4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Chi-square automatic interaction detection (CHAID) algorithm for flood susceptibility assessment in Sardabroud watershed, Iran

Alireza Habibi [1], Mahmoud Reza Delavar *[2], Mohammad Sadegh Sadeghian [3], Borzo Nazari [4]

[1]University of Tehran, College of Engineering, School of Surveying and Geospatial Engineering, Department of GIS, Tehran, Iran
[2]University of Tehran, College of Engineering, School of Surveying and Geospatial Engineering, Center of Excellence in Geomatics Engineering in Disaster Management, Tehran, Iran
[3] Islamic Azad University, Department of Civil Engineering, Central Tehran Branch, Tehran, Iran
[4]University of Tehran, College of Engineering, School of Surveying and Geospatial Engineering, Department of GIS, Tehran, Iran

Keywords
GIS
Artificial intelligence
Machine Learning
Assessment
Natural hazard

Abstract

Flood, as a natural phenomenon, is the most common natural hazard that causes significant damage in the world. It is difficult to predict and identify flood zones due to variable weather conditions and various influencing factors. However, the identification and detection of early flood zones using machine learning techniques is used for smart flood management. In this study, Chi-square automatic interaction detection (CHAID) machine learning model for flood susceptibility map in Sardabroud watershed in north of Iran has been evaluated. For this purpose, a spatial database including 205 present and past flood locations with 8 conditional factors including elevation, slope, landuse, normalized difference vegetation index (NDVI), distance to river, topographic wetness index (TWI), lithology and rainfall are considered. After calculating variance inflation factors (VIF), all of the flood factors were considered for the modeling process. VIF technique uses to quantify multi-collinearity. Receiver operating characteristic (ROC), area under curve (AUC) and accuracy (ACC) metrics were used to evaluate and compare the predictability of the model. The results show that the CHAID model reaches an AUC of 0.939. This model has been proven as an efficient model for detecting flood prone areas in this watershed.

1. Introduction

Flood is known as one of the most frequent and destructive natural disasters in the world among other natural disasters such as earthquake and droughts due to causing great damage to human life and property and lives (Du et al. 2013).

The reasons for urban floods are the weakness of drainage systems and water infiltration into the ground during stormy rains and unhealthy urban growth (Darabi et al. 2019). Monsoon is one of the reasons why southeast Asian countries are most affected by floods and most of their related events (Loo et al. 2015). Iran has experienced a number of floods, especially in the northern parts of the country. For example, Noshahr in 2012, Behshahr in 2013 and Sari in 2015 have suffered from flash floods (Khosravi et al. 2016).

Therefore, optimal, efficient and proper methods should be used to reduce flood damage and losses. In the recent years, the use of artificial intelligence (AI) methods such as machine learning (Ahmadlou et al. 2021; Khosravi et al. 2020; Shahabi et al. 2021; Arora et al. 2021) has been increased.

Nghia et al. (2020) have used the CHAID algorithm to model flash floods in the Luc Yen area of Yen Bai Province in Vietnam, using 10 conditional factors including soil type, land cover, lithology, river density, rainfall, elevation, topographic wetness index (TWI), slope, aspect, and curvature. (Tehrany et al. 2013) have also used this algorithm to model the Kelantan River Basin in northeastern Malaysia by selecting 10 flood factors. However, flood modelings using CHAID decision tree algorithm does not seem to be widely reported in the literature.

The purpose of this study is the flood modeling of Sardabroud watershed in northern Iran using CHAID

* **Corresponding Author**

(habibi.alireza@ut.ac.ir) ORCID ID 0000-0002-7134-149X
*(mdelavar@ut.ac.ir) ORCID ID 0000-0002-9654-6491
(m.sadeghian29@yahoo.com) ORCID ID 0000-0002-3406-9044
(borzoo.nazari@ut.ac.ir) ORCID ID 0000-0003-4064-2617

Cite this study

Habibi, A., Delavar, M. R., Sadeghian, M. S., &Nazari, B. (2022). Chi-square automatic interaction detection (CHAID) algorithm for flood susceptibility assessment in Sardabroud watershed, Iran. 4th Intercontinental Geoinformation Days (IGD), 294-297, Tabriz, Iran

algorithm to produce a flood risk map and undertake its evaluation.

