航測及遙測學刊　第三十一卷　第 3 期　第 161 – 171 頁　民國 115 年 9 月
Journal of Photogrammetry and Remote Sensing
Volume 31, No.3, 2026, pp. 161 – 171
DOI：10.6574/JPRS.202609_31(3).0002

Hydrocarbon Estimation in Northern Taiwan Using the LightGBM Regression Model

Thia Prahesti [1]　Aji Kusumaning Asri [2]　Yu-Ting Zeng [3]　Chih-Da Wu [4,5,6,7,8*]

Abstract

Total Hydrocarbon (THC) is a critical air pollutant composed of various hydrocarbon compounds, commonly emitted from vehicle exhaust, industrial activities, and fuel combustion. Accurate estimation of THC concentrations is essential for understanding pollution patterns and supporting environmental management. This study applies a Light Gradient Boosting Machine (LightGBM) model to estimate the spatiotemporal variation of THC levels in a part of the Northern Air Quality Zone (AQZ) from June 25, 2015, to December 31, 2020, with 2021 used for testing with unseen data. LightGBM, a machine learning algorithm known for its high accuracy and computational efficiency, achieved strong predictive performance with R^2 values of 0.84 for training, 0.82 for validation, and 0.77 for testing. The average THC concentration in the study area was 2.07 ppm. SHAP (SHapley Additive exPlanations) analysis revealed that nitrogen dioxide (NO_2) and nitrogen oxides (NOx) were the most influential predictors, highlighting their close association with combustion-related emissions. Other important variables included population density, land use types such as plantation forests and paddy fields, and wind speed. Spatially, THC concentrations were higher in the western part of the study area, which is more urbanized and industrialized, compared to the more rural eastern region. These results demonstrate the potential of LightGBM to model complex air pollution patterns and support data-driven air quality strategies.

Keywords: Light Gradient Boosting Machine Regression (LightGBM), Spatial Prediction Model, Total Hydrocarbon (THC)

1. Introduction

Total hydrocarbon (THC) concentration refers to the measurement of a group of organic compounds that play a significant role in air pollution and have been linked to adverse health effects (United States Environmental Protection Agency, 2005). These organic compounds primarily originate from sources such as vehicular emissions, industrial activities, and the use of solvents (Bishop & Stedman, 2008; Environment Canada, 2003). Monitoring and predicting THC levels is crucial for effective air quality management, particularly in urban areas where pollution levels are often higher. Studies have shown that the spatiotemporal variation of THC is influenced by various factors, including land use, meteorological conditions, and traffic patterns (Ambrose *et al.*, 2026). Therefore, understanding the distribution and sources of THC is essential for developing strategies to mitigate their impact on public health and the environment.

Light Gradient Boosting Machine (LightGBM) is a fast and accurate tree-based machine learning method that is well-suited for handling large datasets and complex variable interactions (Guo *et al.*, 2023a; Li *et al.*, 2024). For example, Lu *et al.* (2025) applied a space-time LightGBM model to estimate volatile organic compounds (VOCs) in Shanghai, producing high-resolution hourly prediction maps (1 km, next 24 h) with strong performance, achieving R^2 values ranging from 0.527 to 0.938 and MAE values between 41.5 and 126 ppb. Similarly, Thongthammachart *et al.* (2022) integrated LightGBM into a land use regression framework for NO_2 and $PM_{2.5}$ estimation, showing that LightGBM outperformed Random Forest and Extreme Gradient Boosting, particularly in capturing high concentration levels that are critical for health risk assessment.

While previous studies have applied machine learning models such as LightGBM to estimate air

[1] Ph.D Student, Department of Geomatics, National Cheng Kung University
[2] Postdoctoral Research Fellow, Department of Geomatics, National Cheng Kung University
[3] Assistant Researcher, Department of Geomatics, National Cheng Kung University
[4] Professor, Department of Geomatics, National Cheng Kung University
[5] Adjunct Research, National Institute of Environmental Health Sciences, National Health Research Institutes
[6] Adjunct Professor, Innovation and Development Center of Sustainable Agriculture, National Chung Hsing University
[7] Adjunct Research, Research Center for Precision Environmental Medicine, Kaohsiung Medical University
[8] Joint Appointment Research, Chronic Diseases and Health Promotion Research Center, Chang Gung University of Science and Technology
[*] Corresponding Author, E-mail: chidawu@mail.ncku.edu.tw
Received Date: Jul. 16, 2025
Revised Date: Sep. 17, 2025
Accepted Date: Aug. 21, 2026

pollutants (Lu *et al.*, 2025; Thongthammachart *et al.*, 2022), few have explored their application to total hydrocarbon (THC) concentrations, and the spatiotemporal variations of THC, especially in Northern Taiwan, remain underexplored. This study addresses these gaps by applying a LightGBM regression model to estimate THC concentrations using data from multiple monitoring stations, considering that THC primarily originates from sources such as vehicular emissions, industrial activities, and the use of solvents. The study also provides insights for air quality management, highlighting the novelty and significance of this research in advancing previous work on air pollution modeling (Lu *et al.*, 2025; Thongthammachart *et al.*, 2022).

2. Method

This research separates the workflow into three parts, the dataset preprocessing of feature points' density map, the comprehensive deep Q network model for path finding, and the post-processing for smoothing the path, shown as Figure 1 below.

2.1 Study area

Taiwan, historically known as Formosa, is located north of the Philippines, east of China, and south of Japan. For air quality management, the Ministry of Environment (MOENV) divides the island into seven Air Quality Zones (AQZs), one of which is the Northern Area that includes Taipei City, New Taipei City, Keelung, and Taoyuan (Risk Society and Policy Research Center, 2026).

