Estimation of surface $PM_{2.5}$ concentrations from atmospheric gas species retrieved from TROPOMI using deep learning: Impacts of fire on air pollution over Thailand

Rackhun Son[a,d,e], Hyun Cheol Kim[b,c], Jin-Ho Yoon[d] and Dimitris Stratoulias[e,f,*]

[a]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

[b]Air Resources Laboratory, National Oceanic and Atmospheric Administration, College Park, MD, USA

[c]Cooperative Institute for Satellite Earth System Studies, University of Maryland, College Park, MD, USA

[d]School of Earth Science and Environmental Engineering, Gwangju Institute of Science and Technology, Gwangju, Korea

[e]Geospatial Information department, Asian Disaster Preparedness Center (ADPC), Bangkok, Thailand

[f]SERVIR-Mekong, Bangkok, Thailand

* Dimitris Stratoulias: dimitrios.stratoulias@adpc.net

Abstract

Surface $PM_{2.5}$ concentration is routinely observed at limited number of surface monitoring stations. To overcome its limited spatial coverage, space-borne monitoring system has been established. However, it also faces various challenges such as cloud contamination and limited vertical resolution. In this study, we propose a deep learning-based surface $PM_{2.5}$ estimation method using the attentive interpretable tabular learning neural network (TabNet) with atmospheric gas species retrieved from the tropospheric monitoring instrument (TROPOMI). Unlike previous applications that primarily used decision tree-based algorithms, TabNet provides interpretable decision-making steps to identify dominant factors. By incorporating five TROPOMI products (i.e., NO_2, SO_2, O_3, CO, HCHO), we have tested the system's capability to produce surface $PM_{2.5}$ concentration without aerosol optical property, which was used more traditionally. The proposed model successfully captures spatiotemporal variations and its performance surpasses those of other leading machine learning models over Thailand in the period of 2018-2020. The interpretable decision-making steps highlight that carbon monoxide is the most influential chemical component, which relates to the seasonal burning in southeast Asia. It is found that air quality impacts from fire are stronger in the northern part of Thailand and fires in neighboring countries should not be neglected. The proposed method successfully estimates surface PM2.5 concentration without aerosol optical property, implying its potential to advance monitoring air quality over remote regions.

Keywords: $PM_{2.5}$; TROPOMI; deep learning; TabNet

Key Points

1. TabNet successfully estimates surface $PM_{2.5}$ with atmospheric gas compositions
2. CO is highlighted as a key factor in estimating spatiotemporal pattern of $PM_{2.5}$
3. The origin of CO is likely from seasonal fire in Thailand and its bordering countries

1. Introduction

Ambient air pollution is a critical threat to public health, causing more than three million premature fatalities worldwide in 2012 (Organization, 2016) as well as various environmental issues (Gurjar et al., 2010). Among air pollutants, fine particulate matter ($PM_{2.5}$), namely, ambient airborne particulates sized under 2.5 microns, is well known to damage human health seriously. Due to its microscopic size, $PM_{2.5}$ can affect the respiratory and cardiovascular systems, causing or worsening major illnesses such as asthma, lung cancer and heart disease (Weichenthal et al., 2013). Rising public concern about air quality urges not only reductions in air pollutants, but also improvements to air quality monitoring at the ground level to assess the health and socioeconomic impacts.

Annual mean $PM_{2.5}$ concentrations in Thailand reached 21.4 $\mu g/m^3$ in 2020, making it the 34th most polluted country in the world ("World Air Quality Report 2020," n.d.). An estimated 40,000 deaths annually in Thailand are attributable to ambient air pollution (Pinichka et al., 2017), resulting in 0.74 to 1.33 million USD worth of economic costs (Vassanadumrongdee and Matsuoka, 2005). Although air pollution exposure in Thailand temporarily improved during the recent COVID-19 pandemic (Rodríguez-Urrego and Rodríguez-Urrego, 2020; Stratoulias and Nuthammachot, 2020), it remains high due to widespread smoke emissions from agricultural burnings and forest fires (Punsompong et al., 2021). $PM_{2.5}$ emissions from burning crop residue and forest fires are estimated to be 141,000 and 5,000 tons per year, respectively, mostly concentrated in the central and northern regions of Thailand (Junpen et al., 2013; Kanabkaew and Kim Oanh, 2011). However, the insufficient number of in-situ $PM_{2.5}$ measurements, especially for the provinces in the north and northeast of the country (Figure 1a), limits monitoring air quality and establishing a national plan for its management. Given that considerable financial and time resources are required to increase the number of air quality monitoring stations, satellite remote sensing-based $PM_{2.5}$ estimation is an alternative way to increase the limited spatial coverage.

