CIVIL ENGINEERING JOURNAL

Article no. 27

THE CIVIL ENGINEERING JOURNAL 3-2023

$PM_{2.5}$ ESTIMATION IN THE CZECH REPUBLIC USING EXTREMELY RANDOMIZED TREES: A COMPREHENSIVE DATA ANALYSIS

Saleem Ibrahim[1], Martin Landa[1], Eva Matoušková[1], Lukáš Brodský[2] and Lena Halounová[1]*

1. *Department of Geomatics, Faculty of Civil Engineering, Czech Technical University in Prague, 166 29 Prague, Czech Republic, email: saleem.ibrahim@fsv.cvut.cz, martin.landa@fsv.cvut.cz; eva.matouskova@fsv.cvut.cz; lena.halounova@fsv.cvut.cz*
2. *Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, 128 43 Prague, Czech Republic, email: lukas.brodsky@natur.cuni.cz*

ABSTRACT

The accuracy of artificial intelligence techniques in estimating air quality is contingent upon a multitude of influencing factors. Unlike our previous study that examined PM2.5 over whole Europe using unbalanced spatial-temporal data, the focus of this study was on estimating PM2.5 specifically over the Czech Republic using more balanced dataset to train and evaluate the model. Moreover, the spatial autocorrelation between PM2.5 measurements was taken into consideration while building the model. The feature importance while developing the Extra Trees model revealed that spatial autocorrelation had greater significance in comparison to commonly used inputs such as elevation and NDVI. We found that R2 of the 10-CV for the new model was 16% higher than the previous one. Where R2 reached 0.85 with RMSE=5.42 µg/m3, MAE=3.41 µg/m3, and bias=-0.03 µg/m3. The developed spatiotemporal model was employed to generate comprehensive daily maps covering the entire study area throughout the period 2018–2020. The temporal analysis showed that the levels of PM2.5 exceeded recommended limits during the year 2018 in many regions. The eastern part of the country suffered from the highest concentrations especially over Zlín and Moravian-Silesian Regions. Air quality improved during the next two years in all regions reaching promising levels in 2020. The generated dataset will be available for other future air quality studies.

KEY WORDS

Air quality, PM2.5, Artificial intelligence, Spatial autocorrelation, Czech Republic

INTRODUCTION

Atmospheric Particulate Matter (PM) with a diameter smaller than or equal to 2.5 microns (PM2.5) is small enough to be inhaled deeply in the lungs and are able to reach the bloodstream and reduce the immune system's capacities [1]. The exposure of high PM2.5 levels could cause serious health problems especially in densely populated areas that produce enormous amounts of pollution into the atmosphere due to increased combustion sources and human activities [2]. PM has an effect on mortality even at concentrations that are in compliance with the European annual regulation [3]. In Europe, around 300,000 premature deaths are caused by PM annually and more than 330 billion Euros of economic cost, that encouraged the Directive 2008/50/EC to limit the yearly average of PM2.5 to 20 µg/m3 since the first of January 2020 [4].

In this study, we focused on the Czech Republic (CZ). Based on previous studies, CZ suffered from low air quality in some regions throughout last decades. The estimated additional social costs

resulting from the poor air quality in Ostrava city for children aged 0-15 amounted to approximately 20 million Euros per year [5]. In 2012 winter, the mean value of PM2.5 over Ostrava was 159 μg/which caused a smog episode [6]. When studying causes of air pollution in Teplice within the framework of the Teplice Program, initiated around 1970, researchers found that around 70% of PM2.5 fine particles came from the local heating sources that used brown coal with a high SO2 content [7]. As a result of this discovery, the Czech government supported a transition from coal to natural gas for local heating in mining districts in 1994 [7]. The north-eastern part of CZ that shares borders with Poland, which is highly polluted due to its long history of coal mining, heavy industry, traffic infrastructure and the dense population [8]. In 2018, around 1.2% of the CZ's total area, which is home to roughly 6.1% of the population, exceeded 25 μg/m3 [9]. Approximately 20% of households in CZ use individual heating systems that burn solid fuels [10]. During 2013 winter in the residential district of Mladá Boleslav, wood burning was found to be the primary source of PM10 mass, with coal combustion following as the second most significant source [11]. Coal remains a key energy source in CZ, accounting for one-third of the country's total energy supply in 2019 [12]. Coal also accounted for 46% of the country's electricity generation and more than 25% of residential heating [12]. The Czech government is currently exploring strategies for removing coal from its energy mix, including potential timelines for this transition. To support this effort, the government established a Coal Commission in 2019, which presented its recommendations in December 2020. The Commission advised that coal should be phased out no later than 2038 [12]. The data from April 2018 to March 2019 collected in the Moravian-Silesian Region has verified that during the winter season, the inflow of PM cross-border pollution from Poland is a key factor contributing to air pollution levels [13].

In recent decades, numerous studies have utilized the capabilities of artificial intelligence (AI) in estimating PM2.5 concentrations. These studies have focused on developing various types of models to increase the limited spatial coverage that is provided by PM2.5 ground monitors. Covering more auxiliary data as inputs helped to improve the performance of the models when compared to the typical interpolation methods which rely solely on the observations from the monitors [14]. The auxiliary inputs for the models usually include a combination of satellite data, meteorological modeled data, topography, and land cover data. Satellite-based Aerosol Optical Depth (AOD) is a valuable indicator of aerosol levels in the Earth's atmosphere and since PM2.5 is a type of aerosol, there is generally a positive correlation that made AOD a crucial factor in predicting PM2.5 levels [15,16]. Meteorological data such as the planetary boundary layer height (PBLH) that is the vertical extent of the lowest part of the Earth’s atmosphere, Relative Humidity (RH) which represents the total amount of water vapor that exists in the atmosphere relative to the maximum amount water vapor that air can hold at particular temperature, the Total Column Water Vapor (TCWV) that is the measurement of the total amount of water vapor present in the vertical column of the Earth’s atmosphere, Wind Speed (WS), Temperature (T), Total Precipitation (TP), and Evaporation (E) have shown that significance varies depending on the region when PM2.5 is estimated [14,17,18]. Moreover, a few studies considered the Spatial Autocorrelation (SA) of PM2.5 when developing predictive models. Inspired by the first law of geography which proposes that all features present on a geographic surface have a connection with each other, and that geographic entities have a stronger association with nearby entities as compared to those that are located far away [19]. In a study spanning from 1999 to 2016, the yearly average PM2.5 levels in Chinese cities exhibited a typical autocorrelation [20]. In another study, including SA improved the performance of the Random Forest (RF) model and decreased the Root Mean Square Error (RMSE) by ~18% when estimating PM2.5 over Sichuan Basin in 2019 [21]. Adding the spatial lag variable (SLV) as a virtual input in the neural network model for estimating the yearly PM2.5 concentrations increased the coefficient of determination (R2) by ~9% [22].