The study area covers approximately 3,678 km² in northern Taiwan, a region with some of the highest population densities in the country, particularly in Taipei. Nevertheless, motorcycles and private automobiles remain the dominant travel modes and are major sources of THC, thereby intensifying regional air pollution (Bishop & Stedman, 2008).

2.2 Air pollution database

Hourly data on THC concentrations have been recorded since June 26, 2015, at 8 monitoring stations in Northern AQZ managed by MOENV (Figure 1). This study utilizes data spanning from 2015 to 2018, where 80% of the dataset is designated for model training and the remaining 20% is used for validation. Additionally, data from monitoring stations in 2019 were collected for testing to evaluate the model's reliability. The Ministry of Environment Taiwan adheres to standards and uses instruments comparable to those of the United States regulation (Department of Monitoring and Information, Ministry of Environment, Taiwan. (2024). To investigate potential correlations between THC and its co-pollutants, such as PM_{10}, $PM_{2.5}$, O_3, SO_2, NO_2, and NO_x, we also collected measurements for these pollutants from the Ministry of Environment's published database.

2.3 Land-use and geospatial databases

To support the development of the spatial prediction model, a wide array of environmental and geospatial explanatory variables were compiled from meteorological, land-use, topographic, and anthropogenic datasets (Hsu *et al.*, 2022). Meteorological variables including wind speed, wind direction, relative humidity, temperature, ultraviolet (UV) radiation, and precipitation were obtained from the Central Weather Bureau of Taiwan, as these factors significantly influence pollutant dispersion, chemical transformation, and atmospheric behavior of THC.

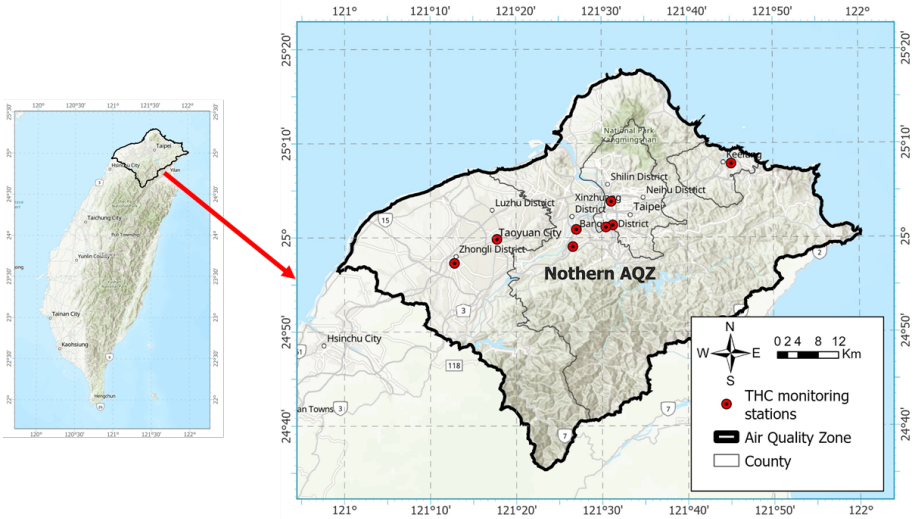

Figure 1 Distribution of Ministry of Environment monitoring stations measuring THC concentrations in Northern Area

Land-use and geospatial data were sourced from the National Land Surveying and Mapping Center (NLSMC), covering the years 2015–2018. These data included classifications of land types such as farmlands, forests, residential and industrial zones, airports, and water bodies. Vegetation characteristics were captured using the MODIS Normalized Difference Vegetation Index (NDVI), with a spatial resolution of 250 m × 250 m. Topographic variables, such as elevation and slope, were derived from a 20 m × 20 m Digital Terrain Model (DTM), which helped account for the influence of terrain on air movement and pollutant accumulation.

To characterize human activities and potential THC emission sources, we collected data from the Kingwaytek landmark database as a Point of Interest (POI), the Ministry of Finance's business tax registry, and the Ministry of Economic Affairs' industrial site records. These anthropogenic activities, including transportation, industrial operations, and fossil fuel combustion (Environment Canada, 2003; Stewart *et al.*, 2021), enable the identification of localized THC emission hotspots for modeling their impact on air quality. Social factors, including the locations of major Chinese temples and restaurants (Wu *et al.*, 2017) and annual township-level population data (2015–2021), were incorporated as predictive variables to capture anthropogenic and cultural influences on THC concentrations. Seasonal variations were also considered by deriving a seasonal variable from the date of each observation, allowing the model to capture changes in pollutant concentrations across different times of the year. Spatial predictor variables were then calculated within multiple circular buffer zones centered at each air quality monitoring station. Specifically, buffer radii included: 50 m, 150 m, 250 m, 500 m, 750 m, 1000 m, 1250 m, 1500 m, 1750 m, 2000 m, 2500 m, 3000 m, 4000 m, and 5000 m. These varying distances allowed for multi-scale analysis of land-use influence, enabling the model to capture both localized and regional factors affecting THC concentrations in Northern AQZ.

2.4 Study framework

To estimate THC concentrations, this study employed a spatiotemporal modeling approach using LightGBM regression as the primary machine learning algorithm. As illustrated in Figure 2, the workflow comprised four main steps: data preparation, model development, model validation, and spatiotemporal prediction of THC concentrations in the Northern Area of Taiwan.

2.5 The proposed LightGBM model approach

LightGBM is a powerful boosting technique tailored for handling large-scale datasets with high predictive performance. It incorporates strategies such as Gradient-based One-Side Sampling (GOSS), which accelerates the training process by focusing on instances with higher prediction errors, and Exclusive Feature Bundling (EFB), which simplifies high-dimensional data by combining features that do not overlap. Unlike conventional boosting methods, LightGBM adopts a leaf-wise tree growth approach, enabling more refined splits that improve prediction accuracy and reduce the risk of overfitting (Guo *et al.*, 2023b; Li *et al.*, 2024).