Two different methods have been developed and widely used to estimate surface $PM_{2.5}$ concentration: Chemical Transport Models (CTMs) and statistical regression models. Based on physicochemical processes and atmospheric conditions, chemical transport models can approximate the quantity of air pollutants with continuous spatiotemporal coverage (Liu et al., 2004; Van Donkelaar et al., 2010). However, uncertainties in emission inventories and limited representation of chemical reactions in the ambient atmosphere remain major concerns (Shin et al., 2020). Among statistical approaches, multiple linear regression has been the most commonly applied in the early stages (Chu et al., 2016). Also, geographically weighted regression, an extension of multiple linear regression, has been proposed to assign distance-based weights to reflect spatial variability and local effects to provide regional estimations (Brunsdon et al., 1998; Jiang et al., 2017; You et al., 2016). Mixed-effect models adopt fixed and random effect terms to separate statistical relationship and variability by time and region (Kloog et al., 2012; Xie et al., 2015). In addition, the generalized additive model has been proposed to consider the nonlinear characteristics between input and target variables (Sorek-Hamer et al., 2013; Zou et al., 2017).

Machine learning (ML) algorithms have recently introduced as innovative developments in the bottom-up approaches to upscale data-driven in-situ models to spatially explicit gridded estimates. Random forest (RF), one of the most frequently applied algorithms, has further improved estimation accuracy and has higher interpretability at both national and regional scales (Chen et al., 2018; Hu et al., 2017; Wei et al., 2019). Elasticnet application has successfully expanded the spatiotemporal dimension with a large number of predictors (Xue et al., 2019). Support vector machine (SVM) can enhance spatial resolution at a 100 m scale by being merged into multiple modeling stages (de Hoogh et al., 2018). Other ML models such as Bayesian maximum entropy (Jiang and Christakos, 2018), gradient boosting machine models (Chen et al., 2019; Wang et al., 2021) and RF combined with ordinary kriging (Han et al., 2022) have also been employed to incorporate satellite-derived products into ground-level observations.

As computing technology and resources have advanced, neural network-based approach has introduced deeper and wider layers, defined as deep learning (DL), and has begun to outperform classical ML models based on decision tree algorithms in various regression tasks (Devlin et al., 2018; He et al., 2016). DL based methods have also recently been attempted in remote sensing due to their high accuracy using large amounts of data (Ghahremanloo et al., 2021; Zhang et al., 2020; Zhu et al., 2020). However, compared with decision trees, the usability of this cutting-edge approach is yet to be explored in-depth for $PM_{2.5}$ satellite-based estimation.

Thus, this study aimed to develop a DL-based model to estimate daily ground-level $PM_{2.5}$ concentrations based in-situ observations in Thailand and satellite-derived atmospheric gas products. Regarding the DL network architecture, we implemented the Attentive Interpretable Tabular Learning neural network (TabNet) (Arık and Pfister, 2021), which is tailored for use with tabular datasets. We evaluated the model’s performance through five different regions in Thailand and compared it with other popular machine learning algorithms such as SVM, RF, XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018). Furthermore, to shed light on the critical characteristics of $PM_{2.5}$ concentration in Thailand, we analyzed the global/local feature selection and reasoning processes as well as fire impacts on $PM_{2.5}$ concentration.

2. Study area and Data

2.1. Study area

Thailand is located at the center of the Indochinese peninsula, has the 10th largest economy in Asia and hosts a population of almost 70 million people ("World Economic Outlook (October 2021)," n.d.). The country is divided into 76 administrative provinces as primary local government units and the capital Bangkok. In this study, Thailand is divided into five regions to analyze the country's regional characteristics: north, northeast, central, east and south (Figure 1a). Despite the low air quality in Thailand, ground monitoring stations are sparse and mostly concentrated in the central region, which contains 29 of the 67 stations used in this study (green points in Figure 1a). For the remaining regions, 15, 5, 11 and 7 stations are distributed in the north (blue), northeast (red), east (yellow) and south (magenta), respectively.

2.2. Ground-level $PM_{2.5}$ observation

The Pollution Control Department in the Air Quality and Noise Management Bureau provides national air quality monitoring records for approximately 84 stations (as of 2021). Considering the consistency of the data availability during the experimental period from January 2018 to June 2021, we selected the daily measurements of $PM_{2.5}$ concentration from 67 stations (Figure 1a) as the target dataset for the model training. The observed $PM_{2.5}$ concentration pattern has an exponential quantile-quantile distribution

(Figure 2a). This asymmetry can hamper model training by blurring the variance in the pollution levels over different input conditions. To transform the data to be closely fitted by a normal distribution, we thus took logs of the $PM_{2.5}$ values after adding one (Figure 2b) and the results showed significantly higher R-squared coefficients (R^2) from 0.744 to 0.996.