2. Study area

Sardabroud watershed with an area about 460 Km² is a narrow basins originates from the snowy heights of Takht-e-Solimansar at a elevation of 4600 meters. Sardabroud watershed is located in the west of Mazandaran province, Iran. The elevation of the watershed ranges from 4800 to -31 meters and the slope ranges from 0 to 78 degrees (Figure 1). The Sardabroud river flows through several mountains including Takht-e Soleiman and Alam Kooh, to the sea (Figure 1). This watershed is one of the tourist attractions places in Kelardasht with high average annual rainfall.

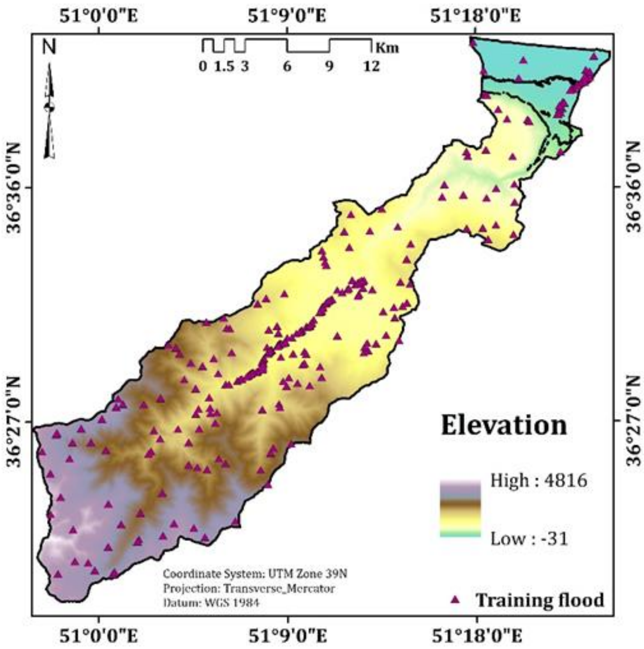

Figure 1. Study area

3. Method

In this study, several influencing factors including elevation, slope, rainfall, landuse, lithology, TWI, distance to river and NDVI have been extracted and collected for flood risk modeling using machine learning.

The CHAID algorithm process is in descending order from top to bottom, dividing large branches into smaller branches, which continue to be grouped according to specific factors. The CHAID is one of the classification decision tree techniques used in regression problems (Althuwaynee et al. 2014). CHAID algorithm has a number of titles such as automatic interaction detection, classification and regression tree (CART) and artificial neural network (ANN). CHAID algorithm uses chi-square statistics as a criterion for data separation and performs dodge separation (Eqs. 1,2, and 3) (Yeon et al. 2010).

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \quad (1)$$

$$n_{ij} = \sum_{n \epsilon D} fnI\ (x_n = i \cap y_n = j) \quad (2)$$

$$m_{ij} = \frac{n_i . n_j}{n_{ij}} \quad (3)$$

where n_{ij} = the observed cells frequency,
m_{ij} = cell frequency for $y_n = j$ and $x_n = i$.

VIF is a powerful statistical technique that detects a strong linear relationship between more than two factors in a multiple regression model (Hong et al. 2020). Accuracy value (ACC) needs to be determined for models accuracy that is calculated based on False Positive (FP), True Negative (TN), False Negative (FN) and True Positive (TP) (FP = non-flood pixels that are incorrectly known as flood pixels, TN = flood pixels that are correctly known as non-flood, FN = non-flood pixels that are incorrectly known as non-flood pixels and TP = flood pixels that are correctly known as flood pixels) (Shahabi et al. 2020).

Therefore, this algorithm has been used for flood risk modeling in this study. The research methodology proposed in this paper is presented in Figure 2. The produced map shows the probability of flooding.

4. Results

The results of calculating VIF (Table 1) for the 8 factors considered for flood modeling, shows that landuse has the lowest and elevation has the highest VIF. Therefore, none of the factors had VIF> 10. Therefore, all of the factors are considered for the modeling process.

Then, out of the total flood pixels, 70% of the data have been used for model training and 30% of the data employed for the model testing process. The AUC values for the testing process of the model are 0.939 (Figure 4).