In this study, we aimed to estimate the concentrations of PM2.5 over the CZ during the years 2018, 2019, and 2020. CZ is a landlocked country that covers an area of 78870 square kilometres located in central Europe bordering Germany, Poland, Slovakia, and Austria.

MATERIALS AND METHODS

Dependent variable and primary independent variables

Daily PM2.5 concentrations for 2018, 2019, and 2020 were collected from the Czech Hydrometeorological Institute (CHMI). The total number of stations and observations after removing the outlier values were 54 and 54,495 respectively. The number of observations per year is 18330 in 2018, 18022 in 2019, and 18144 in 2020.

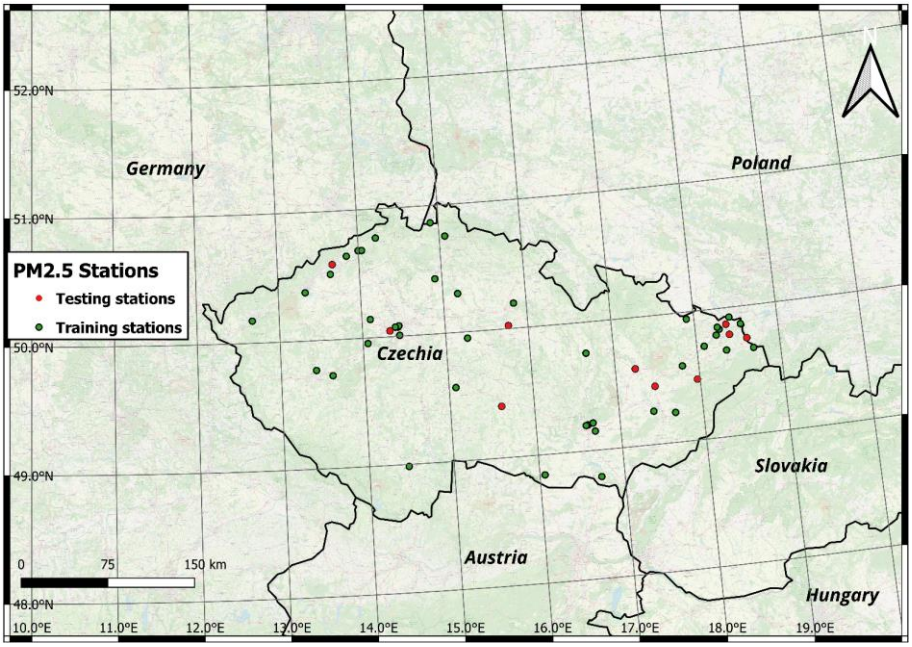

Fig. 1 - Study area with ground stations. The red dots represent the stations that were used to test the model and the green dots represent the stations that were used to train the model

We explored the following data as primary inputs in our study, AOD data over CZ was collected from the Geo-Harmonized Atmospheric Dataset for Aerosols (GHADA) which is a full coverage dataset over Europe with 1 km spatial resolution that was built based on the MCD19A2 MODIS product [23] and modelled AOD from Copernicus Atmosphere Monitoring Service (CAMS) [24]. This dataset showed good results when validated with NASA's Aerosol Robotic Network (AERONET) observations [25]. Meteorological data like PBLH, WS calculated based on the u and v wind components, temperature at 2m (T2m), TP, E, TCWV, and RH were collected from the European Centre for Medium-Range Weather Forecasts ERA5 climate reanalysis [26], and then reprojected to the grid using the bilinear interpolation; monthly NDVI from the MODIS MOD13A3 product [27]; the percentage of artificial surfaces and air pollution resources for each 1km2 cell were calculated from the CORINE Land Cover (CLC) of 2018 which was built based on ortho-rectified satellite images with a spatial resolution ranging from 5-60 m, and were aggregated to 100 m; Open Street Map (OSM) data was processed to calculate the total road lengths (RL) within each cell of the grid; elevation (H) was extracted from the Japan Aerospace Exploration Agency (JAXA) digital surface model [28], and population data was estimated from the monthly Visible Infrared Imaging Radiometer Suite (VIIRS) nighttime lights of 2019 [29]. The linear analysis between the primary inputs and PM2.5 showed that PBLH and T2m were the most negatively correlated variables to PM2.5 with Pearson correlation of -0.25 and -0.22 respectively. NDVI, TCWV, WS, RH, H, and TP also had negative correlations with PM2.5. Whereas, E, AOD, NL, and RL had positive correlations with PM2.5. The following table shows the primary data that was used

in our study. All primary data was reprojected to the European Terrestrial Reference System 1989 (EPSG:3035) with a grid of 1 km2 that covers the study area using bilinear interpolation for meteorological data and the cubic convolution for the elevation model.

Tab. 1 - The primary inputs that were explored in this study

Name		Variable	Unit	Spatial resolution	Source
Aerosol optical depth		AOD	-	1 km	GHADA
Meteorological	Planetary boundary layer height	PBLH	m	0.1°×0.1°	ERA5-Land
	Wind speed	WS	m/s		
	Temperature at 2m	T2m	K		
	Total precipitation	TP	mm		
	Evaporation	E	mm		
	Total column water vapor	TCWV	Kg/m^2		
	Relative humidity	RH	%	0.25°×0.25°	ERA5
Land cover	Normalized Difference Vegetation Index	NDVI	-	1 km	MODIS MOD13A3
	CORINE Land Cover	CLC	-	100 m	Corine LC 2018
	Road length	RL	m	~10 m	Open street maps
Topography		H	m	~30 m	JAXA
Population		NL	$nW/cm^2/sr$	500 m	VIIRS

Model development

A machine learning algorithm was used with feature engineering techniques that were applied to train the PM2.5 predictive model.