2.3. TROPOMI

The Sentinel-5P mission is a precursor satellite measuring atmospheric chemical concentrations at high spatial and radiometric resolutions. The TROPOspheric Monitoring Instrument (TROPOMI) onboard Sentinel-5P is designed to record the reflectance of wavelengths using multispectral sensors. We utilized five TROPOMI products (Borsdorff et al., 2018; De Smedt et al., 2018; Garane et al., 2019; Theys et al., 2017; Van Geffen et al., 2019): the tropospheric NO_2 column (NO_2), SO_2 vertical column density at the ground level (SO_2), total atmospheric column of O_3 (O_3), vertically integrated column of CO (CO) and tropospheric formaldehyde column (HCHO); this was based on 354 of the 388 wavelength pairs. TROPOMI Level 2 products are accessible from the Copernicus Open Access Hub website (https://s5phub.copernicus.eu), and we retrieved a daily Level 3 pre-processed dataset from the Google Earth Engine using the quality assurance values of 0.75 for NO_2 and 0.5 for the other components except for O_3 and SO_2. The Sentinel-5P images were co-located with the ground station data and the values of the pixel encompassing the point location of the ground station were extracted to train the model. When the spatial mapping of $PM_{2.5}$ was inferred, the datasets were resampled to a 10 km grid to incorporate other auxiliary datasets. Subsequently, the variables, except for O_3, were transformed into a logarithmic scale similar to $PM_{2.5}$. Considering that the ranges of each variable varied, specified constants were multiplied and added before the log transform (Figures S1a–d).

2.4. Meteorological dataset

ERA5-Land (Muñoz Sabater, 2019) provides a dataset for land components from ERA5, the fifth-generation climate reanalysis dataset provided by the Copernicus Climate Change Service at the European Centre for Medium-Range Weather Forecasts. Following previous studies (Chen et al., 2018; Wei et al., 2019), we adopted seven meteorological components from the reanalysis dataset: temperature and dew-point temperature at a 2 m height, total evaporation, surface pressure, precipitation and wind components at a 10 m height. We also approximated relative humidity and wind speed using Eqs. (1) and (2):

$$relhumidity = 100 \times \frac{e^{\frac{17.625 \times T_d}{243.04 + T_d}}}{e^{\frac{17.625 \times T}{243.04 + T}}} \quad (1)$$

$$windspeed = \sqrt{U^2 + V^2} \quad (2)$$

where T is temperature, T_d is dew-point temperature, U is the horizontal wind component (U-wind) and V is the meridional wind component (V-wind). For precipitation and wind speed, the scaled log transform was applied as mentioned above (Figures S1e, and f). Furthermore, we considered geographical factors such as elevation from ETOPO1 (Amante and Eakins, 2009) with a 1 arc-minute resolution to integrate the land topography and bathymetry and land cover classifications from GlobCover (Arino, 2010). These were categorized into 22 types based on observations from the ENVISAT satellite mission for 2009 with a spatial resolution of approximately 300 m.

3. Methodology

3.1. TabNet

TabNet is a novel neural network architecture designed to provide an adequate tabular dataset (Arık and Pfister, 2021). Based on an encoder/decoder structure, high-dimensional features can be transformed into a meaningful representation through trainable embedding layers without any pre-processing steps. For instance, the layers can map categorical features into a numerical format as well as handle raw numerical features without normalizing global features. One salient strategy of the TabNet is to employ the sequential attentive transformer architecture to select the importance features in decision steps. In each step, learnable masks search for a subset of the relevant features by quantifying the contribution of the decision.

3.2. Interpretability

The feature attribution mask $\mathbf{M} \in \mathbb{R}^{B\times D}$ provides instance-wise interpretable insights for reasoning; B is the batch size and D is the dimension of the feature. At the i^{th} decision step, the processed features from the preceding step $\mathbf{a}[i\text{-}1]$ are given to a trainable nonlinear processing h_i, composed of a fully connected layer, batch normalization and gated linear unit (Dauphin et al., 2017). The mask is obtained through a sparse regulation function, which we set using *entmax* (Peters et al., 2019), as summarized in Eq. (3):

$$M[i] = entmax\big(P[i-1] * h_i(a[i-1])\big) \tag{3}$$

$\mathbf{P}[i]$ is the prior scale term to regulate the flexibility of feature selection in the multiple steps, as defined in Eq. (4):

$$P[i] = \prod_{j=1}^{i} (\gamma - M[j]) \quad (4)$$

where γ is the coefficient for the feature reselection in the mask. $\mathbf{P}[0]$ is initialized as all ones, $\mathbf{1}^{B \times D}$, indicating that none of the features are used at the beginning. As a feature is considered thoroughly, its scale term is reduced to focus on the other features in the next steps. The weights of the trained mask represent the relative importance of each step in all instances. For example, if $\mathbf{M}_{b,j}[i]$=0, then the j^{th} feature should have no decision contribution in the i^{th} step for the b^{th} sample. Finally, the aggregated weights from the masks allow us to understand the importance of each feature in terms of its global behavior.

