ISSN 2029-6991 / eISSN 2029-7009

VILNIUS TECH
Vilnius Gediminas Technical University

GEODESY and CARTOGRAPHY

2024
Volume 50
Issue 3
Pages 150–155
https://doi.org/10.3846/gac.2024.20237

UDC 528.481

PREDICTION OF BUILDING SUBSIDENCE IN VIETNAM USING MACHINE LEARNING TECHNIQUES BASED ON LEVELING RESULTS

Dinh Trong TRAN, Ngoc Dung LUONG✉, Dinh Huy NGUYEN

Department of Geodesy and Geomatics Engineering, Hanoi University of Cilvil Engineering, Hanoi, Vietnam

Article History:
- received 24 October 2023
- accepted 03 December 2024

Abstract. Vietnam's rapid urbanization and economic growth have led to an increase in high-rise buildings, making building subsidence a significant concern. Monitoring subsidence is crucial for ensuring building safety and reducing potential risks. The leveling method is commonly used in Vietnam to monitor subsidence, providing valuable data for predicting future subsidence behavior. However, traditional prediction methods based on mathematical models have limitations in capturing complex subsidence patterns. Machine learning techniques have shown promise in enhancing subsidence prediction accuracy. In this study, we analyze machine learning methods for predicting building subsidence using leveling results in Vietnam. We utilize a dataset from a subsidence monitoring network in Hoa Binh General Hospital and compare the performance of linear regression, decision tree regression, and random forest regression models. Our results show that the decision tree and random forest models produce consistent predicted subsidence values, aligning with the observed stability of the building. In contrast, the linear regression model fails to capture the diminishing nature of subsidence over time. We discuss the implications of these findings and highlight the advantages of machine learning in accurately forecasting subsidence. The study demonstrates the potential of machine learning in revolutionizing subsidence prediction and enhancing the monitoring and management of building stability and structural integrity in Vietnam.

Keywords: building subsidence, leveling, machine learning, linear regression, decision tree regression, random forest regression.

✉Corresponding author. E-mail: ***dungln@huce.edu.vn***

1. Introduction

Vietnam's rapid urbanization and economic growth have witnessed the proliferation of high-rise buildings that shape the modern skyline. These towering structures serve as symbols of progress and prosperity, accommodating commercial, residential, and institutional activities. However, with the increasing number of high-rise buildings, the issue of building subsidence has emerged as a significant concern. Building subsidence refers to the gradual sinking or settling of a structure into the ground, which can compromise its structural integrity and pose risks to occupants and neighboring buildings (Forth, 2004; Roy & Robinson, 2009).

Monitoring building subsidence is essential to ensure the safety, functionality, and longevity of high-rise buildings in Vietnam. Subsidence can manifest in various ways, such as vertical settlement, tilting, or differential movement of different parts of the structure. These phenomena can lead to structural damage, including cracks in walls, floors, and foundations, impacting the overall stability of the building (Zhang et al., 2021). By monitoring building subsidence, early warning signs can be detected, allowing for timely interventions to mitigate risks and prevent further damage.

The leveling method, which is widely recognized as the most accurate method (Karila et al., 2013; Lyon et al., 2018), is commonly employed in Vietnam to monitor building subsidence. This technique involves periodically measuring monitoring points on a building to track changes in their elevation over time. The collected data offers valuable insights into the magnitude and rate of subsidence.

Based on the leveling results, various prediction methods are commonly used in Vietnam. Predicting building subsidence from the leveling measurements serves two important purposes in practice: firstly, it allows for the estimation of future subsidence behavior, enabling proactive assessment and management of potential risks; secondly, it offers the advantage of reducing the frequency of leveling measurements or even eliminating the need for continuous monitoring. However, these methods often rely on mathematical models, one traditional approach, such as linear functions, polynomial functions (Bui et al., 2016), and exponential functions (Trần & Nguyễn, 2017),... which are considered traditional approaches. While these models utilize historical leveling data to estimate future subsidence trends and make forecasts, they have limitations. These conventional prediction methods often assume a simplistic relationship between time and subsidence, failing to account for the complex interplay of multiple factors contributing to subsidence.

In recent years, machine learning has emerged as a powerful tool for predicting building land subsidence based on leveling results (Li et al., 2023; Shi et al., 2020). Machine learning techniques offer distinct advantages over conventional methods in terms of accuracy, efficiency, and flexibility. By analyzing large volumes of data, machine learning algorithms can identify intricate patterns, correlations, and nonlinear relationships, enabling more accurate and precise predictions of building subsidence. Furthermore, machine learning models can adapt and refine their predictions as new data becomes available, ensuring continuous improvement and enhancing their reliability over time (Tang & Na, 2021).

