Article

Interpolation-Based Fusion of Sentinel-5P, SRTM, and Regulatory-Grade Ground Stations Data for Producing Spatially Continuous Maps of PM_{2.5} Concentrations Nationwide over Thailand

Shinhye Han ^{1,2} , Worasom Kundhikanjana ³, Peeranan Towashiraporn ^{1,4} and Dimitris Stratoulas ^{1,4,*}

¹ Asian Disaster Preparedness Center, SM Tower, 24th Floor, 979/69 Paholyothin Road, Samsen Nai Phayathai, Bangkok 10400, Thailand; gkstlsgp3as@ewhain.net (S.H.); peeranan@adpc.net (P.T.)

² Environmental Science and Engineering, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea

³ United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP), Rajadamnern Nok Avenue, Bangkok 10200, Thailand; worasom@alumni.stanford.edu

⁴ SERVIR-Mekong, SM Tower, 24th Floor, 979/69 Paholyothin Road, Samsen Nai Phayathai, Bangkok 10400, Thailand

* Correspondence: dimitris@adpc.net; Tel.: +66-2298-0681

Citation: Han, S.; Kundhikanjana, W.; Towashiraporn, P.; Stratoulas, D. Interpolation-Based Fusion of Sentinel-5P, SRTM, and Regulatory-Grade Ground Stations Data for Producing Spatially Continuous Maps of PM_{2.5} Concentrations Nationwide over Thailand. *Atmosphere* **2022**, *13*, 161. <https://doi.org/10.3390/atmos13020161>

Academic Editors: Yun Zhu, Jim Kelly, Jun Zhao, Jia Xing and Yuqiang Zhang

Received: 22 November 2021

Accepted: 14 January 2022

Published: 20 January 2022

Corrected: 12 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Atmospheric pollution has recently drawn significant attention due to its proven adverse effects on public health and the environment. This concern has been aggravated specifically in Southeast Asia due to increasing vehicular use, industrial activity, and agricultural burning practices. Consequently, elevated PM_{2.5} concentrations have become a matter of intervention for national authorities who have addressed the needs of monitoring air pollution by operating ground stations. However, their spatial coverage is limited and the installation and maintenance are costly. Therefore, alternative approaches are necessary at national and regional scales. In the current paper, we investigated interpolation models to fuse PM_{2.5} measurements from ground stations and satellite data in an attempt to produce spatially continuous maps of PM_{2.5} nationwide over Thailand. Four approaches are compared, namely the inverse distance weighted (IDW), ordinary kriging (OK), random forest (RF), and random forest combined with OK (RFK) leveraging on the NO₂, SO₂, CO, HCHO, AI, and O₃ products from the Sentinel-5P satellite, regulatory-grade ground PM_{2.5} measurements, and topographic parameters. The results suggest that RFK is the most robust, especially when the pollution levels are moderate or extreme, achieving an RMSE value of 7.11 µg/m³ and an R² value of 0.77 during a 10-day long period in February, and an RMSE of 10.77 µg/m³ and R² and 0.91 during the entire month of March. The proposed approach can be adopted operationally and expanded by leveraging regulatory-grade stations, low-cost sensors, as well as upcoming satellite missions such as the GEMS and the Sentinel-5.

Keywords: spatial interpolation; PM_{2.5}; data fusion; machine learning; Sentinel-5P; air quality; inverse distance weighted; kriging; random forest

1. Introduction

Air quality has become a major global social concern as it is associated with health problems such as stroke, heart disease, lung cancer, and respiratory infection [1,2]. The World Health Organization (WHO) has listed six major related species, namely particulate matter, ground-level ozone, carbon monoxide (CO), sulfur oxides, nitrogen oxides, and lead. Among them, the fine particles with an aerodynamic diameter of less than 2.5 µm (PM_{2.5}) have become the center of interest due to their influence on visibility, human health, and global climate [3]. Several studies have even reported that PM_{2.5} contributes to the spread of viruses such as SARS-CoV-2, which is responsible for the COVID-19 disease [4–6].

In this respect, air quality monitoring and prediction has become a major field of interest in an effort to prevent the consequences caused by air pollution [1,7].

Therefore, countries affected heavily by pollution—such as Thailand which has undergone severe air pollution events due to rapid economic development and urbanization [8]—have developed monitoring systems to address the need for air pollution monitoring and exposure assessments. Three main types of related technology have been normally established. First, the regulatory-grade systems provide the most accurate and robust in-situ observations; nevertheless, they are financially expensive to install and maintain [9]. In order to overcome the consequent low coverage of the aforementioned monitoring stations, low-cost sensors have been deployed in various countries [10,11], of which the quality is still not competitive compared to standard air quality monitoring stations [12]. However, the low cost allows for the deployment of a number of such sensors. Lastly, satellite technology provides broad geographical coverage and consistent data acquisition, hampered only by certain weather conditions [13].

Notwithstanding, since surface PM concentrations measured by regulatory-grade systems are point-based and sparsely distributed, especially in rural areas [10,13], spatial interpolation and remote-sensing-based models to generate estimates of PM concentration in areas where ground stations are not found have been proposed. This approach has gathered attention considering that the data availability could be enhanced and the related costs for establishing additional monitoring stations can be avoided [14,15]. Moreover, interpolation-based estimations in areas with unavailable ground data is important as a support for hot spot and source identification, which is essential for efficient policy planning.

The interpolation of ground-level PM_{2.5} has been demonstrated widely with basic techniques such as the inverse distance weighted (IDW) and the ordinary kriging (OK), which are two commonly compared methods [16–19]. Many studies reckon that IDW generally showed poor performance as a non-geostatistical method, which simply defines spatial relationship by distance [20,21]. However, it is well-functioned as a control method to be compared with other approaches in various studies with recent attempts to estimate PM_{2.5} concentration from ground monitoring stations [18,22,23]. On the other hand, the kriging method is a geostatistical method that provides the best linear unbiased estimate by incorporating spatial correlation defined as empirical variogram [19,24,25]. Although OK has the disadvantages of being computationally demanding [26] and not being suitable for generating sufficient covariate information [24], it has recently been synthesized with RF or support vector machine (SVM) as an approach for interpolating residuals to achieve less sensitivity to data variance and provide geo-referenced features [27–30].

Random forest (RF) is one of the most frequently applied [22] tree-based machine learning (ML) models due to its advantages to achieve high accuracy [1] while handling non-linear relationships [31]. It also enables unbiased estimation of errors by constructing a multitude of trees with randomly chosen parameters and preventing learning from specific features [5]. Moreover, RF blended with OK (RFK) is considered as regression kriging (RK) and is often referred to as RFK or RFOK. The difference between RK and RFK is that RK uses a multiple linear regression (MLR) [29], while RFK is a hybrid method using RF for regression and OK for geostatistical modeling of residuals [22,27–29]. It can overcome the limitation of RF in that it does not account for geo-referenced data [25]. Past studies which have compared the accuracy of RF and RFK [22,30], or RK and RFK [29], concluded that the RFK presented a better ability to map non-linear complex relations.

The current study presents the comparative evaluation of four interpolation methods in an attempt to produce spatially continuous PM_{2.5} concentrations maps. Four interpolation algorithms (IDW, OK, RF, and RFK) are applied and compared, out of which IDW and OK derive estimates solely using regulatory-grade ground data while RF and RFK fuse regulatory-grade ground data, satellite observations, and topographic factors. In particular, we investigate the development of PM_{2.5} concentrations in two test sites with different sources and patterns of air pollution, namely Bangkok and Chiang Mai. The performance is judged using the metrics of the root-mean-squared error (RMSE), scatter index (SI), and

the coefficient of determination (R^2). We showcase the key features of RFK and advocate it as a suitable spatial interpolating technique for Thailand, which has a limited size and distribution of in-situ monitoring stations.

2. Materials and Methods

2.1. Study Area

Thailand is a country located at the center of the Indochinese Peninsula at the coordinates of 15.87° N latitude and 100.99° E longitude. Most regions in Thailand generally exhibit a tropical climate, especially in the south, but present distinct patterns based on seasonal variation and climatic zones. The seasons could be subdivided into the rainy season lasting from mid-May to mid-October, the winter season running from mid-May to mid-October, and the pre-monsoon season lasting from mid-February until mid-May.

