000 – 0001 – 6429 – 5176

Cite this study

Taylan, E. D., Baykal, T., & Terzi, Ö. (2023). Improving of Groundwater Level Estimation Using SMOTE Technique. Intercontinental Geoinformation Days (IGD), 6, 74-77, Baku, Azerbaijan

12' – 38° 12' north parallels. It has a mild and rainy climate in winters. It is bordered by Burdur and Afyon in the east, Aydın and Manisa in the west, Uşak in the north and Muğla in the south. The study region map was given in Figure 1.

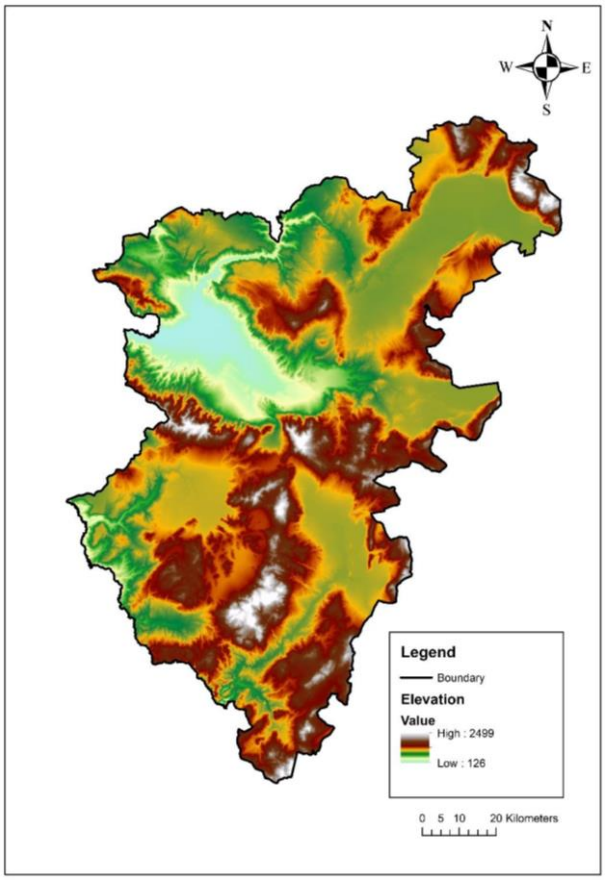

Figure 1. The study region map

Groundwater levels taken from Çıldır (2017) consist of 258 well measurements in Denizli, Türkiye. Monthly precipitation data were obtained from Nasa Power for the years 1981-2021 of Çardak, Buldan, Çivril, Acıpayam, Beyağaç and Çameli districts. Groundwater level and precipitation maps were created with Inverse Distance Weighting (IDW) in ArcGIS environment. Elevation, slope, and curvature data were obtained from SRTM dem data in ArcGIS environment. All maps have pixel dimensions set to 1x1 km.

2.2. Synthetic minority oversampling technique (SMOTE)

SMOTE algorithm, which is an oversampling method, creates new samples from minority classes without repetition. So, it can simply and effectively reduce the phenomenon of imbalance. The SMOTE algorithm creates new instances of the minority class using the Euclidean distance between instances of the minority class (x_i) and obtains k-nearest neighbors. A sampling rate M is calculated to determine the new proportion of various samples in accordance with the imbalance rate of the applied dataset. For x_i, several samples are randomly selected from the k-nearest neighbors, assuming that the chosen nearest neighbor is x_m. For each randomly selected neighbor x_m, a new sample is created from the original sample based on Equation 1.

$$x_{new} = x + rand(0,1) \times |x_i - x|, new \in 1,2,\dots,M \quad (1)$$

Here, rand (0,1) denotes a random number in the range 0 to 1. The above steps are repeated M times to create M new samples. That is, if the total number of the minority category is T, new samples are generated MT. After combining the new samples with several original samples, a new balanced dataset is obtained. (Chen et al. 2022; Geetha et al. 2019; Ishaq et al. 2021). The main processes of the SMOTE algorithm are given in Figure 2.

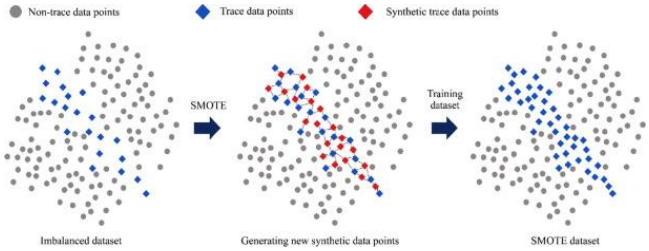

Figure 2. SMOTE algorithm main processes (Chen et al. 2022)

2.3. Extra tree algorithm

Extra trees are one of the decision tree-based ensemble learning methods. Unlike other tree-based ensemble methods, extra tree separates nodes with randomly chosen cut-points and develops trees using the entire learning sample to minimize bias (Geurts et al. 2006). Each decision tree is created from the dependency relationship of the extracted data samples (Kumar et al. 2020; Sharaff and Gupta 2019). At the same time, each decision tree selects the best features based on the mathematical basis of the Gini Index given in Equation 2.

$$gini\ (D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2)$$

where m represents the number of output labels and pi indicates the probability that a sample belongs to the i^{th} output label in dataset D (Du et al. 2023).

