Applying a fusion approach of GAM and XGBoost to predict spatiotemporal distributions of daily elemental $PM_{2.5}$ in Taiwan

Chun-Sheng Huang[1], Kang Lo[1], Yee-Lin Wu[2], Fu-Cheng Wang[3], Yi-Shiang Shiu[4], Chu-Chih Chen[5], Yuan-Chien Lin[6], Cheng-Pin Kuo[7], Tang-Huang Lin[3*], Chang-Fu Wu [18*]

[1] Institute of Environmental and Occupational Health Sciences, School of Public Health, National Taiwan University, Taiwan
[2] Department of Environmental Engineering, College of Engineering, National Cheng Kung University Taiwan
[3] Center for Space and Remote Sensing Research, National Central University, Taiwan
[4] Department of Urban Planning and Spatial Information, College of Construction and Development, Feng Chia University, Taiwan
[5] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Taiwan
[6] Department of Civil Engineering, College of Engineering, National Central University, Taiwan
[7] Green Energy and Environmental Research Laboratories, Industrial Technology Research Institute, Taiwan
[8] Department of Public Health, School of Public Health, National Taiwan University, Taiwan

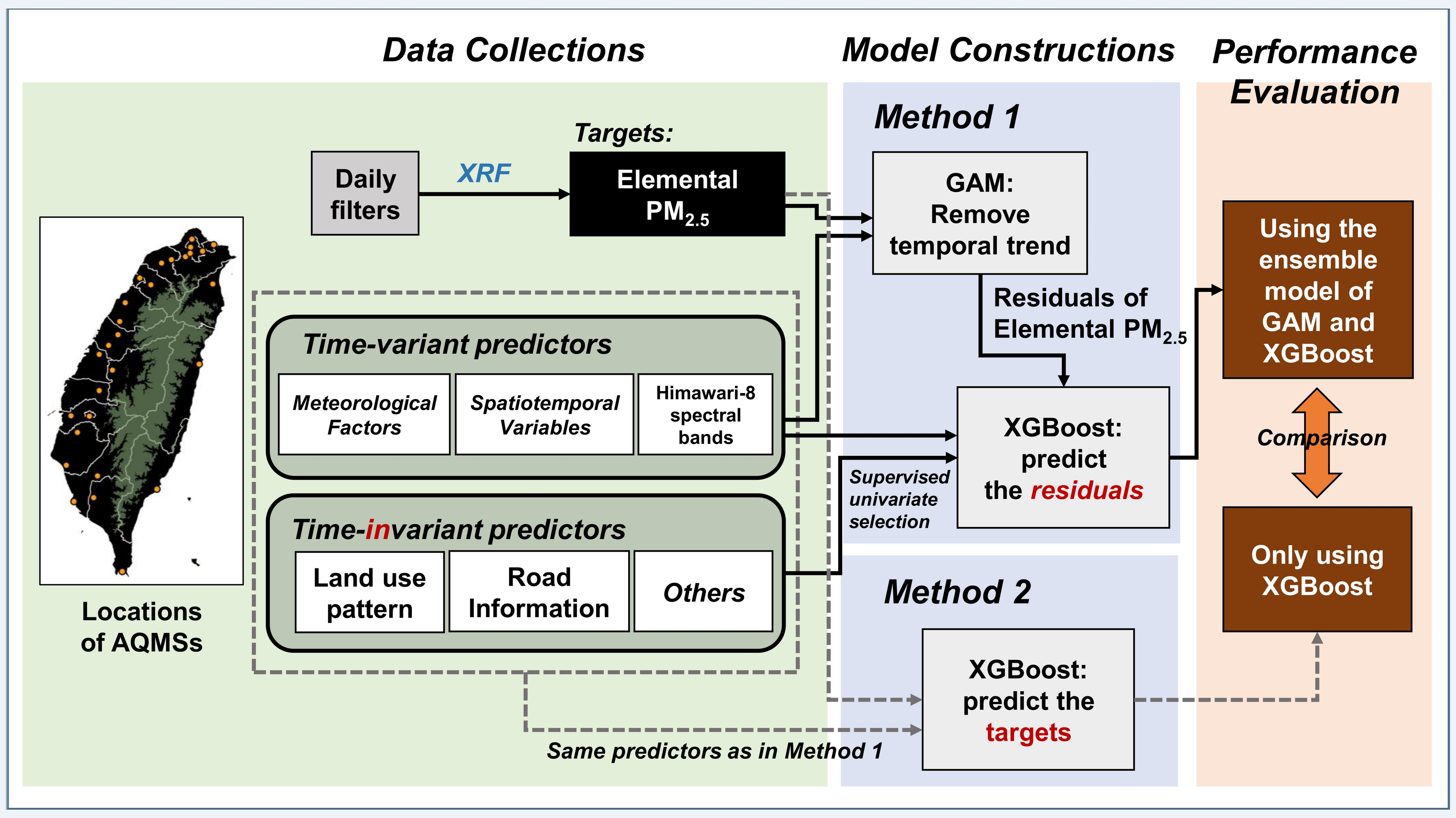

Figure 1. Workflow diagram illustrating the computational procedures

Introduction

1. Fine particulate matter ($PM_{2.5}$) has been linked with **adverse health outcomes**.
2. The toxicity and health effects of $PM_{2.5}$ are primarily due to its **chemical composition**. For example, particles containing specific elements, such as chromium (Cr) or nickel (Ni), are particularly harmful because of their carcinogenic and mutagenic properties.
3. Estimating the **spatiotemporal distributions** of $PM_{2.5}$ compositions is important.
4. **Knowledge gap**: (1) The air pollutant levels from temporal and spatial perspectives were not separately explored; (2) The satellite spectral bands were not utilized in the predictive model of $PM_{2.5}$ chemical compositions.

- **Objective:**

- To **construct** spatiotemporal predictive models for various elemental compositions of $PM_{2.5}$ using an ensemble machine learning model with Himawari-8 spectral bands

Materials and Methods