From a topographical perspective, Thailand is divided into the northern, northeastern, central, eastern, and southern regions (Figure 1), which exhibit different seasonal and climatic patterns. As shown in Table 1, the northern part includes 15 provinces and exhibits mountainous topography and similar climatic features to the central region but receives lower precipitation. The northeastern region is a plateau with high elevations including 20 provinces. It represents the driest conditions among the regions, with the lowest precipitation and relative humidity. The central region is mainly a low-level plain encompassing 18 provinces, while the eastern part consists of plains and small valleys. These regions usually experience rainy, dry, and cool seasons, but are mostly warm because they are located in tropical inland latitudes. Meanwhile, the southern part has low variations in temperature and high relative humidity due to its maritime characteristics [32]. These seasonal variations have an influence on $PM_{2.5}$ concentrations, such as dilution of air pollutants with heavy rain or excessive air pollution due to temperature inversion or anthropogenic activities.

Figure 1. Study area with climatic zones subdivisions and locations of air quality ground stations (yellow points). Each bounding box covers the regions around Chiang Mai (red box; 98.90–100.24° E, 18.09–18.92° N) and Bangkok (blue box; 100.05–101.04° E, 13.45–14.29° N).

Table 1. Climatic and physical characteristics of climatic zones.

	Northern Region	Northeastern Region	Central Region	Eastern Region	Southern Region
Number of provinces	15	20	18	8	10
Prevalent topography	mountainous	A high-level plateau	A low-level plain	Plains and valleys	Peninsula
Average surface temperature * (°C)	23.4/28.1/27.3	24.2/28.6/27.6	26.2/29.7/28.2	26.7/29.1/28.3	26.3/28.2/27.8
Precipitation * (mm)	100.4/187.3/943.2	76.3/224.4/1103.8	127.3/205.4/942.5	178.4/277.3/1433.2	827.9/229.0/680.0
Relative Humidity * (%)	74/63/81	69/66/80	70/68/78	71/75/81	81/78/79

* Aggregation of values based on three seasons: Winter/Summer/Rainy seasons.

Due to the complexity and large size of the country, in the current study, we concentrate on two regions that exhibit heavy pollution events during different times of the year and are attributed to different sources. Bangkok, the capital and most populous city of Thailand, has been developing rapidly as a metropolitan city, and vehicular usage—which is known as an important source of PM_{2.5}—has become excessive. Correspondingly, the concentration of PM_{2.5} over Bangkok tends to follow the daily pattern of vehicular use, which reaches a maximum during ordinary rush hours. Conversely, the region around Chiang Mai province is a predominantly agricultural area where the combustion of agricultural biomass as a form of farm management takes place in the first months of each year. Consequently, this major source contributes to excessive pollution levels in Northern Thailand during winter.

2.2. Data

Data associated with PM_{2.5} concentrations from different sensors, ground and remote alike, were identified and used synergistically in the current study to derive a country-wide representation of spatial continuous maps. More specifically, ground data from the Pollution Control Department (PCD) (<http://air4thai.pcd.go.th/webV2/region.php?region=0>, accessed on 11 March 2021) in Bangkok, Thailand, remotely sensed images from the Sentinel-5 Precursor (Sentinel-5P) satellite, and the digital elevation model (DEM) product from the Shuttle Radar Topography Mission (SRTM) were the main data sources.

The PCD has established a country-wide monitoring network for air quality management consisting of 68 stations as of 2020. After the elimination of corrupted data, a total of 63 out of 68 stations were taken into consideration (Figure 1). The monitoring stations of PCD have provided hourly data of PM_{2.5} (µg/m³), PM₁₀ (µg/m³), O₃ (parts per billion (ppb)), CO (parts per million (ppm)), NO₂ (ppb), and SO₂ (ppb) at 2 m above the ground at hourly intervals; PM_{2.5} was the main attribute used in the current study.

Corresponding satellite data were retrieved from the TROPOspheric Monitoring Instrument (TROPOMI) onboard the Copernicus Sentinel-5P satellite, a mission (<https://s5phub.copernicus.eu/dhus/#/home>, accessed on 24 June 2021) dedicated to monitoring the atmosphere in respect to air quality monitoring, ozone and UV radiation, climate monitoring, and forecasting [33]. The data inherit a high spatial (3.5 × 5.5 km) and temporal (daily) resolution and the usability of these products for monitoring air pollution has been largely demonstrated in the literature e.g., [3,34–41].

Lastly, a secondary source of remotely sensed data was used to retrieve elevation. The DEM data obtained from the SRTM mission (<https://www2.jpl.nasa.gov/srtm/>, accessed on 18 August 2021) were used for this purpose as in similar studies [42–44]. The SRTM was a short 11-day mission during which near-global high-resolution topography data was collected and organized by the National Geospatial-Intelligence Agency (NGA) and the National Aeronautics and Space Administration (NASA). The spatial resolution of the product used in the current study was 90 m.

2.3. Methodology

The overall procedure for estimating ground-level PM_{2.5} is described in Figure 2. The steps consist of (1) data collection and reprojection in order to bring the data sources in

the same coordinate system and time window, (2) implementation of spatial interpolation models on defined scenarios, and (3) model validation and mapping of PM_{2.5} estimates.

Figure 2. Processing workflow of PM_{2.5} spatial interpolation via four models.

2.3.1. Data Collection and Reprojection

The current study integrates the tropospheric NO₂ column, tropospheric O₃ column, total vertical SO₂ column, tropospheric formaldehyde (HCHO) column, total CO column, and aerosol index (AI) layers since several studies have addressed that the TROPOMI atmospheric products are advantageous due to the lower missing proportion in data compared to other AOD products, especially to MODIS which is widely used in estimating ground-level PM_{2.5} concentrations [45–47]. The *p*-values of the correlation of each layer with the ground PM_{2.5} are provided in Table 2. This investigation identifies the most suitable layers based on the correlation coefficient to input into the machine learning models.

Table 2. Pearson correlation coefficients and associated *p*-values for assessing the correlation of variables with PM_{2.5} in respective scenarios.

PCD	Sentinel-5P						SRTM
Correlation with PM _{2.5}	O ₃	SO ₂	NO ₂	HCHO	AI	CO	DEM
8 February–18 February	0.56 ***	0.09 ns	−0.09 ns	0.62 **	0.49 ***	0.74 ***	0.53 ***
1 February–29 February	0.72 ***	0.08 ns	0.10 ns	0.67 ***	0.43 ***	0.72 ***	0.35 **
19 March–29 March	0.36 **	0.22 ns	0.10 ns	0.79 ***	0.88 ***	0.89 ***	0.73 ***
1 March–31 March	0.34 **	0.28 ns	−0.04 ns	0.79 ***	0.87 ***	0.87 ***	0.72 ***
9 July–19 July	0.16 ns	−0.20 ns	0.24 *	0.16 ns	0.27 *	0 ns	−0.34 **
1 July–30 July	0.08 ns	0.02 ns	0.24 *	0.21 ns	0.26 *	−0.15 ns	−0.33 **
5 December–15 December	−0.04 ns	0 ns	0.63 ***	0.65 ***	0.64 ***	0.75 ***	−0.30 *
1 December–31 December	−0.20 ns	0.12 ns	0.61 ***	0.76 ***	0.49 ***	0.74 ***	−0.26 *

*** significant correlation at the 0.001 level, ** significant correlation at the 0.01 level, * significant correlation at the 0.05 level, ns non-significant correlation.

Thereafter, based on the defined time series, the six parameters of NO₂, SO₂, CO, HCHO, O₃, and AI from Sentinel-5P, and the DEM product from SRTM were extracted and

averaged for the model training. Regarding Sentinel-5P data, we selected CO and HCHO due to their similar emission source to that of PM_{2.5} [48–51], SO₂ and NO₂ as precursors for the PM_{2.5} [51–53], and O₃ due to its varying interaction with PM_{2.5} by seasons and regions [54–56]. Lastly, we also included the product of AI as an informative parameter about PM_{2.5} [57], since it captures signals of aerosols emitted from biomass burning or fire events [58], which is another source of particulate matter [59,60]. For the retrieval of data, we accessed the Copernicus Open Access Hub website (<https://scihub.copernicus.eu/>, accessed on 24 June 2021) to retrieve TROPOMI Level 2 products and compared the products with the corresponding pre-processed Google Earth Engine products. After confirming their consistencies, we utilized Google Earth Engine for fetching data with a spatial resolution of 3.5×5.5 km and averaging them based on the described time ranges. In order to remove cloud-contaminated scenes, the quality assurance (QA) value of 0.75 was applied as recommended by Zweers [58].

