4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Assessing the importance of variable selection in land subsidence susceptibility mapping

Sepideh Tavakkoli Piralilou [*1,2], **Golzar Einali** [3], **Shokrolah Kiani** [4], **Khalil Gholamnia** [3]

[1]*University of Salzburg, Department of Geoinformatics—Z_GIS, Salzburg, Austria*
[2]*Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria*
[3]*University of Tabriz, Department of Remote Sensing and GIS, Tabriz, Iran*
[4]*University of Yazd Department of Geography, Yazd, Iran*

Keywords
Remote sensing
Land Subsidence
machine learning (ML)
Random Forest (RF)
Damaneh Plain

Abstract

In the domain of land subsidence risk science, Land Subsidence Susceptibility Mapping (LSSM) aids in spatially identifying regions prone to subsidence. This study used a multi-collinearity analysis through the variance inflation system (VIF) and tolerance (TOL) and a machine learning (ML) model of Random Forest (RF) for LSSM in the Damaneh Plain, Isfahan Province, Iran. The study investigated the importance of the conditioning variables in predicting Land Subsidence occurrences using an ML model. An ML model's prediction capabilities and performance were evaluated using conditioning variables in this paper. Using VIF, we eliminated the least "important" variables related to the LSSM. Conclusively, we found that removing the least "important" variables improves the accuracy of the resulting LSSMs. Based on the results of our study, using VIF could increase the predictive performance of the RF model by three percentage points in the applied accuracy assessment metric.

1. Introduction

Land subsidence is a geological phenomenon that is considered a natural disaster caused by human activities (O. Ghorbanzadeh, Rostamzadeh, Blaschke, Gholaminia, & Aryal, 2018). In most cases, the underground layers compact after natural and manufactured factors work together to cause their downward movement (O. Ghorbanzadeh, Feizizadeh, & Blaschke, 2018). In addition to changing the environment, this phenomenon has significant economic and social repercussions. There have been many causes of land subsidence worldwide, such as earthquakes, extraction of natural gas, mineral exploration, dissolution of limestone, and extraction of groundwater (Tien Bui et al., 2018). In Iran, many land subsidences have occurred on several plains due to the high extraction of groundwater for agriculture and urban consumption during the last t few decades. Land subsidence is one of the most frequently occurring natural disasters that cause significant human casualties and infrastructure destruction in this country. Many problems are associated with land subsidence, including damage to public and private infrastructure, power lines, roads, settlements, sinkholes, and soil erosion (Ranjgar, Razavi-Termeh, Foroughnia, Sadeghi-Niaraki, & Perissin, 2021). Thus, generating the LSSM and modeling variables affecting land subsidence are important approaches that incorporate potential land subsidence locations. The LSSM refers to the likelihood of a land subsidence occurring in a particular region due to several causative factors. Land subsidence risk management is an essential step towards reducing subsidence risk and assists in mapping the spatial distribution of probable manifestations of subsidence (Mohammady, Pourghasemi, & Amiri, 2019).

In land subsidence susceptibility studies, satellite data and geographic information system (GIS) tools and machine learning (ML) models are instrumental in acquiring satisfactory resolution remote sensing and other relevant data, evaluating variables that affect this phenomenon (Omid Ghorbanzadeh, Blaschke, Aryal, & Gholaminia, 2020; Omid Ghorbanzadeh, Crivellari, Ghamisi, Shahabi, & Blaschke, 2021). In the past, researchers have used various models to generate LSSMs using the technologies mentioned above. An adaptive neuro-fuzzy inference system has been evaluated by (O. Ghorbanzadeh, Rostamzadeh, et al., 2018) for LSSM generating for Marand Plain, the East Azerbaijan Province in Iran, using six different membership functions (MF). Two of the six LSSMs, the DsigMF (0.958),

* **Corresponding Author**

*(sepideh.tavakkoli-piralilou@stud.sbg.ac.at) ORCID ID 0000-0002-1188-8290
(Golzar.einali@yahoo.com) ORCID ID 0000-0002-0679-8370
(arashkiani97rs@gmail.com) ORCID ID 0000-0001-9292-3694
(khalil.gh3@gmail.com) ORCID ID 0000-0002-3860-8674

Cite this study

Piralilou, S. T., Einali, G., Kiani, S., & Gholamnia, G. (2022). Assessing the importance of variable selection in land subsidence susceptibility mapping. 4th Intercontinental Geoinformation Days (IGD), 188-191, Tabriz, Iran

and the GaussMF (0.951) yielded very high prediction values based on the calculated areas under the curves (AUC) of the receiver operating characteristic (ROC) analyses. The same model was integrated with two models of imperialist competition algorithms and gray wolf optimization by (Tien Bui et al., 2018) to calculate LSSM for Shahryar County, Iran. The resulting maps were evaluated based on the inventory data set of land subsidence locations and the root mean square error value and the ROC, which showed that the integration with imperialist competitive algorithm had the best accuracy with a 0.932 AUC value.

