Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 146
DOI: 10.11159/icsta25.146

Development of Machine Learning Regression Models for CO_2 Emission Forecasting in Thailand

Pianpool Kamoljitprapa, Piyachat Leelasilapasart
Faculty of Applied Science, King Mongkut's University of Technology North Bangkok
1518 Pracharat 1 Rd.,Wongsawang, Bangsue, Bangkok, Thailand
pianpool.k@sci.kmutnb.ac.th; piyachat.l@sci.kmutnb.ac.th

Abstract **-** Accurate forecasting of CO_2 emissions is vital for developing policies to address climate change and sustainability challenges. In Thailand, understanding emissions across key economic sectors is critical for mitigating the country's carbon footprint. This study evaluates several machine learning models, including Artificial Neural Networks (ANN), Gradient Boosting Machine (GBM), Multiple Linear Regression (MLR), Random Forest (RF), and Support Vector Machines (SVM), using sector-specific data from 2005 to 2024 provided by the Energy Policy and Planning Office. The results indicate that MLR outperformed other models, achieving the lowest *MAE*, *MSE*, and *RMSE*, as well as the highest R^2. While SVM and RF showed moderate performance, GBM and ANN exhibited higher prediction errors, with ANN being particularly unreliable due to extreme deviations. The MLR model was subsequently used to predict CO_2 emissions for 2024, and its predictions closely aligned with actual emissions, confirming its accuracy. These findings highlight MLR as a robust and interpretable model for CO_2 emission forecasting, offering a reliable alternative to more complex models.

Keywords**:** CO_2 emissions, Forecasting, Machine learning, Regression models, Thailand

1. Introduction

Climate change has become a critical global challenge, with carbon dioxide (CO_2) emissions being a significant contributing factor. As countries strive to mitigate their environmental impact, accurate forecasting of CO_2 emissions is essential for informed policy decisions and effective environmental management. In Thailand, rapid industrialization and economic growth have led to a steady rise in greenhouse gas emissions, highlighting the need for reliable predictive models to assess future trends and support sustainability efforts.

Several forecasting methods have been proposed for estimating CO_2 emissions, including traditional time series approaches like the Autoregressive Integrated Moving Average (ARIMA) and the Box-Jenkins methodology. [1-5]. These statistical approaches are widely applied in time series forecasting because of their ability to effectively handling linear data structures. However, they face limitations when addressing complex, curvilinear relationships and high-dimensional structures, which are common in CO_2 emission patterns [6-8]. Moreover, these models require strict assumptions regarding stationarity and struggle to adapt to structural changes in emissions data [9].

In contrast, machine learning (ML) techniques offer more flexibility in capturing intricate patterns and nonlinear dependencies in environmental datasets. Studies have demonstrated that ML-based models, for instance artificial neural networks (ANN), support vector machines (SVM), and ensemble learning methods, outperform traditional statistical models in CO_2 emission forecasting by effectively modelling complex relationships and handling large datasets. [4, 10-11]

This research aims to develop and evaluate machine learning regression models for CO_2 emission forecasting in Thailand, focusing on data from major economic sectors such as power generation, transportation, industrial, and other sectors. In 2023, the power generation sector in Thailand released approximately 89.6 million metric tons of CO_2, while the transportation sector emitted around 79 million metric tons, accounting for 29% of the country's total emissions. In 2016, energy consumption in the industrial sector accounted for about 18% of Thailand's CO_2 emissions, totalling around 80 million metric tons. [12-13]. By utilizing sector-specific data, this study seeks to enhance forecasting accuracy and improve environmental planning and policy formulation. A comparative analysis of different ML algorithms will be conducted to determine the most appropriate approach for CO_2 emission forecasting in Thailand.

Through this research, we aim to contribute to ongoing climate change mitigation efforts by providing a robust, data-driven approach to CO_2 emission forecasting. Leveraging machine learning techniques, this study aspires to improve prediction accuracy and support the development of more effective and sustainable environmental policies.

