CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Civil Engineering

Department of Geomatics

Integrating Satellite Remote Sensing, Machine Learning, and GIS for Fine-Scale Analysis of Air Quality: Aerosol Optical Depth Estimation

DOCTORAL THESIS

Ing. Saleem Ibrahim

Doctoral study program: Civil Engineering

Branch of study: Geodesy and Cartography

Doctoral thesis tutor: Prof. Ing. Lena Halounová

PRAGUE, 2023

DECLARATION

Name of doctoral candidate: Ing. Saleem Ibrahim

Title of the doctoral thesis: Integrating Satellite Remote Sensing, Machine Learning, and GIS for Fine-Scale Analysis of Air Quality: Aerosol Optical Depth Estimation.

I declare that this thesis is my original author's work, which has been composed solely by myself under the guidance of my supervisor Prof. Ing. Lena Halounová. All the literature and other resources from which I drew information are mentioned in the list of used literature and are cited accordingly. The work has not been used to get another or the same title.

Prague, October 2023 …….

Signature

Abstract

This thesis aimed to harness the capabilities of machine learning in generating air quality datasets using various data sources including satellite remote sensing, meteorological inputs, land cover, topography, observations from ground monitors, among others. The research was focused on two key pollutants: aerosol optical depth and $PM_{2.5}$ due to high risks they cause to human health. The research resulted in three datasets: two datasets covered entire Europe providing insights into aerosol optical depth and $PM_{2.5}$ concentrations, while the third dataset focused only on $PM_{2.5}$ over the Czech Republic. To accomplish this, the research employed a combination of GIS techniques, image processing, statistics, data analysis, visualizations, and comprehensive machine learning methods. The research processed big data and used open-source software to develop the spatiotemporal machine learning models which were employed to generate the datasets. To ensure the accuracy of findings, the results were validated using different approaches then compared with relevant studies. The datasets created over Europe were the first with full coverage of 1 km spatial resolution, they were made publicly available and have since been used by fellow researchers to enhance their understanding of air quality across different regions in Europe.

Keywords

Air quality, machine learning, GIS, remote sensing, aerosol optical depth, $PM_{2.5}$

Abstrakt

Cílem této disertační práce bylo využít možnosti strojového učení při tvorbě datových sad kvality ovzduší s použitím různých zdrojů dat včetně vstupních dat družicového dálkového průzkumu Země a meteorologických dat, pokrytí území, topografie a měření z pozemních stanic aj. Výzkum byl zaměřen na dvě klíčová znečištění: optickou hloubku aerosolů a $PM_{2.5}$ kvůli jejich vysokému riziku pro lidské zdraví, které s tím souvisí. Výsledkem výzkumu jsou tři datové soubory: dva datové soubory pokrývají celou Evropu a ukazují optickou hloubku aerosolů a koncentraci $PM_{2.5}$, zatímco třetí datový soubor je zaměřen pouze na $PM_{2.5}$ v České republice. Aby toho bylo dosaženo, byla použita kombinace nástrojů GIS, zpracování obrazu, statistická a datová analýza, vizualizace a obsáhlé metody strojového učení. Pro vývoj časoprostorových modelů byla zpracována "velká data" za použití open-source software. Tyto modely sloužily pro vytvoření těchto datových sad. Výsledky byly pro zajištění přesnosti výstupů validovány různými způsoby a poté byly porovnány s významnými studiemi. Soubory dat vytvořené nad Evropou byly první s plným pokrytím 1km prostorového rozlišení, byly veřejně dostupné a od té doby je používají kolegové výzkumníci k lepšímu pochopení kvality ovzduší v různých regionech Evropy.

Klíčová slova

Kvalita ovzduší, strojové učení, GIS, dálkový průzkum Země, optická tloušťka aerosolů, $PM_{2.5}$

Acknowledgments

I express heartfelt gratitude to my supervisor, Prof. Lena Halounová, for giving me this opportunity and for her unwavering support and guidance throughout this research. Your expertise and encouragement were instrumental in the completion of this thesis and enhancing my research capabilities.

I extend my sincere appreciation to Ing. Jana Rückerová for the support and guidance which helped me secure this opportunity.

I am also thankful to my academic mentors and colleagues in the department of Geomatics, your dedication and resources have helped me to grow as a researcher. I am fortunate to be part of the projects we worked on.

My greatest thanks go to my beloved family. To my father, for continuously pushing me to finish what I started and to my mother, for your sacrifices and constant encouragement. Your endless belief in my abilities has been the driving force behind my success.

To my sisters Odette and Rasha, thank you for being my pillars of support and for always cheering me on. Your love, understanding, assistance, and patience have been invaluable to me in every step of this journey.

To my uncle Wael, thank you for teaching me the invaluable lesson that struggles are temporary and that true winners never quit.

To Hussam, Basem, Nazih, Tia, Jad, and Majd thank you for encouraging me and always reminding me of the end goal.

https://doi.org/10.14311/dis.fsv.2023.017

List of grants:

I am deeply grateful for the invaluable support and trust placed in my work by the granting organizations listed below. Their generous contributions have played a pivotal role in the realization of my research goals.

Grant Agreement Connecting Europe Facility (CEF) Telecom project 2018-EU-IA-0095 by the European Union.

Grant Agency of the Czech Technical University in Prague, grant No. SGS21/054/OHK1/1T/11.

Grant Agency of the Czech Technical University in Prague, grant No. SGS22/047/OHK1/1T/11.

Grant Agency of the Czech Technical University in Prague, grant No. SGS23/050/OHK1/1T/11.

