# Enhancing real estate price prediction using optimized least squares moment balanced machine

*Riqi* Radian Khasani[1*]

[1]Department of Civil Engineering Diponegoro University, Jalan Prof Sudharto SH Semarang, Central Java, Indonesia

**Abstract.** Real estate price prediction is crucial for urban planning and economic forecasting, as property values are influenced by factors such as location and infrastructure. Traditional methods often struggle to capture the complex relationships between these variables, resulting in inefficient land use and suboptimal resource allocation. To address this challenge, this study introduces the Optimized Least Squares Moment Balanced Machine (OLSMBM), an advanced machine learning model designed to enhance the accuracy of real estate price predictions. The model incorporates key features such as transaction date, house age, proximity to MRT stations, number of convenience stores, and geographic location. The OLSMBM was benchmarked against five other machine learning models, including LSSVM, BPNN, ELSIM, Decision Tree, and Linear Regression. The results from a 10-fold cross-validation demonstrate that OLSMBM consistently outperforms other models across five evaluation metrics, including RMSE (6.977), MAE (4.752), MAPE (13.76%), R (0.846), and $R^2$ (0.728). The comprehensive evaluation, summarized by the Reference Index (RI), showed that OLSMBM achieved a perfect RI score of 1.000, highlighting its superior performance. These findings underscore the potential of the OLSMBM as a decision-support tool, enhancing the accuracy of real estate price predictions and reinforcing data-driven strategies in urban planning.

## 1 Introduction

Real estate price prediction plays a crucial role in shaping sustainable urban development. As cities continue to expand, strategic planning becomes essential for balancing economic growth with environmental stewardship [1]. Accurate forecasting of property values provides valuable insights into the evolution of urban areas, enabling planners and policymakers to anticipate growth patterns and align development with sustainability goals. When high-value areas are predicted accurately, city planners can prioritize these locations for high-density mixed-use development [2]. Such developments optimize land use, decrease the need for extensive transportation networks, and help lower the overall carbon footprint of the city.

* Corresponding author: riqi@live.undip.ac.id

Conversely, in regions where property values are expected to decline, perhaps because of environmental risks such as flooding, predictive models can inform decisions on resilient infrastructure investments and safeguard both the urban landscape and its residents from the impacts of climate change [3]. Moreover, real estate price predictions support more effective resource allocation by guiding investments in sustainable infrastructure such as public transportation, green spaces, and renewable energy systems. These insights can help urban planners avoid over-concentration of development in already stressed areas and encourage a more equitable distribution of resources across a city [4]. By integrating machine learning-based real estate price predictions into urban planning, cities can become more resilient, livable, and sustainable, effectively addressing the ongoing challenges posed by economic pressures [5].

The real estate market has long been a vital driver of economic growth, with property values serving as the key indicators of economic and social development. However, as global urbanization accelerates, real estate price predictions become increasingly complex. Traditional valuation methods, often reliant on manual assessments, expert opinions, or basic statistical models, struggle to accurately forecast market trends because of the intricate interactions between multiple variables that influence property prices [6]. Factors such as proximity to transportation, neighborhood amenities, geographical location, and property age are crucial for determining real estate values [7]. However, conventional methods frequently fail to account for these complexities. In response to these challenges, machine learning has emerged as a transformative technology for real-estate price predictions. Machine learning algorithms, particularly those capable of processing large datasets, can simultaneously analyze and identify patterns within a multitude of variables, uncovering relationships that may not be apparent through traditional methods [8]. By incorporating key inputs such as transaction date, house age, distance to transportation hubs (such as MRT stations), and the availability of local amenities (such as convenience stores), machine learning models offer a more data-driven approach to forecasting real estate prices. The ability of these models to integrate dynamic and geographically specific data makes them indispensable for urban planners, real estate developers, and policymakers, who require accurate and nuanced information to make informed decisions in the complex landscape of urban environments.

This study aims to develop Optimized Least Squares Moment Balanced Machine (OLSMBM), an advanced machine learning model designed for predicting real estate prices, focusing on an urban setting characterized by significant variability in property values. The primary objective is to create a model that accurately forecasts real estate prices based on key factors, such as transaction date, house age, distance to transportation hubs, and the availability of local amenities. By utilizing these specific inputs, the model seeks to provide a reliable tool for decision makers in urban planning and development. This study demonstrates the practical application of machine learning in promoting efficient land use and resource allocation within cities. By integrating predictive insights into the urban planning processes, this study contributes to the development of more resilient and sustainable urban environments.