We used the Extra Trees (ET) algorithm which is an ensemble learning method that combines the predictions of several decision trees to make the final prediction [30]. It is an extension of the widely used RF algorithm where in both, the final prediction is the majority of predictions in classification problems and the arithmetic average in regression problems. ET reduces overfitting by introducing additional randomness during the construction of the trees and it uses the entire dataset while training without performing any pruning which decreases the required time for training compared to the RF that applies pruning techniques. A deeper explanation of this algorithm was provided in our previous work [25,31].

Feature engineering and model training

The temporal inputs were represented by the radian day and the year. The radian day will help the model understand the cyclic nature of time and enables it to capture the seasonal patterns in the data. Whereas, adding the year will capture long-term trends that occur over the years of the

study period. The spatial inputs were represented by longitude, latitude, and elevation. Adding the spatial inputs will allow the model to capture the inherent spatial heterogeneity in the data. In addition to the mentioned inputs, SA of the dependent variable was calculated based on the training set. We used the Local Moran Index (LMI) that was based on the foundation of the Moran's I statistic [32]. LMI is a spatial autocorrelation statistic used in geography and other disciplines to identify local clusters or spatial patterns of similar or dissimilar values in a dataset [33]. Positive values for LMI indicate that the observation at the station is a part of a cluster of similar observations from surrounding stations and vice versa, the magnitude of the LMI value represents the strength of SA [34]. For each day of the study period, LMI was calculated for each station considering the closest three neighboring stations using the K-nearest neighbors (KNN) weight matrix with k=3.

$$\boldsymbol{LMI_i} = \frac{\boldsymbol{z_i - \bar{z}}}{\boldsymbol{\sigma^2}} \sum_{\boldsymbol{j=1, i \neq j}}^{\boldsymbol{n}} [\boldsymbol{w_{ij}}(\boldsymbol{z_j} - \boldsymbol{\bar{z}})] \quad (1)$$

Where, Zi is the value of the observation at the location i; $\bar{z}$ is the average value of z with the sample number of n; zj is the value of the observation at all other stations where i≠j; σ2 is the variance of the observation z; and wij is the weight matrix for the locations i and j.

The whole dataset was split into a training set (80% of the dataset) and a test set (20% of the dataset), Figure 1 represents the distribution of the stations. LMI was calculated based on the training set only to assure that the test set remains unseen for the model. The feature importance for each input was calculated and based on that some features were removed to generalize the model and to reduce complexity. CLC, OSM, and population had low importance because these inputs are not real time data. Figure 2 shows feature importance of the primary inputs in the training set.

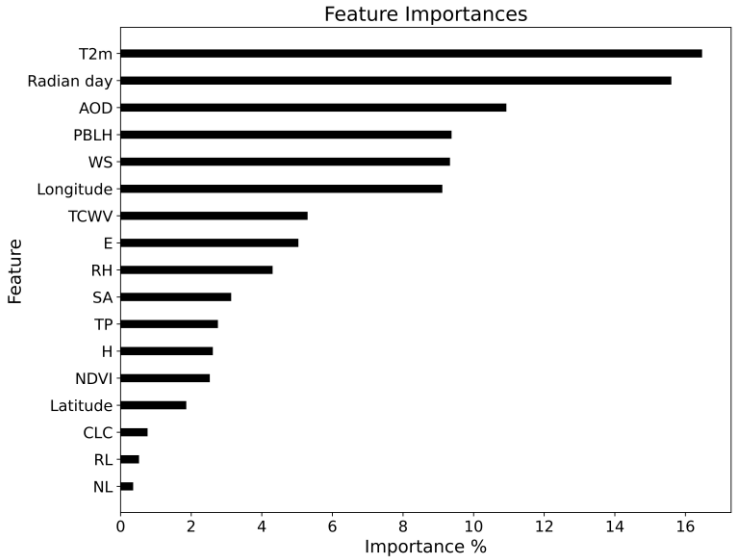

Fig. 2 - Feature importance calculated based on the training data.

The widely used grid search technique with 10-fold Cross Validation (10-CV) was used for hyperparameters tuning. In this process, the training data was split into 10 equal-sized folds, where each fold was used as a validation set while training the model on the remaining 9 folds. We employed R^2, the RMSE, and the Mean Squared Error (MAE) as evaluation matrices. R^2 measures the proportion of variance in $PM_{2.5}$ that can be explained by the model. RMSE quantifies the average difference between the predicted and observed $PM_{2.5}$ values. MAE measures the average absolute difference between the predicted and observed $PM_{2.5}$ values. Utilizing these three metrices together is commonly used in regression problems to provide a comprehensive evaluation of the model. The maximum depth of the trees, the minimum number of samples required to split an internal node and

the minimum number of samples required at a leaf node were the main parameters to fine-tune the model. While applying the 10-CV on the training data, we tested how the performance will drop when excluding some inputs. We found that NDVI did not noticeably affect the performance of the model and it was excluded as well.

Model validation

This section was dedicated to the validation process to assess the reliability and accuracy of our findings.

Validation on the test set

We tested the model on the test set that was taken from the stations in unseen locations for the model. This validation showed the model ability to predict values in new locations that were not used to generate the LMI. The model showed good results when estimating $PM_{2.5}$ in the new locations with R^2 = 0.86, RMSE=5.61 μg/m^3, and MAE=3.37 μg/m^3.

Validation on all data

It is a common approach in $PM_{2.5}$ studies to apply 10-CV of the whole dataset [35–37]. In order to do this validation, we generated LMI based on the data from all stations, then we applied a sample based 10-CV. The model showed similar results compared to the validation on the test set with R^2=0.85, RMSE=5.42 μg/m^3, MAE=3.41 μg/m^3 and, bias=-0.03 μg/m^3. Figure 3 shows the results of the sample based 10-CV.