3.3. Training details

The weather in Thailand has distinct seasonality; the rainy season, which usually lasts from June to October, can significantly affect the $PM_{2.5}$ concentration in the atmosphere (Figure S2). Moreover, the mapping of averaged $PM_{2.5}$ displays a higher concentration in the northern area, above 40, than elsewhere (Figure 1b). Considering these spatiotemporal characteristics, we added the observed month and geographical coordinates (longitude and latitude) of the station as input features. In total, 19 input variables were used in this study: NO_2, SO_2, O_3, CO, HCHO, temperature, dew-point temperature, relative humidity, U-wind, V-wind, wind speed, precipitation, pressure, evaporation, elevation, land cover type, month, longitude and latitude. For the categorical variables, namely, month and land cover type, we set the embedding dimensions to 6 and 17, respectively.

Following convention, we randomly split the data from 2018 to 2020 into training and testing datasets using an 80:20 ratio; the number of samples were 14,069 and 3518, respectively. We also evaluated the functionality of upscaled mapping using a 10 km resolution grid format of the input dataset for 2021. To ensure robust training, a 5-fold cross-validation was set, and the final $PM_{2.5}$ estimation was calculated by averaging the results from the five trained models. The model was implemented using the *pytorch_tabnet* package (https://github.com/dreamquark-ai/tabnet) and trained with the Adam algorithm with weight decay using a 0.01 learning rate and a batch size of 64. Following the guidelines for hyperparameters (Arık and Pfister, 2021), we set the depth and width of TabNet as follows: N_d=N_a=24, N_{steps}=4, γ=1.3 and λ_{sparse}=0.001.

4. Results

4.1. Evaluation of general model performance

Figure 3 presents the accuracy validation results of the estimated $PM_{2.5}$ concentration for Thailand and the five divided regions. For the entire study domain (Figure 3a), three evaluation metrics show 0.873 of R^2, 9.22 of root mean square error (RMSE) and 20.62 of the mean absolute percentage error (MAPE). When these results are compared with other state-of-the-art ML algorithms, R^2 and RMSE of the proposed method show the best scores (Table 1). In terms of the linear relationship between the observations and estimated $PM_{2.5}$ concentrations, all the models show slope coefficient values under 1. These results imply that the ML models tend to underestimate the $PM_{2.5}$ concentration, as is consistently reported in previous studies (Ma et al., 2016; Wei et al., 2019). TabNet can compensate for this bias, as it has the highest value of the slope coefficient (0.84). This improvement is especially noticeable in the extremely high concentration cases of more than 300 μg/m^3 (Figure S3).

When the scores of evaluation metrics are compared by region, the highest value of R^2 (0.884) is observed in the north (Figure 3b). These results are consistent with the mapping of R^2 for each station showing higher than 0.8 of R^2 in all the stations in the north, including the Chiang Mai and Lampang provinces (Figure S4a). On the other hand, the scale of biases is larger than other regions with 13.44 of RMSE, due to its wider range of the $PM_{2.5}$ concentration exceeding 300 μg/m^3 as a maximum (Figures 3b and S4b). Given that the $PM_{2.5}$ concentrations in the north are generally higher (Figure 1b) and extreme cases are more frequent due to agricultural burnings and forest fires (Punsompong et al., 2021), the large errors are typically caused by the underestimation mentioned previously, particularly for high concentration cases. When the regional differences in scale are diminished by considering the ration of the scale between the errors and actual values, some stations in Bangkok and neighbor cities show higher scores of MAPE (Figure S4c). But the south region shows the lowest accuracy with 21.51 of MAPE and 0.507 of R^2 (Figure 3f). The distinctively low slope coefficient in the south represents that its poor performance is mainly caused by underestimation (Figure S4d). Considering that the air quality of southern Thailand is influenced by pollutants from peatland fires in Indonesia during the southwest monsoon (Mahasakpan et al., 2023), our model seems to have limits to estimate air mass transportation from out of the study domain.

4.2. Application on high-coverage mapping

One of the main purposes of employing remote sensing data is to enlarge the spatial coverage of $PM_{2.5}$ monitoring. Figure 4 illustrates the monthly averaged results of the $PM_{2.5}$ estimation for 2021. The mapping results (Figures 4a–f) generally agree with the observations (Figures 4g–l) with respect to seasonal variation by region. In January, the central region of Thailand shows high levels of $PM_{2.5}$ concentrations. In the north, the concentrations significantly increase from January and peak at over 60 μg/m^3 in March. The regional time difference in the peak of air pollutants can be explained by the fact that harvesting and residue burning are carried out in a different season in each region (Kanabkaew and Kim Oanh, 2011). Thereafter, the concentrations decrease in all the regions as the rainy season approaches.