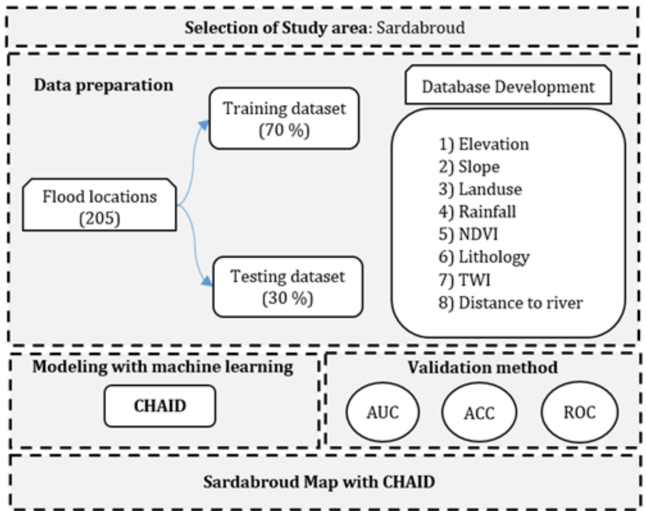

Figure 2. Research methodology

Therefore, according to the flood modeling of this watershed with CHAID algorithm, the value of AUC has been obtained as 0.939. The Accuracy value of the CHAID model for the training and testing process are 0.964 and 0.882, respectively (Table 2). The map has been produced in ArcGIS 10.3 software and illustrated in Figure 3.

Table 1. multi-collinearity analysis

Factors	VIF
Elevation	5.38
Slope	1.89
Rainfall	1.67
Distance to river	1.32
Landuse	1.03
NDVI	3.67
Lithology	1.30
TWI	1.49

Table 2. model performances

Metrics/Model	AUC	ACC
CHAID (training)	0.961	0.964
CHAID (testing)	0.939	0.882

5. Discussion

In this research, a decision tree-based machine learning model (CHAID) has been used to model flood risk in Sardabroud watershed, Iran. Elevation, slope, landuse, NDVI, TWI, rainfall, lithology and distance to river factors were used for the modeling process.

After performing multi-collinearity analysis using VIF methods, the value of VIF factors changes in a range from 1.03 to 5.38, with the highest and lowest values are related to elevation and landuse. Therefore, because VIF values are lower than 10, there is no correlation between the factors.

Therefore, the 8 factors considered for the modeling process were used. According to Figure 3, it is clear that the risk of flood in the north and middle parts of the watershed are more than other areas.

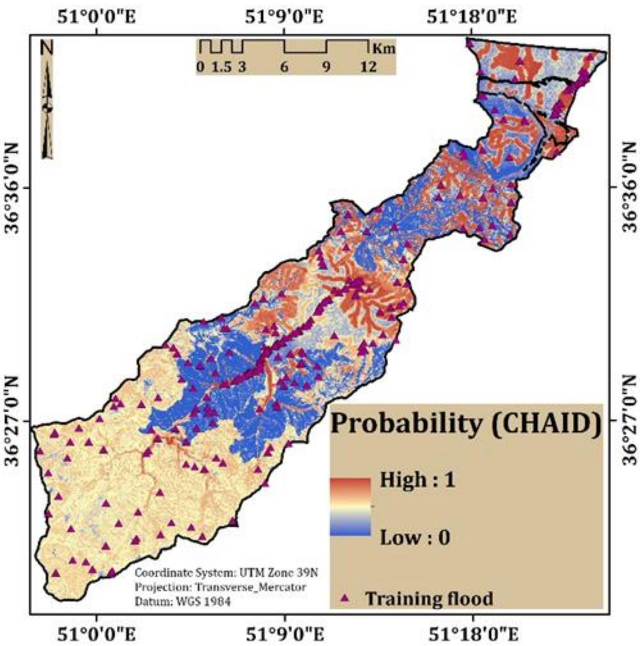

Figure 3. Flood susceptibility mapping

6. Conclusion

The aim of this study was to consider the factors affecting flood risk and to identify the flood susceptible zones in Sardabroud watershed, Iran.

The CHAID employed model is important for future smart flood disaster management decisions because it provides the basic information for controlling and managing flood risk.

Due to the flood damage to urban and agricultural areas, future research needs to focus on selecting and adopting effective flood parameters in the area such as population density and literacy and their relationship with economic processes and other factors affecting floods.

The method adapted in this research can be extended to larger watersheds that are at flood risk, and the accuracy of the models can be compared and evaluated with other basic machine learning models such as support vector machine (SVM) and K-nearest neighbor (KNN) methods.