Table of contents

1. Introduction

Air pollution is one of the most serious problems facing humans nowadays. Many factors contribute and augment this problem, including population inflation and rapid industrial revolution. Air pollution affects the health of humankind and other forms of life. Long term exposure to fine particles produced by combustion is highly correlated with lung cancer [1]. In 2016 around 7.5% of deaths were related to air pollution worldwide [2]. Aerosols are diminutive particles that are regarded as atmospheric contaminants [3]. Aerosol Optical Depth (AOD) is a measure of the columnar atmospheric aerosol content, these particles could absorb or scatter the sunlight preventing it reaching the ground [4]. These small solid or liquid particles are suspended in the atmosphere, and they differ in size, shape, and chemical properties [5]. High levels of AOD have a negative impact on all living things by affecting the respiratory system beside reducing naked eye visibility [6]. The effect of these particles on human health varies according to their size and chemical composition; particle matter (PM) with a diameter of less or equal 10 μm (PM_{10}) can infiltrate the tracheobronchial and such particles become more dangerous as the diameter gets smaller. For example, $PM_{2.5}$ particles are very harmful as they not only cause severe respiratory problems but also reach the systemic circulation [7], further discussed in the text below. Humans are not the main cause of aerosols since they come from many resources like fires, volcanoes, burning of fossil fuels, dust storms and sea drizzles. AOD causes both direct and indirect effects on climate systems according to the lightness or darkness of these particles, in addition to affecting the atmospheric radiation energy balance [8]. Better understanding of aerosol distribution and characteristics is essential for climate change studies [9].

2. Motivation and problem statement

It is not possible to solely rely on ground observations to study AOD or small particles like $PM_{2.5}$ on a large scale due to the considerable number of required monitoring stations and substantial costs and efforts associated to establishing and maintaining them. For this reason, air quality researchers had to find alternative methods to measure these particles.

The rapid development of remote sensing techniques and accurate satellite observations provided solutions to study air quality on regional and global levels. AOD products can be obtained from many satellite sensors. The MOderate Resolution Imaging Spectroradiometer (MODIS) which was focused on in this research, is considered the first satellite plan that provides accurate information of aerosol optical characteristics. Both the Terra and the Aqua

satellite platforms are carrying MODIS instrumentations in a sun-synchronous, near-polar orbits, since 1999 and 2002, respectively [10]. These two satellites can record the earth's surface with 2330 km viewing swath width every 1 to 2 days [11]. MODIS measures 36 spectral bands between 0.4 and 14.4 μm wavelengths at many different spatial resolutions which provides a great opportunity to study the thickness of aerosols and their size characteristics from space with good accuracy, covering the entire world [12,13]. This information helps researchers to estimate AOD loads caused by human-being activities and distinguish it from natural causes [14].

MODIS data has been used to provide useful information on climate changes. Yet, there are many limitations facing satellite aerosol retrieval, including the radiometric calibration, cloud screening, surface reflectance estimation, and aerosol model presumption [15,16]. Several algorithms have been developed by researchers to harness the observed radiances from MODIS in order to derive numerous crucial aerosol products, aiming for improved results. The main purpose of developing and modifying these algorithms is to comply better with the observing instrument specification, properties of aerosols, and nature of clouds [9]. In recently updated products, Quality Assurance (QA) dataset was added, which serves as a check point for certain conditions that are to be met during the retrieval process [17].

Ground measurements are used to validate the results obtained from MODIS sensors. Based on such comparison, MODIS retrieving algorithms could be further improved to reach a satisfactory outcome [18,19]. NASA co-sponsors a global network of ground sensors called the Aerosols Robotic Network (AERONET), which is considered one of the most common and reliable aerosol networks [20]. This network is used to validate satellite retrievals.

While satellite remote sensing can be a valuable source for studying AOD, these data have a great number of gaps due to cloud cover, snow reflectance, and instrument limitations. An analysis of the spatial and temporal distribution of clouds retrieved by MODIS over 12 years of continuous observations from the Terra satellite and over 9 years from the Aqua satellite showed that clouds cover ~67% of the earth's surface worldwide and ~55% over land [21]. Even though many algorithms were designed to remove clouds from satellites-based observations, the AOD retrievals of these clouds have many uncertainties [22] and hence the available daily retrievals are used to calculate the average AOD at annual, seasonal, and monthly time scales. Other methods were used to overcome the low spatial coverage in satellite based AOD products, by simply combining multiple products together or using more advanced methods like artificial intelligence algorithms.

The latter methods resulted in developing various models that fill the gaps either by removing the clouds [23], applying spatiotemporal interpolation [24], or merging different sources of data to predict gaps-free images [25]. These techniques aided air quality researchers in exploring more detrimental pollutants through the analysis of satellite data, such as fine Particulate Matter (PM) with diameters less than 10 micrometers (PM_{10}) or less than 2.5 micrometers ($PM_{2.5}$). $PM_{2.5}$ can penetrate deep into the lungs and may reach the blood circulation causing dangerous diseases such as cardiovascular problems, diabetes, prenatal disorder and even mortality [26–29]. Numerous techniques were used to increase $PM_{2.5}$ spatial coverage provided by ground-based monitors, in other words, to estimate the pollutant concentrations in the areas where no monitors do exist. Examples of that are interpolation techniques that count only on the ground stations [30, 31].