To evaluate the temporal variation of the $PM_{2.5}$, we compare the daily variations in the observed and estimated $PM_{2.5}$ over the five regions of Thailand (Figure 5). The northern area shows the highest performance scores (0.83, 12.58 and 19.81 for R^2, RMSE and MAPE, respectively). The value of slope coefficient is almost 1 representing a significant improvement in the underestimation for extreme levels of $PM_{2.5}$, with high accuracy for peak days during March and April. The other regions, except for the south, also show good performances according to the evaluation metrics. Although the south region has smaller scale of error (4.69 of RMSE), the underestimation on high concentration days, particularly on those days with values above 60 μg/m^3, has scope for further exploration and improvement for long-range transport effects across neighboring countries.

5. Discussion

5.1. Model interpretation

Interpretability makes it possible for us to understand model's behavior at each learning steps and to point out important processes, which can be translated into more practical way. However, there isn't a perfect method to interpret ML and DL approaches, which is well recognized as a potential limitation. A major advantage of the TabNet is its attentive transformer structure, which provides post-hoc explanations by assessing the contribution of each feature from both global and local perspectives. First, the global importance of each feature is illustrated in Figure 6. The observed month displays the highest ratio of contribution with approximately 40% of importance, which is expected according to the seasonality of $PM_{2.5}$ in Thailand (see Figure S2). Geographical features such as land cover type, coordination and elevation follow next, demonstrating their importance. In terms of chemical components, NO_2 and SO_2, which are commonly known as precursors in the secondary formation of $PM_{2.5}$ (Baker and Scheff, 2007; Tucker, 2000), rank relatively low among all the features; SO_2 shows almost zero contribution to the estimation. Instead, CO accounts for about 20% of the contribution. Considering that CO is a by-product of carbon-containing fuel combustion, these results agree with the scenario that vehicular emissions and fires have a greater impact on the variation in air quality in Thailand than industrial emissions (ChooChuay et al., 2020).

Figure S5 illustrates the top five important features on each decision step as the aspect of local feature importance. Consistent with the global perspective, the observed month, CO and land cover type are ranked as the most determining factors in all the steps, regardless of season. Interestingly, the second step displays different composition of importance, especially for meteorological features, by season. The importance of wind speed and relative humidity are relatively lower for dry season ranking fourth and fifth (Figure S5f), while they are selected as the second and third most important features in wet season (Figure S5j). Some other meteorological factors, such as pressure, evaporation and dew-point temperature, are also displayed in other steps, in spite of their low contribution (less than 5%). Considering that windy and humid weather can reduce pollution levels, the trained model locally employs weather information to identify the ideal conditions for lower $PM_{2.5}$ concentrations.

5.2. Impacts of fire on PM2.5 concentration in Thailand

To investigate the impact of fire on the air quality in Thailand, we analyze spatial distribution of fire radiative power (FRP) from the Global Fire Assimilation System (GFAS) in the Copernicus Atmosphere Monitoring Service (CAMS) and chemical components (Figure 7) for a period when all the sub-regions show the rise of PM2.5 concentration (from February 25th to March 2nd 2021, red columns in Figure 5). During this period, high levels of FRP were mainly observed in the central region and the border areas in the north and east adjacent to neighboring countries, such as Myanmar, Laos and Cambodia. The concentrations of $PM_{2.5}$ and major chemical components also increased nearby fire hotspots, especially when the highest FRP was observed in the central west of Thailand on March 1st. This causal relationship between FRP and the concentrations can be seen in all the regions during the dry season for the year 2021. FRP shows statistically significant correlations with $PM_{2.5}$ in the north, northeast and east regions as well as with other chemical components in the north (Figures S6-10). Considering that those regions generally have higher concentrations of $PM_{2.5}$, the results demonstrate that the frequency and duration of fire can significantly influence on the level of air quality in Thailand. Besides, fires in the neighbor countries can also be another factor to cause considerable increase of $PM_{2.5}$ concentration. For instance, when many hotspots were detected in the territory of Myanmar and Laos on February 26th and Marth 1st, the chemical and $PM_{2.5}$ concentrations distinctly increased in the north and northeast parts of Thailand the next day.

5.3. Potential further improvements

Traditionally, aerosol optical depth (AOD) has been played as an essential factor to estimate the surface level of $PM_{2.5}$ concentrations. However, the presence of cloud and snow along with the limited vertical resolution causes unfeasibility for archive reliable AOD limiting its spatial coverage (Hsu et al., 2013; Levy et al., 2007). Our approach based on atmospheric gas composition offers a viable alternative to address the spatial limits. We also tested the skill of estimation including aerosol index and its results do not show any clear difference in the metric scores (Figure S11), proving that $PM_{2.5}$ concentration can be accurately estimated only with atmospheric trace gases. Although the spatial constraints still remain in this study due to excluding data below the quality threshold, future work will focus on the development of a model to handle the low-quality data aiming to achieve reliable full coverage for $PM_{2.5}$ estimation.