To enhance model performance while reducing the risk of overfitting, hyperparameter tuning was performed on key parameters, including learning rate, number of leaves, maximum tree depth, feature fraction, and bagging fraction (Tırınk, 2025; Vertes, 2020). The model's generalization ability was evaluated using an 80/20 train–test split and k-fold cross-validation (GeeksforGeeks, 2025). Additionally, regularization and controls on tree complexity, such as limiting tree depth and the number of leaves, were applied to prevent overly complex models (Bahmani, 2020).

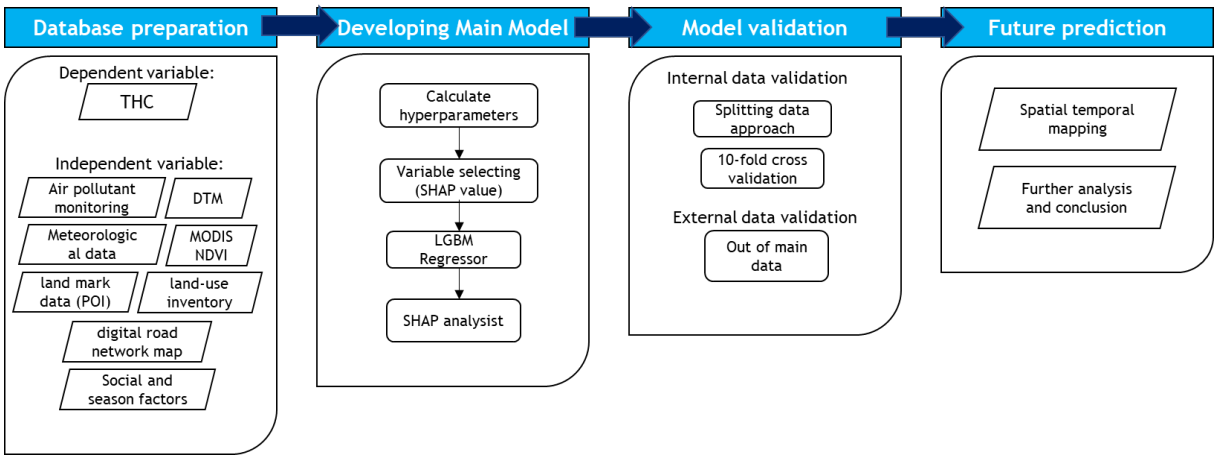

Figure 2 Study framework

This study developed prediction models to estimate monthly THC concentrations in Taiwan by machine learning algorithms. We applied machine learning algorithm (LightGBM) to enhance model performance, given their ability to capture non-linear relationships between THC levels and predictor variables. The process begins with hyperparameter tuning to identify the most effective combination of model parameters. Next, feature selection is performed to eliminate unimportant or redundant variables, ensuring the model focuses only on the most relevant inputs. With the optimal parameters and selected features, the model is then trained to achieve the best performance. Finally, SHAP (SHapley Additive exPlanations) analysis is used to interpret the model by showing how each feature contributes to the predictions. To simplify the hyperparameter tuning process, we focused on algorithms with automated tuning features. Model performance was evaluated by calculating the R^2 values from the 80/20 data split, cross-validation, and external validation. Each algorithm was tuned to optimize predictive accuracy, and the best-performing model was chosen to generate final predictions.

2.6 Model validation

We employed three categories of validation methods, encompassing three distinct approaches, to assess model reliability. These included an 80%/20% data split, 10-fold cross-validation and external validation (using 2019 data). In the 10-fold cross-validation procedure, the dataset was partitioned into ten equal subsets, with each subset used once as a validation set while the remaining nine subsets served as the training data. We recalculated R^2 and other metrics to confirm model stability. To evaluate model accuracy, we used correlation coefficient (r), R^2, and adjusted R^2 values, alongside mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE) to measure prediction accuracy, defined as follows (1-6). These metrics were used complementarily to provide a comprehensive evaluation of model performance, and no single metric was considered the sole indicator of model accuracy.

$$\boldsymbol{r} = \frac{\left(\sum(y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})\right)}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \hat{\bar{y}})^2}} \quad \text{.....(1)}$$

$$R^2 = 1 - \left(\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}\right) \text{.....(2)}$$

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \quad \text{.....(3)}$$

$$MSE = \left(\frac{1}{n}\right)\sum(y_i - \hat{y}_i)^2 \quad \text{.....(4)}$$

$$MAE = \left(\frac{1}{n}\right)\sum|y_i - \hat{y}_i| \quad \text{.....(5)}$$

$$RMSE = \sum\left(\left(\frac{1}{n}\right)\sum(y_i - \hat{y}_i)^2\right) \quad \text{.....(6)}$$

Where y_i is the observed THC concentration, $\hat{y}_i$ is the predicted THC concentration, $\bar{y}$ is the mean of the observed THC concentrations, $\hat{y}$ is the mean of the predicted THC concentrations, n denotes the number of observations, p denotes the number of predictor variables, and i represents the index of each observation.

2.7 Geographical mapping of the spatiotemporal variations of THC

Using the model with the best explanatory power, we generated spatiotemporal maps of THC concentrations at a 50 m × 50 m grid resolution. First, we converted the selected geographic predictors from raster format to ASCII format. Then, the ensemble model was applied to estimate predictions for each 50 m grid center using the Python platform. In the final step, the ASCII-format predictions were converted back into raster format to create the THC concentrations prediction map. For spatial data analysis, we utilized ArcGIS Pro, statistical analyses were performed in R 4.3.2, and machine learning models were developed using Python.