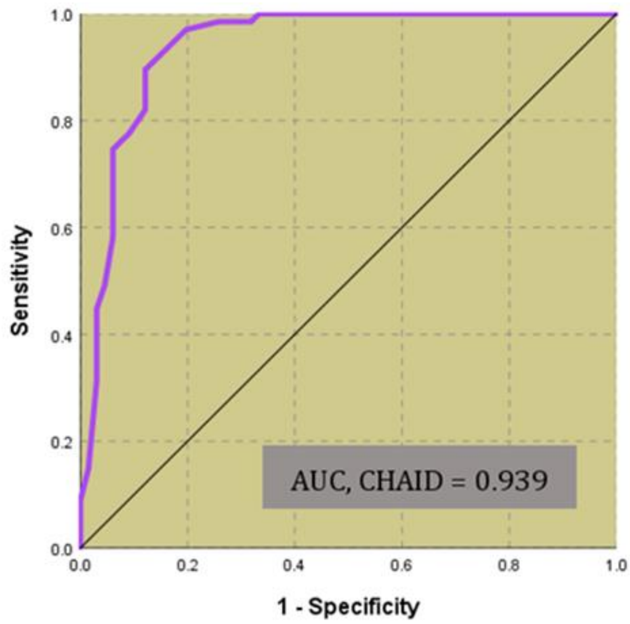

Figure 4. Plot of the ROC curve

References

Ahmadlou, M., Al-Fugara, A., Al-Shabeeb, A. R., Arora, A., Al-Adamat, R., Pham, Q. B., & Sajedi, H. (2021). Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks. *Journal of Flood Risk Management*, *14*(1). https://doi.org/10.1111/jfr3.12683

Althuwaynee, O. F., Pradhan, B., & Ahmad, N. (2014). *Landslide susceptibility mapping using decision-tree based CHi-squared automatic interaction detection (CHAID) and Logistic regression (LR) integration*. *20*, 012032. https://doi.org/10.1088/17551315/20/1/012032

Arora, A., Arabameri, A., Pandey, M., Siddiqui, M. A., Shukla, U. K., Bui, D. T., & Bhardwaj, A. (2021). Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India. *Science of The Total Environment*, *750*, 141565. https://doi.org/10.1016/j.scitotenv.2020.141565

Darabi, H., Choubin, B., Rahmati, O., Torabi Haghighi, A., Pradhan, B., & Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. *Journal of Hydrology*, *569*, 142–154. https://doi.org/10.1016/j.jhydrol.2018.12.002

Du, J., Fang, J., Xu, W., & Shi, P. (2013). Analysis of dry/wet conditions using the standardized precipitation index and its potential usefulness for drought/flood monitoring in Hunan Province, China. *Stochastic Environmental Research and Risk Assessment*, *27*(2), 377–387. https://doi.org/10.1007/s00477-012-0589-6

Hong, H., Liu, J., & Zhu, A.-X. (2020). Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. *Science of The Total Environment*, *718*, 137231. https://doi.org/10.1016/j.scitotenv.2020.137231

Khosravi, K., Nohani, E., Maroufinia, E., & Pourghasemi, H. R. (2016). A GIS-based flood susceptibility assessment and its mapping in Iran: A comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. *Natural Hazards*, *83*(2), 947–987. https://doi.org/10.1007/s11069-016-2357-2

Khosravi, K., Panahi, M., Golkarian, A., Keesstra, S. D., Saco, P. M., Bui, D. T., & Lee, S. (2020). Convolutional neural network approach for spatial prediction of flood hazard at national scale of Iran. *Journal of Hydrology*, *591*, 125552. https://doi.org/10.1016/j.jhydrol.2020.125552

Loo, Y. Y., Billa, L., & Singh, A. (2015). Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia. *Geoscience Frontiers*, *6*(6), 817–823. https://doi.org/10.1016/j.gsf.2014.02.009

Nghia, N., Yariyan, P., Amiri, M., Tran, A., Pham, T. D., Phuong, D., & Bui, D. (2020). A New Modeling Approach for Spatial Prediction of Flash Flood with Biogeography Optimized CHAID Tree Ensemble and Remote Sensing Data. *Remote Sensing*, *12*, 1373. https://doi.org/10.3390/rs12091373

Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J. J., & Ahmad, A. (2020). Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier. *Remote Sensing*, *12*(2), 266. https://doi.org/10.3390/rs12020266

Shahabi, H., Shirzadi, A., Ronoud, S., Asadi, S., Pham, B. T., Mansouripour, F., & Bui, D. T. (2021). Flash flood susceptibility mapping using a novel deep learning model based on deep belief network, back propagation and genetic algorithm. *Geoscience Frontiers*, *12*(3), 101100. https://doi.org/10.1016/j.gsf.2020.10.007

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2013). Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology*, *504*, 69–79. https://doi.org/10.1016/j.jhydrol.2013.09.034

Yeon, Y.-K., Han, J.-G., & Ryu, K. H. (2010). Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Engineering Geology*, *116*(3), 274–283. https://doi.org/10.1016/j.enggeo.2010.09.009