While there have been prior attempts to apply DL-based modeling to estimate $PM_{2.5}$, its performance is lower than that of other algorithms (Chen et al., 2022; Pu and Yoo, 2021; Wong et al., 2021). Importantly, these results could be linked to their simple model structures, which mostly consist of a series of fully connected hidden layers with nonlinear activation functions. In the current study, we showcase how by fusing TROPOMI data with other geospatial sources and incorporating an advanced DL algorithm to provide an accurate representation of $PM_{2.5}$ concentration; consequently, an air pollution indicator can be developed. Previous studies have reported the potential of DL algorithms such as CNN and LSTM to improve estimation performance (Chen et al., 2021; Lu et al., 2021), and our results also support this by adopting a state-of-the-art DL algorithm. Numerous advanced DL methods have recently been developed and have achieved remarkable progress in several fields (Devlin et al., 2018; He et al., 2016); however, applying DL to estimate $PM_{2.5}$ concentration has not yet been widely explored. Thus, monitoring air quality by implementing DL approaches has considerable room for improvement.

6. Conclusion

To estimate ground-level $PM_{2.5}$ concentration across Thailand, we develop a novel method based on DL algorithm, namely TabNet, with atmospheric gas from the TROPOMI on the Sentinel-5P. Our model shows more robust performance than other state-of-the-art ML algorithms, with an R^2 and RMSE (MAPE and slope coefficient) of 0.873 and 9.22 (20.62 and 0.84), respectively. The interpretable decision processes in TabNet indicate that monthly variation is the most significant feature in $PM_{2.5}$ estimation. Geospatial characteristics such as land cover type and latitude also provide a notable contribution from a global perspective. Among the chemical components from the TROPOMI, CO shows higher ratio of importance than the others. These results suggest that emissions from biomass burning influence air quality in Thailand considerably. In the low-level $PM_{2.5}$ concentration scenarios, humid and windy weather conditions are also highlighted in the local decision processes. Based on its robust performance, the model is applied to generate grid format mapping of $PM_{2.5}$ concentrations. We find that it can capture the temporal variation in and uneven spatial distribution of $PM_{2.5}$ concentrations using a 10 km grid. The enhanced estimation ability and its application are expected to not only boost other air quality studies, but also contribute to air quality management by providing advanced monitoring and evaluation techniques.

Acknowledgements

This research was conducted as a part of the Int'l Meteorological Expert Training Program supported by the Korea Meteorological Administration.

References

Amante, C., Eakins, B.W., 2009. ETOPO1 arc-minute global relief model: procedures, data sources and analysis.

Arino, O., 2010. GlobCover 2009.

Arık, S.O., Pfister, T., 2021. Tabnet: Attentive interpretable tabular learning, in: AAAI. pp. 6679–6687.

Baker, K., Scheff, P., 2007. Photochemical model performance for PM2. 5 sulfate, nitrate, ammonium, and precursor species SO2, HNO3, and NH3 at background monitor locations in the central and eastern United States. Atmos. Environ. 41, 6185–6195.

Borsdorff, T., Aan de Brugh, J., Hu, H., Aben, I., Hasekamp, O., Landgraf, J., 2018. Measuring carbon monoxide with TROPOMI: First results and a comparison with ECMWF-IFS analysis data. Geophys. Res. Lett. 45, 2826–2832.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. D (The Stat. 47, 431–443.

Chen, B., Song, Z., Pan, F., Huang, Y., 2022. Obtaining vertical distribution of PM2. 5 from CALIOP data and machine learning algorithms. Sci. Total Environ. 805, 150338.

Chen, B., You, S., Ye, Y., Fu, Y., Ye, Z., Deng, J., Wang, K., Hong, Y., 2021. An interpretable self-adaptive deep neural network for estimating daily spatially-continuous PM2. 5 concentrations across China. Sci. Total Environ. 768, 144724.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L.D., Abramson, M.J., Guo, Y., 2018. Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. Environ. Pollut. 242, 605–613.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Chen, Z.-Y., Zhang, T.-H., Zhang, R., Zhu, Z.-M., Yang, J., Chen, P.-Y., Ou, C.-Q., Guo, Y., 2019. Extreme gradient boosting model to estimate PM2. 5 concentrations with missing-filled satellite data in China. Atmos. Environ. 202, 180–189.

ChooChuay, C., Pongpiachan, S., Tipmanee, D., Suttinun, O., Deelaman, W., Wang, Q., Xing, L., Li, G., Han, Y., Palakun, J., 2020. Impacts of PM2. 5 sources on variations in particulate chemical compounds in ambient air of Bangkok, Thailand. Atmos. Pollut. Res. 11, 1657–1667.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., 2016. A review on predicting ground PM2. 5 concentration using satellite aerosol optical depth. Atmosphere (Basel). 7, 129.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks, in: International Conference on Machine Learning. PMLR, pp. 933–941.

de Hoogh, K., Héritier, H., Stafoggia, M., Künzli, N., Kloog, I., 2018. Modelling daily PM2. 5 concentrations at high spatio-temporal resolution across Switzerland. Environ. Pollut. 233, 1147–1154.