1. **Data collections:** $PM_{2.5}$ filter samples were collected from 28 air quality monitoring stations on two days for each month from June 2021 to May 2022.
2. **Chemical analysis**: Filters were used for Energy-Dispersive X-Ray Fluorescence spectrometer analysis to determine 12 trace elements, namely Mg, Si, S, Ca, V, Cr, Mn, Fe, Ni, Cu, Zn, and Pb (**prediction target**)
3. Information on **model predictors/features**:

Categories	Model predictors	Sources
Himawari-8 Spectral Bands	Spectral band 1 – band 16 (Top of atmospheric reflectance and Temperature brightness)	JAXA's P-Tree system
Meteorological Factors	Temperature, Humidity, Wind speed …	Ministry of Environment of Taiwan
Spatiotemporal Variables	Sampling month, air quality region	Ministry of Environment of Taiwan
Land Use Pattern	Industrial building, natural greenness, farm ...	Satellite images composited of Himawari-8, SPOT-7, SPOT-8, Landsat-2, and Sentinel-8
Road Information	Road length, distance to the nearest road …	Ministry of Transportation and Communications
Others	Number of emission sources, temples….	Taiwan Emission Data System, Kingwaytek Tech Co.,Ltd…

4. **Model descriptions**:

- **Generalized Additive Model (GAM):** To remove the temporal trend in elemental $PM_{2.5}$ derived from meteorological factors and spatiotemporal variables:

$$y_{s,t} = f(TEMPERATURE) + f(HUMIDITY) + f(WINDSPEED) + f(WINDDIREC\_SIN) + f(WINDDIREC\_COS) + (MONTH) + r(AQRs) + \underline{f(H8BAND1-16)} + \eta_{s,t}$$

$$s=1,\cdots,28;\ t=1,\cdots,24$$

The spectral bands are sensitive to complex atmospheric conditions such as cloud and aerosol detections

- ***eXtreme Gradient Boosting (XGBoost)***: To predict the spatial variations in temporal trend-removed **residuals** of elemental $PM_{2.5}$ retrieved from GAM

(1) All available samples were applied for model training (N = 648 site-day)
(2) The **5-fold cross validation** method was used for hyperparameter tuning
(3)The R packages 'xgboost', 'mlr3', 'mlr3tuning', and 'paradox' were utilized for main algorithm implementation, hyperparameter optimization and setting up machine learning tasks.
(4) LOOCV was utilized to evaluate the model performance.

