Land subsidence susceptibility mapping using machine learning in the Google Earth Engine platform

Van Anh Tran [1,4][0000-0003-2311-7351] Thanh Dong Khuc [2][0000-0001-7890-1050], Trung Khien Ha [2][0009-0002-4171-7481], Hong Hanh Tran [1,4][0000-0002-8771-8351], Thanh Nghi Le [1,4][0000-0003-0536-472X], Thi Thanh Hoa Pham [1][0009-0000-8820-8161 id], Dung Nguyen [1][0000-0003-4843-1160], Hong Anh Le [1][0000-0002-0483-3195], and Quoc Dinh Nguyen [3] [0000-0003-3844-091X]

[1] Hanoi University of Mining and Geology, 18 Vien Street, Hanoi 100000, Vietnam

tranvananh@humg.edu.vn, tranhonghanh@humg.edu.vn, lethanhnghi@humg.edu.vn, phamthithanhhoa@humg.edu.vn
nguyenthimaidung@humg.edu.vn, lehonganh@humg.edu.vn

[2] Hanoi University of Civil Engineering, 55 Giai Phong Street, Hanoi 100000, Vietnam
dongkt@huce.edu.vn, khienht@huce.edu.vn

[3] Phenikaa University, Nguyen Trac, Yen Nghia, Ha Đong, Hanoi 100000, Vietnam
dinh.nguyenquoc@phenikaa-uni.edu.vn

[4] Geomatics in Earth sciences Group (HUMG, Vietnam)

Abstract.
This study aims to compare the effectiveness of two predictive models, CART regression and Random Forest in mapping land subsidence susceptibility. The analysis is supported by the Google Earth Engine cloud computing platform. The study focuses on Camau province, located in the Mekong Delta, where significant land subsidence occurs annually. Eight variables were considered in the models, including elevation, slope, aspect, land cover, NDVI, soil map, geology, and groundwater level. Land subsidence points, obtained through the PSInSAR method, were used in the study, comprising a total of 989 points. These points were divided into a 70% training dataset and a 30% testing dataset for both models. The results produced a land subsidence sensitivity map categorized into five levels: very low, low, moderate, high, and very high. The performance of the models was evaluated using ROC curve and the area under the curve (AUC). The AUC values for the Random Forest (RF) model are 0.86 and 0.87 for the training and validation datasets, respectively. In comparison, the CART model achieves AUC values of 0.79 and 0.73 for the training and validation datasets, respectively. The research findings demonstrate an 7% superior performance of the RF model compared to the CART method. Therefore, the RF model is chosen as the final model for land subsidence susceptibility mapping in Camau.

Keywords: Camau, CART, Random Forest, subsidence.

1 Introduction

The Mekong Delta in Vietnam has been experiencing serious land subsidence in recent years due to various natural and artificial causes. Over the past few decades, large-scale land use changes have occurred due to rapid population growth, urbanization, and the increase in agricultural and aquacultural production. These have contributed to the subsidence of land and exacerbated the severity of flooding. One of the provinces most affected by land subsidence is Camau. Located in the southernmost part of Vietnam, Camau is facing the dangers of land subsidence, sea level rise, flooding, and saltwater intrusion. A meticulous study by Erban et al [1] demonstrated that the Camau Peninsula and the entire Mekong Delta are subsiding at a rate of several centimeters per year, exceeding the current absolute sea level rise by a significant margin. Meanwhile, Minderhoud's research has shown a close correlation between land use and the rate of land subsidence [2].

In order to study land subsidence and predict the risk of land subsidence effectively, the problem of input data and algorithms used for prediction are extremely important factors. Omid Rahmati [3] employed two machine learning algorithms, the maximum entropy (MaxEnt) and genetic algorithm rule-set production (GARP), to construct a subsidence assessment model in Kashmar, Iran. The model incorporated various data such as land use, lithology, distances to groundwater extraction sites and afforestation projects, distances to fault locations, and groundwater level reductions. The research findings indicate that the GARP algorithm outperforms the MaxEnt algorithm in terms of accuracy and performance. Both algorithms provide reliable subsidence prediction.