The accuracy of these interpolations is highly related to the spatial distribution of the stations; although they can have good estimations in the areas that are surrounded by the network stations, they will probably fail to have good estimations where there is a lack of the stations [30]. A positive correlation between satellite based AOD and surface particulate matter was found [32, 33]. In the last few decades, artificial intelligence models have been applied to estimate $PM_{2.5}$ and were found to give a better description of the complex non-linear relationship between $PM_{2.5}$, AOD and other independent variables [34], based on the usage of machine learning algorithms [25, 35, 36], or deep neural networks [37, 38]. These algorithms utilize satellite observations, various modelled meteorological variables like planetary boundary layer height (PBLH), wind speed (WS), relative humidity (RH), and temperature (T), in addition to other data like population, land use, land cover, etc. to estimate $PM_{2.5}$. The importance of the inputs differs from one area to another but generally, they can enhance the AOD-PM correlation and provide better estimations since counting solely on AOD to estimate near-surface particulate matter values is not sufficient [39]. AOD without other variables was not enough to provide good $PM_{2.5}$ estimations over Europe [40].

Considering the aforementioned points, the objective of this research was to enhance the current state-of-the-art methods for estimating AOD and $PM_{2.5}$ levels across Europe with the ultimate aim of creating a reliable mapping of these pollutants.

3. Goals of the Thesis: Objectives and Milestones

The author aimed in this research to achieve the following goals:

- The first goal was to acquire a deeper understanding of MODIS algorithms, including their strengths and limitations when applied to AOD products.

- Increase the spatial coverage of MODIS AOD products.
- Establish the first AOD dataset with full coverage of 1 km spatial resolution over Europe.
- Analyze the effects of the COVID-19 lockdowns on AOD levels over Europe.
- Establish the first $PM_{2.5}$ dataset with full coverage of 1 km spatial resolution over Europe.
- Analyze the effects of the COVID-19 lockdowns on $PM_{2.5}$ levels over Europe.
- Include the spatial autocorrelation of the ground based $PM_{2.5}$ observations while developing the machine learning model.
- Compare the performance of two machine learning models: one trained on data from all of Europe, and the other trained exclusively on data from the Czech Republic.

4. Research methods used

The integration of the following tools contributed to a comprehensive and robust methodology, enabling a thorough investigation of the research objectives. We used open-source software and open data in our research.

4.1. Data sources

- Remote sensing data: utilized to retrieve spatial information and satellite observations. We used various MODIS data to extract AOD and NDVI and Visible Infrared Imaging Radiometer Suite (VIIRS) to extract population data.
- Ground-based observations: employed to gather real-time, on-site measurements for training the models, validation, and comparison. Two parameters were used, AOD from AERONET stations and $PM_{2.5}$ from various monitors across Europe.
- Modelled data: integrated to enhance predictive capabilities and provide comprehensive insights. To overcome the low spatial coverage provided by satellite data for AOD, we used Copernicus Atmosphere Monitoring Service (CAMS) modelled AOD with high temporal resolution as inputs in the machine learning model we developed to generate full coverage AOD dataset. Moreover, we used the modelled meteorological variables provided by the European Centre for Medium-Range Weather Forecasts reanalysis (ECMWF) as auxiliary data while developing the predictive models for $PM_{2.5}$ and for other analysis.
- Land cover and topography data: land cover data were extracted from the 2018 CORINE Land Cover (CLC) and the Japan Aerospace Exploration Agency (JAXA) digital surface model was used as auxiliary data while developing the machine learning models.

4.2. Data management and analysis

- Spatial database: employed to efficiently store, manage, and analyze geospatial data. This allowed for streamlined querying, filtering, and manipulation of spatial information, enhancing the accuracy and depth of our analysis.

4.3. Geospatial analysis

We used two open-source software, QGIS and GRASS GIS for geospatial visualization, manipulation, and interpretation. Both software are user-friendly, providing many tools and plugins besides the possibility of scripting for customization.

4.4. Quantitative Analysis

- Statistics: utilized in data preprocessing and cleaning especially when we had to deal with big data while developing the AOD models, identifying the outliers that significantly deviate from the open data we used to develop the $PM_{2.5}$ predictive model, and to derive meaningful insights supporting the formulation of conclusions and recommendations.

4.5. Programming languages

- Python: leveraged for data processing, analysis, visualization, and implementing machine learning algorithms.
- SQL: construct queries to extract specific subsets of data from spatial databases, optimizing data retrieval and analysis.
- Bash scripts: developed to automate routine processes especially while processing NETCDF and HDF files, ensuring efficiency and consistency.

5. Outputs of the dissertation thesis

In this section we will present the outcome of our research. Basically, the publications and the datasets we have released and made publicly available.

5.1. Publications

We have published four papers focusing on air quality remote sensing, GIS, and machine learning. In this section we will present the publications as they were published in the mentioned journals.

5.1.1. Statistical Study of MODIS Algorithms in Estimating Aerosol Optical Depth over the Czech Republic

Abstract

As a result of the rapid development of remote sensing techniques and accurate satellite observations, it has become customary to use these technologies in ecological and aerosols studies on a regional and global level. In this paper, we analyse the performance of three Moderate Resolution Imaging Spectroradiometer (MODIS) algorithms in estimating Aerosol Optical Depth (AOD) in the Czech Republic to gain knowledge about their accuracy and uncertainty. The Dark Target (DT), the Deep Blue (DB), and the merged algorithm (DTB) of the MODIS latest collection 6.1 Level 2 aerosol products (MOD04_L2) were tested by comparing its results with the measurements of Aerosol Robotic Network (AERONET) Level 3 Version 2.0 ground station at Brno airport. The DT algorithm is compatible the best with AERONET observations with a correlation coefficient (R = 0.823), retrievals falling within the EE envelope (EE% = 82.67%), root mean square error (RMSE = 0.059), and mean absolute error (MAE = 0.044). The DTB algorithm provided close results of the DT algorithm but with less accuracy, on the other hand the DB algorithm has the lowest accuracy between all, but this algorithm was able to provide a bigger sample size than the other two algorithms.