A negative bias indicates that, on average, the model tends to underpredict $PM_{2.5}$ values. However, a value of -0.03 appears to align reasonably well with the characteristics of the data where the values range between 2 and 200 with an average of 17 μg/m^3.

R^2 values indicate that the model explains around 86% and 85% of the variance in $PM_{2.5}$ values, which suggests that the model is performing well and generalizing reasonably to unseen data.

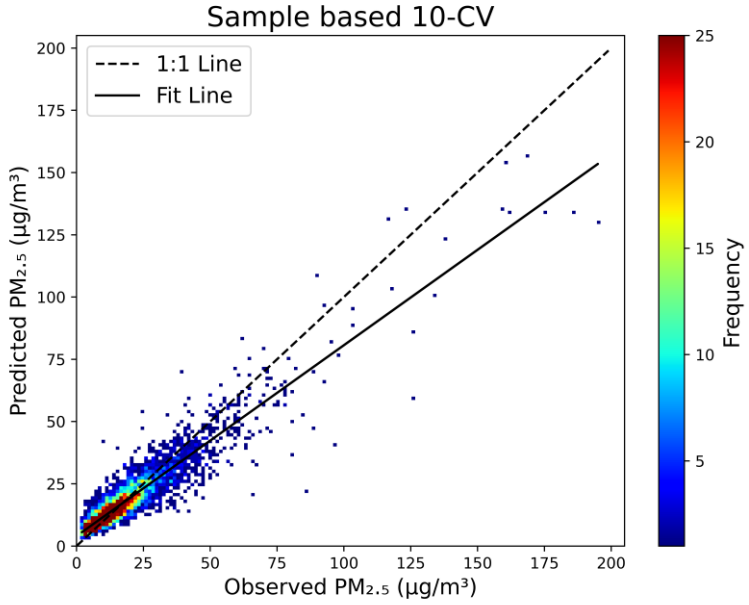

Fig 3 - Density scatter plot for the 10-CV applied on all data.

Results

Model deployment

We utilized the model to generate daily full coverage $PM_{2.5}$ maps over CZ. To validate the deployment of the model we extracted values of the estimated $PM_{2.5}$ at station locations and compared their temporal profiles with observed values. Figure 4 represents the temporal profile for three stations with high, normal, and low $PM_{2.5}$ levels.