## 2 Methodology

### 2.1 Model Structure

In computational modeling, the OLSMBM framework employs Backpropagation Neural Networks (BPNN) to assign weights to each data point while incorporating the principles of Least Squares Support Vector Machines (LSSVM) to determine the optimal moment hyperplane [9]. As shown in Eq. (1), a specific weight is assigned to each data point within

the dataset used for regression analysis. The objective function of the OLSMBM model is designed to minimize the moment required to achieve a balanced condition, as shown in Eq. (2). To enhance the predictive accuracy, the Optical Microscope Algorithm (OMA) is utilized to optimize the parameters of the OLSMBM, ensuring that the model delivers optimal predictions [10].

$$D = \{(x_1, y_1, F_1), (x_2, y_2, F_2), \dots, (x_i, y_i, F_i)\} \in \mathbb{R}^n \tag{1}$$

$$\begin{aligned} \text{Minimize} \quad & J(w,d) = \frac{1}{2}||w||^2 + \gamma \frac{1}{2}\sum_{k=1}^{N} m \\ & \text{Where: } m = F_k {d_k}^2 \\ \text{Subject to} \quad & y_k = w.\varphi(x_k) + b + d_k \end{aligned} \tag{2}$$

In the LSMBM model, $x_k$ and $y_k$ denote the input and output, respectively. The weight vector is represented by $(w)$, while $(\gamma)$ acts as the regularization constant that balances model complexity with generalization. The term $(m)$ indicates the moments generated by each data point, $(d_k)$ corresponds to the error term. The total number of data points is represented by $(n)$, and $(F_k)$ refers to the weight assigned to the data point $(x_k)$, as determined by the initial prediction using the BPNN algorithm.

A systematic framework for developing OLSMBM to predict real estate prices, organized into five key stages, as shown in Fig. 1. Stage (1) begins with the construction of a comprehensive real-estate price database. This involves collecting and organizing essential data, such as transaction dates, house ages, distances to the nearest MRT stations, the number of nearby convenience stores, and geographical coordinates. This well-structured dataset forms the foundation for the modeling process, ensuring that the subsequent stages are based on accurate and relevant information. Stage (2) involves determining the input and output variables for the model. Inputs include various factors that influence property values, such as proximity to transportation and neighborhood amenities, while the output variable is the house price per unit area. Defining these variables is crucial for setting the scope of the model and ensuring that all the significant factors are accounted for in the prediction process. Stage (3) focuses on data pre-processing, which prepares the dataset for model training. This stage includes normalizing the input variables to ensure consistency, ensuring that the inputs are correctly scaled and ready for analysis. Stage (4) is dedicated to the development of a predictive model. OLSMBM is employed to create a model capable of accurately forecasting real estate prices. This stage involves selecting the appropriate algorithm, training the model on the preprocessed data, and optimizing the model parameters for the best performance. Finally, Stage (5) evaluates the performance of the model and analyzes the prediction results. Key performance metrics, such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient correlation (R), and coefficient determination ($R^2$) are used to assess the accuracy of the predictions. In addition, the Reference Index (RI) is calculated to rank the overall performance of the model relative to the other models. The Reference Index (RI) was employed as a deterministic metric, integrating the results of all measurements by assigning equal weights to each, thereby offering a comprehensive summary of the performance outcomes [11]. The final stage ensures that the model is accurate, robust, and reliable for practical applications. The equations for the performance metrics used in this study are presented in Table 1.