2. Materials and Methods

2.1. Data Collection and Preprocessing

This study employs sector-specific CO_2 emissions data from Thailand's key economic sectors, including power generation, transportation, industry, and others. The dataset, obtained from the Energy Policy and Planning Office, Ministry of Energy [14], covers the period from January 2005 to December 2024, providing a comprehensive historical record for analysing emission trends and improving forecasting accuracy.

To develop a CO_2 emission forecasting model, machine learning methodologies−including artificial neural networks (ANN), gradient boosting machine (GBM), multiple linear regression (MLR), random forest (RF) and support vector machines (SVM)−are employed. The analysis is conducted using R software, and the dataset, covering the period from January 2005 to December 2023, is divided into two subsets: a training set (70%) consisting of data from 2005 to 2020 for model development and a test set (30%) containing data from 2021 to 2023 for performance evaluation. The trained models are validated using multiples performance metrics, including Mean Absolute Error (*MAE*), Mean Squared Error (*MSE*), Root Mean Squared Error (*RMSE*), and R-squared (R^2).

To assess the reliability and effectiveness of the forecasting models, the best-performing model is applied to predict CO_2 emissions for 2024, and the results are compared against actual emission data from 2024 to verify its predictive accuracy and practical applicability.

2.2. Machine Learning Methods

2.2.1. Artificial Neural Networks (ANN)

Artificial neural networks are computing models consisting of multiple layers of connected neurons, inspired by the structure of the human brain. The network learns through training to map input features to a desired output. The model is defined as shown in Eq. (1):

$$y = f\sum_{i=1}^{n} w_i x_i + b, \tag{1}$$

where y is the output (CO_2 emission prediction).

f is the activation function.

x_i are the input features (economic sectors).

w_i are the weights of the connections.

b is the bias term.

2.2.2. Gradient Boosting Machine (GBM)

Gradient boosting machine is an ensemble method that constructs decision trees in sequence, with each new tree learning from and correcting the errors of the previous one. The model is represented by Eq. (2):

$$f(x) = \sum_{m=1}^{M} \alpha_m h_m(x), \tag{2}$$

where $f(x)$ is the predicted CO_2 emission.

$h_m(x)$ is the m-th decision tree.

α_m is the weight of the m-th tree.

M is the total number of trees in the ensemble.

2.2.3. Multiple Linear Regression (MLR)

Multiple linear regression is a statistical method for examining the association between a dependent variable (CO_2 emissions), and multiple independent variables. (economic sectors: power generation, transportation, industry and other sectors). The model is expressed as shown in Eq. (3):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon, \tag{3}$$

where Y is the predicted CO_2 emission.

$X_1, X_2, \ldots, X_n$ are the independent variables.

$\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients.

ε is the error term.

2.2.4. Random Forest (RF)

Random forest utilizes an ensemble learning technique that constructs a set of decision trees, each trained on a randomly selected subset of data. For regression tasks, the last prediction is obtained by averaging the outputs of all individual trees. The model is represented in Eq. (4):

$$f(x) = \frac{1}{M} \sum_{t=1}^{M} h_t(x), \tag{4}$$

where $f(x)$ is the predicted CO_2 emission.

$h_t(x)$ is the prediction of the t-th decision tree.

M is the total number of trees in the random forest.

The summation $\sum_{t=1}^{M} h_t(x)$ aggregates predictions from all trees, and the division by M ensures the final output is the average of individual tree predictions.

2.2.5. Support Vector Machines (SVM)

Support vector machines (SVM) are supervised learning algorithms capable of addressing both classification and regression problems by identifying the optimal hyperplane to separate data. When applied to regression, support vector regression (SVR) is used to determine a function that best captures the relationship between predictor variables and the target variable, with minimizing errors within a specified margin. The regression model is shown in Eq. (5):

$$f(x) = \langle w, x \rangle + b, \tag{5}$$

where $f(x)$ is the predicted CO_2 emission.