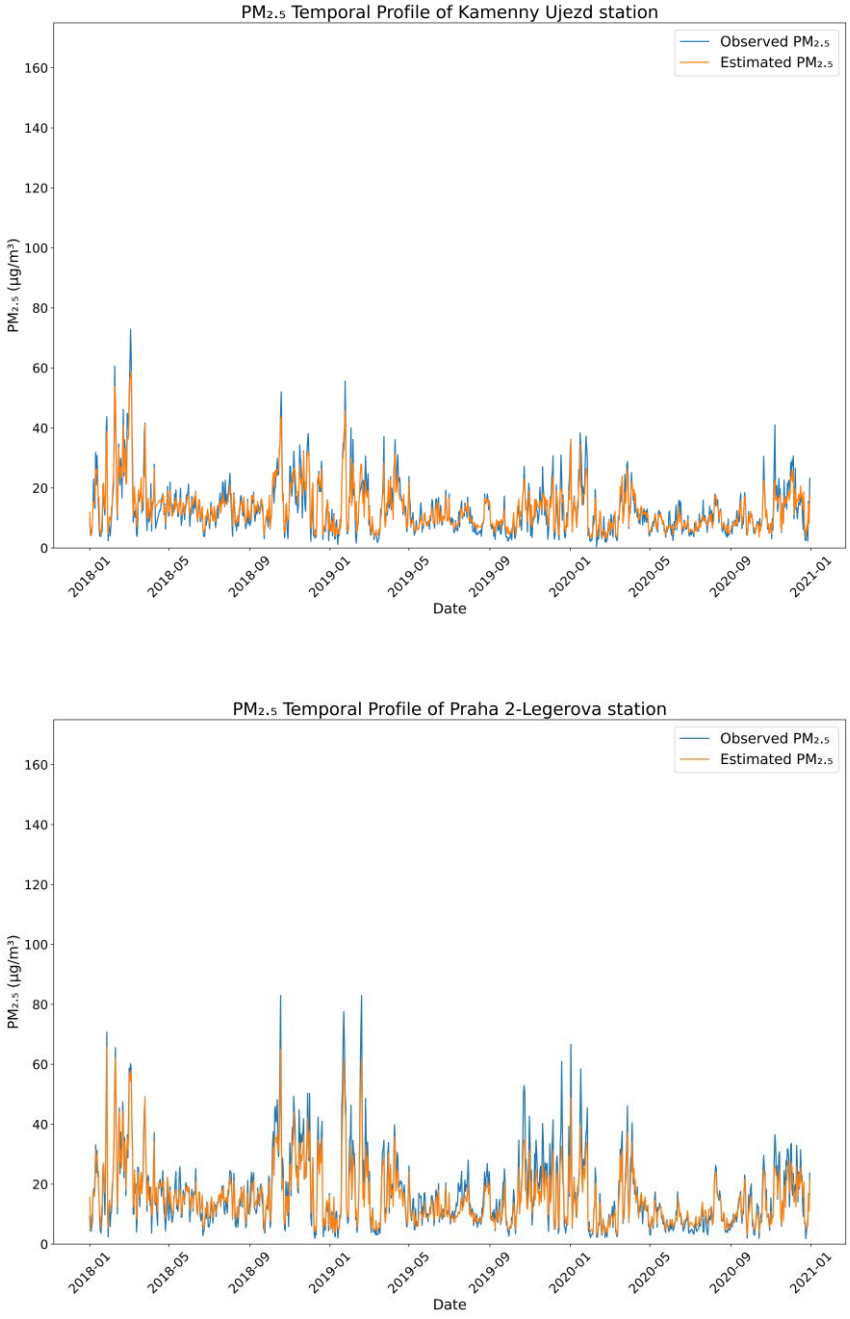

Fig. 4 - $PM_{2.5}$ temporal profile over three stations

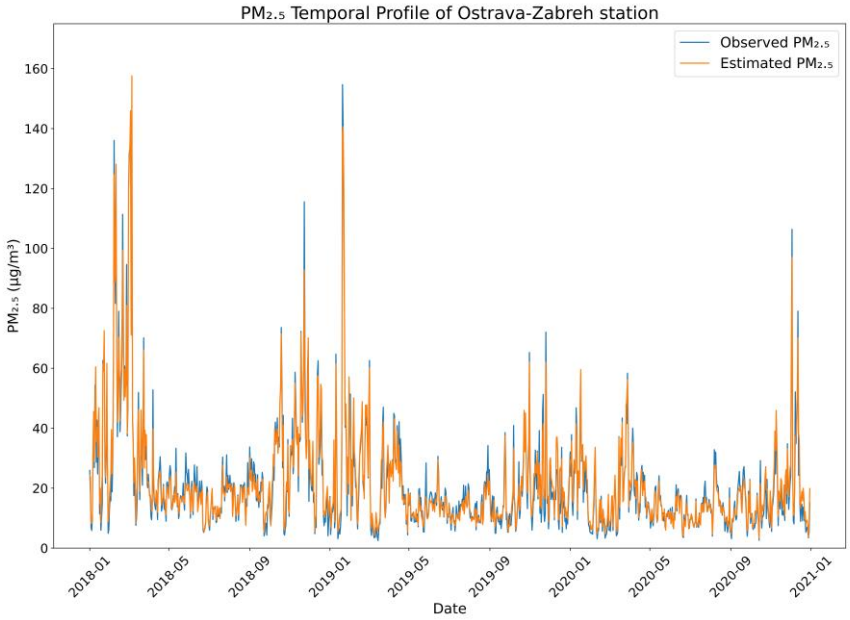

Fig. 4 - $PM_{2.5}$ temporal profile over three stations: Kamenny Ujezd station, Praha 2-Legerova station, and Ostrava-Zabreh station.

The results in all stations show nearly perfect overlap, which confirms not only high general accuracy of the model but also temporal clarity of the predictions. They also show slight bias of the model in the peaks' predictions, small underestimation in high values and slight underestimation in down-peaks. It can be noticed that $PM_{2.5}$ values are higher during winter compared to other seasons in the three chosen stations.

Temporal and regional analysis

We calculated the average $PM_{2.5}$ levels for each year during the study period. In Figure 5 we show the yearly average levels. $PM_{2.5}$ decreased gradually throughout the study period. The eastern part of CZ had the highest $PM_{2.5}$ levels. The Moravia-Silesian Region was the most polluted region with an average $PM_{2.5}$ level of 25.2 µg/m^3 in 2018, 18 µg/m^3 in 2019 and 15.8 µg/m^3 in 2020. Karlovy Vary Region had the lowest $PM_{2.5}$ values with 16.4 µg/m^3 in 2018, 11.1 µg/m^3 in 2019, and 10.2 µg/m^3 in 2020. Besides, the Moravia-Silesian Region, $PM_{2.5}$ values exceeded 20 µg/m^3 in Zlín and Olomouc Regions with average values of 22.7 µg/m^3 and 22.2 µg/m^3 respectively during 2018. Good $PM_{2.5}$ levels <= 12 µg/m^3 were found in six regions in 2020, these regions are Plzeň, Karlovy Vary, Southern Bohemia, Vysočina, Central Bohemia, and Liberec.