3. Results

3.1 Descriptive statistics of THC concentrations

Figure 3(a) presents the histogram of THC concentrations measured across all stations from 2015 to 2018, showing the overall distribution of observed values over the study period. Figure 3(b) illustrates the average THC concentration at each monitoring station, highlighting spatial differences in pollutant levels across the study area. As a result, the overall THC concentration in the dataset (n = 8,369) exhibited a mean of 2.10 ppm and a median of 2.07 ppm, indicating a relatively symmetric distribution. The concentrations ranged from a minimum of 1.6 ppm to a maximum of 4.23 ppm, with a standard deviation of 0.19 ppm, reflecting moderate variability across the study period.

At the station level, the average THC concentrations varied, with Keelung showing the lowest mean of 1.97 ppm and Zhongli the highest at 2.27 ppm. Other stations recorded intermediate values, including 2.05 ppm at Guting, 2.07 ppm at Tucheng, 2.09 ppm at Zhongshan, 2.10 ppm at Taoyuan, 2.12 ppm at Yonghe, and 2.15 ppm at Banqiao, highlighting spatial heterogeneity in THC levels across the study area.

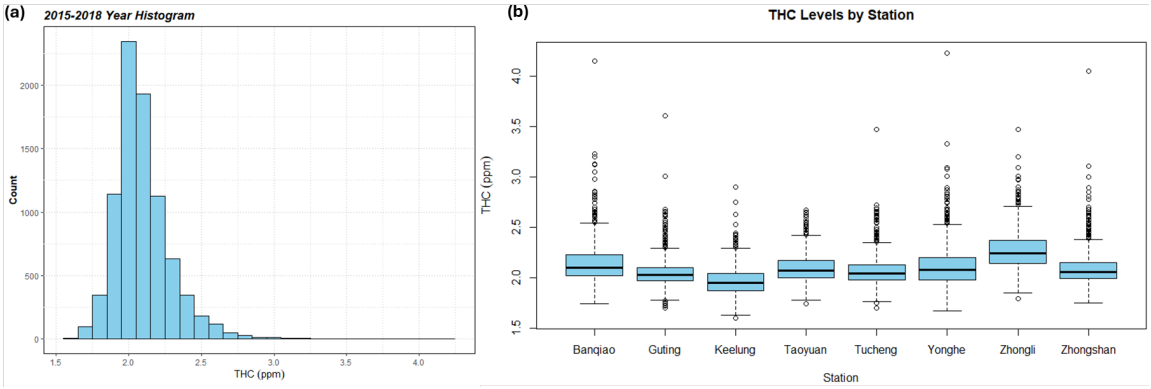

Figure 3 (a) histogram of THC concentrations from 2015 to 2018, (b) average concentration at each station

3.2 Explanatory variables of the LightGBM model and model performance

Figure 4 presents the statistical values, including R^2 for LightGBM. The model showed strong predictive performance across different evaluation stages. During training, the model achieved an R^2 of 0.84, indicating a good fit to the training data. In validation, the R^2 was slightly lower at 0.82, demonstrating the model's ability to generalize well. External testing with unseen data resulted in an R^2 of 0.77, confirming reasonable predictive accuracy on independent datasets. Overall, the combined total R^2 was 0.84, reflecting consistent performance. Additionally, cross-validation yielded an average R^2 of 0.80, supporting the model's robustness and stability.

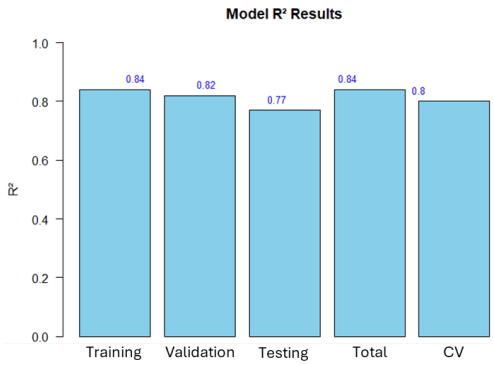

Figure 4 Coefficient of determination (R^2) of the LightGBM model for the training, validation, testing, total, and cross-validation datasets

Figure 5 presents the SHAP analysis from the LightGBM spatiotemporal model for THC mapping. Panel (A) shows the mean absolute SHAP values, representing the overall importance of each variable in the model, while Panel (B) shows the impact of each variable on the LightGBM model for THC mapping based on individual SHAP values. In Panel (B), positive SHAP values indicate that a variable contributes to higher predicted THC concentrations, whereas negative SHAP values indicate a contribution toward lower predicted THC concentrations. The color gradient represents the feature values, with blue indicating lower values and red indicating higher values.

The mean absolute SHAP values are used to rank the variables according to their relative contribution to the model predictions. NO_2 (ppm) shows the highest importance, with a mean absolute SHAP value of 0.055, indicating its strong effect on model predictions. This is followed by NOx (ppm) at 0.041, and Population (1750m) at 0.032, highlighting the relevance of air pollutants and population density in estimating THC levels. Plantation forest (4000m) and Wind Speed (m/s) also contribute significantly, reflecting the role of land use and meteorological conditions. Other important variables include Paddy fields (2000m), Plantation forest (2500m), and Park (1750m), underscoring the spatial influence of surrounding land use types. Overall, these findings demonstrate that both environmental pollutants and land characteristics are key drivers in the model's predictions.