$\langle w, x \rangle$ is the inner product between the weight w vector and the input feature vector x.

b is the bias term.

To determine the optimal weight vector w and bias b, the following objective function is minimized as shown in Eq. (6):

$$\mathrm{w} = \frac{1}{2} \| w \|^2, \tag{6}$$

subject to the ε-insensitive loss function constraints:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon. \tag{7}$$

The goal is to ensure that most predictions lie within an ε-tube, while minimizing the model complexity $\| w \|^2$ to prevent overfitting. In real-world cases where some points may fall outside the ϵ-tube, slack variables ξ and ξ^* are introduced to handle these deviations:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i, \tag{8}$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*. \tag{9}$$

$$\xi_i, \xi_i^* \geq 0.$$

The modified optimization objective becomes

$$\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*). \tag{10}$$

Here, C represents a regularization parameter that balances model complexity and the allowance for deviations.

y_i is the actual CO_2 emission value for the i-th data point.

ϵ is a margin within which predictions are considered acceptable, meaning small deviations from actual values are ignored.

2.3. Machine Learning Framework

The proposed framework for forecasting CO_2 emissions in Thailand follows a structured sequence of steps, ensuring systematic data processing and model development [15-16].

2.3.1. Data collection:

The initial step involves gathering a comprehensive dataset that includes CO_2 emissions records alongside key contributing factors. These variables provide the necessary inputs for identifying trends and patterns, enabling the model to make informed predictions.

2.3.2. Data processing:

Before analysis, the dataset undergoes preprocessing to improve data integrity and consistency. This includes detecting and addressing missing or erroneous values while standardizing input features to maintain a uniform scale. These refinements help optimize the performance of machine learning models.

2.3.3. Dataset splitting:

For effective model training and evaluation, the dataset is divided into two subsets: a training set (70%) covering data from 2005 to 2020, used for model learning, and a test set (30%) spanning 2021 to 2023, reserved for performance assessment. Cross-validation is implemented during this phrase to improve the model's performance to generalize across different data distribution.

2.3.4. Model training:

During the training phrase, machine learning algorithms are applied to the training dataset. This step involves selecting the most suitable predictive models, fine-tuning hyperparameters, and optimizing key components such as kernel functions, weight assignments, and bias terms, cross-validation assures that the models are well-fitted without overfitting the training data.

2.3.5. Model testing:

Once trained, the models are tested using the reserved test dataset. The model generates CO_2 emission predictions based on input factors, and these predictions are subsequently compared against actual recorded values to evaluate their accuracy and reliability.

2.3.6. Performance evaluation:

In the final step, the predictive performance of each model is evaluated using various assessment metrics, such as mean absolute error (*MAE*), mean squared error (*MSE*), root mean squared error (*RMSE*), and R-squared (R^2). These metrics offer valuable insights into the accuracy and efficiency of the models, aiding in the selection of the most suitable approach for CO_2 emission forecasting.

2.4. Model Evaluation Criteria

To assess the predictive performance of machine learning models for CO_2 emission forecasting, a range of statistical metrics are applied. These comprise mean absolute error (*MAE*), mean squared error (*MSE*), root mean squared error (*RMSE*), and R-squared (R^2). Each of these metrics provides important insights into the precision and dependability of the predictions [17].

Mean absolute error (*MAE*) estimates the mean error between predicted and actual values, as computed in Eq. (11).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{11}$$

where y_i represents the actual CO_2 emissions.

$\hat{y}_i$ is the predicted value.

n is the total number of observations.

Mean squared error (*MSE*) evaluates the average squared deviations between predicted and actual values, giving more weight to larger errors due to the squaring process, as represented in Eq. (12):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 . \tag{12}$$

A lower *MAE* and lower *MSE* indicate better model performance, as it signifies lower deviation between predictions and actual values.