De Smedt, I., Theys, N., Yu, H., Danckaert, T., Lerot, C., Compernolle, S., Van Roozendael, M., Richter, A., Hilboll, A., Peters, E., 2018. Algorithm theoretical baseline for formaldehyde retrievals from S5P TROPOMI and from the QA4ECV project. Atmos. Meas. Tech. 11, 2395–2426.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Prepr. arXiv1810.04805.

Garane, K., Koukouli, M.-E., Verhoelst, T., Lerot, C., Heue, K.-P., Fioletov, V., Balis, D., Bais, A., Bazureau, A., Dehn, A., 2019. TROPOMI/S5P total ozone column data: global ground-based validation and consistency with other satellite missions. Atmos. Meas. Tech. 12, 5263–5287.

Ghahremanloo, M., Lops, Y., Choi, Y., Yeganeh, B., 2021. Deep Learning Estimation of Daily Ground-Level NO2 Concentrations From Remote Sensing Data. J. Geophys. Res. Atmos. 126, e2021JD034925.

Gurjar, B.R., Molina, L.T., Ojha, C.S.P., 2010. Air pollution: health and environmental impacts. CRC press.

Han, S., Kundhikanjana, W., Towashiraporn, P., Stratoulias, D., 2022. Interpolation-Based Fusion of Sentinel-5P, SRTM, and Regulatory-Grade Ground Stations Data for Producing Spatially Continuous Maps of PM2. 5 Concentrations Nationwide over Thailand. Atmosphere (Basel). 13, 161.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hsu, N.C., Jeong, M., Bettenhausen, C., Sayer, A.M., Hansell, R., Seftor, C.S., Huang, J., Tsay, S., 2013. Enhanced Deep Blue aerosol retrieval algorithm: The second generation. J. Geophys. Res. Atmos. 118, 9296–9315.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944.

Jiang, M., Sun, W., Yang, G., Zhang, D., 2017. Modelling seasonal GWR of daily PM2. 5 with proper auxiliary variables for the Yangtze River Delta. Remote Sens. 9, 346.

Jiang, Q., Christakos, G., 2018. Space-time mapping of ground-level PM2. 5 and NO2 concentrations in heavily polluted northern China during winter using the Bayesian maximum entropy technique with satellite data. Air Qual. Atmos. Heal. 11, 23–33.

Junpen, A., Garivait, S., Bonnet, S., 2013. Estimating emissions from forest fires in Thailand using MODIS active fire product and country specific data. Asia-Pacific J. Atmos. Sci. 49, 389–400.

Kanabkaew, T., Kim Oanh, N.T., 2011. Development of spatial and temporal emission inventory for crop residue field burning. Environ. Model. Assess. 16, 453–464.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Kloog, I., Nordio, F., Coull, B.A., Schwartz, J., 2012. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2. 5 exposures in the Mid-Atlantic states. Environ. Sci. Technol. 46, 11913–11921.

Levy, R.C., Remer, L.A., Mattoo, S., Vermote, E.F., Kaufman, Y.J., 2007. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. J. Geophys. Res. Atmos. 112.

Liu, Y., Park, R.J., Jacob, D.J., Li, Q., Kilaru, V., Sarnat, J.A., 2004. Mapping annual mean ground-level PM2. 5 concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States. J. Geophys. Res. Atmos. 109.

Lu, X., Wang, J., Yan, Y., Zhou, L., Ma, W., 2021. Estimating hourly PM2. 5 concentrations using Himawari-8 AOD and a DBSCAN-modified deep learning model over the YRDUA, China. Atmos. Pollut. Res. 12, 183–192.

Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., Liu, Y., 2016. Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004–2013. Environ. Health Perspect. 124, 184–192.

Mahasakpan, N., Chaisongkaew, P., Inerb, M., Nim, N., Phairuang, W., Tekasakul, S., Furuuchi, M., Hata, M., Kaosol, T., Tekasakul, P., 2023. Fine and ultrafine particle- and gas-polycyclic aromatic hydrocarbons affecting southern Thailand air quality during transboundary haze and potential health effects. J. Environ. Sci. 124, 253–267.

Muñoz Sabater, J., 2019. ERA5-Land hourly data from 1981 to present. Copernicus Clim. Chang. Serv. Clim. Data Store 10.

Organization, W.H., 2016. Ambient air pollution: A global assessment of exposure and burden of disease.

Peters, B., Niculae, V., Martins, A.F.T., 2019. Sparse sequence-to-sequence models. arXiv Prepr. arXiv1905.05702.