Keywords: AERONET, AOD, DB, DT, DTB, MODIS, Remote sensing

---

1. INTRODUCTION

Aerosol Optical Depth (AOD) is a measure of the columnar atmospheric aerosol content, these particles could absorb or scatter the sunlight and prevent it reaching the ground [1]. These small solid or liquid particles are suspended in the atmosphere, and they differ in size, shape, and

chemical adaptation [2]. Studying of AOD is obtaining more interest day by day, due to its negative impact on all living things by affecting the respiratory system beside reducing naked eye visibility [3]. Humans are not the main cause of aerosols. Aerosols come from many resources like fires, volcanoes, burning of fossil fuels, dust storms and sea drizzles. AOD causes both direct and indirect effects on climate systems according to the lightness or darkness of these particles, in addition to affect the atmospheric radiation energy balance [4]. Deeper and better understanding of aerosol distribution and characteristics is essential for climate change studies [5]. It is not possible to solely rely only on ground observations in estimating AOD, since this process requires a great number of such stations in order to cover all areas, which requires high costs and efforts. For this reason, researches focused on climate changes had to find alternative methods to measure AOD. One of these effective techniques is the Moderate Resolution Imaging Spectroradiometer (MODIS), which is considered the first satellite plan that can provide accurate information of aerosol optical characteristics. Both the Terra and Aqua satellite platforms are carrying MODIS instrumentations in a sun-synchronous polar orbits, since the year 1999 and 2002, respectively [6]. They are able to record earth's surface with 2330 km viewing swath width every 1 to 2 days [7]. MODIS measures 36 spectral bands between 0.4 and 14.4 μm wavelengths at many different spatial resolutions that provides a great opportunity to study aerosols thickness and parameters characterizing aerosol size from space with good accuracy and on a world-wide scale [8,9], this information helps researchers to estimate AOD loads caused by human-being activities and distinguish it from natural causes [10]. MODIS data has been used to provide useful information on climate changes. Yet, there are many limitations facing satellite aerosol retrieval, including the radiometric calibration, cloud screening, surface reflectance estimation, and aerosol model presumption [11,12]. To get better results from MODIS, several algorithms were designed and developed to use the observed radiances for deriving many important aerosol products. The main purpose of modifying these algorithms is to comply better with the observing instrument specification, properties of aerosols, and nature of clouds [6]. Updated versions of operational aerosol products have been made available over the years, and because of the improvements of these products, we have new datasets collections continuously, starting with collection 4 (C4) to C5, C6, and the latest collection (C6.1) which was released in July 2017.

MODIS Characterization Support Team (MCST) has produced the C6.1 aerosol products, based on the new updated Level 1B calibrated radiance products [13]. Additionally, NASA Ocean Biology Processing Group (OBPG) developed more calibration corrections, and these improvements were applied to the MCST top of atmosphere (TOA) products starting with C5

[14,15]. MODIS C6.1 aerosol products have major improvements in both radiometric calibration and all aerosol retrieval algorithms.

MODIS products include many scientific data sets (SDS). In recent updated products, Quality Assurance (QA) dataset is added, which serves as a check point for certain conditions that are to be met during the retrieval process [16]. At the end of the process, QA dataset will provide confidence level; 0 = no retrieval, 1 = poor quality, 2 = moderate quality and 3 = good quality [17]. Since the launch of Terra and Aqua satellites, the Dark Target (DT) algorithm which was proposed by [2] has been applied to the MODIS data. There are two distinct DT algorithms for retrieving AOD, one for retrieving AOD over ocean and the second for retrieving AOD over land. Many improvements were applied to the latest algorithm especially of estimating the model for main urban surfaces [18]. The most common used SDS for the DT algorithm is "Optical-Depth-Land-And-Ocean" it contains only filtered values of AOD retrievals which meet the quality assurance (QA ≥ 1 over ocean and QA = 3 over land) to provide beneficial retrievals over dark areas [19]. By contrast, this algorithm has disadvantages over bright surfaces. For this reason, another algorithm called the Deep Blue (DB) was developed in order to retrieve AOD over bright surfaces like deserts and arid areas [20,21]. Since the releasing of C6, DB has been improved to work affectively over vegetated land surfaces, brighter deserts and urban areas [15]. In the latest C 6.1 DB algorithm was developed from collection 6. It has the following advantages over land, the ability to detect thick smoke, efficient modeling for terrains, and many bug fixes, among others mentioned elsewhere [13]. Beside DB and DT products, there is a merged dataset consisting of both DT and DB algorithms (DTB). This merged algorithm works based on the Normalized Difference Vegetation Index (NDVI). According to this methodology, if NDVI > 0.3 then the DT algorithm will be applied on the retrievals, if NDVI < 0.2 then the DB algorithm will be applied, and if NDVI value is between 0.2 and 0.3 then the combined algorithm of both DT and DB will be applied. DTB dataset offers better spatial coverage especially for low vegetated areas [19].

To validate the results obtained from MODIS or other satellite sensors, data is usually compared with the measured aerosol parameters of ground-AERONET. A similar regional study by Zawadzka and Markowicz compared the Spinning Enhanced Visible Infrared Radiometer (SEVIRI) data with AERONET observations in Poland and their study showed a good correlation with a root mean square error (RMSE) equals to 0.05 [22]. Based on such comparison, MODIS retrieving algorithms could be further improved to reach a satisfactory outcome [23,24].