Recently, the study by Ata Allah Nadiri [4] introduced a method for assessing land subsidence susceptibility using the ALPRIFT framework and various artificial intelligence models, including Sugeno Fuzzy Logic (SFL), Support Vector Machine (SVM), Artificial Neural Network (ANN), and the Group Method of Data. The research results indicate that the combination of multiple artificial intelligence models can improve the accuracy of determining the susceptibility to land subsidence in the studied area.

Another study conducted in the Marand District of Tehran Province, Iran [5] utilized the adaptive-fuzzy inductive inference system (ANFIS) method with six categories of input data, including subsidence distance from borehole and faults, elevation, distance to roads, rivers and streams, groundwater depth, slope, and land use. ROC curve validation indicated that the Gauss MF and Dsig MF methods had high accuracy and were comparable. Lee' research [6] applied the ANN method to forecast the risk of land subsidence associated with Korean coal mines and achieved an accuracy of approximately 98.95%. In another study, the authors employed a combination of FR, logistic regression (LR), and ANN models, and it was found that the combined method had higher accuracy than using any single model alone [7]. In 2018, a study by D. Tien Bui utilized four models, Bayesian logic regression (BLR), support vector machine (SVM), logistic model tree (LMT), and intermittent decision tree (IDT), to determine the susceptibility of land subsidence [8]. Eight input factors including slope, distance to the nearest fault, density of faults, geology, distance to the nearest road, density of roads, land use, and rock mass rating were used. Evaluation of the four models showed that BLR was the most accurate method (0.941) for mapping the susceptibility of land

subsidence. Additionally, some studies have shown a correlation between floods and land subsidence [9].

The study by [10] evaluated the land subsidence susceptibility (LSS) in the Gharabolagh Plain in Fars Province, Iran based on factors such as changes in groundwater, distance to rivers, streams, distance to faults, elevation, slope, aspect, terrain wetness index (TWI), Landuse, and Lithology using the Google Earth Engine (GEE) platform, and with two probabilistic models, the belief function proof (EBF) and Bayesian theory (BT).

Overall, Camau in the Mekong Delta has a dense network of rivers and canals, but has been experiencing water shortages in recent years. This is due to upstream hydropower dams blocking water flow, leading to water shortages downstream, resulting in more frequent saline intrusion and droughts. Agricultural cultivation depends mainly on groundwater, which is being overexploited and depleted, causing land subsidence throughout the region. Due to its low elevation, with an average height of around 1m above sea level, flooding is a common occurrence when sea level rises [11]. Therefore, the need for predicting land subsidence risk is becoming increasingly urgent. Our study aims to provide an overview of the land subsidence susceptibility for the Camau Peninsula based on existing data sources and using the cloud computing platform Google Earth Engine. The study aims to test two machine learning methods, CART and Random Forest for subsidence predicted modeling for this area.

2 Study area

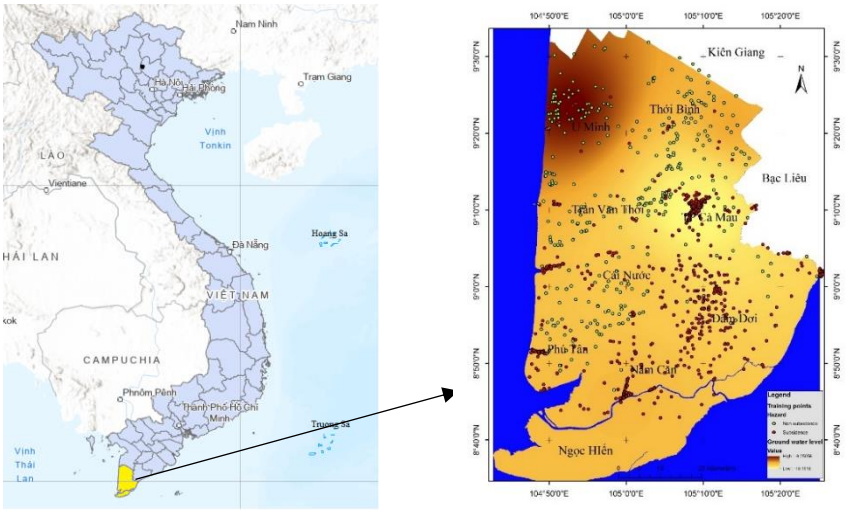

Fig. 1. Distribution map of subsidence and non-subsidence sample points in Camau area and its location on the map of Vietnam