Table 1. Hyperparameter search space configuration for XGBoost

Hyperparameter	Range of value	
	Lower bound	Upper bound
Maximum depth	3	6
Learning rate (eta)	0.01	0.03
Number of boosting rounds	100	300
Minimum child weight	3	6
Subsample	0.5	0.7
Column sample by tree	0.5	0.7
Column sample by level	0.5	0.7
L1 regularization term on weights	30	60

Modeling results

Table 2. Comparison of predictive model performance for elemental $PM_{2.5}$ concentrations

Elements	Leave-one-out cross validation (LOOCV)			
	Method 1: ensemble model		Method 2: XGBoost only	
	NRMSE	R^2	NRMSE	R^2
Mg	0.54	0.73	0.49	0.77
Si	0.60	0.67	0.52	0.75
S	0.69	0.59	0.59	0.68
Ca	0.49	0.77	0.43	0.83
V	0.75	0.52	0.92	0.33
Cr	0.89	0.29	0.82	0.44
Mn	0.72	0.55	0.65	0.62
Fe	0.58	0.69	0.49	0.77
Ni	0.73	0.51	0.82	0.44
Cu	0.86	0.40	0.75	0.52
Zn	0.82	0.43	0.69	0.58
Pb	0.76	0.50	0.69	0.58

1. Two methods (**Method 1: ensemble model of GAM and XGBoost** vs. **Method 2: XGBoost only**) were compared.
2. In Table 2, Method 2 generally showed a slightly higher LOOCV R^2 (average = 0.06) than Method 1. However, the **difference in LOOCV R^2** and NRMSE between two methods were **not significant** (p-value > 0.05) with Wilcoxon signed-rank test.
3. **V and Ni** exhibited **higher LOOCV R^2** and lower NRMSE in Method 1

Table 3. Comparison of the highest occurrences of 15 features based on the SHAP values

Order	Method 1: ensemble model		Method 2: XGBoost only	
	Feature	Frequency	Feature	Frequency
1	TEMPERATURE	10	TEMPERATURE	12
2	WINDDIREC_COS	10	MONTH	11
3	WINDSPEED	10	H8BAND_8	9
4	**ALLROADLEN_1000**	8	H8BAND_9	8
5	WINDDIREC_SIN	8	WINDDIREC_SIN	8
6	HUMIDITY	6	H8BAND_10	7
7	H8BAND_7	5	**POINT_N _5000**	6
8	**HALFNATURAL_300**	5	**INDUSTRY_5000**	5
9	**INDUSTRY_5000**	4	HUMIDITY	5
10	**POINT_N_5000**	4	WINDSPEED	5
11	**DISTCOAST**	3	AQR_1	4
12	**NONINDUSTRY_5000**	3	**DISTCOAST**	4
13	**RESTAURANT_500**	3	H8BAND_7	4
14	**TEMPLE_500**	3	AQR_4	3
15	**FARM_1000**	2	**FARM_1000**	3

[a] Shapley Additive exPlanations (SHAP) values were calculated to evaluate the important featrues; [b] The ten predictors with the highest SHAP values for each 12 elements were integrated for the calculation; [c] The time-invariant features are highlighted in bold; The number after the feature code denotes the buffer size (unit: meter)

1. **Important time-invariant predictors** can be highlighted by the **Method 1**, which can improve the **interpretability of spatial variation** (Table 3)
2. The procedure of temporal-trend removal enables the **capture of potential contributing sources**, which helps focus on air pollution hotspots in environmental policy regulations.