2. Data description

2.1. MODIS Data

Two worldwide products are included in the MODIS level-2 daily swath, MxD04-L2 at 10 km resolution and MxD04-3k at 3 km resolution, whereas: x = O for Terra, and x = Y for Aqua. In this study we use the level-2 daily product at 10 km resolution MOD04_L2 of the TERRA satellite, during the period of 18 months (Jun 2017- Dec 2018) over the Czech Republic. Three AOD subset products; DT, DB, and the merged DTB at 550 μm, are generated from the MODIS latest collection C 6.1. All data are publicly available and were downloaded from https://ladsweb.modaps.eosdis.nasa.gov/.

Table. 1: Scientific dataset of MODIS used in this study.

Product	(SDS) name	Contents	Spatial resolution
MOD04-L2 C6.1	Optical-Depth-Land-And-Ocean	DT over land (QA=3)	10 Km
	Deep-Blue-Aerosol-Optical-Depth-Land-Best-Estimate	DB over land (QA ≥2)	
	AOD-550-Dark-Target-Deep-Blue-Combined	DTB over land and ocean	

2.2. AERONET Data

NASA co-sponsors a global network of ground sensors called the Aerosols Robotic Network (AERONET), which is considered one of the most common and reliable aerosol networks [25]. It is a multi-channel instrument that takes automatic measurements for both direct solar irradiance and sky radiance at the Earth's surface. AERONET takes observations of the solar radiation at seven wavelengths (380, 440, 500, 675, 870, 936 and 1020 nm) around every 15 minutes with low uncertainty ranging between (0.01-0.02) under cloud-free conditions [26]. The AOD is retrieved from these channels to provide high accuracy and quick results. The latest version of AERONET is version three (V3) level two (L2.0) which is computed for three data quality levels: Level 1.0 (unscreened), Level 1.5 (cloud-screened and quality controlled), and Level 2.0 (quality-assured). Inversions, precipitable water, and other AOD-dependent products are derived from these levels [27]. In the Czech Republic there is only one AERONET station. This AERONET CIMEL instrument has approximately 1.2º full angle field of view (FOV) and it is installed on the roof of the administrative building in Brno Airport (Fig. 1) at the following coordinates: latitude 49.15647º N, longitude 16.68333º E, and with an elevation of 238 m above sea level, this station

can observe and process the data automatically, and it is calibrated yearly to provide the best results, and to avoid offsets occurrence in the radiance measurements [28].

In this study, we present data from level 2.0 of the data quality assurance. AERONET AOD measurements at 440 μm and 675 μm from Brno Airport station during the period (June 2017 – December 2018). These observations were interpolated to 550 nm, in order to compare it with MODIS retrievals, using the Angstrom exponents (440 – 675 μm) provided in the AERONET datasets according to the Angstrom's turbidity equation [29] represented in Equation (1).

$$\tau_a(\lambda) = \beta\lambda^{-\alpha} \qquad (1)$$

the AOD values at two different wavelength values $\lambda 1$, $\lambda 2$ are related by Eq (2).

$$\tau_a(\lambda_1) = \tau_a(\lambda_2) * \left(\frac{\lambda_1}{\lambda_2}\right)^{-\alpha} \qquad (2)$$

where $\tau_a(\lambda)$ is the AOD at a wavelength λ in microns, α is the Angstrom wavelength exponent, and β is the Angstrom's turbidity coefficient.

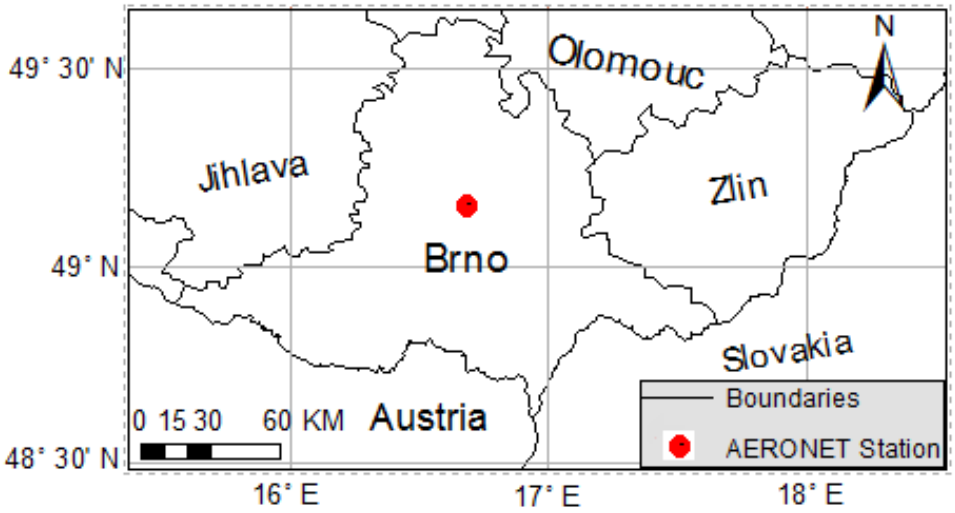


Fig. 1 – Geographical boundaries of the area of study. The red dot represents the location of the AERONET station.