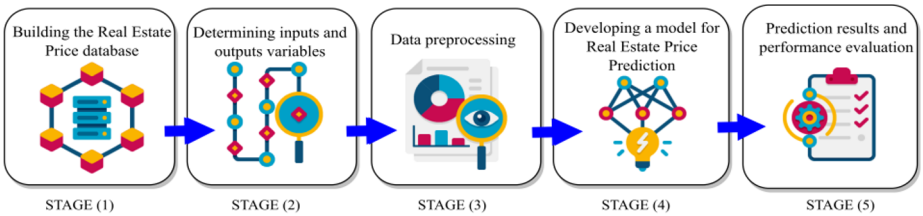


**Fig. 1.** Framework used to develop the real estate prices model.

**Table 1.** Performance evaluation.

| Performance evaluation | Formula |
|---|---|
| RMSE | $\sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - f_i)^2}$ |
| MAPE | $\frac{100}{n}\sum_{i}^{n}\frac{\lvert y_i - f_i\rvert}{y_i}$ |
| MAE | $\frac{1}{n}\sum_{i}^{n}\lvert y_i - f_i\rvert$ |
| R | $\frac{n\sum_i^n y_i f_i - (\sum_i^n y_i)(\sum_i^n f_i)}{\sqrt{n(\sum_i^n {y_i}^2) - (\sum_i^n y_i)^2}\sqrt{n(\sum_i^n {f_i}^2) - (\sum_i^n fi)^2}}$ |
| R2 | $1 - \frac{\sum_i^n (y_i - f_i)^2}{\sum_i^n (y_i - \overline{y})^2}$ |
| RI | $\frac{R_{norm} + R^2_{norm} + (1 - RMSE_{norm}) + (1 - MAE_{norm}) + (1 - MAPE_{norm})}{5}$ |

## 2.2 Model Adaptation

The flowchart in the Fig.2. outlines the primary processes involved in adapting the OLSMBM model. The adaptation begins with the collection and organization of relevant real estate price data, which are then subjected to data preprocessing. This step is essential to ensure that the data is ready for use in the machine learning model. Preprocessing involves normalizing the data, a technique that scales numerical features to a standardized range, between 0 and 1. Following pre-processing, the dataset was divided into training and testing subsets using a 10-fold cross-validation technique. This method partitions the dataset into ten equal parts, with the model being trained on nine parts and tested on the remaining part. This process was repeated ten times, with each fold serving as the testing set once, ensuring that the performance of the model was thoroughly evaluated and generalized across different subsets of the data. The training dataset was then input into the OLSMBM model, which was subjected to fitness evaluation. This involves iteratively adjusting the model parameters to optimize performance. The Optical Microscope Algorithm (OMA) was employed to fine-tune the parameters of the model (γ and σ), improving the accuracy of the model. The OMA algorithm adopts the magnification process of an optical microscope and enhances the fine details necessary for optimizing the model. The process continues until a termination

criterion is satisfied, which occurs when the number of iterations (NI) exceeds a predetermined maximum iteration (max_iter). Once this criterion is satisfied, the model is optimized, and the final OLSMBM parameters are established. The optimized model was then tested on the testing dataset, which is a separate subset of data not used during training. This evaluation of independent data is critical for assessing how well the model generalizes to new and unseen cases. The final prediction results provide a clear indication of the performance of the model, demonstrating its capability to predict real estate prices accurately based on input variables. This structured approach, from data preprocessing to parameter optimization and testing, ensures that the OLSMBM model is both robust and reliable, and capable of delivering accurate predictions for real estate pricing.

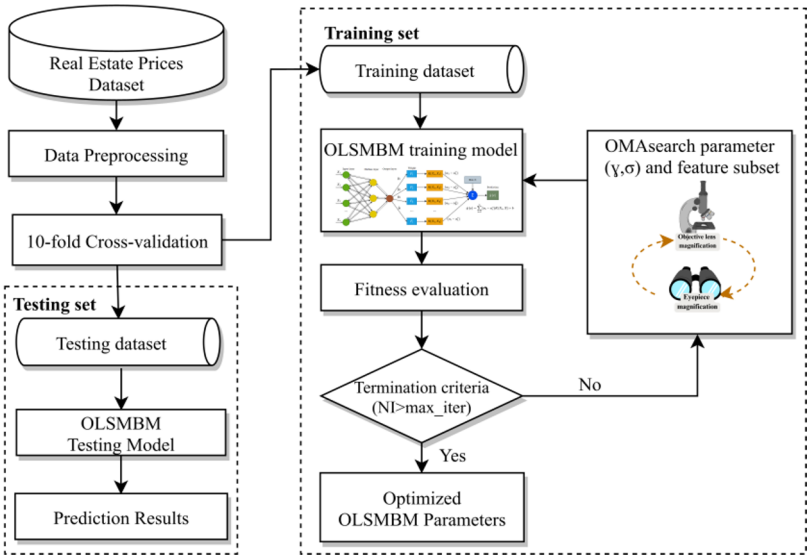


**Fig. 2.** Framework used to develop the real estate prices model.

## 3 Results and Discussion

### 3.1 Data collection and processing

The dataset used in this study was sourced from the public database of the Ministry of Interior of Taiwan, which covers real estate transactions between June 2012 and May 2013. Data was collected from two districts in Taipei City and two districts in New Taipei City, with a particular focus on the Xindian District in New Taipei City, Taiwan. Comprising 414 records, this dataset provides detailed information on various factors influencing property prices and was originally adopted from the study by Yeh & Hsu (2018) [12]. This historical market data serves as the foundation for developing and evaluating the machine learning model used to predict house prices per unit area, ensuring that the analysis is well grounded in real-world transactions and market conditions. The Table 2. presents a comprehensive overview of the key factors influencing house-price predictions, highlighting their importance in the valuation process. The transaction date is recorded in fractional years to reflect the specific timing of the sale, accounting for the market conditions at that time. For instance, 2013.2 represents a transaction in March 2013, while 2012.5 corresponds to June 2012, ensuring that temporal market fluctuations are accurately captured in the model. House age is another