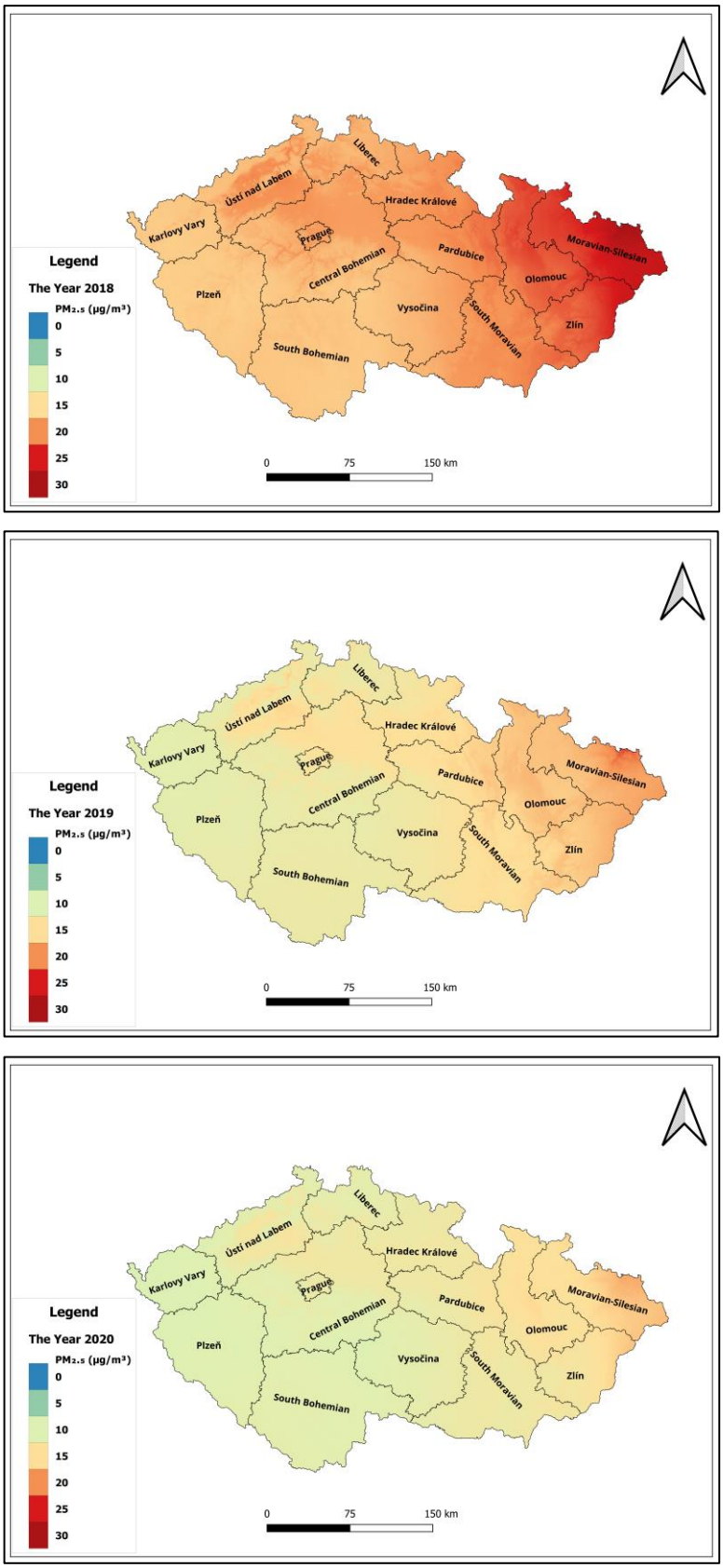

Fig. 5 - The average $PM_{2.5}$ levels over the Czech Republic in the years 2018, 2019, and 2020.

Seasonal analysis

In this analysis, we delved into the seasonal patterns of $PM_{2.5}$ concentrations of 2018–2020. By examining the fluctuations across different seasons and analyzing the variations in $PM_{2.5}$ levels over time, we aimed to gain valuable insights into the underlying factors influencing pollution levels during specific seasons of the study period. Winter was represented by January, February, and December; summer encompasses June, July, and August; spring spans from March through May; and autumn extends from September to November. We calculated the average $PM_{2.5}$ levels for each region in CZ in the different seasons. Figure 6 shows the results we conducted.

The average $PM_{2.5}$ levels in summer are relatively consistent for each year across the entire country. $PM_{2.5}$ concentrations exhibit significant variations during winter seasons. In winter, the average $PM_{2.5}$ was the highest in all regions except two in 2018 where Prague had the highest values during autumn and Karlovy Vary had the highest levels during spring. The eastern part of CZ was highly polluted during 2018 winter with average values of 30 μg/m^3 over Olomouc Region, 31 μg/m^3 over Zlín Region, and 35 μg/m^3 over the Moravian-Silesian Region. Pardubice, Karlovy Vary, and South Moravian Regions also had average concentrations higher than 25 μg/m^3 during this season. In 2019, only the eastern part of CZ had an average concentration higher than 25 μg/m^3. Air quality improved throughout the study period; the Moravian-Silesian Region recorded the highest average value of 20 μg/m^3 in 2020 winter.

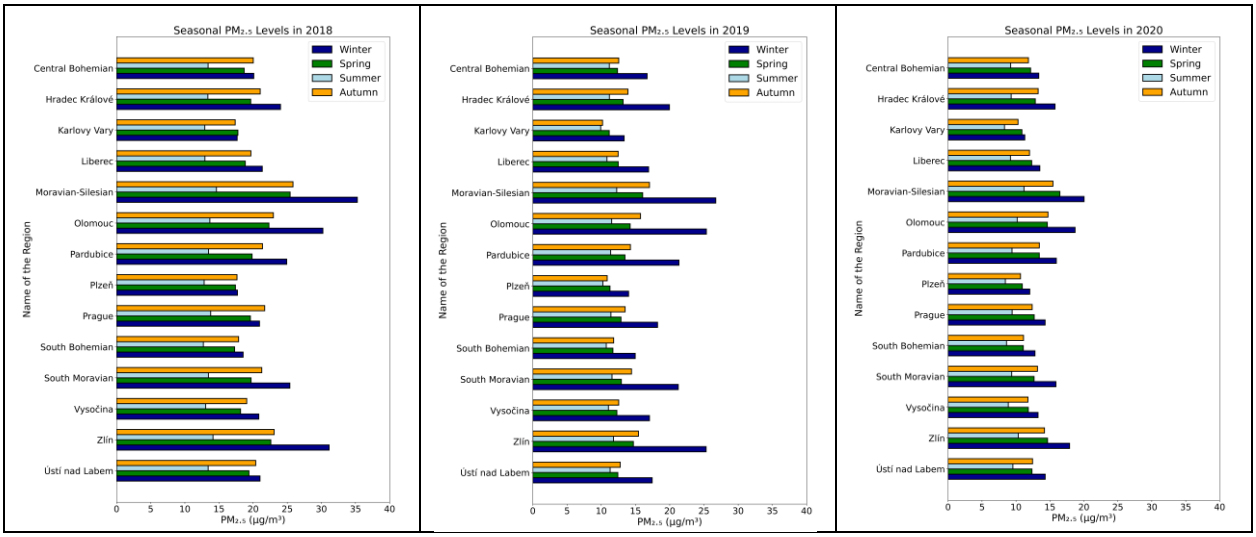

Fig. 6 - $PM_{2.5}$ seasonal analysis over the Czech Republic in 2018, 2019, and 2020.

DISCUSSION

In this study, we used a tree-based machine learning algorithm called the Extra Trees to estimate PM2.5 over CZ with a high spatial resolution of 1 km during 2018–2020. In contrast to our prior study, which concentrated on the entire Europe [31], we discovered that incorporating more balanced data in terms of spatial and temporal distribution enhances the overall accuracy of the model and simplifies the modeling approach. The R2 obtained from the 10-fold cross-validation of the model developed specifically for CZ was 0.85, whereas the corresponding R2 for the model developed for the entire European region was 0.69 [31]. Dividing the data according to stations, ensured that the model can accurately forecast the absent PM2.5 values in new locations, achieving a high R2 of 0.86 and a low RMSE of 5.61 μg/m3.

The spatial autocorrelation we calculated based on the Local Moran Index had higher feature importance than other spatial independent variables like elevation. Calculating the Local Moran Index can give different results due to factors like the K value and the data's distribution, which are important to consider when using it in machine learning models. It should be noted that

the spatial autocorrelation must be generated from the training data only without including the test data, so the test set remains totally unseen to the model to evaluate its performance in an unbiased way.