2.4. Evaluate metrics

Cohen's Kappa coefficient (k) represents a measure of how many samples are classified in the machine learning model (Equation 3).

$$K = \frac{(p0-pe)}{(1-pe)} \quad (3)$$

Here, p0 and pe show the total accuracy and random accuracy of the model, respectively (Vujović 2021). Precision is obtained by dividing the true positive (TP) predictions by the total number of true positive and false positive (FP) predictions (Equation 4). The recall given in Equation 5 is obtained by dividing the correct positive predictions by the total number of positives.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Here FN is false negative. Accuracy is one of the most popular criteria in multiclass classification and is calculated directly from the confusion matrix (Equation 6)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

F1-Score, which collects precision and recall measures under the concept of harmonic mean, is given in Equation 7:

$$F1 - Score = 2 \times \left(\frac{precision \times recall}{precision+recall}\right) \quad (7)$$

The best value for the F1 score is obtained at 1 and the worst value at 0 (Grandini et al. 2020).

3. Results and Discussion

In this section, the modeling results for the estimation of the groundwater level are presented. By using precipitation, elevation, slope and curvature data, groundwater level was estimated with the extra tree (ET) algorithm. Then, SMOTE algorithm on data were used to improve the performance of the ET algorithm. While developing the model, the data were randomly divided into 80% training and 20% testing. Cohen's Kappa, Accuracy, Recall, Precision and F1-Score metrics were used to evaluate the model result.

Groundwater level is divided into five classes as very low, low, medium, high, and very high. Before the ET model was developed, the ratio of each class to the total data was examined. It is seen that the very high class constitutes only 4% of the total data and the data is unbalanced distributed.

In the study, firstly, the ET model was developed by ignoring the data unbalance. The accuracy value of the developed model was calculated as 0.53 and the Cohen's Kappa value as 0.36. Precision, recall and F1-score values calculated for each class in the test set are given in Table 1, and confusion matrix is given in Figure 3.

Table 1. Evaluation metric results of ET model test set

Class	Precision	Recall	F1 score
Very low	0.63	0.65	0.64
Low	0.51	0.61	0.56
Moderate	0.46	0.42	0.44
High	0.51	0.42	0.46
Very high	0.35	0.13	0.19

When the accuracy and Cohen's Kappa values calculated for the ET model and Table 1 are examined, it is seen that this model is not sufficient for estimating the groundwater level. Then, the SMOTE algorithm was used to increase the success of the ET model. The ratio of each class belonging to the groundwater level to the total number of data was balanced with the SMOTE algorithm.

After each class was equalized with the SMOTE algorithm, the ET model was reapplied to the new data set. The accuracy value of the reconstructed SMOTE-ET model was 0.69 and the Cohen's Kappa value was 0.62. Precision, recall and F1-score values calculated for each class in the test set are given in Table 2, and confusion matrix is given in Figure 4.

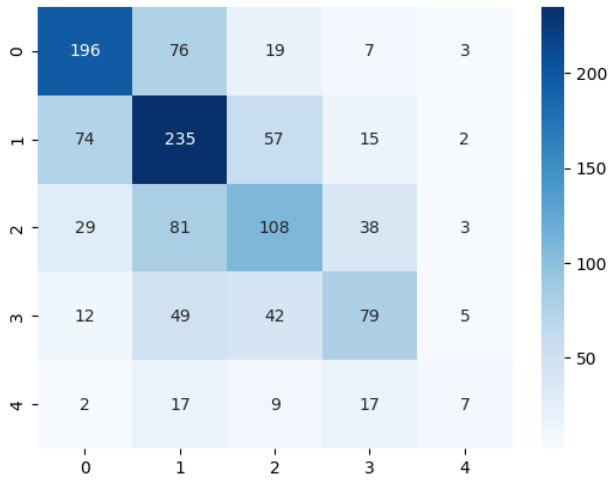

Figure 3. Confusion matrix of ET model test set (0-Very low, 1-Low, 2-Moderate, 3-High and 4-Very high)