Root mean squared error (*RMSE*) is derived from the square root of *MSE*, offering a comprehensible measure of prediction error, expressed in the same units as the target variable, as shown in Eq. (13):

$$RMSE = \sqrt{MSE}\ . \tag{13}$$

A lower *RMSE* indicates a better-fitting model. It is especially useful when errors of greater magnitude need to be considered, as it penalizes larger deviations more than *MAE*.

R-squared (R^2) Coefficient, or coefficient of determination, represents the proportion of variance in the dependent variable that is captured by the model, as shown in Eq. (14).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{14}$$

where $\bar{y}$ is the mean of actual CO_2 emissions.

An R^2 value closer to 1 indicates a strong relationship between the predicted and actual values, whereas a value near 0 suggests poor predictive performance.

3. Results

The performance of five machine learning models - artificial neural networks (ANN), gradient boosting machine (GBM), multiple linear regression (MLR), random forest (RF) and support vector machines (SVM) - was evaluated based on four key metrics: mean absolute error (*MAE*), mean squared error (*MSE*), root mean squared error (*RMSE*), and R-squared (R^2). The results are visualized in Table 1 and Fig. 1.

Table 1: Model performance.

Methods	***MAE***	***MSE***	***RMSE***	R^2
ANN	1180.28	2126834.83	1458.37	N/A
GBM	556.12	516011.70	718.34	0.4833
MLR	**132.51**	**27634.04**	**166.23**	**0.9699**
RF	427.59	270636.77	520.23	0.8594
SVM	388.81	214745.20	463.41	0.8283

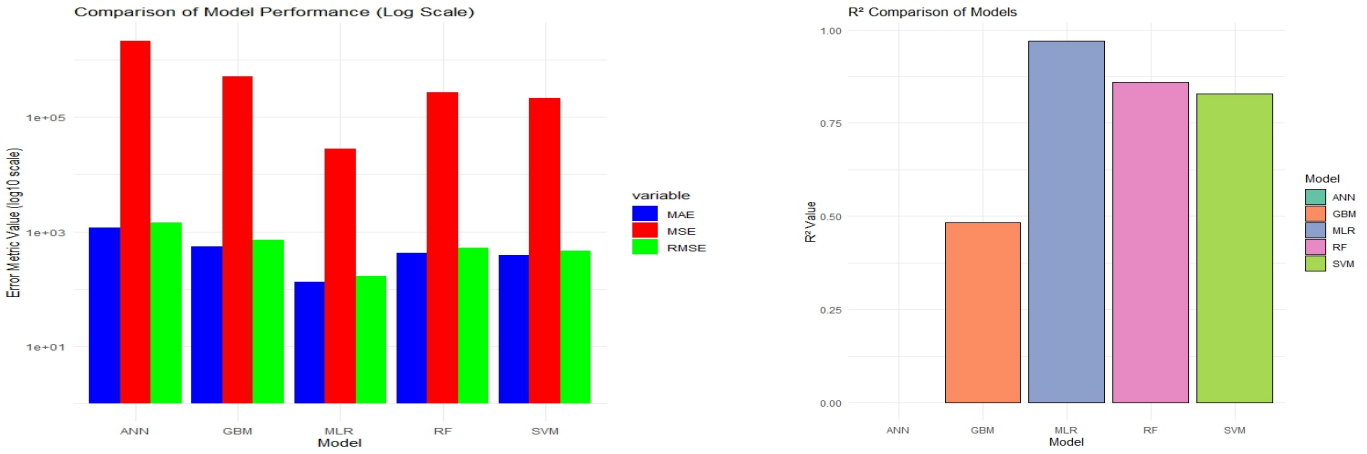

Fig. 1: Comparison of the model's performance.

As shown in Table 1 and Fig. 1, MLR achieved the lowest *MAE* (132.51), *MSE* (27634.04) and *RMSE* (166.23), demonstrating the highest predictive accuracy among all models. Additionally, MLR had the highest R^2 (0.97), indicating strong alignment with actual CO_2 emissions. SVM and RF exhibited moderate performance, with R^2 of 0.83 and 0.86, respectively. GBM and ANN, however, showed significantly higher prediction errors. ANN had the highest *MAE* (1180.28) and *RMSE* (1458.37), and its R^2 was N/A, likely due to extreme prediction deviations, rendering the model unreliable for this task. GBM also performed poorly, with an R^2 of 0.48, suggesting limited predictive capability.