factor that indicates the number of years since a house was built. Typically, newer properties command higher values owing to modern construction standards and lower maintenance requirements, making this a crucial determinant of the price. The distance to the nearest MRT station is also considered, reflecting the accessibility of the property to public transportation. Measured as the straight-line distance from the house to the closest MRT station, this factor shows that shorter distances generally lead to higher property values owing to the convenience of transportation. In addition to accessibility, the number of convenience stores within a certain radius of the house was also included, as this represents the availability of local amenities. A higher number of nearby convenience stores suggests a more developed neighborhood, which can positively impact property values. The latitude and longitude of the house location are used to provide precise geographical positioning, helping to identify the specific location of the property on the Earth's surface. These coordinates are particularly important in urban areas where locations play a critical role in valuation. Finally, house price per unit area is the primary outcome variable, representing the predicted value per unit area of the property. This value is expressed in thousands of New Taiwan Dollars (NTD) per Ping, with one Ping equaling approximately 3.3 square meters. The model seeks to predict this value based on the interaction of all aforementioned factors, ensuring a comprehensive and accurate valuation process.

**Table 2.** The description of input and output variables.

| Code | Factor | Unit | Description |
|---|---|---|---|
| X1 | Transaction date | - | Represents the transaction date in fractional years. For example, 2013.2 refers to March 2013 and June 2012 is presented as 2012.5. |
| X2 | House age | year | The age of the house at the time of transaction, measured in years. |
| X3 | Distance to the nearest MRT station | meter | The straight-line distance from the house to the nearest MRT station. |
| X4 | Number of convenience store | integer | The number of convenience stores located within a certain radius of the house. |
| X5 | Latitude | degree | The geographical latitude of the house location |
| X6 | Longitude | degree | The geographical longitude of the house location, measured in degrees. |
| Y | House price of unit area | 1000 NTD/ Ping | The predicted price of the house per unit area, measured in thousands of New Taiwan Dollars (NTD) per Ping (1 Ping = 3.3 square meters). |

### 3.2 Model Application

The predictive performance of the OLSMBM model was evaluated in comparison with four other machine learning models, including Evolutionary Least Square Support Vector Machine Inference Model (ELSIM) [13], Least Square Support Vector Machine (LSSVM), Backpropagation Neural Network (BPNN), Decision Tree (DT), and Linear Regression (LR). A comparative evaluation of various models using several key performance metrics is presented in Table 3. These metrics measure the accuracy of the models in predicting real estate prices with the Reference Index (RI) used to rank the overall performance. A higher

RI reflects better overall performance, with an RI of 1.000, representing the benchmark standard. OLSMBM topped the rankings as the best performing model, achieving the highest RI of 1,000. It demonstrated superior accuracy with the lowest RMSE (6.977), MAE (4.752), and MAPE (13.76%). Additionally, the R value of 0.846 indicate strong correlation, and $R^2$ of 0.728 suggests that the model fits the data well, accurately capturing the relationship between the predictors and the outcome, establishing OLSMBM as the most reliable model in this comparison. ELSIM followed closely in second place, with a slightly lower RI of 0.920. It has higher RMSE (7.162) and MAE (4.882) values compared to OLSMBM and a MAPE of 14.17%, indicating slightly less reliable predictions, ELSIM still performs well overall. The R value of 0.841 and $R^2$ of 0.718 support its good performance, although it is less optimal than that of OLSMBM. LSSVM is ranked third, with an RI of 0.684. The RMSE (7.554) and MAE (5.290) reflect reduced accuracy, whereas the MAPE of 15.81% and lower R (0.821) and $R^2$ (0.685) values suggest it less effective than both OLSMBM and ELSIM. The BPNN ranks fourth, exhibiting a significant drop in performance, with an RI of 0.491. The RMSE of the model (8.001) and MAE (5.559) were notably higher, indicating lower accuracy. The MAPE of 16.57% further highlights its limitations in providing reliable predictions. DT and LR are the lowest-ranked models with RI of 0.272 and 0.006, respectively. RMSE of DT (8.798) and MAE (5.836) demonstrated poor accuracy, whereas LR performed the worst overall, with an RMSE of 8.744, MAE of 6.268, and MAPE of 19.05%. Both models had the lowest $R^2$ values (0.629 for DT and 0.586 for LR), indicating the least suitable for this predictive task.