3. Methodology

The comparison takes place between the average of Brno AERONET observations in the period (± 30 minutes) of the Terra satellite passing over this station (approximately 10:30 am), and the mean value of AOD retrievals at 550 μm of nine-pixel sample centered on this AERONET station, at least three pixels should be available and have the required quality assurance, QA=2,3 for DB, and QA=3 for DT and DTB. Considering that AODAERONET represents the true value [30]. To determine the uncertainty of retrieving algorithms with a sample size (N) versus AERONET

measurements, we calculate the Pearson product-moment correlation coefficient (R), RMSE, shown in Equation (3), and the Mean Absolute Error (MAE) as presented in Equation (4), to find which algorithm is compatible the best with the ground observations. Moreover, we will use the Expected Error (EE) equation for retrieving AOD over land at the 10 km spatial resolution [31] to determine the quality of retrievals, the EE equation is represented in Equation (5). Retrievals falling within the EE envelopes must meet Equation (6).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum(AOD_{AERONET} - AOD_{MODIS})^2} \quad (3)$$

$$\text{MAE} = \frac{1}{N}\sum|\,AOD_{AERONET} - AOD_{MODIS}\,| \quad (4)$$

$$\text{EE} = \pm\,(0.05 + 0.15 \times \text{AODAERONET}) \quad (5)$$

$$\text{AODAERONET} - |\text{EE}| \le \text{AODMODIS} \le \text{AODAERONET} + |\text{EE}| \quad (6)$$

4. RESULTS AND DISCUSSION OF VALIDATION AND COMPARISON WITH AERONET OBSERVATIONS

After downloading and processing MODIS data, only data satisfying QA requirements corresponding to each algorithm in question were used during the study analysis.

Fig 2 shows the validations of Terra C6.1 DB, DT, and DTB retrievals compared with AERONET AOD measurements at the Brno Airport site from June 2017 to December 2018 (18 months). During the retrieval process, we noticed that the least number of retrievals were obtained from winter months due to thick clouds and snow coverage. According to data analysis, the C6.1 DT AOD retrievals agrees the best with AERONET AOD measurements (R = 0.823), and the percentage of retrievals falling within the EE envelope is remarkably high (82.67%), with an average Mean Absolute Error (MAE = 0.044) and the smallest root mean square error compared to the other algorithms (RMSE = 0.059). DB has the lowest correlation coefficient (R = 0.765), also the error was noticeably high with (RMSE = 0.069 and MAE = 0.052). On the other hand, the DB has a slightly better percentage of data samples that fell within the EE envelope than the DTB retrievals with EE (80.85%) and (80%) respectively. Moreover, DTB retrievals show better results than DB retrievals (R = 0.819), and the error is slightly higher than that of DT retrievals (MAE = 0.047 and RMSE = 0.063). Figure 3 shows the linear regression between each MODIS algorithm retrievals and AERONET observations, it also shows the real error ($\tau_{MODIS} - \tau_{AERONET}$) for each pair of AOD. According to Fig 3, we found that the errors of all three algorithms have normal distribution on both sides of the 1:1 line with close proportions. Besides that, almost all retrievals of the three algorithms with low values of AOD (AOD < 0.1) have small errors. Based on obtained results, we found that the DTB (Fig 3c) was more influenced by the DT (Fig 3b) than DB (Fig 3a).

Besides, the sample size for both algorithms was the same (N = 75) since the required QA value for both the DT and DTB algorithms is 3. DT algorithm alone gave good results. This is of no surprise as the DT algorithm is known to be suitable for highly vegetated areas, such as the Czech Republic. According to Wie et al, the DT is more suitable for highly vegetated and low AOD loading areas in all Europe, which is consistent with our findings [13]. However, one drawback for this algorithm might be the sample size as larger sample size and probably larger coverage area can be obtained by the DB algorithm due to lower QA requirement (QA = 2 or 3). One challenge that faced us during this study is the fact that there is only one AERONET station in the Czech Republic located in Brno. Even this station was under calibration and data from three months (June – August 2018) were missing. However, by merging the data from the years 2017 and 2018 we were able to have MODIS AOD retrievals from the four seasons and increase the reliability of the validation.