In Vietnam, machine learning has found effective applications in various areas such as forecasting soil compression (Le et al., 2020), flooding (Ngo et al., 2020; H. D. Nguyen et al., 2022), and more. These studies have demonstrated the effectiveness of machine learning in enhancing the accuracy and efficiency of predictions, complementing traditional methods. However, the application of machine learning specifically in subsidence prediction remains limited and primarily focused on research. For example, Q. L. Nguyen et al. (2021) conducted a study where they applied a multilayer feed-forward artificial neural network along with the back-propagation algorithm to forecast ground subsidence caused by underground coal mining in Mong Duong. Their research highlighted the potential of machine learning in predicting building subsidence accurately. The application of machine learning in building subsidence prediction holds promising opportunities to revolutionize the field, offering more robust and reliable forecasts.

In this paper, we provide a thorough analysis of machine learning techniques applied to predict building subsidence using leveling results in Vietnam. Our study utilizes a dataset of leveling results obtained from the subsidence monitoring network established for the Oncology and Rehabilitation building at Hoa Binh General Hospital in Vietnam. Through our analysis, we demonstrate the superior performance of machine learning in accurately predicting building subsidence. The results obtained highlight the advantages of machine learning in this context. Furthermore, we discuss the implications of these findings and their significance for the monitoring of building subsidence.

2. Background theory

In this section, we provide an overview of the background theory behind the machine learning models used in this study for subsidence prediction, along with the key evaluation metrics chosen for assessing the model performance.

Linear Regression:

Linear regression is a widely used supervised machine learning algorithm for predicting continuous target variables. It establishes a linear relationship between the input features and the target variable by fitting a straight line to the data. The algorithm assumes that the relationship between the features and the target variable is linear, and it aims to find the best-fitting line that minimizes the difference between the predicted values and the actual target values. Linear regression is a simple yet powerful algorithm that provides interpretability and is widely used in various fields. Brief Outline of the Linear Regression Algorithm (Montgomery et al., 2021):

1. Data Preparation: Gather the dataset consisting of input features (independent variables) and their corresponding target values (dependent variable).
2. Model Representation: Represent the linear regression model as a linear equation, where the target variable is predicted as a linear combination of the input features.
3. Cost Function: Define a cost function, typically the MSE. The goal is to minimize this cost function.
4. Parameter Estimation: Estimate the parameters (coefficients) of the linear regression model that minimize the cost function.
5. Predictions: Utilize the trained linear regression model to make predictions on new, unseen data by plugging in the input features into the linear equation.

Decision Tree Regression:

The Decision Tree Regression algorithm is a supervised machine learning algorithm used for predicting continuous target variables. It utilizes a tree-like model to make predictions by partitioning the feature space into distinct regions. Each region represents a leaf node in the tree, and the target value is estimated by taking the average (or any other statistical measure) of the target values in that region. Decision trees are versatile and widely used due to their interpretability and ability to handle both numerical and categorical features. Brief Outline of the Decision Tree Regression Algorithm (Hastie et al., 2009):

1. Select the target variable: Determine the variable to be predicted (continuous target variable) and the features (independent variables) that will be used for prediction.
2. Splitting criteria: Choose a splitting criterion to determine the optimal feature and value to split the data at each node. Common criteria include MSE or variance reduction.
3. Build the tree: Recursively partition the data based on the selected splitting criteria until a stopping condition is met. This condition can be the maximum depth of the tree, minimum number of samples required to split a node, or other pre-defined conditions.
4. Assign a prediction value: At each leaf node, assign a prediction value based on the target values of the samples within that region. The most common approach is to take the average of the target values,

but other statistical measures can be used as well.

5. Predictions: Traverse the tree to predict the target variable for new data points by following the splitting conditions until reaching a leaf node. The prediction is the assigned value at that leaf node.

Random Forest Regression:

The Random Forest Regression algorithm is an ensemble learning method that combines multiple decision trees to create a predictive model for regression tasks. It works by constructing a multitude of decision trees during the training phase and making predictions based on the average or majority vote of the individual tree predictions. Here is a brief outline of the Random Forest Regression algorithm (Breiman, 2001):

1. Randomly select a subset of the training data.
2. Construct a decision tree based on the selected subset by recursively partitioning the data based on the feature that provides the best split.
3. Repeat steps 1 and 2 to create a collection of decision trees.
4. For prediction, pass the test data through each decision tree and obtain a prediction from each tree.
5. Aggregate the predictions from all the decision trees to obtain the final prediction. For regression tasks, this can be done by taking the average of the individual tree predictions.

