機械学習による $PM_{2.5}$ 濃度高精度推定手法の開発

環境エネルギー工学専攻　環境工学コース
共生環境評価領域　嶋寺研究室　　Supitcha SUKPRASERT

1. Introduction

Thailand has faced severe $PM_{2.5}$ pollution for over a decade, mainly due to Biomass Burning (BB) emissions from agricultural burning and forest fires. Long-term exposure to $PM_{2.5}$ is linked to health risks, such as respiratory and cardiovascular diseases. To effectively study the health impacts, researchers need complete-coverage $PM_{2.5}$ data. However, Thailand's air quality monitoring (AQM) network is limited, and expanding it is costly, making modeling a more feasible solution. The Chemical Transport Model (CTM) and Machine Learning (ML) models are common models for $PM_{2.5}$ predictions. Tree-based ML models, such as Light Gradient Boosting Machine (LightGBM), are preferred for training speed and efficiency over deep learning models, such as Neural Networks. However, CTMs require complex atmospheric settings and high-quality inputs, while ML models lack atmospheric process integration. Thus, researchers have integrated both models to enhance the performance, but no study has yet decoupled single emissions from CTMs for ML applications. This research aims to develop a LightGBM model incorporating BB-decoupled CTM predictors to improve daily $PM_{2.5}$ predictions in Central Thailand during 2019 (see **Figure 1**).

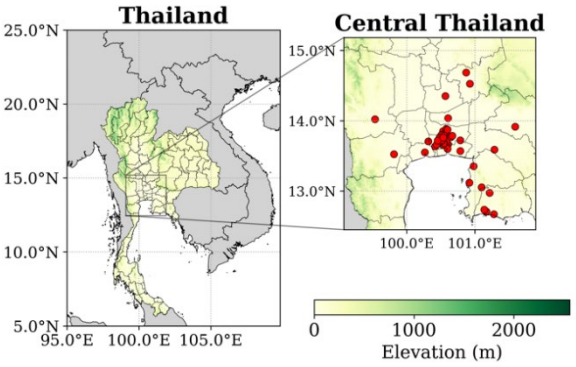

Figure 1 The study area (red dots are AQM stations)

2. Methodology

This study used 1-year $PM_{2.5}$ data from 55 AQM stations provided by the Pollution Control Department and Bangkok Metropolitan Administration.

2.1 The LightGBM model's predictors: There are two types of predictors. (a) CTM predictors: The CTM was used to simulate $PM_{2.5}$ concentrations under two scenarios. The first scenario includes all emission sources (baseline CTM predictor), and the second excludes BB emissions (BB-excluded CTM predictor). The difference between these scenarios yields a BB-contributed CTM predictor. BB-excluded and BB-contributed collectively form the BB-decoupled CTM predictors (see **Figure 2**). (b) Control predictors: They include population density, elevation, land use, and meteorological variables simulated from the weather model.

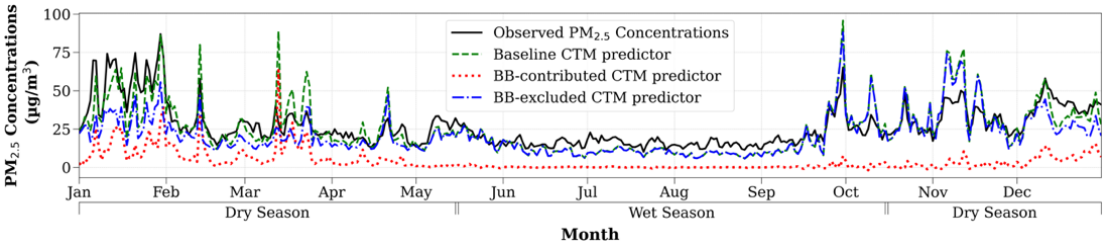

Figure 2 Daily average of observed and CTM-simulated $PM_{2.5}$ concentrations across all AQM stations

2.2 The LightGBM model development: Two LightGBM models were developed (baseline and decoupled model). The baseline model used the baseline CTM predictor, and the decoupled model used BB-decoupled CTM predictors. Both used identical control predictors. The model hyperparameters were adopted directly from Thongthammachart et al. [1], who studied the same area and dataset. The model evaluation employed five-fold cross-validation (CV) across three aspects: overall, spatial, and temporal, splitting the data by samples, monitoring sites, and dates, respectively. Finally, a one-tailed paired t-test was used to compare the model performance.

3. Results and discussion

Table 1 The baseline and decoupled models' performance

Model	CV aspects	Annual		Wet season		Dry season	
		R^2	RMSE*	R^2	RMSE*	R^2	RMSE*
Baseline model	Overall	0.82	6.68	0.76	4.71	0.77	7.82
	Spatial	0.70	8.61	0.51	6.80	0.65	9.74
	Temporal	0.71	8.46	0.69	5.41	0.62	10.14
Decoupled model	Overall	0.86	5.92	0.79	4.46	0.83	6.80
	Spatial	0.73	8.18	0.52	6.69	0.69	9.13
	Temporal	0.75	7.83	0.70	5.30	0.68	9.27

* The unit of RMSE is μg/m^3

Table 1 presents the performance scores of the baseline and decoupled models evaluated across three CV aspects. The incorporation of decoupled predictors has been statistically proven to improve the performance scores of the LightGBM model across all CV aspects (p-value < 0.01). The largest improvement is observed in the dry season, which is the period with the intense BB activities.

The spatial map further highlights the decoupled model's advantage in capturing the spatial variations of $PM_{2.5}$ on the high BB Day (see **Figure 3**). With an R^2 of 0.64 and RMSE of 7.86 μg/m^3, it outperforms the baseline model (R^2=0.51, RMSE=9.16 μg/m^3). The improvement is likely due to the decoupling process, which creates two CTM predictors with unique variations (see **Figure 2**). This could help the Light model to better learn the relationship between source-specific $PM_{2.5}$ and other predictors, enabling it to correct CTM biases more effectively.

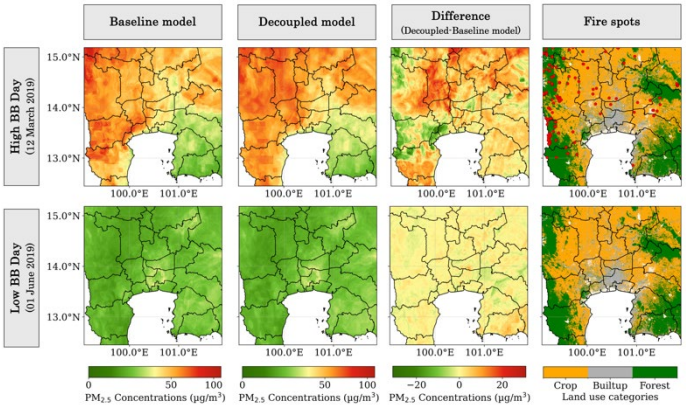

Figure 3 Spatial distribution of predicted $PM_{2.5}$ concentrations on high BB Day (12 March 2019) and low BB Day (01 June 2019). Red dots in the rightmost plot represent fire spots.

4. Conclusion

Applying the BB-decoupled predictors significantly improves the LightGBM model performance in all CV aspects, demonstrating the effectiveness of the decoupling process in refining the $PM_{2.5}$ predictions. However, other ML models, such as deep learning algorithms, may yield different results. Further exploration is needed to determine whether the decoupling process enhances performance in other types of ML models.

5. Reference

[1] T. Thongthammachart et al., *Atmos. Environ.*, **297**, 119595, (2023).