Additionally, with regard to the SRTM processing, the tiles covering the whole Thailand were downloaded from the Earth Explorer website, stitched together, and reprojected to the same coordinate system as the other data (i.e., the World Geodetic System (WGS) 1984 (EPSG 4326)). Finally, the mosaiced image was subset to the area of study (defined as the entire administrative area of Thailand) and spatially resampled to a 3.5×5.5 km grid in order to approximate the spatial resolution of the Sentinel-5P data and allow for the synergistic use of the two remotely sensed satellite products.

The whole dataset was split according to the settled time schemes, and the analysis of pollution patterns in Thailand was adopted to define the time window as shown in Figure 3. To compare the pollution trend between Bangkok and Chiang Mai, the hourly data from 63 ground stations were averaged to daily PM_{2.5} concentrations while grouped by districts. The different pollution sources and patterns for each region led to differing patterns observed; in February, a time window with moderate pollution is observed in both regions, in March and for most of the month there is excessive air pollution in Chiang Mai, while this is not the case for Bangkok. June to September is a long period with the lowest pollution for both regions, and lastly in December, moderate pollution existed in Bangkok while Chiang Mai had relatively lowered pollution. Thereafter, data from 8 February to 18 February, 19 March to 29 March, 7 July to 17 July, and 5 December to 15 December of the year 2020 were extracted and considered as four scenarios with the aforementioned 10-day and monthly averages as the time periods for the scenarios based on which the analysis took place. The total number of data used in this study was 182,952 from 63 stations and the number of layers utilized for the model training was seven for each scenario, which were six satellite images of air pollutants from Sentinel-5P, and the DEM.

Figure 3. Daily averaged PM_{2.5} concentrations over Bangkok (blue curve) and Chiang Mai (orange curve) for the year 2020. The light and dark grayed areas correspond to the monthly and 10-day time windows respectively selected in this study as representative of the four unique scenarios for the PM_{2.5} temporal distribution.

2.3.2. Spatial Interpolation Modeling

The current experiment considered two types of interpolation methods to produce spatially continuous data of PM_{2.5}, two basic interpolation techniques and two ML techniques. Among the four models, IDW is widely used for spatial interpolation, and simply calculates estimates using distance-based weights based on the formula below [61]

$$Z_{(x_0)} = \sum_{i=1}^n \lambda_i Z_{(x_i)} \quad (1)$$

where $Z_{(x_0)}$ is the estimated PM_{2.5} concentration at the target position x_0 , $Z_{(x_i)}$ is the PCD's ground-level PM_{2.5} data at sample location x_i , and i indicates the weighting value, which is calculated by the equation [61]

$$\lambda_i = \left[d_{(x_i, x_0)} \right]^p / \sum_{i=1}^n \left[d_{(x_i, x_0)} \right]^p \quad (2)$$

where d is the Euclidean distance between the location of the target and sample data, while p indicates a power parameter, which is set at 2 in this study.

Unlike the IDW interpolation method, the OK technique is based on statistical information from the spherical, exponential, Gaussian, and Matern covariance models. These options provide additional features in calculating weights, which could be derived from the semi-variogram. The variogram calculates values from all pairs of the sample locations, x_i and x_j , using the formula below [62]

$$\gamma_{(x_i, x_j)} = \frac{1}{2} \text{Var} \left[Z_{(x_i)} - Z_{(x_j)} \right] \quad (3)$$

where Var indicates the variance. Since the variogram is established on the basis of the assumption that the closer data tend to have higher correlations, and the four approximation expressions embedded in the R project could provide the correlation formula, we needed to select the formula that best represents the variogram. Finally, the chosen formula, which was a spherical model in this study, was utilized for estimating unknown PM_{2.5} concentrations.

Judging the major difference between OK and RFK, the RFK relies on the RF method, which belongs to a supervised machine learning algorithm based on decision trees [63]. Conversely, IDW and OK only utilize PM_{2.5} data from PCD, RFK and RF employed the whole dataset to apply the model for the spatial interpolation. The Sentinel-5P data and the DEM product at the sample location were extracted from each raster data to calculate the Pearson correlation coefficient as presented in Table 2, and the RF approach was implemented to appropriately produce interpolation results with the existence of unrelated covariables. Despite the fact that both approaches incorporated the same RF model ($N_{\text{tree}} = 500$ and $m_{\text{try}} = 2$, $\text{nodesize} = 5$, $\text{maxnodes} = \text{null}$; limited by nodesize) and used the whole stacked raster data as the input data, RFK required the additional process to correct estimates by applying the OK method on residuals between actual and approximate value.

2.3.3. Model Validation

The processed dataset was used as the input data for the four spatial prediction models considered in this study, namely the IDW, OK, RF, and RFK techniques. The dataset was split into two parts, out of which approximately 10% (6 out of 63 stations) was used as a testing set, by randomly extracting one station from northern, northeastern, eastern, and southern regions, and two stations from central regions based on their uneven spatial distribution. The remaining data corresponding to 90% (57 stations) was used as a training set, while the split was randomly implemented 10 times for 10-fold cross-validation [64–66]

to exclude the impact of extreme values in each division, especially in the case of adopting data with large variance.

Therefore, the interpolation results and their validation metrics based on test sets were averaged in the evaluation step. The metrics used for the assessment were the R^2 , the Pearson's correlation coefficient and associated p -values, and the SI, which is derived from the division of RMSE with the mean of $PM_{2.5}$ data multiplied by 100 to diminish the influence of the variance. The R statistical programming language [67] was used for the data analysis.

3. Results

3.1. Spatial Distribution Maps of $PM_{2.5}$ over Thailand

The $PM_{2.5}$ interpolation maps over Thailand from the four models are presented in Figures 4–7 for each distinct time window corresponding to the four scenarios. Each sub-figure demonstrates results based on aggregated data from 10 and 30 consecutive days for the months of February, March, July, and December. The ground-level observations of $PM_{2.5}$ are indicated as colored-coded point data representing concentrations based on the categorical thresholds set by PCD.

Figure 4. Interpolated $PM_{2.5}$ concentrations from different spatial interpolation methods (a–d) from 1–29 February and (e–h) from 8–18 February.

Figure 5. Interpolated $PM_{2.5}$ concentrations from different spatial interpolation methods (a–d) from 1–31 March and (e–h) from 19–29 March.

Figure 6. Interpolated $PM_{2.5}$ concentrations from different spatial interpolation methods (a–d) from 1–30 July and (e–h) from 9–19 July.

Figure 7. Interpolated $PM_{2.5}$ concentrations from different spatial interpolation methods (a–d) from 1–31 December and (e–h) from 5–15 December.

In February, a monthly dataset of which $PM_{2.5}$ concentration ranged from 11.06 to 86.02 $\mu\text{g}/\text{m}^3$, presented high concentration ($>50 \mu\text{g}/\text{m}^3$) around Mae Hong Son, and southern regions showed the lowest pollution under 25 $\mu\text{g}/\text{m}^3$, which is considered as very good air quality according to the criteria set by PCD. Additionally, the central and northeastern regions were moderately polluted according to Figure 4, demonstrating concentration from 26 to 37 $\mu\text{g}/\text{m}^3$, while RF and RFK estimated $PM_{2.5}$ concentration in northeastern areas over 38 $\mu\text{g}/\text{m}^3$. RF and RFK tended to illustrate a spatial variation of PM concentration over the regions, and IDW and OK showed zoning distributions around the monitoring stations. The 10-day dataset, which had minimum and maximum values of 8.21 and 79.69 $\mu\text{g}/\text{m}^3$ respectively, also accompanied similar spatial distribution as shown with the 10-day dataset, but with overall lower concentrations.