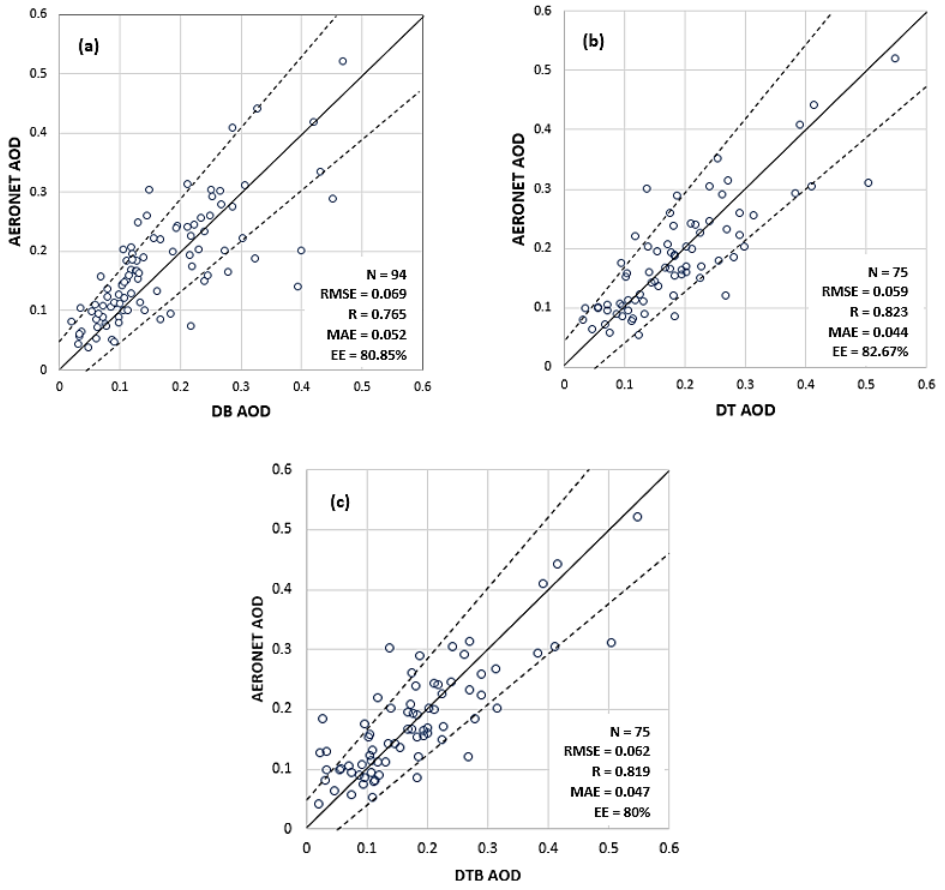


Fig. 2 – Scatter plots of Terra MODIS C6.1 DB (a), DT (b) and DTB (c) AOD retrievals against AERONET AOD observations from June 2017 to December 2018. The solid line indicates the 1:1 line, and the dashed lines indicate the envelopes of the expected error (EE). The sample size (N), correlation coefficient (R), mean absolute error (MAE), and root-mean-square error (RMSE) are also given. EE represents the percentages (%) of retrievals falling within the EE envelopes.

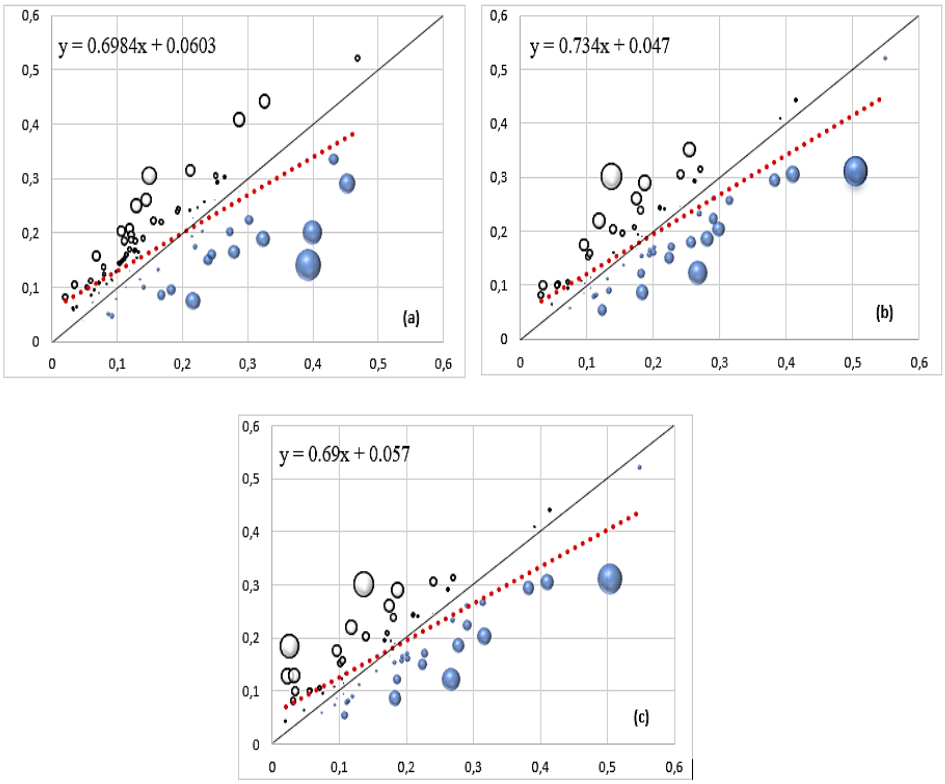


Fig. 3 – The linear regression between MODIS C6.1 DB (a), DT (b), and DTB (c) AOD retrievals against AERONET AOD observations. The X axis represents MODIS retrievals, and the Y axis represents AERONET observations. The solid line indicates the 1:1 line. Each circle represents one pair of MODIS/AERONET AOD, and its size is based on the value of the real error. The blue circles represent the pairs of AOD when ($\tau_{MODIS} > \tau_{AERONET}$) and white circles represent the pairs when ($\tau_{AERONET} > \tau_{MODIS}$).

5. Summary and conclusion

Three AOD products; DB, DT, and DTB, generated from MODIS C6.1, were compared and validated over land at Brno AERONET station (version 3 Level 2) in the Czech Republic during the period (June 2017 till the end of 2018). We investigated the accuracy and uncertainty of the three algorithms in order to draw recommendations. Based on our results, the DT algorithm gave the closest estimations to the real AOD values observed at Brno AERONET station, with a correlation coefficient (R = 0.823), root mean square error (RMSE = 0.059), and with a high percentage of retrievals falling within the EE envelope (EE = 82.67%). The combined algorithm, DTB, failed to bring better estimations than the DT algorithm alone, yet it was found to be more suitable than the use of the DB algorithm solely. The accuracy of the DB was lower than the other two algorithms, yet still acceptable for estimating AOD as 80.85% of retrievals fell within the expected error envelope. We also found that the MODIS coverage is highly affected by NDVI,

among other factors like snow surfaces and cloud density, and thus we recommend testing the coverage of the three MODIS algorithms above all the Czech Republic first and then use the results of the current study to reach an optimal methodology to estimate the AOD over the whole country. Another recommendation would be using the AERONET data of 2019 when it is fully available to investigate whether a longer period influences the results of the current statistics study.