3.3 Model validation

Table 1 show the validation of the LightGBM model includes eleven different validation processes based on several statistical values. These include the coefficient (r), R^2, and adjusted R^2 values, as well as MSE, MAE, and RMSE to measure prediction accuracy from 80%/20% data split (training and validation), 10-fold cross-validation and external validation. The model performs well across different stages. In training, it explains 84% of the variation in the data (R^2 = 0.84) with low errors (RMSE = 0.10, MAE = 0.06). Validation results are similar, showing the model generalizes well (R^2 = 0.82). Testing with unseen data is slightly lower (R^2 = 0.77), but still indicates good accuracy. Overall, the combined results show consistent performance (R^2 around 0.83-0.84). Cross-validation confirms the model's stability with an average R^2 of 0.80 and only a small increase in error (RMSE = 0.11). This means the model is reliable and effective for predicting the target variable.

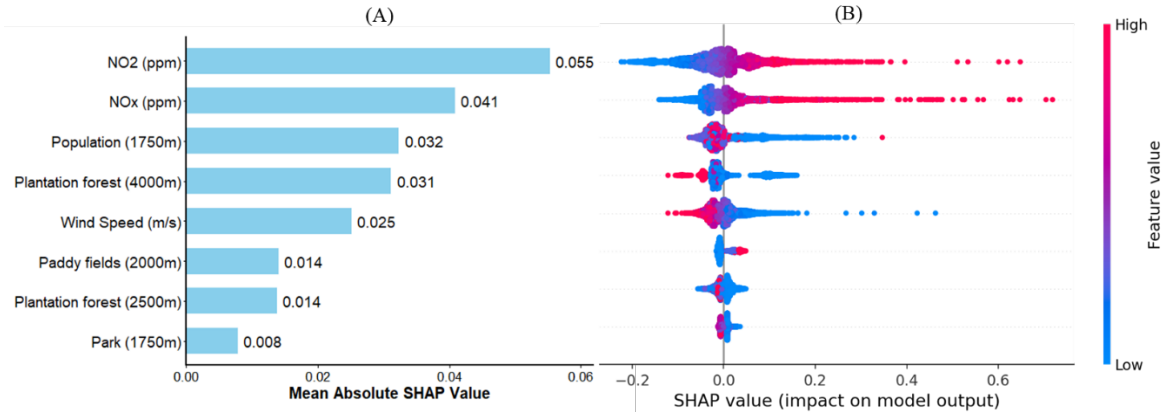

Figure 5 Impact of each variable on the LightGBM model for THC mapping : (a) mean absolute SHAP values, and (b) SHAP summary plot showing the direction and magnitude of each variable's contribution

Table 1 The validation of the LightGBM model

Result	R^2	adj R^2	RMSE	MSE	MAE
Training	0.84	0.84	0.10	0.01	0.06
Validation	0.82	0.82	0.10	0.01	0.06
Testing	0.77	0.77	0.10	0.01	0.07
Total	0.84	0.83	0.10	0.01	0.06
CV	0.80	0.80	0.11	0.01	0.06

3.4 Spatio temporal distribution of THC concentration

Figure 6 presents the prediction maps of THC concentrations estimated using the LightGBM model. Subfigure (a) shows the average THC concentrations for 2015–2018, while subfigure (b) presents seasonal maps for spring, summer, fall, and winter. The maps use dark green to indicate low THC concentrations and red to represent high concentrations. The predicted average THC concentration over the four years is 2.07 ppm. Among the seasons, spring exhibited the highest average concentration at 2.12 ppm, while summer had the lowest at 2.05 ppm. Notably, higher THC concentrations were observed in the western part of the Northern AQZ and lower concentrations in the eastern area.

4. Discussion

The LightGBM spatiotemporal model demonstrated robust predictive performance in estimating THC concentrations across different evaluation stages. During the training phase, the model achieved a high R^2 value of 0.84, indicating that it was able to explain 84% of the variability in the THC data. This strong fit was maintained during validation phase, with an R^2 of 0.82, which suggests that the model generalizes well to unseen data without substantial overfitting. The external testing dataset, which represents truly independent samples, yielded a slightly lower R^2 of 0.77, reflecting the expected reduction in predictive accuracy when applied to new data but still indicating reliable performance. The overall total R^2 of 0.84, along with a consistent RMSE of approximately 0.10, reinforces the model's robustness and stability in capturing the complex spatiotemporal dynamics of THC concentrations. Cross-validation further supported these findings, with an average R^2 of 0.80, confirming the model's consistent behavior across multiple subsets of data.

When comparing the performance of our LightGBM spatiotemporal model with previous studies, it is evident that our model achieves competitive accuracy in estimating THC concentrations. For instance, Su *et al.* (2010) utilized Land Use Regression (LUR) techniques to capture the spatial distribution of volatile organic compounds, including THC, in Toronto, reporting moderate R^2 values between 0.66 and 0.68. Similarly, Doris *et al.* (2024) applied LUR models to evaluate THC exposure near natural gas production areas in northeastern British Columbia, achieving a wider range of R^2 from 0.73 up to 0.91 depending on the specific site and time period. In contrast, our model consistently attained R^2 values of 0.84 during training and 0.82 in validation, with testing R^2 of 0.77, situating its predictive power well within or above the range of these previous LUR studies. This suggests that the integration of advanced machine learning techniques like LightGBM, combined with a comprehensive set of environmental, pollutant, and land use variables, can provide improved accuracy and robustness in THC estimation compared to traditional regression approaches. Such enhanced performance is particularly

valuable for capturing the complex spatiotemporal patterns of THC in diverse environments, ultimately supporting more precise exposure assessments and better-informed air quality management decisions (Lu *et al.*, 2025; Thongthammachart *et al.*, 2022).