For evaluating the performance of the subsidence prediction models, we employ the following metrics:

Mean Squared Error (MSE):

It measures the average squared difference between the predicted and actual subsidence values. *MSE* is calculated using the formula (Wang & Bovik, 2009):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \tag{1}$$

where n is the number of validation samples, y_i represents the actual subsidence value, and $\hat{y}_i$ is the predicted subsidence value.

MSE quantifies the overall prediction error, with lower values indicating better accuracy.

Mean Absolute Error (MAE):

It measures the average absolute difference between the predicted and actual subsidence values. *MAE* is calculated using the formula (Chai & Draxler, 2014):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{2}$$

MAE provides a measure of the average magnitude of errors, regardless of their direction.

R-squared (R^2) Score:

It represents the proportion of the variance in the target variable that can be explained by the model. R^2 is calculated using the formula (Lewis-Beck & Skalaban, 1990):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \tag{3}$$

where $\bar{y}_i$ the mean of the actual subsidence value.

R^2 ranges from 0 to 1, with higher values indicating a better fit of the model to the data. It provides an indication of how well the model captures the variability in the target variable.

We chose these models and evaluation metrics for our study due to their effectiveness in subsidence prediction tasks and their interpretability. Linear regression provides a simple and transparent model, allowing us to understand the individual impact of input features. Decision tree regression and random forest regression capture non-linear relationships and handle complex datasets well. Support vector regression is known for its ability to handle non-linear data and complex feature spaces. The selected evaluation metrics provide a comprehensive assessment of the models' performance in terms of accuracy and explanatory power.

By employing these models and evaluation metrics, we aim to develop accurate subsidence prediction models for buildings in Vietnam. Through careful comparison and analysis of the model results, we can determine the most suitable model for subsidence prediction in this specific context and gain valuable insights into the factors influencing subsidence behavior.

3. Data and method

3.1. Study data

The data used in this study focuses on the subsidence values of monitoring point M1. The dataset includes the measurement times and corresponding subsidence values in millimeters (mm) for monitoring point M1 (Table 1). The data points are as follows:

Table 1. The data points

Measurement time	Subsidence value (mm)
18/06/2021	–1.4
18/07/2021	–2.1
18/08/2021	–2.11
28/03/2022	–4.62
18/04/2022	–5.0
18/05/2022	–5.4
18/06/2022	–6.0
18/07/2022	–6.1
18/08/2022	–7.0
18/09/2022	–7.5
18/10/2022	–7.8
18/11/2022	–8
18/12/2022	–8.1
18/01/2023	–8.2
18/02/2023	–8.21
18/03/2023	–8.20
18/04/2023	–8.19
18/05/2023	–8.20

The subsidence value at a measurement time is determined as the difference between the height at that specific time and the height at the previous time. These heights were obtained using leveling techniques and are part of a subsidence monitoring network consisting of nine points. The monitoring network is specifically established for the Oncology and Rehabilitation building, which is part of the "Expansion of Hoa Binh General Hospital" project. The foundation of the building utilizes pile technology with a length of 18 meters.

The objective of this research paper is to employ machine learning methodologies to predict subsidence values for monitoring point M1, providing valuable insights for the assessment and management of the building's stability and structural integrity.

3.2. Study method

In this study, we employed machine learning techniques to predict subsidence values for a building in Vietnam. The methods used in this study are described in Figure 1:

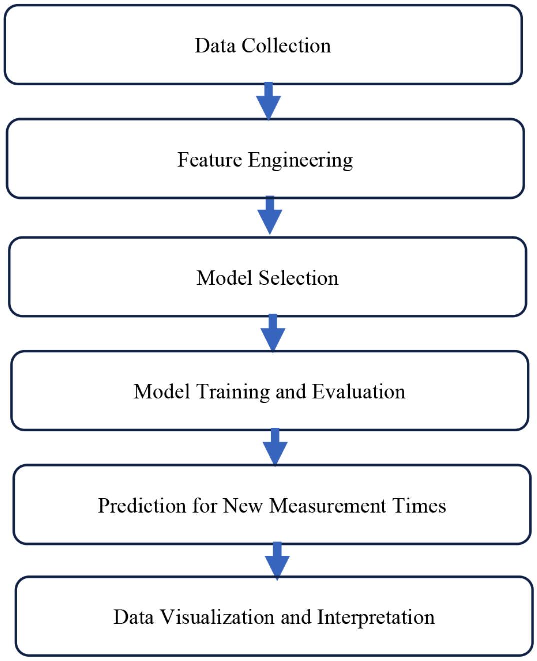
Data Collection
Feature Engineering
Model Selection
Model Training and Evaluation
Prediction for New Measurement Times
Data Visualization and Interpretation