Camau province is located in the southernmost part of the country, surrounded by the sea on three sides. Its east coast extends 107 km and borders the South China Sea, while its west and south coasts extend 147 km and meet the Gulf of Thailand. The North borders Bac Lieu and Kien Giang provinces. This area is characterized by low-lying terrain that is often flooded. The average altitude in Camau is only about 1m above sea level [2]. Camau has 5 main types of soil: acid, peat, alluvial, saline and canals.
Camau's ecosystem is home to a unique type of coastal mangrove forest extending 254 km along the coast. In addition, the province also has a melaleuca forest ecosystem located deep in the mainland in the districts of U Minh, Tran Van Thoi, Thoi Binh with an area of 35,000 ha. Mangroves cover 77% of the total area of mangroves in the Mekong Delta. Fig. 1 shows the location of Camau on the map of Vietnam and the administrative boundaries of the province.

3 Materials and methods

3.1 Methods

Classification and Regression Tree – CART

The Classification and Regression Trees (CART) algorithm is a widely used supervised machine learning technique for predicting a categorical target variable, creating a classification tree, or a continuous target variable, creating a regression tree. The CART classification requires a binary tree, which is a combination of an initial root node, decision nodes, and leaf nodes. The root node and each decision node represent a feature and the threshold value of that feature. Due to its easy-to-understand and straightforward nature, CART is one of the most commonly used machine learning methods today [12]. The schematic of the CART algorithm is presented in Fig. 2.

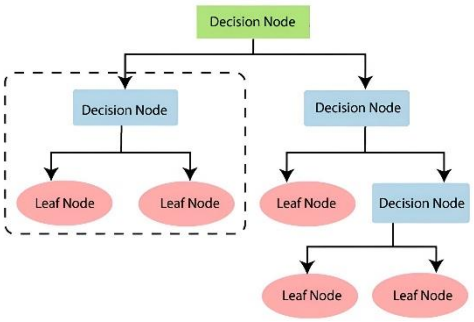

Fig.2 Description of the CART algorithm

The CART algorithm for Classification and Regression Trees require the tree to be classified in the best possible way. In the CART algorithm, the Gini index is used to evaluate whether the split at the condition nodes is accurate or not. To find the best classification, the total weight of the Gini index for all branch nodes is calculated, and then the part with the lowest Gini index is taken as the part with the best classification accuracy.

Mathematically, when analyzing a feature, different threshold divisions will lead to different classification results, and there may be cases where the same threshold for that feature leads to different classification results. Therefore, the Gini index is used to determine noise in the dataset. Assuming the dataset is classified into two classes A and B, the Gini index is determined as follows [12]:

$$Gini\ Index = 1 - \sum_{i=1}^{n}(P_i)^2 = 1 - [(P_A)^2 + (P_B)^2] \quad (1)$$

Where, P_A is the probability of data belonging to class A, P_B is the probability of data belonging to class B. The above formula uses probability to determine the Gini Index value on each characteristic branch, determining which branch most likely to occur.

Random Forest

Random Forest (RF) is an algorithm comprising of many single decision trees that act like unions. The algorithm uses random features to create a tree. The method of bootstrapping is used to create training samples and each selected feature is randomly sampled by replacing N (the size of the original training set). Finally, the final prediction result is obtained by combining multiple decision trees [12].