**Table 3.** Testing model performance evaluation.

| Model | RMSE | MAE | MAPE | R | R2 | RI |
|---|---|---|---|---|---|---|
| OLSMBM | 6.977 | 4.752 | 13.76 | 0.846 | 0.728 | 1.000 |
| ELSIM | 7.162 | 4.882 | 14.17 | 0.841 | 0.718 | 0.920 |
| LSSVM | 7.554 | 5.290 | 15.81 | 0.821 | 0.685 | 0.684 |
| BPNN | 8.001 | 5.559 | 16.57 | 0.808 | 0.660 | 0.491 |
| DT | 8.798 | 5.836 | 16.53 | 0.785 | 0.629 | 0.272 |
| LR | 8.744 | 6.268 | 19.05 | 0.758 | 0.586 | 0.006 |

The linear correlation between the actual and predicted real estate prices using the OLSMBM model is shown in Fig.3. The left graph represents the training results, and the right graph shows the test results for fold 1. These results indicate that the OLSMBM effectively captures the relationship between the input variables and predicted real estate prices. The alignment of data points along the diagonal line suggests a strong correlation, demonstrating the ability of the model to predict real estate prices accurately based on the provided dataset. The close proximity of the predicted values to the actual values in both the training and testing phases highlights the robustness and reliability of the model for forecasting property prices.

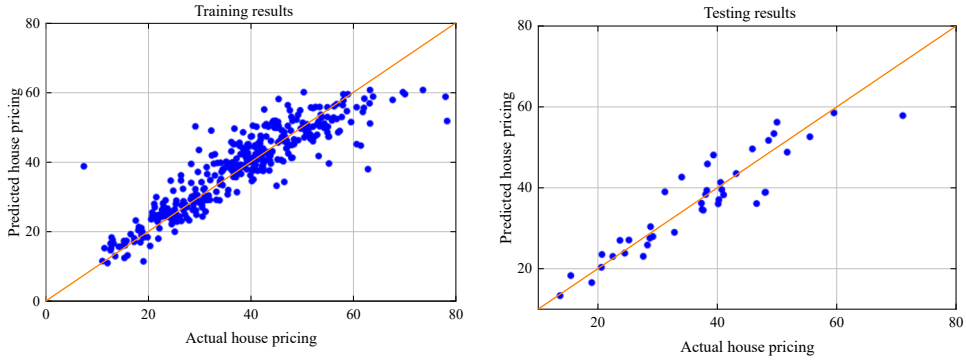


**Fig. 3.** Actual and predicted results for Fold 1.

### 3.3 Discussion

The primary contribution of this study lies in the development and application of the Optimized Least Squares Moment Balanced Machine (OLSMBM) model for real estate price prediction. The OLSMBM model integrates the principles of Least Squares Support Vector Machine (LSSVM) to determine the optimal moment hyperplane, and employs Backpropagation Neural Networks (BPNN) to assign appropriate weights to each data point. This combination provides a distinct advantage by enhancing the predictive accuracy of the model and robustness in handling complex real estate datasets. To thoroughly evaluate the OLSMBM model, a rigorous benchmarking process was conducted, comparing its performance against five other established machine learning models, including ELSIM, LSSVM, BPNN, DT, and LR. The predictive capabilities of the model were assessed using various performance metrics, including RMSE, MAE, MAPE, R and $R^2$. OLSMBM consistently demonstrated superior accuracy across all metrics, particularly in its ability to minimize error and maximize correlation with actual real estate prices. The 10-fold cross-validation technique was employed to validate the models, ensuring a robust assessment of their performance during both the training and testing phases. The results of this comprehensive evaluation revealed that OLSMBM achieved the best scores across all performance metrics, consistently outperforming the other models. Specifically, the OLSMBM model recorded an RMSE of 6.977, an MAE of 4.752, a MAPE of 13.76%, R value of 0.846 and $R^2$ value of 0.728, highlighting its effectiveness in predicting real estate prices with high precision. Furthermore, the performance of the model was encapsulated by the Reference Index (RI) score, where OLSMBM achieved the highest rank, underscoring its superiority in real estate price prediction. The findings from this study suggest that the OLSMBM offers significant potential as a decision-support tool, providing valuable insights for policymakers and urban developers that can inform sustainable urban development strategies. This study reinforces the importance of integrating advanced machine learning models, such as OLSMBM, in real estate prediction, supporting data-driven approaches to urban planning and sustainability.