Confirming the findings from our earlier study, the independent variables which exhibit a high degree of invariance over the duration of the study, like land cover data or the length of the roads in every 1 km of the grid, will have a lower importance on the model. Unlike other studies that included all input features regardless to their importance in generating the model [38], we showed that excluding these inputs will better generalize the model leading to improved estimations. We believe that the inclusion of temporally varying data will enhance the training process of the model, resulting in increased accuracy. For instance, including road traffic intensity yields more refined estimations compared to relying solely on static factors such as the length of roads. For each year during the study period, the yearly averages were computed by taking a simple average of all the available values per pixel.

The results showed that PM2.5 levels were above the recommended limits in many regions of CZ in 2018. The eastern part suffered from the highest values especially during the winter season where the concentrations reached unhealthy levels with values higher than 30 μg/m3. The part located on the Czech-Polish border is characterized as a significant industrial zone with abundant coal deposits and a long-standing presence of factories involved in power generation and manufacturing of coal specifically used for steel-making purposes. PM2.5 levels found to exceed the limits over Polish cities in winter seasons [39], airborne transport facilitate the inflow of particulate matter from Poland across borders, making it a crucial factor in contributing to elevated air pollution levels in the eastern part of CZ. The average concentrations of PM2.5 during summer season were almost consistent for all regions each year and lower than average concentrations during winter, which indicates high effects of heating on PM2.5 levels of especially over the regions that count on burning coal as the main heating source. The measures that were taken by the government to reduce the usage of coal played an important role in improving air quality in recent years. Moreover, the COVID-19 lockdown had a positive effect on PM2.5 levels in the year 2020 due to decreased industrial activities and reduced transportation emissions [31]. The concentrations of PM2.5 in 2020 were less than 20 μg/m3 in all regions except the Moravian-Silesian Region during winter months. The yearly average PM2.5 concentrations calculated over CZ during 2018–2020 in this study align well with our previous findings [31], this serves as validation for the reliability of the dataset we generated using open PM2.5 data for conducting air quality studies throughout Europe. Even though the western part of the country had low concentrations of PM2.5, we recommend augmenting the number of ground monitors in this part to establish a more extensive network that can be utilized for subsequent analysis. We strongly encourage the ongoing reduction of coal usage for local heating, acknowledging the progress that has already been made in this regard. Besides using green energy especially in the eastern part of the country where the highest concentrations were found.

CONCLUSION

In this study, we estimated daily PM2.5 concentration over the Czech Republic with a high spatial resolution of 1 km throughout 2018-2020. A comprehensive data analysis was applied to tune and generalize the spatiotemporal PM2.5 predictive model. The model achieved high accuracy in estimating missing PM2.5 values with R2 of 0.85, RMSE of 5.42 μg/m3, MAE of 3.41 μg/m3, and bias of -0.03 μg/m3. Leveraging machine learning techniques and incorporating auxiliary data in model construction can enhance our comprehension of both the temporal and spatial fluctuations in PM2.5 concentrations. Based on our findings, the eastern part of the country suffered from the highest concentrations especially over Zlín and Moravian-Silesian Regions where the values for 2018 winter, reached risky average concentrations of 30 μg/m3 and 35 μg/m3 respectively. In contrast to 2018, PM2.5 levels dropped over the whole Czech Republic during the next two years reaching acceptable levels that are less than 20 μg/m3 in almost all regions during

the year 2020. The COVID-19 lockdown played a role in improving air quality due to reduced human activities. The generated dataset can be used to obtain a better understanding of the regional and seasonal PM2.5 concentrations throughout the study period.

FUNDING

This work is co-financed by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/050/OHK1/1T/11 and by the Grant Agreement Connecting Europe Facility (CEF) Telecom project 2018-EU-IA-0095 by the European Union.

ACKNOWLEDGEMENT

The authors sincerely thank the Czech Hydrometeorological Institute for providing PM2.5 observations, NASA EOSDIS for providing the daily MCD19A2 product that is available from the Land Processes Distributed Active Archive Centre (LPDAAC), the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing global reanalysis of atmospheric composition, and the Japan Aerospace Exploration Agency (JAXA) for providing the digital surface model used in this study.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

1. Martins NR, Carrilho da Graça G. Impact of PM2.5 in indoor urban environments: A review. Sustain Cities Soc. 2018;42. doi:10.1016/j.scs.2018.07.011
2. Baklanov A, Molina LT, Gauss M. Megacities, air quality and climate. Atmos Environ. 2016;126. doi:10.1016/j.atmosenv.2015.11.059
3. Pascal M, Falq G, Wagner V, et al. Short-term impacts of particulate matter (PM10, PM10-2.5, PM2.5) on mortality in nine French cities. Atmos Environ. 2014;95. doi:10.1016/j.atmosenv.2014.06.030
4. European Commission. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe (OJ L 152, 11.6.2008, Pp. 1-44).; 2008. https://environment.ec.europa.eu/topics/air/air-quality/eu-air-quality-standards_en
5. Tóthová D. Respiratory diseases in children and air pollution - The cost of - Illness assessment in Ostrava City. Cent Eur J Public Policy. 2020;14(1):43-56. doi:10.2478/CEJPP-2020-0003
6. Mikuška P, Křůmal K, Večeřa Z. Characterization of organic compounds in the PM2.5 aerosols in winter in an industrial urban area. Atmos Environ. 2015;105:97-108. doi:10.1016/J.ATMOSENV.2015.01.028
7. Sram RJ. Impact of Air Pollution on the Health of the Population in Parts of the Czech Republic. Int J Environ Res Public Heal 2020, Vol 17, Page 6454. 2020;17(18):6454. doi:10.3390/IJERPH17186454
8. Seibert R, Nikolova I, Volná V, Krejčí B, Hladký D. Air Pollution Sources' Contribution to PM2.5 Concentration in the Northeastern Part of the Czech Republic. Atmos 2020, Vol 11, Page 522. 2020;11(5):522. doi:10.3390/ATMOS11050522
9. Hůnová I. Erratum: Hůnová, I. Ambient Air Quality in the Czech Republic: Past and Present. Atmosphere 2020, 11, 214. Atmos 2021, Vol 12, Page 720. 2021;12(6):720. doi:10.3390/ATMOS12060720
10. Horák J, Hopan F, Šyc M, et al. Estimation of selected pollutant emissions from solid-fuel combustion in small heating appliances. Chem Sheets. 2011;105(11):851-855. Accessed June 11, 2023. http://www.chemicke-listy.cz/ojs3/index.php/chemicke-listy/article/view/1028
11. Hovorka J, Pokorná P, Hopke PK, Křůmal K, Mikuška P, Píšová M. Wood combustion, a dominant source of winter aerosol in residential district in proximity to a large automobile factory in Central Europe. Atmos Environ. 2015;113:98-107. doi:10.1016/J.ATMOSENV.2015.04.068
12. IEA. Czech Republic 2021 – Analysis - IEA. Published 2021. Accessed June 11, 2023. https://www.iea.org/reports/czech-republic-2021