RF model is very effective for image classification and prediction because it mobilizes hundreds of smaller models inside with different rules to make the final decision. Each sub-model can be different and weak, but according to the" wisdom of the crowd" principle, the classification result will be more accurate than using any single model. Algorithm details can refer to [12]

Model quality assessment

To assess the quality of a predicting model, the ROC curve and the area under the ROC curve (AUC) are used. The ROC curve describes the relationship between pairs of the true positive rate (TPR) and false positive rate (FPR) for land subsidence and non subsidence positions. Reference points with good results will have a high true positive rate and a low false positive rate, and vice versa. TPR and FPR values are usually calculated with different thresholds to evaluate the model's effectiveness. The AUC is a comprehensive performance evaluation index of land subsidence prediction models. The closer the AUC value is to 1, the more effective the model.

The statistical parameters: positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR) are showed in the equations below.

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

$$FPR = \frac{FP}{FP+TN} \quad (3)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

3.2 Materials

Inventory of land subsidence

The inventory of land subsidence is a critical component in assessing the susceptibility of subsidence in the study area [13] . The dataset for the inventory comprises 989 sample points of land subsidence obtained from Sentinel-1 Radar data using the PSInSAR method. Fig. 1 shows the distribution of the subsidence points (red color) and non-subsidence points (blue color) randomly taken from the set of subsidence PS points from the results in the Mekong Delta area of the Copernicus website [14].

Input factors of subsidence susceptibility and Tools

In order to identify the factors that contribute to land subsidence susceptibility in the study area, we refer to some researches in this region [2][15] to understand the subsidence patterns in the region. Through this analysis, eight key factors were identified, including terrain elevation, slope, aspect, land cover, NDVI, soil, geology, and groundwater depth map.

The land cover map is derived from the ESA's landcover 10m 2021 product. The NDVI is computed from Sentinel-2 images, which have been averaged for the entire year of 2021. The geological map of the Camau area is sourced from a 1:100,000 scale map provided by the Vietnam Institute of Geosciences and Minerals. Groundwater level data represents the average water level observed during the years 2020, 2021, and 2022. The soil map of the Camau area is based on a 1:50,000 scale map provided by the Camau Department of Natural Resources and Environment. Additionally, the DEM map is obtained from the ALOS World 3D - 30m (AW3D30) dataset.

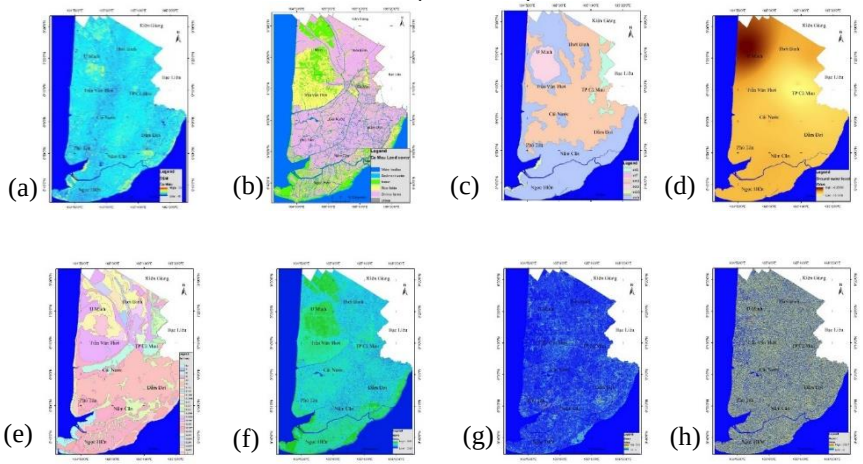

Fig. 3. Input factors of subsidence susceptibility. (a): Elevation map, (b): Land cover map, (c): Geology map, (d): Ground water level, (e): Soil map, (f): NDVI map, (g): Slope map, (h): Aspect map

The Google Earth Engine (GEE) cloud computing platform is utilized in this study to take advantage of its ability to gather data from various sources in the cloud [16] . By doing so, we minimize the need for desktop data preparation. Multiple sources of data, including DEM elevation digital model maps, Land use land cover map, and land subsidence inventory data to train and evaluate the model. The data sets are summarized in Fig. 3. The Sentinel-2 satellite image with 10m and 20m resolution is processed using the GEE platform to determine the NDVI vegetation index. To ensure consistency in the model, all data is set to 30x30 meter resolution. These land subsidence inventory data are then divided into training and testing sets at a ratio of 70:30 (Fig.4). The training set is comprised over 692 randomly points to extract values of elevation, slope, aspect, NDVI, LULC, soil, groundwater level, geology, location of land subsidence

with values of 1 (land subsidence) and 0 (non subsidence). Flow chart illustrating the process of image processing and the construction of predictive models using two methods, CART and RF (Fig.4).