The dataset in March showed drastic concentration variability over Thailand, by having maximum and minimum values of 12.13 and 193.54 $\mu\text{g}/\text{m}^3$ respectively. Similar to the results in February, the southern region indicated the best air quality, quantified at under 25 $\mu\text{g}/\text{m}^3$, and northern regions were represented as hotspots of pollution, while their coverage became enlarged. The zoning distribution of IDW became excessive, and OK showed an oversimplified gradual increase toward the north. RF and RFK showed detailed spatial variation along with extreme hotspots in the southern and northern areas. Moreover, the pollution that occurred in the northeastern part of the country, as presented in Figure 5, is depicted with a more realistic representation by RF and RFK. The dataset for

10 days displayed a similar pattern to that of the monthly dataset, while the data ranges from 10.90 to 214.05 $\mu\text{g}/\text{m}^3$.

The dataset in July produced a rather uniform distribution among the predefined scenarios, which can be attributed to the small variance of the data ranging from 3.65 to 21.64 $\mu\text{g}/\text{m}^3$. The pollution pattern is stable as the country enters the monsoon season, which resulted in a uniform color map, and the eastern and central regions showed relatively higher concentrations of $\text{PM}_{2.5}$ than other regions. The dataset for 10 days ranged from 3.55 $\mu\text{g}/\text{m}^3$ to 21.62 $\mu\text{g}/\text{m}^3$, with a similar distribution for the monthly intervals. The missing data in the TROPOMI atmospheric products resulted in a few blank spaces in the interpolated maps produced from the RF and RFK model, which incorporated satellite images as covariates.

Moreover, in December the pollution was indicated with $\text{PM}_{2.5}$ concentrations over 37 $\mu\text{g}/\text{m}^3$ occurring around the central area, spreading out to the northern part. The data ranged from 5.04 $\mu\text{g}/\text{m}^3$ to 47.98 $\mu\text{g}/\text{m}^3$ for a month and from 5.74 $\mu\text{g}/\text{m}^3$ to 58.77 $\mu\text{g}/\text{m}^3$ for 10 days. Figure 7 showed relatively higher concentrations with Figure 7e–g. The same pattern was observed, with IDW and OK showing a distribution largely depending on the values of the monitoring stations, while RF and RFK presented a larger and heterogeneous spatial variability.

3.2. Ten-Fold Cross-Validation

We conducted 10-fold cross-validation to evaluate the model performance for each scenario and presented the results in Figure 8. In February, OK exhibited the lowest RMSE of 4.267 $\mu\text{g}/\text{m}^3$ and 5.362 $\mu\text{g}/\text{m}^3$ for 10 days and a month, and relatively high R^2 as 0.90 and 0.84 for 10 days and a month respectively, followed by RF. RF attained RMSE as 5.57 $\mu\text{g}/\text{m}^3$ and R^2 as 0.88 for the 10-day dataset, and RMSE as 6.34 $\mu\text{g}/\text{m}^3$ and R^2 as 0.85 for a monthly dataset. Moreover, in March, all models showed degraded performance in terms of RMSE due to the high variance of the data, while R^2 became the best performing among the four scenarios. A comparison between the models indicates that OK had the lowest RMSE of 9.27 $\mu\text{g}/\text{m}^3$ for 10 days and 6.91 $\mu\text{g}/\text{m}^3$ for a month, while RF and RFK showed relatively high R^2 of 0.91 and 0.9 for 10 days, and 0.93 and 0.91 for a month respectively.

Figure 8. Ten-fold cross-validation results of (a) RMSE, (b) R^2 , and (c) SI from different spatial interpolation methods for each scenario.

The dataset in July, which had very low $PM_{2.5}$ concentration and limited variation, demonstrated good performance in terms of RMSE which ranged from $3.16 \mu\text{g}/\text{m}^3$ to $3.65 \mu\text{g}/\text{m}^3$, and very poor performance in terms of R^2 which ranged from 0.15 to 0.25. However, RMSE could be highly influenced by the mean observation. Therefore, we integrated the SI to further validate the performance. With the objective of evaluating the validity of RMSE with respect to mean observation, we calculated the SI by dividing RMSE with the average of $PM_{2.5}$ concentrations. The SI excluded the misleading influence of the data average in July, demonstrating values higher than other scenarios, which indicates poor performance.

In December, even though R^2 was reduced to an average of around 0.6, the estimates generally matched the observed concentrations with RMSEs being 8.36, 8.51, 7.93, and $7.84 \mu\text{g}/\text{m}^3$ for IDW, OK, RF, and RFK respectively with a 10-day dataset. On the other hand, they were improved with a monthly dataset, reaching 6.90, 7.07, 6.93, and $6.91 \mu\text{g}/\text{m}^3$ for each model for the latter case. As a comparison between the models, RFK showed the highest R^2 and lowest RMSE, followed by RF.

3.3. Data Range of Interpolated Estimates

Figure 9 presents the box plots representing the distribution, minimum, and maximum values of interpolation results from each model. IDW and OK produced results with a range that is almost identical to that of the ground data. RF tended to generalize the variability of $PM_{2.5}$ concentrations, which was excessive with the dataset in March, by having $163.44 \mu\text{g}/\text{m}^3$ and $151.17 \mu\text{g}/\text{m}^3$ as a maximum when the observed value had a maximum value of $214.05 \mu\text{g}/\text{m}^3$ and $193.54 \mu\text{g}/\text{m}^3$. Lastly, RFK produced a broader range of data especially in March, demonstrating data range from 1.02 to $249.26 \mu\text{g}/\text{m}^3$ with a 10-day dataset, and from 2.17 to $255.28 \mu\text{g}/\text{m}^3$ with a monthly dataset, when the original was from 10.90 to 214.05 , and from 12.12 to 193.54 respectively. Overall, the 10-day dataset and monthly dataset produced a similar distribution of estimates.

3.4. Station-Based Comparison between Observations and Estimates from Test Sets Observing $PM_{2.5}$ Concentration Exceeding $50 \mu\text{g}/\text{m}^3$

Table 2 presents the results of the correlation analysis between $PM_{2.5}$ and covariates. The correlation with $PM_{2.5}$ was relatively high for the variables of HCHO and CO, while NO_2 and SO_2 showed lower correlations, which conformed with the fact that $PM_{2.5}$ [48,49], HCHO [50], and CO [51] are highly associated with fuel combustion. On the other hand, SO_2 and NO_2 , which are important precursors in the process of $PM_{2.5}$ formation [51–53], presented low or negative correlations with $PM_{2.5}$. O_3 manifested high coefficient values in February and March, and negative coefficients in December due to the reversed influence in different seasons [54–56]. Lastly, AI had a higher correlation with $PM_{2.5}$ especially in March, when agricultural biomass burning is prevalent since AI and $PM_{2.5}$ are influenced by the aerosol formation due to the agricultural practices prevailing in SE Asia during this season [44,57–60].

Since the RF and RFK are primarily based on the RF model, we derived significant features used for the models as shown in Figure 10. The importance is represented by the value of the mean decrease in MSE, which indicates the total decrease in model performance when each variable is neglected from the model. According to the results, during the most polluted period in March, overall feature importance was very high compared to other scenarios, with CO, AI, and HCHO being the most important factors to estimate $PM_{2.5}$ with the monthly dataset, whereas DEM surpassed HCHO with the 10-day dataset. SO_2 and NO_2 were the least influential throughout the scenarios for both datasets. In February, CO, O_3 , and NO_2 were influential parameters for a month, whereas CO, DEM, and NO_2 became more significant for the 10-day dataset. The feature importance in July and December indicated that most covariates were insignificant for the RF model. These results matched the correlation of covariates with $PM_{2.5}$ overall except NO_2 being one of the influential

parameters in March. In comparison with the results in Table 2, Figure 10 demonstrates that highly correlated parameters tend to be essential factors for the RF model.

Figure 9. Box plots of data distribution of observations and estimates from four models with (a) monthly and (b) 10-day datasets. IDW, OK, RF, and RFK indicates inverse distance weighted, ordinary kriging, random forest, and random forest blended with ordinary kriging, respectively.