Beyond traditional performance metrics, the SHAP (SHapley Additive exPlanations) analysis provided valuable insights into the contribution of individual features to the model's predictions. The analysis identified NO_2 (ppm) as the most influential variable, underscoring the critical role of nitrogen dioxide concentrations in driving THC levels. NOx (ppm) was also found to have a significant impact, highlighting the broader importance of nitrogen oxides in air pollution chemistry relevant to hydrocarbons (Bishop & Stedman, 2008; Geddes *et al.*, 2021; Guaman *et al.*, 2022; Environment Canada, 2003). Population density within a 1750-meter radius emerged as another important predictor, suggesting that human activities and urbanization strongly affect local THC concentrations (Coffin *et al.*, 2021). Land use variables such as plantation forests and paddy fields at varying distances (e.g., 2500 m to 4000 m) were also influential, indicating that the type and extent of vegetation and land cover affect the dispersion and accumulation of hydrocarbons (Kansal, 2009; Zhou *et al.*, 2023). Plantation forest at the 4000 m buffer showed a higher ranking than at 2500 m because a larger buffer of point density captures broader spatial patterns, which better represent pollutant dispersion and regional variation, while smaller buffers mainly reflect localized effects with lower explanatory power. Additionally, meteorological factors, particularly wind speed, played a notable role, consistent with its known effect on pollutant transport and dilution Terrain variables, including higher elevations and steeper slopes, were generally associated with lower THC concentrations due to enhanced airflow and pollutant dispersion, whereas valleys and low-lying areas tended to accumulate more THC (Bose & Chowdhury, 2023; Cheng, 2002).

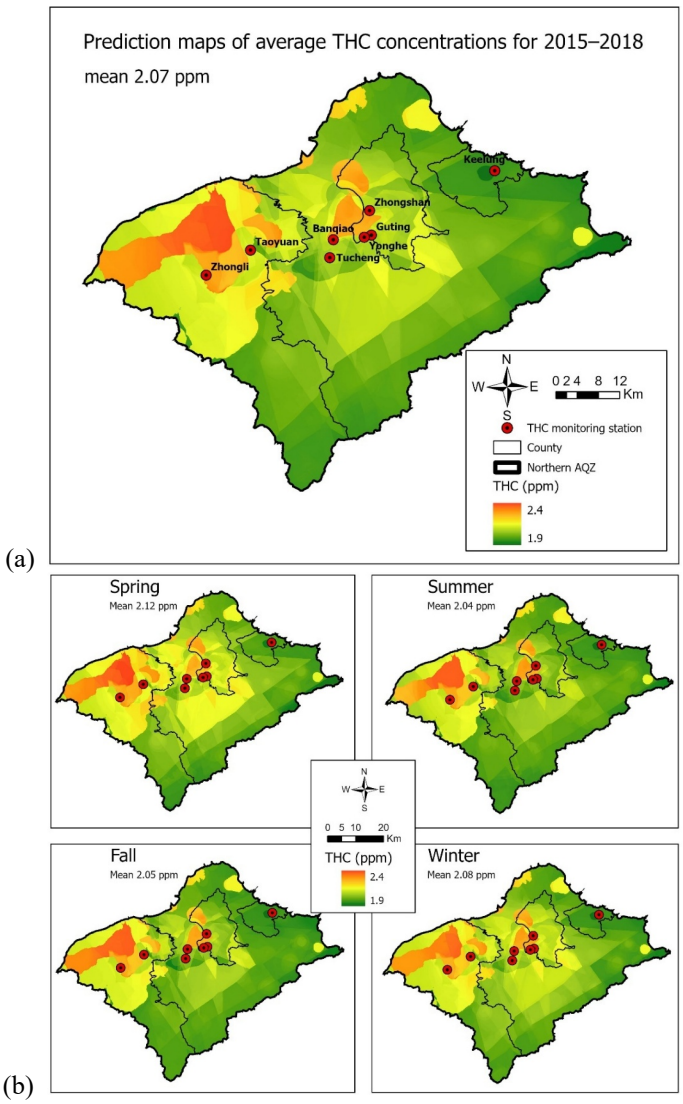

Figure 6 (a) Prediction maps of average THC concentrations for 2015–2018 using LightGBM, (b) the maps for each season (spring, summer, fall, and winter)

These combined results emphasize that THC concentrations are governed by a complex interplay of both anthropogenic emissions and environmental conditions. The ability of the model to integrate air pollutant measurements, population data, land use characteristics, and meteorological variables allows for a more comprehensive and accurate estimation of THC distribution in space and time. This holistic approach is critical for developing effective air quality management strategies and for understanding the sources and dynamics of hydrocarbon pollution in the studied region. Future research could further improve model performance by incorporating additional variables such as traffic intensity or industrial emissions, as well as by exploring temporal trends with finer resolution data.

Seasonal variations in predicted THC concentrations are influenced by meteorological conditions that affect pollutant dispersion (Bose & Chowdhury, 2023; Li *et al.*, 2021). In Northern Taiwan, spring showed the highest average THC concentration (2.12 ppm) due to relatively stable atmospheric conditions and lower rainfall, which limit pollutant dispersion. In contrast, summer exhibited the lowest concentration (2.05 ppm) as higher temperatures, stronger winds, and more frequent precipitation enhance atmospheric mixing and pollutant washout. Fall and winter presented intermediate values, reflecting the combined effects of local emissions and seasonal weather patterns.

From a regulatory and health perspective, THC concentrations are not regulated as a criteria air pollutant in Taiwan, and no specific ambient air quality or health-based standard has been established for THC. Hydrocarbon-related compounds are instead addressed within the broader VOC framework through source-specific emission controls (Ministry of Environment, 2023). Because THC represents a mixture of different hydrocarbon compounds with varying toxicological properties, a single health-based threshold cannot be directly applied. Therefore, the THC concentrations estimated in this study are interpreted as an indicator of atmospheric hydrocarbon levels rather than against a specific regulatory or health-based threshold.