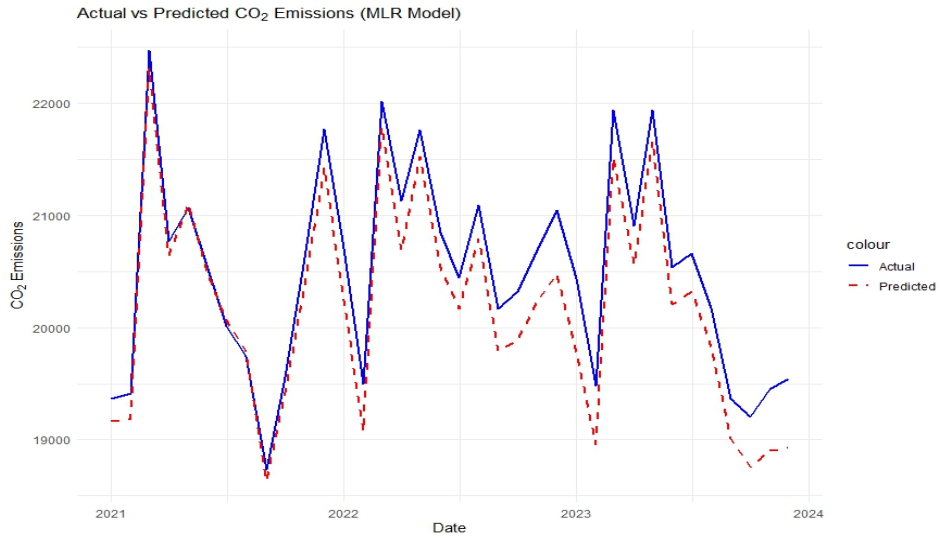

Fig. 2: Comparison of the actual data and prediction using MLR model.

The MLR model, being the best-performing approach, was applied to predict CO_2 emissions for 2024. Fig. 2 illustrates the comparison between actual and predicted values, demonstrating a close fit and further confirming MLR's reliability for forecasting emissions in Thailand. These findings emphasize that while complex models may capture intricate patterns, simpler and more interpretable models like MLR can still provide robust and accurate predictions.

4. Conclusion

This study evaluated machine learning models for CO_2 emission forecasting in Thailand using real economic sectors specific data. MLR achieved the highest accuracy with the lowest *MAE* (132.51), *MSE* (27634.04) and *RMSE* (166.23), while GBM and ANN exhibited higher errors, indicating potential overfitting. SVM and RF performed moderately but did not surpass MLR.

The R^2 for ANN was N/A, likely due to extreme prediction errors, making it unreliable. These findings suggest that while advanced models capture complex patterns, MLR remains effective and interpretable. Future work should refine model tuning and explore hybrid approaches for enhanced predictive performance.

Acknowledgements

We would like to express our sincere gratitude to the Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology (KMUTNB) for their generous sponsorship and support in registering this presentation. Their assistance has been invaluable in making this work possible.

References

[1] P. Kamoljitprapa and S. Sookkhee, "Forecasting models for carbon dioxide emissions in major economic sectors of Thailand," in *Journal of Physics: Conference Series*, vol. 2346, 2022, 012001, doi: 10.1088/1742-6596/2346/1/012001.

[2] O. Polsen and P. Kamoljitprapa, "Time series models for reservoir inflows in Thailand," in *Proceedings of the 2024 9th International Conference on Information and Education Innovations (ICIEI)*, Verbania, Italy, 2024, pp. 93-98, doi: 10.1145/3664934.3664938.