Pinichka, C., Makka, N., Sukkumnoed, D., Chariyalertsak, S., Inchai, P., Bundhamcharoen, K., 2017. Burden of disease attributed to ambient air pollution in Thailand: A GIS-based approach. PLoS One 12, e0189909.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.

Pu, Q., Yoo, E.-H., 2021. Ground PM2. 5 prediction using imputed MAIAC AOD with uncertainty quantification. Environ. Pollut. 274, 116574.

Punsompong, P., Pani, S.K., Wang, S.-H., Pham, T.T.B., 2021. Assessment of biomass-burning types and transport over Thailand and the associated health risks. Atmos. Environ. 247, 118176.

Rodríguez-Urrego, D., Rodríguez-Urrego, L., 2020. Air quality during the COVID-19: PM2. 5 analysis in the 50 most polluted capital cities in the world. Environ. Pollut. 266, 115042.

Shin, M., Kang, Y., Park, S., Im, J., Yoo, C., Quackenbush, L.J., 2020. Estimating ground-level particulate matter concentrations using satellite-based data: a review. GIScience Remote Sens. 57, 174–189.

Sorek-Hamer, M., Strawa, A.W., Chatfield, R.B., Esswein, R., Cohen, A., Broday, D.M., 2013. Improved retrieval of PM2. 5 from satellite data products using non-linear methods. Environ. Pollut. 182, 417–423.

Stratoulias, D., Nuthammachot, N., 2020. Air quality development during the COVID-19 pandemic over a medium-sized urban area in Thailand. Sci. Total Environ. 746, 141320.

Theys, N., De Smedt, I., Yu, H., Danckaert, T., van Gent, J., Hörmann, C., Wagner, T., Hedelt, P., Bauer, H., Romahn, F., 2017. Sulfur dioxide retrievals from TROPOMI onboard Sentinel-5 Precursor: algorithm theoretical basis. Atmos. Meas. Tech. 10, 119–153.

Tucker, W.G., 2000. An overview of PM2. 5 sources and control strategies. Fuel Process. Technol. 65, 379–392.

Van Donkelaar, A., Martin, R. V, Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ. Health Perspect. 118, 847–855.

Van Geffen, J., Eskes, H.J., Boersma, K.F., Maasakkers, J.D., Veefkind, J.P., 2019. TROPOMI ATBD of the total and tropospheric NO2 data products. DLR Doc.

Vassanadumrongdee, S., Matsuoka, S., 2005. Risk perceptions and value of a statistical life for air pollution and traffic accidents: evidence from Bangkok, Thailand. J. Risk Uncertain. 30, 261–287.

Wang, Y., Yuan, Q., Li, T., Tan, S., Zhang, L., 2021. Full-coverage spatiotemporal mapping of ambient PM2. 5 and PM10 over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions. Sci. Total Environ. 793, 148535.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution PM2. 5 concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221.

Weichenthal, S.A., Godri Pollitt, K., Villeneuve, P.J., 2013. PM2. 5, oxidant defence and cardiorespiratory health: a review. Environ. Heal. 12, 1–8.

Wong, P.-Y., Lee, H.-Y., Chen, Y.-C., Zeng, Y.-T., Chern, Y.-R., Chen, N.-T., Lung, S.-C.C., Su, H.-J., Wu, C.-D., 2021. Using a land use regression model with machine learning to estimate ground level PM2. 5. Environ. Pollut. 277, 116846.

World Air Quality Report 2020 [WWW Document], n.d. URL https://www.greenpeace.org/static/planet4-romania-stateless/2021/03/d8050eab-2020-world_air_quality_report.pdf

World Economic Outlook (October 2021) [WWW Document], n.d. URL https://www.imf.org/external/datamapper/datasets/WEO (accessed 2.7.22).

Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., Bai, Y., 2015. Daily estimation of ground-level PM2. 5 concentrations over Beijing using 3 km resolution MODIS AOD. Environ. Sci. Technol. 49, 12280–12288.

Xue, T., Zheng, Y., Tong, D., Zheng, B., Li, X., Zhu, T., Zhang, Q., 2019. Spatiotemporal continuous estimates of PM2. 5 concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. Environ. Int. 123, 345–357.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., Wang, W., 2016. National-scale estimates of ground-level PM2. 5 concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. Remote Sens. 8, 184.

Zhang, X., Han, Liangxiu, Han, Lianghao, Zhu, L., 2020. How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery? Remote Sens. 12, 417.

Zhu, L., Huang, L., Fan, L., Huang, J., Huang, F., Chen, J., Zhang, Z., Wang, Y., 2020. Landslide susceptibility prediction modeling based on remote sensing and a novel deep learning algorithm of a cascade-parallel recurrent neural network. Sensors 20, 1576.

Zou, B., Chen, J., Zhai, L., Fang, X., Zheng, Z., 2017. Satellite based mapping of ground PM2. 5 concentration using generalized additive modeling. Remote Sens. 9, 1.