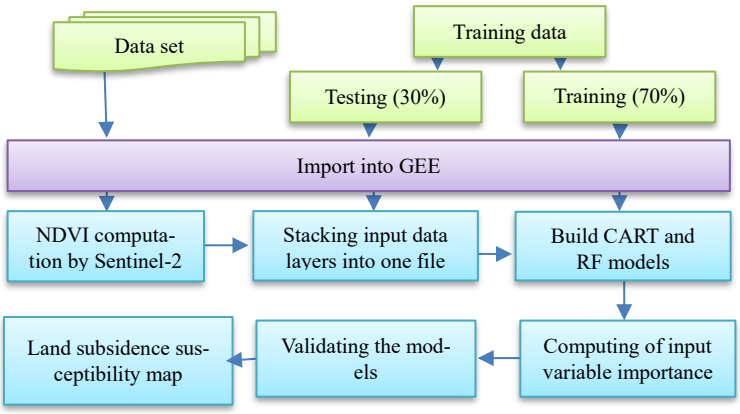

Fig.4. Flow chart of image processing and building predictive models by two methods CART and RF by GEE.

4 Results and Discussions

By utilizing eight variables to assess the susceptibility to land subsidence mentioned above, the two models exhibit clear disparities. Fig. 5 provides insight into the significance of variables in the application of two machine learning models: CART and RF. In the CART model, out of these input variables, up to four layers have no influence on the model, whereas all eight inputs contribute to the predicting process in the RF model. Fig. 5 depicts the importance of the input variables for the two models. Both models highlight the substantial impact of the water level decline data layer on predicting the risk of land subsidence. By referring to Fig. 1, we observe that the distribution of settlements aligns with the groundwater depth map, enabling easy identification of the considerably high decline in groundwater level within the vicinity of Camau city, leading to concentrated settlements surpassing those in the surrounding areas. Conversely, the northern part of Camau province, which displays the lowest settlement density, exhibits the least decline in water level. This northern region encompasses the expansive U Minh mangrove area with a small population and minimal groundwater extraction compared to the surrounding areas.

To evaluate model performance, the Receiver Operating Characteristics (ROC) curve is a graphical tool used to describe the relationship between false-positive rates and sensitivity across different thresholds [17]. This technique is widely used to evaluate the performance of probabilistic models. By adjusting the decision threshold, we can generate an ROC curve by plotting different combinations of the True Positive Rate (TPR) and the ratio of the False Positive Rate (FPR) [18]. The AUC value represents the area under the ROC curve, providing quantitative confirmation of the overall

performance of land subsidence models [19]. Higher AUC values indicate superior performance of settlement models and can be classified into different grades: excellent (0.9–1), very good (0.8–0.9) , good (0.7–0.8), moderate (0.6–0.7) and poor (0.5–0.6) [20].

After analyzing the training and test sets, it is clear that both the CART and RF models perform well, exhibiting high accuracy. The CART method achieved an AUC value of 0.79 for training set and 0.73 for testing set, while the RF method outperformed with an AUC value of 0.87 for training set and 0.86 for testing set. The AUC values for both machine learning methods are in excess of 0.7, confirming their effectiveness in predicting the susceptibility of land subsidence in the study area however the RF model gives very good performance which AUC is higher and more suitable. These findings strongly support the view that the RF model is reliable for such predictions.