Figure 1. Workflow of building subsidence prediction using machine learning

Data Collection: We collected subsidence data from the monitoring point M1. The data consists of measurement times and corresponding subsidence values in millimeters (mm).

Feature Engineering: We engineered the features required for training the machine learning models. In this study, we focused on a single feature, namely the measurement time. We transformed the measurement time into an ordinal representation, which is a numerical format suitable for model training.

Model Selection: We explored different machine learning algorithms to predict subsidence values. The models used in this study include linear regression, decision tree regression, and random forest regression. These models were selected based on their ability to capture different types of relationships between the input feature (measurement time) and the target variable (subsidence value). The models were implemented using the scikit-learn library in Python.

Model Training and Evaluation: We divided the data into training and validation sets to train and evaluate the performance of the selected models. The training set was used to fit the models to the data, while the validation set allowed us to assess the models' predictive capabilities. We used evaluation metrics such MSE, MAE, and R^2 score to quantify the accuracy and goodness of fit of each model.

Prediction for New Measurement Times: After training the models, we utilized them to predict subsidence values for new measurement times. We provided a sample set of measurement times for future reference and used the trained models to generate predictions.

Data Visualization and Interpretation: To visualize the results, we created plots showcasing the actual subsidence measurements, model predictions, and the predicted subsidence values for the new measurement times. These visual representations aided in understanding the patterns and trends in the subsidence data and the performance of the trained models.

By employing these study methods and analyzing the collected data, we aimed to develop accurate and reliable predictions of subsidence values for the building in Vietnam, facilitating effective monitoring and decision-making in the context of structural stability and integrity.

4. Result and discussion

The dataset was divided into a training set consisting of 14 data points and a validation set containing 4 data points. Three different models, namely Decision Tree, Random Forest, and Linear Regression, were trained and evaluated on the data. The performance metrics for each model on the validation set are presented in Table 2.

Table 2. Evaluation metrics for each model

Evaluation metrics	Decision Tree	Random Forest	Linear Regression
MSE	0.36855	1.05472	0.23799
MAE	0.50500	0.85995	0.42129
R^2	0.93088	0.80219	0.95536

Among the models, Linear Regression shows the lowest errors (MSE and MAE) and the highest R^2 value, indicating a strong linear relationship between the measurement time and subsidence values. However, it may not accurately capture the diminishing nature of subsidence over time. The Decision Tree and Random Forest models, although slightly less accurate according to the evaluation metrics, provide predictions that align with the observed stability of the building and may be more suitable for capturing non-linear trends in subsidence.

Using the trained models, subsidence values were predicted for three new measurement times: 18/06/2023, 18/07/2023, and 18/08/2023. Table 3 displays the predicted subsidence values, while Figure 2 showcases the predicted subsidence lines for each model.

Table 3. Predicted subsidence of monitoring point M1 for each model

Measurement time	Decision Tree (mm)	Random Forest (mm)	Linear Regression (mm)
2023-06-18	–8.2000	–8.1979	–9.4743
2023-07-18	–8.2000	–8.1979	–9.7771
2023-08-18	–8.2000	–8.1979	–10.0900

The Decision Tree and Random Forest models produce identical predicted subsidence values for all three measurement times, suggesting a consistent subsidence value of –8.2 mm. These models capture the observed stability of the building where no significant subsidence has been observed since November 2022. The predictions align with the fact that no further substantial subsidence is expected in the future. In contrast, the Linear Regression model predicts a decreasing trend in subsidence values over time. However, the predicted subsidence values (–9.474272, –9.777098, –10.090019) still indicate subsidence, which does not accurately reflect the gradual reduction observed in the actual data. The linear trend implied by the Linear Regression model fails to capture the diminishing nature of subsidence over time.

Although the Linear Regression model exhibits superior performance in terms of evaluation metrics, caution should be exercised when using it to predict subsidence in this particular scenario. Its predictions may lead to misleading results as the model assumes a linear relationship between time and subsidence, which contradicts the diminishing subsidence observed in the data.