5. Conclusions

In conclusion, the LightGBM spatiotemporal model demonstrated strong predictive capability in estimating THC concentrations, with R^2 values of 0.84 for training, 0.82 for validation, and 0.77 for testing with unseen data, indicating reliable and consistent performance across datasets. The average THC concentration in the study area was 2.07 ppm. SHAP analysis identified NO_2 (ppm), NOx (ppm), and population density as the most influential variables driving THC levels, alongside land use factors such as plantation forests and paddy fields, and meteorological variables like wind speed. These findings highlight the combined effects of air pollution sources, human activity, and environmental conditions in shaping THC distribution. Overall, the model provides a robust tool for accurately capturing the spatial and temporal variation of THC, which is essential for effective air quality monitoring and management strategies.

References

Ambrose, I.S., Offiong, N.-A.O., Ayua, T.J., Ebong, G.A., Ambrose, E.I., Silas, I.I., and Saturday, U.U., 2026. Seasonal dynamics of atmospheric heavy metals and total hydrocarbon content at the Edonwhii Beach, Nigeria, Discover Environment, 4: 200, DOI:10.1007/s44274-026-00712-5.

Bahmani, M.J., 2020. Understanding LightGBM parameters (and how to tune them), Available at: http://dev.to/kamil_k7k/understanding-lightgbm-parameters-and-how-to-tune-them-14n0, Assessed March 11, 2026.

Bishop, G.A., and Stedman, D.H., 2008. A decade of on-road emissions measurements, Environmental Science & Technology, 42(5): 1651–1656, DOI:10.1021/es702413b.

Bose, A., and Chowdhury, I.R., 2023. Investigating the association between air pollutants' concentration and meteorological parameters in a rapidly growing urban center of West Bengal, India: A statistical modeling-based approach, Modeling Earth Systems and Environment, 9(2): 2877–2892, DOI:10.1007/s40808-022-01670-6.

Cheng, W.-L., 2002. Ozone distribution in coastal central Taiwan under sea-breeze conditions, Atmospheric Environment, 36(21): 3445–3459, DOI:10.1016/S1352-2310(02)00307-2.

Coffin, A.W., Ouren, D.S., Bettez, N.D., Borda-de-Água, L., Daniels, A.E., Grilo, C., Jaeger, J.A.G., Navarro, L.M., Preisler, H.K., and Rauschert, E.S.J., 2021. The ecology of rural roads: Effects, management, and research (Issues in Ecology Report No. 23), Ecological Society of America (ESA).

Department of Monitoring and Information, Ministry of Environment, Taiwan, 2024. Air quality monitoring daily average values (air pollutants), Ministry of Environment Environmental Information Open Platform, Available at: https://data.moenv.gov.tw/en/dataset/detail/AQX_P_19, Assessed March 17, 2026.

Doris, M., Daley, C., Zalzal, J., Chesnaux, R., Minet, L., Kang, M., Caron-Beaudoin, É., MacLean, H. L., and Hatzopoulou, M., 2024. Modelling spatial & temporal variability of air pollution in an area

of unconventional natural gas operations, Environmental Pollution, 348: 123773, DOI:10.1016/j.envpol.2024.123773.

Environment Canada, 2003. Future Canadian measures for reducing emissions of volatile organic compounds (Report No. En14-153/2003E), Available at: https://publications.gc.ca/collections/collection_2014/ec/En14-153-2003-eng.pdf, Assessed March 17, 2026.

Geddes, J.A., Wang, B., and Li, D., 2021. Ozone and nitrogen dioxide pollution in a coastal urban environment: The role of sea breezes, and implications of their representation for remote sensing of local air quality, Journal of Geophysical Research: Atmospheres, 126(18): e2021JD035314, DOI:10.1029/2021JD035314.

GeeksforGeeks, 2025. Cross-validation and hyperparameter tuning of LightGBM Model, Available at: https://www.geeksforgeeks.org/machine-learning/cross-validation-and-hyperparameter-tuning-of-lightgbm-model/, Assessed March 11, 2026.

Guaman, M., Roberts-Semple, D., Aime, C., Shin, J., and Akinrem, A., 2022. Traffic density and air pollution: Spatial and seasonal variations of nitrogen dioxide and ozone in Jamaica, New York, Atmosphere, 13(12): 2042, DOI:10.3390/atmos13122042.

Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L., and Yang, L., 2023a. Prediction of heating and cooling loads based on light gradient boosting machine algorithms, Building and Environment, 236: 110252, DOI:10.1016/j.buildenv.2023.110252.

Guo, X., Gui, X., Xiong, H., Hu, X., Li, Y., Cui, H., Qiu, Y., and Ma, C., 2023b. Critical role of climate factors for groundwater potential mapping in arid regions: Insights from random forest, XGBoost, and LightGBM algorithms, Journal of Hydrology, 621: 129599, DOI: 10.1016/j.jhydrol.2023.12959.

Hsu, C.-Y., Xie, H.-X., Wong, P.-Y., Chen, Y.-C., Chen, P.-C., and Wu, C.-D., 2022. A mixed spatial prediction model in estimating spatiotemporal variations in benzene concentrations in Taiwan, Chemosphere, 301: 134758, DOI:10.1016/j.chemosphere.2022.134758.

Kansal, A., 2009. Sources and reactivity of NMHCs and VOCs in the atmosphere: A review, Journal of Hazardous Materials, 166(1): 17–26, DOI:10.1016/j.jhazmat.2008.11.048.

Li, S., Dong, X., Ma, D., Dang, B., Zang, H., and Gong, Y., 2024. Utilizing the LightGBM algorithm for operator user credit assessment research, arXiv preprint arXiv:2403.14483, DOI:10.48550/arXiv.2403.14483.