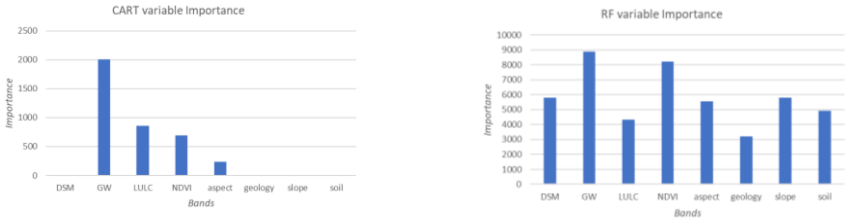

Fig.5 The importance of the input variables, CART (left). RF (right)

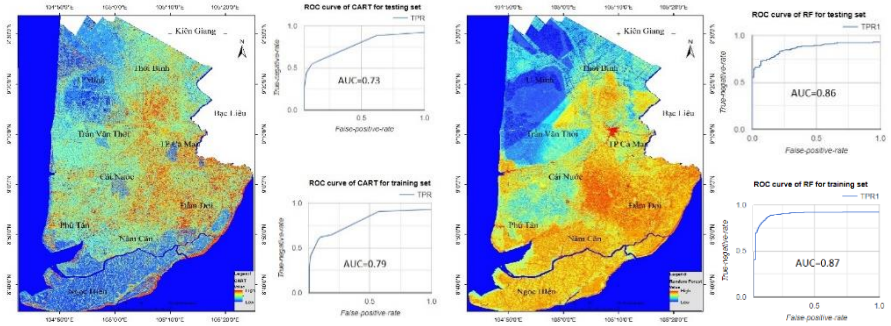

Fig.6. The land subsidence susceptibility maps generated using the CART model, along with the ROC curve and AUC value of CART (on the left), and the RF model, along with the ROC curve and AUC value of RF (on the right).

Fig. 7 depicts a map showing the susceptibility of land subsidence in the Camau region. This map was created using the CART and RF models in GEE, along with multiple data layers from various sources. The values on the subsidence susceptibility map range from 0 to 1, indicated by shades of blue to red. These colors represent areas with low to high levels of subsidence, which are determined based on factors such as elevation, slope, aspect, land cover, NDVI, soil, geology, and groundwater depth. The regions with the highest susceptibility to land subsidence are concentrated in Camau city, followed by the southern districts of Camau, namely Dam Doi, and Nam Can. The

northern districts of Camau province, which have extensive forest coverage and a low population such as U Minh, and Thoi Binh, experience a lower rate of subsidence.

5 Conclusions

The study aims to generate subsidence susceptibility maps in the Camau area of the Mekong Delta using GEE cloud computing and a multi-source dataset, employing two machine learning methods: CART and Random Forest (RF). The resulting land subsidence sensitivity map showcases the potential of utilizing free data sources and cloud-based algorithms. Regarding land subsidence sensitivity prediction in Camau, Vietnam, the RF machine learning model demonstrated superior performance compared to the CART model, displaying better accuracy.

These findings provide valuable insights into the land subsidence susceptibility map, offering useful information for managers and planners in devising strategies to mitigate this issue and facilitate rational land use conversion. For future research, it is recommended to expand the study by considering additional input variables that influence land subsidence, aiming to further enhance the accuracy of machine learning models on the GEE platform.

Acknowledgement
This research was funded by Scientific Research project of the Ministry of Education and Training of Vietnam, code: B2022-MDA-13.

References

1. Laura E Erban, Steven M Gorelick and Howard A Zebker: Groundwater extraction, land subsidence, and sea-level rise in the Mekong Delta, Vietnam. Environmental Research Letters 9, 8 (2014).
2. P S J Minderhoud, G Erkens, V H Pham, V T Bui, L Erban, H Kooi and E Stouthamer: Impacts of 25 years of groundwater extraction on subsidence in the Mekong delta, Vietnam. Environmental Research Letters 12, 6 (2017).
3. Omid Rahmati, Ali Golkarian, Trent Biggs, Saskia Keesstra, Farnoush Mohammadi,N. Daliakopoulos: Land subsidence hazard modeling: Machine learning to identify predictors and the role of human activities. Journal of Environmental Management 236, 466-480 (2019).
4. Ata Allah Nadiri, Iraj Habibi, Maryam Gharekhani, Sina Sadeghfam, Rahim Barzegar, Sadra Karimzadeh: Introducing dynamic land subsidence index based on the ALPRIFT framework using artificial intelligence techniques. Earth Science Informatics 15, 1007–1021 (2022).
5. Omid Ghorbanzadeh, Hashem Rostamzadeh, Thomas Blaschke, Khalil Gholaminia & Jagannath Aryal: 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 94, 497–517 (2018).
6. Saro Lee, Inhye Park & Jong-Kuk Choi: Spatial Prediction of Ground Subsidence Susceptibility Using an Artificial Neural Network. Environmental Management 49, 347–358 (2012).