[3] O. Polsen and P. Kamoljitprapa, "Time series analysis of particulate matter in Bangkok, Thailand," in *Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Tenerife, Canary Islands, Spain, 2023, pp. 1-6, doi: 10.1109/ICECCME57830.2023.10253226.
[4] T. Alam and A. AlArjani, "A comparative study of CO_2 emission forecasting in the Gulf countries using autoregressive integrated moving average, artificial neural network, and Holt-Winters exponential smoothing models," *Advances in Meteorology*, [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/8322590
[5] P. Kamoljitprapa, O. Polsen and U. K. Abdullahi, "Forecasting of Thai international imports and exports using Holt-Winters' and autoregressive integrated moving average models," *Journal of Applied Science and Emerging Technology (JASET)*, vol. 22, no. 3, e252955, 2023, doi: 10.14416/JASET.KMUTNB.2023.03.002.
[6] P. Kamoljitprapa, F. M. Baksh, A. De Gaetano, O. Polsen and P. Leelasilapasart. "Statistical study design for analyzing multiple gene loci correlation in DNA sequences," *Mathematics,* vol. 11, no. 23, 4710, 2023, https://doi.org/10.3390/math11234710.
[7] P. Kamoljitprapa and P. Leelasilapasart, "Nonlinear models for influenza patients for different age groups in Thailand," in *Proceedings of the 2024 9th International Conference on Information and Education Innovations (ICIEI)*, Verbania, Italy, 2024, pp. 109-112, doi: 10.1145/3664934.3664941.
[8] P. Kirdwichai, "Estimation and use of correlation in multiple hypothesis testing with high dimensional data," in *Proceedings of the 2nd International Conference on Mathematics and Statistics,* Prague, Czech Republic, 2019, pp. 36-39, doi: 10.1145/3343485.3343489.
[9] P. Linardatos, V. Papastefanopoulos, T. Panagiotakopoulos and S. Kotsiantis, "CO_2 concentration forecasting in smart cities using a hybrid ARIMA–TFT model on multivariate time series IoT data," *Sci Rep.*, vol. 13, 17266, 2023, doi: 10.1038/s41598-023-42346-0.
[10] Y. Liu, H. Chen, L. Zhang, X. Wu and X. Wang, "Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China," *Journal of Cleaner Production*, vol. 272, 122542, 2020, doi: 10.1016/j.jclepro.2020.122542.
[11] V. S. Pendyala and S. Podali, "An Overview of Carbon Footprint Mitigation Strategies. Machine Learning for Societal Improvement, Modernization, and Progress," *Machine Learning for Societal Improvement, Modernization, and Progress*, pp.1-26 2022, doi: 10.4018/978-1-6684-4045-2.ch001.
[12] A. Walderich, "Carbon dioxide emissions from energy consumption in Thailand in 2023, by sector," *Statista*, 2024, [Online]. Available: https://www.statista.com/statistics/1296621/thailand-co2-emissions-from-energy-consumption-by-sector/.
[13] Asian Transport Outlook (2024). Transport and Climate Profile: Thailand [Online]. Available: https://asiantransportobservatory.org/analytical-outputs/countryprofiles/.
[14] CO_2 Statistic. (2025, February 28). Energy Policy and Planning Office Ministry of Energy [Online]. Available: http://www.eppo.go.th/index.php/en/.
[15] S. Uddin, S. Ong and H. Lu, "Machine learning in project analytics: a data-driven framework and case study," *Sci Rep*, vol. 12, 15252, 2022, doi: 10.1038/s41598-022-19728-x.
[16] P. Kadam and S. Vijayumar, "Prediction Model: CO_2 emission using machine learning," in *Proceedings of 2018 3rd International Conference for Convergence in Technology (12CT)*, Pune, India, 2018, pp. 1-3, doi: 10.1109/I2CT.2018.8529498.
[17] A. K. Sen, P. Kayal and M. Maiti, "Machine learning approaches for modelling water futures," *Development and Sustainability in Economics and Finance*, vol. 2-4, 100029, 2024, doi: 10.1016/j.dsef.2024.100029.