On the one hand, although ML models show acceptable performance and have high accuracy, they are dependent on a precise land subsidence inventory map, which is somewhat challenging in Iran (Arabameri et al., 2020). On the other hand, the ML models are sensitive regarding the input data set for training. A selection of important ones is essential to get the best accuracy for the modeling (Saha et al., 2021). This matter is more evident for modeling and mapping the susceptibility of land subsidence because there are two leading causes of land subsidence: natural and human activities that disrupt the underground layers (Arabameri et al., 2021; Ranjgar et al., 2021). However, land subsidence can be attributed to a variety of geological and hydrological factors in different parts of the world.

We aim to use the most current ML models and analyze the importance of each variable prior to training. Thus, it is possible to examine how each of the land subsidence relevant variables can affect the performance of the ML models for having the best prediction. This approach in research can lead to the use of variables that contribute more to the modeling goal and avoid data duplication that complicates computation and results in ambiguous conclusions.

2. Study area and Data set

We studied the Damaneh Plain in Faridan County, Isfahan Province, Iran. Figure 1 shows the boundaries of the study area, which is located between longitudes ´30 ˚50 to ´35 ˚50 and latitudes ˚50 to ´35 ˚50. The plain covers an area of more than 2651057 ha. The climate in Damaneh plain is mild in spring and summer and cold in winter. The coldest month of the year is January with an average of -3.1˚ C. As July approaches, the temperature rises to an average of 23˚ C, making it the hottest month of the year. (http://www.esfahanmet.ir/). The study area receives an average annual rainfall of 317.4 mm and average annual temperatures of 10.4 °C. This plain sees about 48% of its total rainfall in winter, the wettest season of the year (https://www.esrw.ir/). There have been significant land subsidence issues in Damaneh plain, especially in Damaneh city, resulting in most houses being cracked and destroyed, forcing local people to leave their homes.

2.1. Land subsidence inventory data set

To predict and evaluate the LSSMs, 2667 land subsidence occurrence locations were collected through a comprehensive field survey. 70% and 30% of the points in this inventory data set have been used to train and validate the models, respectively.

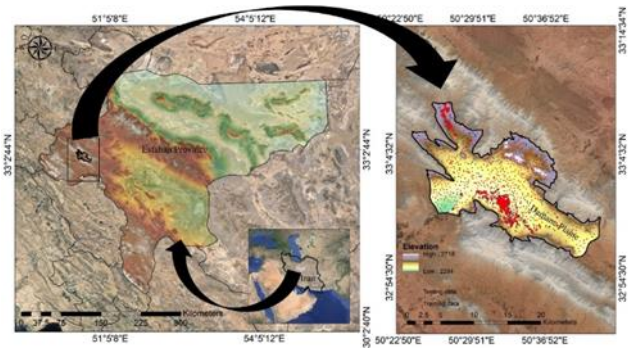

Figure 1. The location of the Damaneh Plain in Faridan County, Isfahan Province, Iran

2.2. Conditioning variables

Fourteen inter-related variables have been selected based on the literature review in the present land subsidence study. The prepared conditional variables are distinct from rivers, Landover, the topographic wetness index (TWI), distance to villages, slope, aspect, catchment slope, rainfall, elevation, groundwater depth, soil, lithology, distance to roads, and well density. These variables are derived from different sources, which are fully explained by (Arabameri et al., 2021; Omid Ghorbanzadeh et al., 2020).

3. Methods

All prepared variables were once used to train the ML model of RF. Following are the results of evaluating the fixed and dependent variables in the multi co-linearity analysis among the factors was done through the variance inflation system (VIF) and tolerance (TOL). For the training of the models, we selected variables with a VIF value greater than two for the ML training process. We evaluated the validity of the LSSMs from each model using the ROC curve. The applied multivariate linear regression (MLR), and the VIF and RF models are explained as follows.