7. Inhye Park, Jiyeong Lee & Lee Saro: Ensemble of ground subsidence hazard maps using fuzzy logic. Central European Journal of Geosciences 6, 207–218 (2014).
8. Dieu Tien Bui, Himan Shahabi, Ataollah Shirzadi, Kamran Chapi, Biswajeet Pradhan, Wei Chen, Khabat Khosravi, Mahdi Panahi, Baharin Bin Ahmad, Lee Saro: Land Subsidence Susceptibility Mapping in South Korea Using Machine Learning Algorithms. Sensors 18, 2464 (2018).
9. Jie Yin, Dapeng Yu, Rob Wilby: Modelling the impact of land subsidence on urban pluvial flooding: A case study of downtown Shanghai, China. Science of The Total Environment 544, 744-753 (2016).
10. Zeynab Najafi, Hamid Reza Pourghasemi, Gholamabbas Ghanbarian & Seyed Rashid Fallah Shamsi: Land-subsidence susceptibility zonation using remote sensing, GIS, and probability models in a Google Earth Engine platform. Environmental Earth Sciences 79, 491 (2020).
11. Tran, V. A., Le, T. L., Nguyen, N. H., Le, T. N., & Tran, H. H. Monitoring Vegetation Cover Changes by Sentinel-1 Radar Images Using Random Forest Classification Method. Inżynieria Mineralna, (2021).
12. Breiman L, Friedman J, Olshen R, Stone C: Cart. Classification and regression trees. Wadsworth and Brooks/Cole: Monterey, CA, USA (1984).
13. Anh, T. V.,. Monitoring Subsidence in Ca Mau City and Vicinities using the Multi Temporal Sentinel-1 Radar Images. In 4th Asia Pacific Meeting on Near Surface Geoscience & Engineering (Vol. 2021, No. 1, pp. 1-5). EAGE Publications BV. (2021)
14. https://emergency.copernicus.eu, https://emergency.copernicus.eu/mapping/list-of-components/EMSN062, last accessed 2019/02/26.
15. P S J Minderhoud, H Middelkoop, G Erkens, E Stouthamer: Groundwater extraction may drown mega-delta: projections of extraction-induced subsidence and elevation of the Mekong delta for the 21st century. Environmental Research Communications 2, 1 (2020).
16. Van Anh, T., Hanh, T. H., Nga, N. Q., Nghi, L. T., Quang, T. X., Dong, K. T., & Anh, T. T. . Determination of Illegal Signs of Coal Mining Expansion in Thai Nguyen Province, Vietnam from a Combination of Radar and Optical Imagery. In Advances in Geospatial Technology in Mining and Earth Sciences: Selected Papers of the 2nd International Conference on Geo-spatial Technologies and Earth Resources. Cham: Springer International Publishing. 225-242 (2022).
17. W. Thuiller, M. B. Araújo, S. Lavorel: Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. J. Veg. Sci.14(5), 669–680 (2003).
18. C. Baeza, N. Lantada, and J. Moya: Validation and evaluation of two multivariate statistical models for predictive shallow landslide susceptibility mapping of the Eastern Pyrenees (Spain). Environmental Earth Sciences 61, 507–523 (2010).
19. B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, M. B. Dholakia: A comparativestudy of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environ . Environmental Modelling & Software 84, 240-250 (2016).
20. Tien Bui, D., Tuan, T. A., Hoang, N. D., Thanh, N. Q., Nguyen, D. B., Van Liem, N., & Pradhan, B. (2017). Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. Landslides, 14, 447-458.