3.5. Analysis of Important Features and Correlation of Covariates

Next, we analyzed the difference between observed and interpolated values to evaluate the model performance at the stations which recorded $\text{PM}_{2.5}$ concentrations over $50 \mu\text{g}/\text{m}^3$, which coincides with the threshold set by the PCD. The analysis was run for the test sets in February and March, when atmospheric pollution is at the highest levels. As a result, the selected test stations were distributed mainly in the northern and northeastern regions as shown in Figure 11. According to the results presented in Tables 3–6, IDW and OK had a strong tendency to underestimate values, with the error rate of IDW ranging from -23 to $+5\%$ for the monthly dataset, and -23 to $+6\%$ for the 10-days dataset, while that of OK ranged from -13 to $+8\%$ for the monthly dataset, and -22 to $+7\%$ for the 10-days

dataset. Despite the fact that both IDW and OK achieved a low error rate and RMSE, they still presented generally negative rates from -2 to -14% and -10 to $+3\%$ for the monthly and 10-day datasets, and -14 to $+2\%$ respectively.

Figure 10. Feature importance of RF model for (a) a monthly dataset and (b) 10-day dataset.

Figure 11. Location of stations used in station-based comparison with the dataset (a) in March and (b) in February.

Table 3. Test station-based comparison between observations and estimates from interpolation models with observed concentration over $50 \mu\text{g}/\text{m}^3$ from a monthly dataset in March.

March (Monthly)		IDW		OK		RF		RFK	
Station ID	Obs.	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)
39t	56.590	59.618	5.350	61.296	8.316	67.097	18.567	58.073	2.620
70t	94.528	75.322	−20.318	83.495	−11.672	115.955	22.667	82.959	−12.239
67t	68.601	71.376	4.045	73.554	7.220	75.356	9.846	88.475	28.970
46t	50.302	38.427	−23.607	43.719	−13.086	53.699	6.754	55.668	10.667
35t	91.029	79.408	−12.766	83.815	−7.925	83.575	−8.189	87.761	−3.590
38t	56.961	58.294	2.341	60.105	5.521	71.671	25.826	56.987	0.047
Average RMSE		10.51646		6.753721		12.23814		9.750932	

Table 4. Test station-based comparison between observations and estimates from interpolation models with observed concentration over $50 \mu\text{g}/\text{m}^3$ from a 10-day dataset in March.

March (10-Day)		IDW		OK		RF		RFK	
Station ID	Obs.	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)
39t	60.631	62.802	3.580	65.054	7.295	78.084	28.786	66.182	9.156
70t	95.967	80.808	−15.796	92.471	−3.642	120.197	25.249	115.140	19.979
67t	69.676	74.421	6.810	73.792	5.906	80.126	14.997	82.751	18.764
46t	50.302	38.427	−23.607	43.719	−13.086	53.699	6.754	55.668	10.667
35t	114.466	93.681	−18.158	89.176	−22.095	91.627	−19.953	108.798	−4.952
38t	60.884	61.079	0.321	57.831	−5.014	71.236	17.003	61.102	0.358
Average RMSE		11.76217		11.11301		16.53919		10.24966	

The results from the application of RF indicated relatively high error rates in March and low in February, with the lowest RMSE in February and the highest RMSE in March, which led to high variability in model performance. RFK, on the other hand, showed a reversed estimation tendency to IDW and OK, demonstrating positively biased ranges of -12 to $+28\%$ for the month dataset, and -4 to 19% for the 10-days datasets, with the lowest and the second-lowest RMSE occurring in March. Furthermore, RFK presented

improvements in error rates in February, ranging from -3 to 13% and 1 to 11% for the monthly and 10-day datasets respectively.

Table 5. Test station-based comparison between observations and estimates from interpolation models with observed concentration over $50 \mu\text{g}/\text{m}^3$ from a monthly dataset in February.

February (Monthly)		IDW		OK		RF		RFK	
Station ID	Obs.	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)
70t	60.775	52.245	-14.035	51.691	-14.947	55.539	-8.616	64.708	6.471
67t	57.262	51.190	-10.605	51.095	-10.771	59.107	3.221	64.708	13.003
35t	55.991	53.350	-4.717	54.028	-3.506	53.869	-3.791	53.898	-3.738
38t	51.162	50.028	-2.218	52.585	2.781	51.488	0.636	52.155	1.940
Average RMSE		5.429112		5.622299		2.976325		4.366747	

Table 6. Test station-based comparison between observations and estimates from interpolation models with observed concentration over $50 \mu\text{g}/\text{m}^3$ from a 10-day dataset in February.

February (10-Day)		IDW		OK		RF		RFK	
Station ID	Obs.	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)	Est.	Error Rate (%)
70t	52.716	47.757	-9.407	49.715	-5.693	54.308	3.019	53.894	2.235
67t	50.563	45.326	-10.358	45.055	-10.893	50.778	0.426	55.477	9.719
35t	53.473	55.100	3.041	54.184	1.329	54.807	2.494	59.639	11.530
38t	51.162	50.028	-2.218	52.585	2.781	51.488	0.636	52.155	1.940
Average RMSE		3.477375		3.678844		0.882139		4.722686	

4. Discussion

The current study analyzed four spatial interpolation methods, two of which are simple traditional interpolation approaches and the other two are machine learning based methods. According to the results from the 10-fold cross-validation, OK demonstrated the best performance, especially when pollution occurred at critical levels to affect human health ($>50 \mu\text{g}/\text{m}^3$), while RF and RFK demonstrated overall high performance throughout the given scenarios and IDW exhibited the worst performance from all. Overall, their performance presented variability depending on differing temporal and regional conditions as also discussed in several other studies e.g., [68–70]. Therefore, the period undergoing monsoon season, which is best represented by the month of July, resulted in low accuracy in estimates, while months with considerably higher actual variance of $\text{PM}_{2.5}$ exhibited high RMSE as well as the highest R^2 , which in return achieved the best accuracy in February, followed by March.

The models also represented varied results depending on regions, which was reflected in the point-based comparative analysis of outputs. Since the extremely high and low concentrations would influence the performance, and we aimed to achieve appropriate estimation even with high contamination, we compared estimates with the original ground-level data as shown in Tables 3–6. According to the latter, IDW and OK tended to underestimate $\text{PM}_{2.5}$ concentration, which would be problematic in apprehending the severity of pollution. Meanwhile, RFK was inclined to produce higher estimations, which we considered as a less significant problem than an underestimation, in accordance with its lower RMSE compared to the rest of the models.

The data range of interpolated estimates also indicated different explanations of model performance. IDW and OK best matched the observation data range since the models are highly dependent on close monitoring sites, which adversely caused excessive zoning dis-

tribution around the ground-level stations. As also demonstrated by Sajjadi et al. [71], IDW produced worse estimates when the sample range cannot be represented by sample points of data, by having lower performance with RMSE of $7.19 \mu\text{g}/\text{m}^3$ and $7.01 \mu\text{g}/\text{m}^3$ with 10-day and monthly dataset in March respectively. RF produced oversimplified variability in $\text{PM}_{2.5}$ concentrations by having a narrower range with the 10-day dataset in February and March, which consequently had higher variance compared to other datasets. RFK, which is a combination of OK and RF, showed the most appropriate range covering the data not represented by the ground observations. This is a very distinctive and important feature in deriving estimates out of the data range of original observations. The different conclusions from cross-validation statistics and data range comparison suggest that traditional cross-validation results cannot solely determine model performance as also advocated by Zou et al. [72].

The spatial distribution map of $\text{PM}_{2.5}$ advocates the advantage of implementing RFK as a spatial interpolation method. IDW and OK, which are solely based on regulatory-grade monitoring stations, were unlikely to illustrate variances resulting from geographical and anthropogenic factors, especially in the northern and southern regions where monitoring stations are scarce. On the other hand, RF and RFK produced more realistic $\text{PM}_{2.5}$ spatial distributions with accumulated auxiliary data over larger geographical areas. This implies that these methods, and especially RFK, could be useful for countries that have limited in-situ monitoring stations but plenty of auxiliary data, such as Thailand, while becoming more reliable as more datasets are integrated, especially when extreme contamination has occurred.