### 3.4 Limitations

The OLSMBM demonstrated significant accuracy and robustness, marking a significant advancement in real estate price predictions. Despite these contributions, this study has

several limitations. First, the model relies on a specific set of input variables, such as transaction date, house age, and proximity to MRT stations, which may not fully capture all the factors influencing real estate prices. Other variables such as economic indicators, social dynamics, and environmental conditions were not included in this study and could potentially improve the accuracy of the model. Another limitation is the geographical scope of the dataset, which was confined to certain districts in Taipei City. The applicability of the model to other regions or different real estate markets may require further validation or adaptation, as real estate dynamics can vary significantly across different urban settings. Finally, this study primarily compared the performance of the OLSMBM model with several other machine learning models, including ELSIM, LSSVM, BPNN, DT and LR. Future research could be enhanced by incorporating an even broader range of models for comparison.

## 4 Conclusion

This study successfully developed Optimized Least Squares Moment Balanced Machine (OLSMBM), an advanced machine learning model designed for precise real estate price prediction. The OLSMBM model integrates the principles of Least Squares Support Vector Machine (LSSVM) to determine the optimal moment hyperplane, and employs Backpropagation Neural Networks (BPNN) to assign weights to each data point. This combination significantly enhances the predictive accuracy and robustness of the model, making it a valuable tool for urban planning and decision-making. The OLSMBM model demonstrated strong performance across the key metrics, achieving a Root Mean Squared Error (RMSE) of 6.977, Mean Absolute Error (MAE) of 4.752, Mean Absolute Percentage Error (MAPE) of 13.76%, correlation coefficient (R) of 0.846, and coefficient of determination ($R^2$) value of 0.728. The effectiveness of the model was further highlighted by its Reference Index (RI) score of 1.000, indicating superior predictive accuracy compared to other models. The OLSMBM model provides a robust decision-support tool for urban planners and policymakers, offering reliable predictions that can guide strategic land-use and investment decisions. By accurately forecasting real estate prices, the model supports data-driven approaches to urban planning, helps to optimize resource allocation, and promotes sustainable development. Although the OLSMBM model demonstrated strong performance, future research should address several areas for improvement. Expanding the input variables to include economic indicators and environmental conditions could enhance the accuracy. The application of the model to different regions tests its generalizability across various real estate markets. Additionally, comparing the OLSMBM with a broader range of machine learning models and incorporating climate risk data could further refine its predictive capabilities and effectiveness in urban planning.

## References

1. Y. Xu, R. Keivani, and A. J. Cao, Impact Assess. Proj. Apprais. **36**(4), 308–322 (2018)
2. S. Lehmann, Futur. Cities Environ. **2**, 8 (2016)
3. H. A. Sander and C. Zhao, Land use policy. **42**, 194–209 (2015)
4. M. Artmann, M. Kohler, G. Meinel, J. Gan, and I.-C. Ioja, Ecol. Indic. **96**(2), 10–22 (2019)
5. B. Tang, W. K. O. Ho, and S. W. Wong, Sustain. Dev. **29**(4), 708–718 (2021)
6. P. Jafary, D. Shojaei, A. Rajabifard, & T. Ngo, Habitat Int. **148**, 103075 (2024)
7. A. Aziz, M. M. Anwar, and M. Dawood, GeoJournal. **86**, 1915–1925 (2021)

8. L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, Neurocomputing. **237**, 350–361 (2017)
9. M. Y. Cheng and R. R. Khasani, Journal of Information Technology in Construction. **29**, 503–524 (2024)
10. M. Y. Cheng and R. R. Khasani, J. Comput. Civ. Eng. **38**(6), 1–15 (2024)
11. M. Y. Cheng and R. R. Khasani, Constr. Build. Mater. **441**, 137482 (2024)
12. I.-C. Yeh and T.-K. Hsu, Appl. Soft Comput. **65**, 260–271 (2018)
13. M.-Y. Cheng, N.-D. Hoang, L. Limanto, and Y.-W. Wu, Knowledge-Based Syst. **71**, 314–321 (2014)