13. Pavlíková I, Hladký D, Motyka O, Vergel KN, Strelkova LP, Shvetsova MS. Characterization of PM10 Sampled on the Top of a Former Mining Tower by the High-Volume Wind Direction-Dependent Sampler Using INNA. Atmos 2021, Vol 12, Page 29. 2020;12(1):29. doi:10.3390/ATMOS12010029
14. Lee HJ. Advancing Exposure Assessment of PM2.5 Using Satellite Remote Sensing: A Review. Asian J Atmos Environ. 2020;14(4). doi:10.5572/ajae.2020.14.4.319
15. Wang J, Christopher SA. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies. Geophys Res Lett. 2003;30(21):2095. doi:10.1029/2003GL018174
16. Liu Y, Park RJ, Jacob DJ, et al. Mapping annual mean ground-level PM2.5 concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States. J Geophys Res Atmos. 2004;109(D22):1-10. doi:10.1029/2004JD005025
17. Liu B, Ma X, Ma Y, et al. The relationship between atmospheric boundary layer and temperature inversion layer and their aerosol capture capabilities. Atmos Res. 2022;271. doi:10.1016/j.atmosres.2022.106121
18. Li X, Feng YJ, Liang HY. The Impact of Meteorological Factors on PM2.5 Variations in Hong Kong. In: IOP Conference Series: Earth and Environmental Science. Vol 78. ; 2017. doi:10.1088/1755-1315/78/1/012003
19. Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ Geogr. 1970;46:234. doi:10.2307/143141
20. Wang H, Chen Z, Zhang P. Spatial Autocorrelation and Temporal Convergence of PM2.5 Concentrations in Chinese Cities. Int J Environ Res Public Heal 2022, Vol 19, Page 13942. 2022;19(21):13942. doi:10.3390/IJERPH192113942
21. Zhang Y, Zhai S, Huang J, et al. Estimating high-resolution PM2.5 concentration in the Sichuan Basin using a random forest model with data-driven spatial autocorrelation terms. J Clean Prod. 2022;380:134890. doi:10.1016/J.JCLEPRO.2022.134890
22. Wang W, Zhao S, Jiao L, et al. Estimation of PM2.5 Concentrations in China Using a Spatial Back Propagation Neural Network. Sci Reports 2019 91. 2019;9(1):1-10. doi:10.1038/s41598-019-50177-1
23. Lyapustin A, Wang Y, Laszlo I, et al. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. J Geophys Res Atmos. 2011;116(3). doi:10.1029/2010JD014986
24. Inness A, Ades M, Agustí-Panareda A, et al. The CAMS reanalysis of atmospheric composition. Atmos Chem Phys. 2019;19(6). doi:10.5194/acp-19-3515-2019
25. Ibrahim S, Landa M, Pešek O, Pavelka K, Halounová L. Space-time machine learning models to analyze COVID-19 pandemic lockdown effects on aerosol optical depth over europe. Remote Sens. 2021;13(15). doi:10.3390/rs13153027
26. Muñoz-Sabater J, Dutra E, Agustí-Panareda A, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst Sci Data. 2021;13(9). doi:10.5194/essd-13-4349-2021
27. Didan K. MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006. NASA EOSDIS L Process DAAC. Published online 2015.
28. Tadono T, Ishida H, Oda F, Naito S, Minakawa K, Iwamoto H. Precise Global DEM Generation by ALOS PRISM. ISPRS Ann Photogramm Remote Sens Spat Inf Sci. 2014;II-4. doi:10.5194/isprsannals-ii-4-71-2014
29. Elvidge CD, Zhizhin M, Ghosh T, Hsu FC, Taneja J. Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. Remote Sens 2021, Vol 13, Page 922. 2021;13(5):922. doi:10.3390/RS13050922
30. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3-42. doi:10.1007/s10994-006-6226-1
31. Ibrahim S, Landa M, Pešek O, Brodský L, Halounová L. Machine Learning-Based Approach Using Open Data to Estimate PM2.5 over Europe. Remote Sens 2022, Vol 14, Page 3392. 2022;14(14):3392. doi:10.3390/RS14143392
32. Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950;37(1/2):17. doi:10.2307/2332142
33. Anselin L. Local Indicators of Spatial Association—LISA. Geogr Anal. 1995;27(2):93-115. doi:10.1111/J.1538-4632.1995.TB00338.X
34. Zhang C, Luo L, Xu W, Ledwith V. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci Total Environ. 2008;398(1-3):212-221. doi:10.1016/J.SCITOTENV.2008.03.011

35. Schneider R, Vicedo-Cabrera AM, Sera F, et al. A satellite-based spatio-temporal machine learning model to reconstruct daily PM2.5 concentrations across great britain. Remote Sens. 2020;12(22). doi:10.3390/rs12223803
36. Li T, Shen H, Zeng C, Yuan Q, Zhang L. Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment. Atmos Environ. 2017;152. doi:10.1016/j.atmosenv.2017.01.004
37. Wei J, Huang W, Li Z, et al. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens Environ. 2019;231. doi:10.1016/j.rse.2019.111221
38. Stafoggia M, Bellander T, Bucci S, et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ Int. 2019;124:170-179. doi:10.1016/J.ENVINT.2019.01.016
39. Czernecki B, Marosz M, Jędruszkiewicz J. Assessment of Machine Learning Algorithms in Short-term Forecasting of PM10 and PM2.5 Concentrations in Selected Polish Agglomerations. Aerosol Air Qual Res. 2021;21(7):200586. doi:10.4209/AAQR.200586