Table 2. Evaluation metric results of SMOTE-ET model test set

Class	Precision	Recall	F1 score
Very low	0.68	0.76	0.72
Low	0.59	0.49	0.54
Moderate	0.63	0.61	0.62
High	0.70	0.72	0.71
Very high	0.83	0.88	0.86

Figure 4. Confusion matrix of SMOTE-ET model test set (0-Very low, 1-Low, 2-Moderate, 3-High and 4-Very high)

When the SMOTE-ET model accuracy and Cohen's Kappa value are examined, it is seen that it gives better results than the ET model. In addition, when Tables 1 and 2 are compared, it is seen that although the recall and F1-Score values of the Low class in the SMOTE-ET model are lower than the recall and F1-Score values of the Low class in the ET model, it gives better results in other classes.

4. Conclusion

In this study, ET algorithm was used to estimate the groundwater level of Denizli, Türkiye. In order to

increase the success of the estimation results obtained with ET algorithm, the imbalanced data in the groundwater level are eliminated with the SMOTE algorithm. Thus, more accurate groundwater level estimates were obtained.

References

Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., & Esau, T. (2020). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. Water, 12(1), 5. https://doi.org/10.3390/w12010005

Chen, J., Huang, H., Cohn, A. G., Zhang, D., & Zhou, M. (2022). Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning. International Journal of Mining Science and Technology, 32(2), 309-322. https://doi.org/10.1016/j.ijmst.2021.08.004

Çıldır, M. A. (2017). measurement of groundwater levels in Denizli and mapping with geographic information systems (Master's thesis, Pamukkale University Institute ff Science).

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. Journal of hydrology, 309(1-4), 229-240. https://doi.org/10.1016/j.jhydrol.2004.12.001

Dash, N. B., Panda, S. N., Remesan, R., & Sahoo, N. (2010). Hybrid neural modeling for groundwater level prediction. Neural Computing and Applications, 19, 1251–1263. https://doi.org/10.1007/s00521-010-0360-1

Du, Y., Liu, Y., Yan, Y., Fang, J., & Jiang, X. (2023). Risk management of weather-related failures in distribution systems based on interpretable extra-trees. Journal of Modern Power Systems and Clean Energy. https://doi: 10.35833/MPCE.2022.000430.

Geetha, R., Sivasubramanian, S., Kaliappan, M., Vimal, S., & Annamalai, S. (2019). Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. Journal of medical systems, 43, 1-19. https://doi.org/10.1016/j.ijmst.2021.08.004

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42. https://doi.org/10.1007/s10994-006-6226-1

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756. https://doi.org/10.48550/arXiv.2008.05756

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE access, 9, 39707-39716. https://doi: 10.1109/ACCESS.2021. 3064084.

Kumar, P., Singh, S. N., & Dawra, S. (2022). Software component reusability prediction using extra tree classifier and enhanced Harris hawks optimization algorithm. International Journal of System Assurance Engineering and Management, 13(2), 892-903. https://doi.org/10.1007/s13198-021-01359-6

Poursaeid, M., Poursaeid, A. H., & Shabanlou, S. (2022). A comparative study of artificial intelligence models and a statistical method for groundwater level prediction. Water Resources Management, 36(5), 1499-1519. https://doi.org/10.1007/s11269-022-03070-y

Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. IAES International Journal of Artificial Intelligence, 9(1), 73. https://doi:10.11591/ijai.v9.i1.pp73-80

Sahoo, S., & Jha, M. K. (2013). Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. Hydrogeology Journal, 21(8), 1865. https://doi: 10.1007/s10040-013-1029-5

Sharaff, A., & Gupta, H. (2019). Extra-tree classifier with metaheuristics approach for email classification. In Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018, 189-197, Springer Singapore.

Tang, T., Jiao, D., Chen, T., & Gui, G. (2022). Medium-and long-term precipitation forecasting method based on data augmentation and machine learning algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1000-1011. https://doi: 10.1109/JSTARS.2022.3140442.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606. https://doi:10.14569/IJACSA.2021.0120670

Wang, Y., Wu, X., Chen, Z., Ren, F., Feng, L., & Du, Q. (2019). Optimizing the Predictive Ability of Machine Learning Methods for Landslide Susceptibility Mapping Using SMOTE for Lishui City in Zhejiang Province, China. International Journal of Environmental Research and Public Health, 16(3), 368. MDPI AG. Retrieved from http://dx.doi.org/10.3390/ijerph16030368

Yadav, B., Gupta, P. K., Patidar, N., & Himanshu, S. K. (2020). Ensemble modelling framework for groundwater level prediction in urban areas of India. Science of the Total Environment, 712, 135539. https://doi.org/10.1016/j.scitotenv.2019.135539