Li, Z., Ho, K.-F., Chuang, H.-C., and Yim, S.H.L., 2021. Development and intercity transferability of land-use regression models for predicting ambient PM_{10}, $PM_{2.5}$, NO_2 and O_3 concentrations in northern Taiwan, Atmospheric Chemistry and Physics, 21(6): 5063–5078, DOI:10.5194/acp-21-5063-2021.

Lu, B., Zhang, S., Liu, C., Ganbat, G., Herrmann, H., Li, X., and Zhao, Y., 2025. Machine learning-enabled estimation and high-resolution forecasting of atmospheric VOCs, Atmospheric Environment, 358: 121364, DOI:10.1016/j.atmosenv.2025.121364.

Ministry of Environment, 2023. Volatile Organic Compounds Air Pollution Control and Emission Standards, Available at: https://oaout.moenv.gov.tw/laW/LawContent.aspx?id=FL015377, Assessed August 18, 2026.

Risk Society and Policy Research Center, 2026. Air pollution governance and its distance with people, National Taiwan University, Available at: https://rsprc.ntu.edu.tw/web/research/research_in.jsp?lang=en&rp_id=RP1729156478027, Assessed March 17, 2026.

Stewart, G.J., Nelson, B.S., Drysdale, W.S., Acton, W.J.F., Vaughan, A.R., Hopkins, J.R., Dunmore, R.E., Hewitt, C.N., Nemitz, E., Mullinger, N., Langford, B., Shivani, Reyes-Villegas, E., Gadi, R., Rickard, A.R., Lee, J.D., and Hamilton, J.F., 2021. Sources of non-methane hydrocarbons in surface air in Delhi, India, Faraday Discussions, 226: 409–431, DOI:10.1039/d0fd00087f.

Su, J.G., Jerrett, M., Beckerman, B., Verma, D., Arain, M.A., Kanaroglou, P., Stieb, D., Finkelstein, M., and Brook, J., 2010. A land use regression model for predicting ambient volatile organic compound concentrations in Toronto, Canada. Atmospheric Environment, 44(29): 3529–3537, DOI:10.1016/j.atmosenv.2010.06.015.

Thongthammachart, T., Araki, S., Shimadera, H., Matsuo, T., and Kondo, A., 2022. Incorporating Light Gradient Boosting Machine to land use regression model for estimating NO_2 and $PM_{2.5}$ levels in Kansai region, Japan, Environmental Modelling & Software, 155: 105447, DOI:10.1016/j.envsoft.2022.105447.

Tırınk, S., 2025. Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM, PLOS One, 20(10): e0334252, DOI:10.1371/journal.pone.0334252.

United States Environmental Protection Agency, 2005. Conversion Factors for Hydrocarbon Emission Components (EPA420-R-05-015 / NR-002c), US EPA Report.

Vertes, D., 2020. Beyond grid search: Using Hyperopt, Optuna, and Ray Tune to hypercharge hyperparameter tuning for XGBoost and LightGBM, Available at: https://druce.ai/2020/10/hyperparameter-tuning-with-xgboost-ray-tune-hyperopt-and-optuna, Assessed March 11, 2026.

Wu, C.D., Chen, Y.C., Pan, W.C., Zeng, Y.T., Chen, M.J., Guo, Y.L., and Lung, S.C.C., 2017. Land-use regression with long-term satellite-based greenness index and culture-specific sources to model $PM_{2.5}$ spatial-temporal variability, Environmental Pollution, 224: 148–157, DOI:10.1016/j.envpol.2017.01.074.

Zhou, Z.-Y., Zhang, X.-X., Wang, F., Zhu, C.-D., and Pan, J.-J., 2023. Vertical pollution characteristics and source analysis of soil PAHs in different land use types, Environmental Science, 44(3): 1583–1592, DOI:10.13227/j.hjkx.202204071.

台灣北部地區使用 LightGBM 回歸模型進行碳氫化合物估計

帕斯蒂[1] 古曼寧[2] 曾于庭[3] 吳治達[4,5,6,7,8*]

摘要

總碳氫化合物(THC)是重要的大氣污染物，主要來自交通、工業活動及燃料燃燒。本研究採用 LightGBM 模型估計 2015 年 6 月至 2020 年 12 月台灣北部空氣品質區的 THC 濃度，並以 2021 年資料進行外部驗證。模型訓練、測試及外部驗證的 R^2 分別為 0.84、0.82 及 0.77，研究區域平均 THC 濃度為 2.07 ppm。SHAP 分析顯示，NO_2 與 NOx 是主要影響因子，其次為人口密度、林地、水田及風速。空間分析發現，西部 THC 濃度高於東部，可能與都市化及工業活動的空間差異有關。本研究顯示，LightGBM 結合 SHAP 可有效估計 THC 濃度並辨識重要影響因子，為空氣品質評估及環境管理提供參考。

關鍵詞：Light Gradient Boosting Machine Regression(LightGBM)、空間預測模型、總碳氫化合物(THC)

[1] 國立成功大學測量及空間資訊學系 博士生
[2] 國立成功大學測量及空間資訊學系 博士後研究員
[3] 國立成功大學測量及空間資訊學系 研究助理
[4] 國立成功大學測量及空間資訊學系 教授
[5] 國家衛生研究院國家環境醫學研究所 兼任研究員
[6] 國立中興大學永續農業創新發展中心 兼任教授
[7] 高雄醫學大學精準環境醫學研究中心 兼任研究員
[8] 長庚科技大學慢性疾病暨健康促進研究中心 合聘研究員
[*] 通訊作者, E-mail: chidawu@mail.ncku.edu.tw

收到日期：民國 114 年 07 月 16 日
修改日期：民國 114 年 09 月 17 日
接受日期：民國 115 年 08 月 21 日