However, there are a few attempts to be made to constructively reinforce the conclusion that RFK could be the most suitable method among the four models discussed in the current paper. Firstly, the results from our work could be further improved by integrating additional geostationary satellite data such as the anticipated Global Environmental Monitoring System (GEMS), which has a higher temporal resolution. The TROPOMI satellite system has a daily revisit, therefore, incorporation of additional satellite data with a more frequent scanning cycle could enhance the correlation analysis and modelling of non-linear relations. Secondly, despite the fact that this study excluded meteorological data due to their relatively low spatial resolution, a potential high-spatial-resolution weather data set representing the dynamic climatic conditions and the interaction with $\text{PM}_{2.5}$ could improve the performance of the models. This could be materialized by integrating data from numerical weather prediction (NWP) models or satellite images. Lastly, the model accuracy could be enhanced by applying the gap-filling method on atmospheric products of TROPOMI. Even though TROPOMI products contain more abundant information on atmospheric constituents compared to MODIS AOD products, which have been traditionally used in $\text{PM}_{2.5}$ estimation, several studies have also shown that filling the gaps improved the model accuracy [73–75]. Therefore, the model performance of RFK would be improved by integrating more auxiliary data and implementing advanced preprocessing techniques. In the current study, we demonstrated the key features of RFK and suggest it to be the appropriate model for $\text{PM}_{2.5}$ estimation for the cases of scarce distribution and limited number of the in-situ monitoring systems. This novel approach was demonstrated and materialized by integrating satellite- and ground-based information of predominantly chemical composites as covariates.

5. Summary

The current study presented the application and inter-comparison of four spatial interpolation models in an attempt to estimate ground-level $\text{PM}_{2.5}$ over Thailand. Overall, the basic interpolation models of IDW and OK could simply calculate the estimates from observations of PCD ground stations, which resulted in a smooth and unrealistic representation of $\text{PM}_{2.5}$ concentration. On the other hand, machine learning approaches, namely RF and RFK, could successfully derive maps of the spatial distribution of $\text{PM}_{2.5}$ with a finer spatial resolution of $3.5 \times 5.5 \text{ km}$.

This research demonstrated the appropriateness of RFK as a spatial interpolation method for PM_{2.5} by conducting 10-fold cross-validation of test sets, and analyzing differences in data range of estimates, station-based estimation, and spatial distribution depicted in maps as additional approaches to evaluate the models. Therefore, we argue that the synergistic use of ground stations and satellite data leads to improvement of accuracy in spatially estimating the ground-level PM_{2.5} concentrations. These fusion techniques are anticipated to provide more robust results if more data are integrated and our research will focus in the future on fusing additional information into the workflow, such as the anticipated GEMS, Sentinel-4, and Sentinel-5 satellites as well as low-cost sensors. Last but not least, the presented approach can be reproduced for large geographical regions with low computational demands and therefore assist authorities seeking to set up high-spatial-resolution nowcasting monitoring systems for PM_{2.5} concentrations. In particular, our findings can act as a reference for comparing traditional and newly adopted spatial interpolation methods, and primarily RFK. We suggest that the latter can be used as a robust technique to derive spatially continuous maps of PM_{2.5} concentrations, even in the event of high atmospheric contamination, despite the limited atmospheric monitoring networks.

Author Contributions: Conceptualization, D.S.; Methodology, S.H. and W.K.; Formal analysis, S.H.; Investigation, S.H.; Data curation, S.H. and D.S.; Writing—original draft preparation, S.H. and D.S.; Writing—review and editing, S.H., D.S. and W.K.; Supervision, D.S.; Funding acquisition, P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data used in this study, including satellite and ground data, are from sources providing the data freely available through the internet.

Acknowledgments: Shinhye Han was supported by the Int'l Meteorological Expert Training Program scholarship awarded by the Korea Meteorological Administration. We would also like to acknowledge the support received from the Korean Meteorological Agency, the Asian Disaster Preparedness Center, and the reviewers' fruitful feedback.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The table below lists the abbreviations and acronyms used in the paper in alphabetical order.

Abbreviation	Definition
AI	aerosol index
CO	carbon monoxide
DEM	digital elevation model
GEMS	Global Environmental Monitoring System
HCHO	formaldehyde
IDW	inverse distance weighted
ML	machine learning
NASA	National Aeronautics and Space Administration
NGA	National Geospatial-Intelligence Agency
NWP	numerical weather prediction
NO ₂	nitrogen dioxide
O ₃	ozone
OK	ordinary kriging
PCD	Pollution Control Department
PM	particulate matter
PM _{2.5}	particulate matter with an aerodynamic diameter of less than 2.5 µm
R ²	coefficient of determination
RF	random forest

RFK	random forest combined with ordinary kriging
RK	regression kriging
RMSE	root-mean-squared error
Sentinel-5P	Sentinel 5 Precursor
SI	scatter index
SO ₂	sulfur dioxide
SRTM	Shuttle Radar Topography Mission
SVM	support vector machine
TROPOMI	TROPOspheric Monitoring Instrument

References

- Lee, M.; Lin, L.; Chen, C.Y.; Tsao, Y.; Yao, T.H.; Fei, M.H.; Fang, S.H. Forecasting air quality in Taiwan by using machine learning. *Sci. Rep.* **2020**, *10*, 4153. [\[CrossRef\]](#)
- Ghorani-Azam, A.; Riahi-Zanjani, B.; Balali-Mood, M. Effects of air pollution on human health and practical measures for prevention in Iran. *J. Res. Med. Sci.* **2016**, *21*, 65.
- Wang, J.; Ogawa, S. Effects of meteorological conditions on PM_{2.5} concentrations in Nagasaki, Japan. *Int. J. Environ. Res. Public Health* **2015**, *12*, 9089–9101. [\[CrossRef\]](#) [\[PubMed\]](#)
- Domingo, J.L.; Marquès, M.; Rovira, J. Influence of airborne transmission of SARS-CoV-2 on COVID-19 pandemic. A review. *Environ. Res.* **2020**, *188*, 109861. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cazzolla Gatti, R.; Velichevskaya, A.; Tateo, A.; Amoroso, N.; Monaco, A. Machine learning reveals that prolonged exposure to air pollution is associated with SARS-CoV-2 mortality and infectivity in Italy. *Environ. Pollut.* **2020**, *267*, 115471. [\[CrossRef\]](#)
- Comunian, S.; Dongo, D.; Milani, C.; Palestini, P. Air pollution and COVID-19: The role of particulate matter in the spread and increase of COVID-19's morbidity and mortality. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4487. [\[CrossRef\]](#)
- Iskandaryan, D.; Ramos, F.; Trilles, S. Air Quality Prediction in smart cities using machine learning technologies based on sensor data: A Review. *Appl. Sci.* **2020**, *10*, 2401. [\[CrossRef\]](#)
- Vichit-Vadakan, N.; Vajanapoom, N. Health Impact from Air Pollution in Thailand: Current and Future Challenges. *Environ. Health Perspect.* **2011**, *119*, A197. [\[CrossRef\]](#)
- Yu, H.; Russell, A.; Mulholland, J.; Odman, T.; Hu, Y.; Chang, H.H.; Kumar, N. Cross-comparison and evaluation of air pollution field estimation methods. *Sustain. Cities Soc.* **2018**, *179*, 49–60. [\[CrossRef\]](#)
- Chen, L.J.; Ho, Y.H.; Lee, H.C.; Wu, H.C.; Liu, H.M.; Hsieh, H.H.; Huang, Y.; Lung, S.C.C. An open framework for participatory PM_{2.5} monitoring in smart cities. *IEEE Access* **2017**, *5*, 14441–14454. [\[CrossRef\]](#)
- Kim, S.; Park, S.; Lee, J. Evaluation of performance of inexpensive laser based PM_{2.5} sensor monitors for typical indoor and outdoor hotspots of South Korea. *Appl. Sci.* **2019**, *9*, 1947. [\[CrossRef\]](#)
- Mak, H.W.L.; Lam, Y.F. Comparative assessments and insights of data openness of 50 smart cities in air quality aspects. *Sustain. Cities Soc.* **2021**, *69*, 102868. [\[CrossRef\]](#)
- Park, S.; Lee, J.; Im, J.; Song, C.K.; Choi, M.; Kim, J.; Lee, S.; Park, R.; Kim, S.M.; Yoon, J.; et al. Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models. *Sci. Total Environ.* **2020**, *713*, 136516. [\[CrossRef\]](#) [\[PubMed\]](#)
- Janssen, S.; Dumont, G.; Fierens, F.; Mensink, C. Spatial interpolation of air pollution measurements using CORINE land cover data. *Atmos. Environ.* **2008**, *42*, 4884–4903. [\[CrossRef\]](#)
- Pearce, J.L.; Rathbun, S.L.; Aguilar-Villalobos, M.; Naeher, L.P. Characterizing the spatiotemporal variability of PM_{2.5} in Cusco, Peru using kriging with external drift. *Atmos. Environ.* **2009**, *43*, 2060–2069. [\[CrossRef\]](#)
- Li, J.; Heap, A.D. Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.* **2014**, *53*, 173–189. [\[CrossRef\]](#)
- Karydas, C.G.; Gitas, I.Z.; Koutsogiannaki, E.; Lydakis-Simantiris, N.; Silleos, G.N. Evaluation of spatial interpolation techniques for mapping agricultural topsoil properties in Crete. *EARSeL EProceedings* **2009**, *8*, 26–39.
- Zhang, G.; Rui, X.; Fan, Y. Critical review of methods to estimate PM_{2.5} concentrations within specified research region. *ISPRS Int. J. Geo Inf.* **2018**, *7*, 368. [\[CrossRef\]](#)
- Lu, G.Y.; Wong, D.W. An adaptive inverse-distance weighting spatial interpolation technique. *Comput. Geosci.* **2008**, *34*, 1044–1055. [\[CrossRef\]](#)
- Kravchenko, A.; Bullock, D.G. A comparative study of interpolation methods for mapping soil properties. *Agron. J.* **1999**, *91*, 393–400. [\[CrossRef\]](#)
- Deligiorgi, D.; Philippopoulos, K. Spatial interpolation methodologies in urban air pollution modeling: Application for the greater area of metropolitan Athens, Greece. *Adv. Air Pollut.* **2011**, *17*, 341–362.
- Li, J.; Heap, A.D.; Potter, A.; Daniell, J.J. Application of machine learning methods to spatial interpolation of environmental variables. *Environ. Model. Softw.* **2011**, *26*, 1647–1659. [\[CrossRef\]](#)
- Chen, P.C.; Lin, Y.T. Exposure assessment of PM_{2.5} using smart spatial interpolation on regulatory air quality stations with clustering of densely-deployed microsensors. *Environ. Pollut.* **2021**, *292*, 118401. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wackernagel, H. *Multivariate Geostatistics: An Introduction with Applications*; Springer: Berlin, Germany, 1998.