Considering the limitations of the Linear Regression model, the Decision Tree and Random Forest models are more suitable for predicting the subsidence of the building in this study. These models effectively capture the absence of significant subsidence in recent data and provide predictions that align with the observed stability of the building. By considering various features and their interactions, the Decision Tree and Random Forest models offer more flexibility and robustness in capturing the non-linear behavior of subsidence over time.

5. Conclusions

This study utilized machine learning models, Decision Tree, Random Forest, and Linear Regression, to predict subsidence values for a building in Vietnam. Performance metrics and the ability to capture observed subsidence behavior were used to evaluate the models, which were trained and validated using a specific time period's subsidence measurements.

The analysis revealed that Linear Regression had the lowest errors and highest R^2 value but failed to capture the diminishing subsidence trend in the data. In contrast, the Decision Tree and Random Forest models provided predictions aligned with the building's observed stability, despite slightly higher errors and lower R^2 values. These models demonstrated the ability to capture non-linear

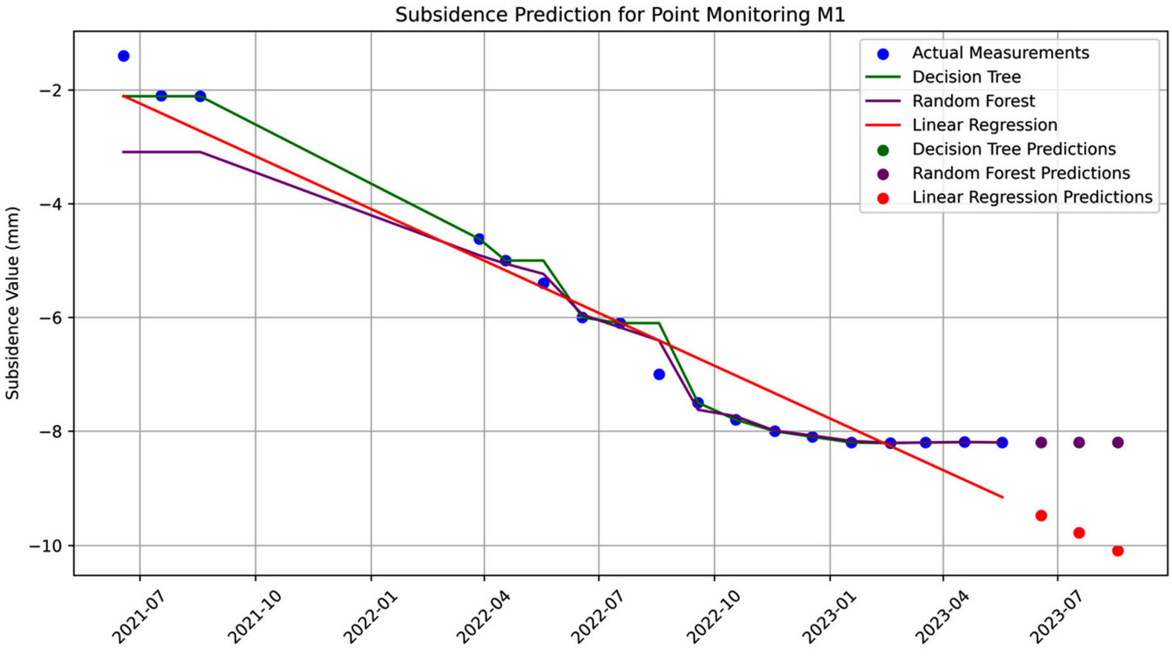

Figure 2. Predicted subsidence lines of monitoring M1 for each models

relationships and adapt to complex data patterns, making them more suitable for subsidence prediction in this context.

This study demonstrates the potential of machine learning models–Decision Tree and Random Forest–in predicting subsidence values for buildings. Despite slightly lower accuracy metrics than Linear Regression, these models align with observed stability and capture non-linear trends.