3.1. Multi-collinearity analysis

The probability estimation model uses this criterion to exclude co-linearity and select influential variables. In particular, the multi-collinearity among the conditional variables is more likely to be error-prone if the VIF is greater than five and the TOL is less than 0.1 (Lei et al., 2020).

3.2. Random Forest (RF)

Breiman and Cutler created the RF ML model, developed by Ho (1994) (Ho, Hull, & Srihari, 1994). RF estimates the final result based on most votes instead of relying on a decision tree. Using this approach will enhance accuracy and prevent overfitting. In this work, in both the 11 and 14 variable models, the maximum number of variables is designed upon the square root of the total number of variables.

4. Results

In this study, 14 conditional variables were generated for the LSSM. An essential part of this study was identifying the key variables, which are illustrated in the following "Table 1". To evaluate the relationship between the conditional variables, we used the multi-collinearity test considering VIF and the TOL. The results show that this test assigns more than 5 VIF values to variables of distance to villages, aspect, and distance to roads.

Table 1. Collinearity results among the variables. The red highlighted rows with more than 5 VIF resulting values represent the least "important" related variables to the LSSM.

Conditional variables	TOL	VIF
Distinct to rivers	0.52	1.92
Landover	0.29	3.44
TWI	0.22	4.54
Distance to villages	0.17	5.87
Slope	0.38	2.63
Aspect	0.16	6.25
Catchment slope	0.79	1.26
Rainfall	0.52	1.92
Elevation	0.59	1.69
Depth of GW	0.45	2.22
Soil	0.54	1.85
Lithology	0.49	2.04
Distance to roads	0.18	5.37
Well Density	0.39	2.56

The resulting LSSM from the RF model is classified into five classes, very low class with 65.94%, low class with 14.33%, medium class with 8.41%, high class with 5.48% and very high class with 5.84% of the area (see "Figure 2"). Model RF-VIF, with a very low-class area of 51.81%, a high-class area of 18.55%, a medium class area of 11.53%, a high-class area of 9.62%, and a very high-class area of 8.48%, contained a larger area and focused more on the center and types of land use, primarily agricultural land. We evaluated the performance of the applied ML model using a standard accuracy assessment method. The calculated areas under the curves (AUC) of the receiver operating characteristic (ROC) analyses discovered that the LSSM predicted by the RF-VIF got a higher AUC value of 92.29%. In comparison, that of the RF model was 89.89%.

5. Discussion

Considering a wide variety of land subsidence triggers, many variables are involved in LSSM generating. This variation can affect modeling and lead to errors (Lei et al., 2020). Modeling should therefore avoid these problems to the extent possible. Therefore, using the multi-collinearity analysis, we eliminated the least "important" three variables and trained the RF model with 11 remaining conditional variables. According to the present study, land subsidence in Damaneh Plain is significantly related to the river's distance, landcover, TWI, slope, catchment slope, rainfall, elevation, depth of ground water, soil lithology, and well density. There are 572 wells in the study area, 200 illegally drilled. Moreover, all the region slopes have been cultivated for agriculture, and water consumption has increased.