25. Sekulić, A.; Kilibarda, M.; Heuvelink, G.B.M.; Nikolić, M.; Bajat, B. Random forest spatial interpolation. *Int. J. Remote Sens.* **2020**, *12*, 1687. [CrossRef]
26. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.M.; Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ.* **2018**, *8*, e5518. [CrossRef]
27. Bozán, C.; Takács, K.; Körösparti, J.; Laborczi, A.; Túri, N.; Pásztor, L. Integrated spatial assessment of inland excess water hazard on the Great Hungarian Plain. *Land Degrad. Dev.* **2018**, *29*, 4373–4386. [CrossRef]
28. Szatmári, G.; Pásztor, L. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma* **2019**, *337*, 1329–1340. [CrossRef]
29. Laborczi, A.; Bozán, C.; Körösparti, J.; Szatmári, G.; Kajári, B.; Túri, N.; Kerecsi, G.; Pásztor, L. Application of Hybrid Prediction Methods in Spatial Assessment of Inland Excess Water Hazard. *ISPRS Int. J. Geo Inf.* **2020**, *9*, 268. [CrossRef]
30. Mammadov, E.; Nowosad, J.; Glaesser, C. Estimation and mapping of surface soil properties in the Caucasus Mountains, Azerbaijan using high-resolution remote sensing data. *Geoderma Reg.* **2021**, *26*, e00411. [CrossRef]
31. Araki, S.; Shima, M.; Yamamoto, K. Spatiotemporal land use random forest model for estimating metropolitan NO₂ exposure in Japan. *Sci. Total Environ.* **2018**, *634*, 1269–1277. [CrossRef]
32. Climatological Group; Meteorological Development Bureau; Meteorological Department. The Climate of Thailand. 2015. Available online: https://www.tmd.go.th/en/archive/thailand_climate.pdf (accessed on 20 December 2021).
33. Veefkind, J.P.; Aben, I.; McMullan, K.; Förster, H.; de Vries, J.; Otter, G.; Claas, J.; Eskes, H.J.; de Haan, J.F.; Kleipool, Q.; et al. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sens. Environ.* **2012**, *120*, 70–83. [CrossRef]
34. Lary, D.J.; Lary, T.; Sattler, B. Using machine learning to estimate global PM_{2.5} for environmental health studies. *Environ. Health Insights* **2020**, *9s1*, EHI-S15664. [CrossRef]
35. Lary, D.J.; Faruque, F.S.; Malakar, N.; Moore, A.; Roscoe, B.; Adams, Z.L.; Eggelston, Y. Estimating the global abundance of ground level presence of particulate matter (PM_{2.5}). *Geospat. Health* **2014**, *8*, S611–S630. [CrossRef] [PubMed]
36. Schulte, N.; Li, X.; Ghosh, J.K.; Fine, P.M.; Epstein, S.A. Responsive high-resolution air quality index mapping using model, regulatory monitor, and sensor data in real-time. *Environ. Res. Lett.* **2020**, *15*, 1040a7. [CrossRef]
37. Theys, N.; de Smedt, I.; Yu, H.; Danckaert, T.; van Gent, J.; Hörmann, C.; Wagner, T.; Hedelt, P.; Bauer, H.; Romahn, F.; et al. Sulfur dioxide retrievals from TROPOMI onboard Sentinel-5 Precursor: Algorithm theoretical basis. *Atmos. Meas. Tech.* **2017**, *10*, 119–153. [CrossRef]
38. Sharma, S.; Zhang, M.; Anshika; Gao, J.; Zhang, H.; Kota, S.H. Effect of restricted emissions during COVID-19 on air quality in India. *Sci. Total Environ.* **2020**, *728*, 138878. [CrossRef] [PubMed]
39. Stratoulas, D.; Nuthammachot, N. Air quality development during the COVID-19 pandemic over a medium-sized urban area in Thailand. *Sci. Total Environ.* **2020**, *746*, 141320. [CrossRef]
40. Fan, C.; Li, Y.; Guang, J.; Li, Z.; Elnashar, A.; Allam, M.; de Leeuw, G. The impact of the control measures during the COVID-19 outbreak on air pollution in China. *Int. J. Remote Sens.* **2020**, *12*, 1613. [CrossRef]
41. Bauwens, M.; Compennolle, S.; Stavrou, T.; Müller, J.-F.; van Gent, J.; Eskes, H.; Levelt, P.F.; van der A, R.; Veefkind, J.P.; Vlietinck, J.; et al. Impact of coronavirus outbreak on NO₂ pollution assessed using TROPOMI and OMI observations. *Geophys. Res. Lett.* **2020**, *47*, e2020GL087978. [CrossRef]
42. Gitahi, J.; Hahn, M.; Ramirez, A. High-resolution urban air quality monitoring using sentinel satellite images and low-cost ground-based sensor networks. *E3S Web Conf.* **2020**, *3*, 102–111. [CrossRef]
43. Wang, Y.; Wang, M.; Huang, B.; Li, S.; Lin, Y. Estimation and analysis of the nighttime PM_{2.5} concentration based on LJ1-01 images: A case study in the Pearl River Delta urban agglomeration of China. *Remote Sens.* **2021**, *13*, 3405. [CrossRef]
44. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1 km resolution PM_{2.5} estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. [CrossRef]
45. Choi, W.; Lee, H.; Kim, D.; Kim, S. Improving spatial coverage of satellite aerosol classification using a random forest model. *Remote Sens.* **2021**, *13*, 1268. [CrossRef]
46. Li, T.; Wang, Y.; Yuan, Q. Remote sensing estimation of regional NO₂ via space-time neural networks. *Remote Sens.* **2020**, *12*, 2514. [CrossRef]
47. Wang, Y.; Yuan, Q.; Li, T.; Tan, S.; Zhang, L. Full-coverage spatiotemporal mapping of ambient PM_{2.5} and PM₁₀ over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions. *Sci. Total Environ.* **2021**, *793*, 148535. [CrossRef] [PubMed]
48. Zhou, L.; Zhou, C.; Yang, F.; Che, L.; Wang, B.; Sun, D. Spatio-temporal evolution and the influencing factors of PM_{2.5} in China between 2000 and 2015. *J. Geogr. Sci.* **2019**, *29*, 253–270. [CrossRef]
49. Huang, R.J.; Zhang, Y.; Bozzetti, C.; Ho, K.F.; Cao, J.J.; Han, Y.; Daellenbach, K.R.; Slowik, J.G.; Platt, S.M.; Canonaco, F.; et al. High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **2014**, *514*, 218–222. [CrossRef] [PubMed]
50. Luecken, D.J.; Napelenok, S.L.; Strum, M.; Scheffe, R.; Phillips, S. Sensitivity of ambient atmospheric formaldehyde and ozone to precursor species and source types across the U.S. *J. Environ. Sci. Technol.* **2018**, *52*, 4668. [CrossRef] [PubMed]