Acknowledgements

The authors gratefully thank Institute of Geodesy in Construction – Hanoi University of Civil Engineering, for providing the subsidence data of Hoa Binh General Hospital project.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bui, D. C., Nguyen, N. D., Bui, V. K., Nguyen, P. T., Vu, T. H., Nguyen, V. K., & Tran, A. V. (2016). Research on establishing a program to process monitoring data and forecast construction settlement. *Journal of Geodesy and Cartography*, (29), 53–58. https://doi.org/10.54491/jgac.2016.29.193 (in Vietnamese)

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions*, *7*(1), 1525–1534. https://doi.org/10.5194/gmdd-7-1525-2014

Forth, R. A. (2004). Groundwater and geotechnical aspects of deep excavations in Hong Kong. *Engineering Geology*, *72*(3–4), 253–260. https://doi.org/10.1016/j.enggeo.2003.09.003

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7

Karila, K., Karjalainen, M., Hyyppä, J., Koskinen, J., Saaranen, V., & Rouhiainen, P. (2013). A comparison of precise leveling and persistent scatterer SAR interferometry for building subsidence rate measurement. *ISPRS International Journal of Geo-Information*, *2*(3), 797–816. https://doi.org/10.3390/ijgi2030797

Le, H.-A., Nguyen, T.-A., Nguyen, D.-D., & Prakash, I. (2020). Prediction of soil unconfined compressive strength using Artificial Neural Network model. *Vietnam Journal of Earth Sciences*, *42*(3), 255–264. https://doi.org/10.15625/0866-7187/42/3/15342

Lewis-Beck, M. S., & Skalaban, A. (1990). The *R*-squared: Some straight talk. *Political Analysis*, *2*, 153–171. https://doi.org/10.1093/pan/2.1.153

Li, F., Liu, G., Tao, Q., & Zhai, M. (2023). Land subsidence prediction model based on its influencing factors and machine learning methods. *Natural Hazards*, *116*(3), 3015–3041. https://doi.org/10.1007/s11069-022-05796-9

Lyon, T. J., Filmer, M. S., & Featherstone, W. E. (2018). On the use of repeat leveling for the determination of vertical land motion: Artifacts, aliasing, and extrapolation errors. *Journal of Geophysical Research: Solid Earth*, *123*(8), 7021–7039. https://doi.org/10.1029/2018JB015705

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Ngo, T. P. T., Ngo, L. H., Nguyen, K. Q., Bui, T. T., Tran, P. V., Nhu, H. V., & Nguyen, Y. H. T. (2020). Applying Random Forest approach in forecasting flash flood susceptibility area in Lao Cai region. *Journal of Mining and Earth Sciences*, *61*(5), 30–42. https://doi.org/10.46326/JMES.2020.61(5).04

Nguyen, H. D., Quang-Thanh, B., Nguyen, Q.-H., Nguyen, T. G., Pham, L. T., Nguyen, X. L., Vu, P. L., Thanh Nguyen, T. H., Nguyen, A. T., & Petrisor, A.-I. (2022). A novel hybrid approach to flood susceptibility assessment based on machine learning and land use change. Case study: A river watershed in Vietnam. *Hydrological Sciences Journal*, *67*(7), 1065–1083. https://doi.org/10.1080/02626667.2022.2060108

Nguyen, Q. L., Nguyen, Q. M., Tran, D. T., & Bui, X. N. (2021). Prediction of ground subsidence due to underground mining through time using multilayer feed-forward artificial neural networks and back-propagation algorithm – case study at Mong Duong underground coal mine (Vietnam). *Mining Science and Technology (Russia)*, *6*(4), 241–251. https://doi.org/10.17073/2500-0632-2021-4-241-251

Roy, D., & Robinson, K. E. (2009). Surface settlements at a soft soil site due to bedrock dewatering. *Engineering Geology*, *107*(3–4), 109–117. https://doi.org/10.1016/j.enggeo.2009.05.006

Shi, L., Gong, H., Chen, B., & Zhou, C. (2020). Land subsidence prediction induced by multiple factors using machine learning method. *Remote Sensing*, *12*(24), Article 4044. https://doi.org/10.3390/rs12244044

Tang, L., & Na, S. (2021). Comparison of machine learning methods for ground settlement prediction with different tunneling datasets. *Journal of Rock Mechanics and Geotechnical Engineering*, *13*(6), 1274–1289. https://doi.org/10.1016/j.jrmge.2021.08.006

Trần, N. Đ., & Nguyễn, C. C. (2017). Establish a model of construction foundation submission according to leveling of settlement monitoring. *Journal of Building Science and Technology*, *1*, 54–62 (in Vietnamese).

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, *26*(1), 98–117. https://doi.org/10.1109/MSP.2008.930649

Zhang, J., Liu, H., Sun, X., & Liu, S. (2021). Processing of building subsidence monitoring data based on fusion Kalman filtering algorithm. *Alexandria Engineering Journal*, *60*(3), 3353–3360. https://doi.org/10.1016/j.aej.2021.02.002