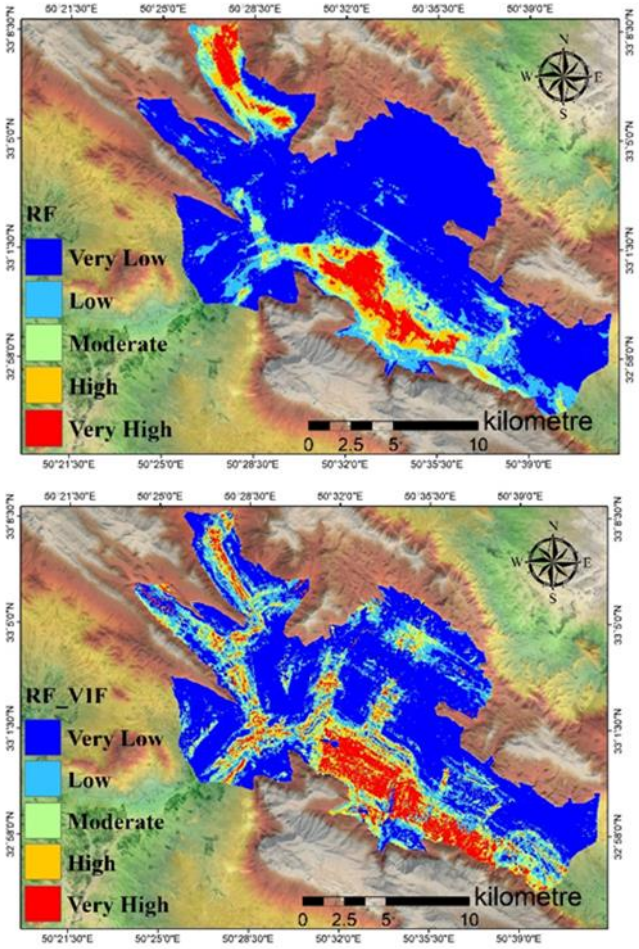

Figure 2. The resulting LSSMs predicted using the RF model based on all conditional variables and based on the selected variables derived from the VIF for Damaneh Plain.

6. Conclusion

To conclude, the results overall showed the effectiveness of the RF model in LSSM generation and the multi-collinearity method for assessing the importance of the conditioning variables. The integrated approach can be reproduced and applied to other land subsidence susceptible plains with different environmental indicators.

Acknowledgement

The authors are grateful for the support of the University of Salzburg, and the Institute of Advanced Research in Artificial Intelligence (IARAI) GmbH, Vienna, Austria.

References

Arabameri, A., Pal, S. C., Rezaie, F., Chakrabortty, R., Chowdhuri, I., Blaschke, T., & Ngo, P. T. T. (2021). Comparison of multi-criteria and artificial intelligence models for land-subsidence susceptibility zonation. *Journal of Environmental Management*, *284*, 112067.

Arabameri, A., Saha, S., Roy, J., Tiefenbacher, J. P., Cerda, A., Biggs, T., … Collins, A. L. (2020). A novel ensemble computational intelligence approach for the spatial prediction of land subsidence susceptibility. *Science of The Total Environment*, *726*, 138595.

Ghorbanzadeh, O., Feizizadeh, B., & Blaschke, T. (2018). An interval matrix method used to optimize the decision matrix in AHP technique for land subsidence susceptibility mapping. *Environmental Earth Sciences*, *77*(16). https://doi.org/10.1007/s12665-018-7758-y

Ghorbanzadeh, O., Rostamzadeh, H., Blaschke, T., Gholaminia, K., & Aryal, J. (2018). A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping. *Natural Hazards*. https://doi.org/10.1007/s11069-018-3449-y

Ghorbanzadeh, Omid, Blaschke, T., Aryal, J., & Gholaminia, K. (2020). A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. *Journal of Spatial Science*, *65*(3), 401–418.

Ghorbanzadeh, Omid, Crivellari, A., Ghamisi, P., Shahabi, H., & Blaschke, T. (2021). A comprehensive transferability evaluation of U-Net and ResU-Net for landslide detection from Sentinel-2 data (case study areas from Taiwan, China, and Japan). *Scientific Reports*, *11*(1), 1–20.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1), 66–75.

Lei, X., Chen, W., Avand, M., Janizadeh, S., Kariminejad, N., Shahabi, H., … Mosavi, A. (2020). GIS-based machine learning algorithms for gully erosion susceptibility mapping in a semi-arid region of Iran. *Remote Sensing*, *12*(15), 2478.

Mohammady, M., Pourghasemi, H. R., & Amiri, M. (2019). Land subsidence susceptibility assessment using random forest machine learning algorithm. *Environmental Earth Sciences*, *78*(16), 1–12.

Ranjgar, B., Razavi-Termeh, S. V., Foroughnia, F., Sadeghi-Niaraki, A., & Perissin, D. (2021). Land subsidence susceptibility mapping using persistent scatterer sar interferometry technique and optimized hybrid machine learning algorithms. *Remote Sensing*, *13*(7), 1326.

Saha, S., Arabameri, A., Saha, A., Blaschke, T., Ngo, P. T. T., Nhu, V. H., & Band, S. S. (2021). Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Science of the Total Environment*, *764*, 142928.

Tien Bui, D., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., … Saro, L. (2018). Land subsidence susceptibility mapping in south korea using machine learning algorithms. *Sensors*, *18*(8), 2464.