51. Fu, H.; Zhang, Y.; Liao, C.; Mao, L.; Wang, Z.; Hong, N. Investigating PM_{2.5} responses to other air pollutants and meteorological factors across multiple temporal scales. *Sci. Rep.* **2020**, *10*, 15639. [CrossRef]
52. Liu, X.; Pan, X.; Wang, Z.; He, H.; Wang, D.; Liu, H.; Tian, Y.; Xiang, W.; Li, J. Chemical characteristics and potential sources of PM_{2.5} in Shahe city during severe haze pollution episodes in the winter. *Aerosol Air Qual. Res.* **2020**, *20*, 2741–2753. [CrossRef]
53. Eskes, H.J.; Eichmann, K.U. S5P MPC Product Readme Nitrogen Dioxide; 2019; 1.5, S5P-MPC-KNMI-RPF-NO2. Available online: http://www.tropomi.eu/sites/default/files/files/publicSentinel-5P-Nitrogen-Dioxide-Level-2-Product-Readme-File_20191105.pdf (accessed on 21 November 2021).
54. Eatough, D.J.; Caka, F.M.; Farber, R.J. The conversion of SO₂ to sulfate in the atmosphere. *Isr. J. Chem.* **1994**, *34*, 301–314. [CrossRef]
55. Khoder, M.I. Atmospheric conversion of sulfur dioxide to particulate sulfate and nitrogen dioxide to particulate nitrate and gaseous nitric acid in an urban area. *Chemosphere* **2002**, *49*, 675–684. [CrossRef]
56. Zhu, J.; Chen, L.; Liao, H.; Dang, R. Correlations between PM_{2.5} and Ozone over China and Associated Underlying Reasons. *Atmosphere* **2019**, *10*, 352. [CrossRef]
57. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM_{2.5} Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ. Sci. Technol.* **2016**, *50*, 4712–4721. [CrossRef] [PubMed]
58. Zweers, S. TROPOMI ATBD of the UV Aerosol Index; 2021; 2.0, S5P-KNMI-L2-0008-RP. Available online: <https://sentinel.esa.int/documents/247904/2476257/Sentinel-5P-TROPOMI-ATBD-UV-Aerosol-Index> (accessed on 21 November 2021).
59. Sritong-aon, C.; Thomya, J.; Kertpromphan, C.; Phosri, A. Estimated effects of meteorological factors and fire hotspots on ambient particulate matter in the northern region of Thailand. *Air Qual. Atmos. Health* **2021**, *14*, 1857–1868. [CrossRef]
60. Weichenthal, S.; Kulka, R.; Lavigne, E.; van Rijswijk, D.; Brauer, M.; Villeneuve, P.J.; Stieb, D.; Joseph, L.; Burnett, R.T. Biomass Burning as a Source of Ambient Fine Particulate Air Pollution and Acute Myocardial Infarction. *Int. J. Epidemiol.* **2017**, *28*, 329–337. [CrossRef]
61. Roberts, E.A.; Sheley, R.L.; Lawrence, R.L. Using sampling and inverse distance weighted modeling for mapping invasive plants. *West. N. Am. Nat.* **2004**, *64*, 8–27.
62. Cressie, N. Geostatistics. *Am. Stat.* **1989**, *43*, 197–202.
63. Schneider, R.; Vicedo-Cabrera, A.M.; Sera, F.; Masselot, P.; Stafoggia, M.; de Hoogh, K.; Kloog, I.; Reis, S.; Vieno, M.; Gasparrini, A. A satellite-based spatio-temporal machine learning model to reconstruct daily PM_{2.5} concentrations across Great Britain. *Int. J. Remote Sens.* **2020**, *12*, 3803. [CrossRef] [PubMed]
64. Jiang, T.; Chen, B.; Nie, Z.; Ren, Z.; Xu, B.; Tang, S. Estimation of hourly full-coverage PM_{2.5} concentrations at 1-km resolution in China using a two-stage random forest model. *Atmos. Res.* **2021**, *248*, 105146. [CrossRef]
65. Jung, C.R.; Chen, W.T.; Nakayama, S.F. A national-scale 1-km resolution PM_{2.5} estimation model over Japan using MAIAC AOD and a two-stage random forest model. *Remote Sens.* **2021**, *13*, 3657. [CrossRef]
66. Zhang, T.; He, W.; Zheng, H.; Cui, Y.; Song, H.; Fu, S. Satellite-based ground PM_{2.5} estimation using a gradient boosting decision tree. *Chemosphere* **2021**, *268*, 128801. [CrossRef]
67. R Core Team. R: The R Project for Statistical Computing. 2020. Available online: <https://www.r-project.org/> (accessed on 7 January 2022).
68. Lee, C.; Lee, K.; Kim, S.; Yu, J.; Jeong, S.; Yeom, J. Hourly ground-level PM_{2.5} estimation using geostationary satellite and reanalysis data via deep learning. *Int. J. Remote Sens.* **2021**, *13*, 2121. [CrossRef]
69. Shen, H.; Li, T.; Yuan, Q.; Zhang, L. Estimating regional ground-level PM_{2.5} directly from satellite top-of-atmosphere reflectance using deep belief networks. *J. Geophys. Res. Atmos.* **2018**, *123*, 13875–13886. [CrossRef]
70. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **2019**, *130*, 104909. [CrossRef]
71. Sajjadi, S.A.; Zolfaghari, G.; Adab, H.; Allahabadi, A.; Delsouz, M. Measurement and modeling of particulate matter concentrations: Applying spatial analysis and regression techniques to assess air quality. *MethodsX* **2017**, *4*, 372–390. [CrossRef] [PubMed]
72. Zou, B.; Luo, Y.; Wan, N.; Zheng, Z.; Sternberg, T.; Liao, Y. Performance comparison of LUR and OK in PM_{2.5} concentration mapping: A multidimensional perspective. *Sci. Rep.* **2015**, *5*, 8698. [CrossRef]
73. Guo, B.; Zhang, D.; Pei, L.; Su, Y.; Wang, X.; Bian, Y.; Zhang, D.; Yao, W.; Zhou, Z.; Guo, L. Estimating PM_{2.5} concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017. *Sci. Total Environ.* **2021**, *778*, 146288. [CrossRef] [PubMed]
74. Liang, F.; Xiao, Q.; Huang, K.; Yang, X.; Liu, F.; Li, J.; Lu, X.; Liu, Y.; Gu, D. The 17-y spatiotemporal trend of PM_{2.5} and its mortality burden in China. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 25601–25608. [CrossRef]
75. Meng, X.; Liu, C.; Zhang, L.; Wang, W.; Stowell, J.; Kan, H.; Liu, Y. Estimating PM_{2.5} concentrations in Northeastern China with full spatiotemporal coverage, 2005–2016. *Remote Sens. Environ.* **2021**, *253*, 112203. [CrossRef]