References

[1] Zhang, L., M. Zhang, and Y.B. Yao., 2019. Multi-Time Scale Analysis of Regional Aerosol Optical Depth Changes in National-Level Urban Agglomerations in China Using Modis Collection 6.1 Datasets from 2001 to 2017. Remote Sensing, 11(2).

[2] Kaufman, Y.J., et al., 1997. Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect. Journal of Geophysical Research-Atmospheres, 102(D14): 16815-16830.

[3] Cheng, A.Y.S., M.H. Chan, and X. Yang., 2006. Study of aerosol optical thickness in Hong Kong, validation, results, and dependence on meteorological parameters. Atmospheric Environment, 40(24): 4469-4477.

[4] IPCC (2007), Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by S. Solomon et al., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[5] Poschl, U., 2005. Atmospheric aerosols: Composition, transformation, climate and health effects. Angewandte Chemie-International Edition, 44(46): 7520-7540.

[6] Remer, L.A., et al., 2005. The MODIS aerosol algorithm, products, and validation. Journal of the Atmospheric Sciences, 62(4): 947-973.

[7] Sun, L., et al., 2016. A Universal Dynamic Threshold Cloud Detection Algorithm (UDTCDA) supported by a prior surface reflectance database. Journal of Geophysical Research-Atmospheres, 121(12): 7172-7196.

[8] Tanre, D., M. Herman, and Y.J. Kaufman., 1996. Information on aerosol size distribution contained in solar reflected spectral radiances. Journal of Geophysical Research-Atmospheres, 101(D14): 19043-19060.

[9] Tanre, D., et al., 1997. Remote sensing of aerosol properties over oceans using the MODIS/EOS spectral radiances. Journal of Geophysical Research-Atmospheres, 102(D14): 16971-16988.

[10] Kaufman, Y.J., D. Tanre, and O. Boucher., 2002. A satellite view of aerosols in the climate system. Nature, 419(6903): 215-223.

[11] Li, Z., et al., 2009. Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: a review and perspective. Annales Geophysicae, 27(7): 2755-2770.

[12] Bilal, M., et al., 2013. A Simplified high resolution MODIS Aerosol Retrieval Algorithm (SARA) for use over mixed surfaces. Remote Sensing of Environment, 136: 135-145.

[13] Wei, J., et al., 2019. MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison. Atmospheric Environment, 201: 428-440.

[14] Jeong, M.J., et al., 2011. Impacts of Cross-Platform Vicarious Calibration on the Deep Blue Aerosol Retrievals for Moderate Resolution Imaging Spectroradiometer Aboard Terra. Ieee Transactions on Geoscience and Remote Sensing, 49(12): 4877-4888.

[15] Sayer, A.M., et al., 2014. MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations. Journal of Geophysical Research-Atmospheres, 119(24): 13965-13989.

[16] Levy, R.C., et al., 2010. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. Atmospheric Chemistry and Physics, 10(21): 10399-10420.

[17] King, M.D., et al., 2013. Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. Ieee Transactions on Geoscience and Remote Sensing, 51(7): 3826-3852.

[18] Gupta, P., et al., 2016. A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in MODIS Dark Target retrieval algorithm. Atmospheric Measurement Techniques, 9(7): 3293-3308.

[19] Levy, R.C., et al., 2013. The Collection 6 MODIS aerosol products over land and ocean. Atmospheric Measurement Techniques, 6(11): 2989-3034.

[20] Hsu, N.C., et al., 2006. Deep blue retrievals of Asian aerosol properties during ACE-Asia. Ieee Transactions on Geoscience and Remote Sensing, 44(11): 3180-3195.

[21] Hsu, N.C., et al., 2004. Aerosol properties over bright-reflecting source regions. Ieee Transactions on Geoscience and Remote Sensing, 42(3): 557-569.

[22] Zawadzka, O. and Markowicz, K., 2014. Retrieval of aerosol optical depth from optimal interpolation approach applied to SEVIRI data. Remote Sensing, 6, 7182-7211; doi:10.3390/rs6087182

[23] Chu, D.A., et al., 2002. Validation of MODIS aerosol optical depth retrieval over land. Geophysical Research Letters, 29(12).

[24] Remer, L.A., et al., 2002. Validation of MODIS aerosol retrieval over ocean. Geophysical Research Letters, 29(12).

[25] Holben, B.N., et al., 1998. AERONET - A federated instrument network and data archive for aerosol characterization. Remote Sensing of Environment, 66(1): 1-16.

[26] Holben, B.N., et al., 2001. An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET. Journal of Geophysical Research-Atmospheres, 106(D11): 12067-12097.

[27] Goddard Space Flight Center. Aerosol Robotic Network 2019. Available online at https://aeronet.gsfc.nasa.gov/ (accessed on 10th of May, 2019).

[28] Dubovik, O., et al., 2000. Accuracy assessments of aerosol optical properties retrieved from Aerosol Robotic Network (AERONET) Sun and sky radiance measurements. Journal of Geophysical Research-Atmospheres, 105(D8): 9791-9806.

[29] Liu, Y., et al., 2004. Validation of multiangle imaging spectroradiometer (MISR) aerosol optical thickness measurements using aerosol robotic network (AERONET) observations over the contiguous United States. Journal of Geophysical Research-Atmospheres, 109(D6).

[30] Wei, J. and L. Sun, 2017. Comparison and Evaluation of Different MODIS Aerosol Optical Depth Products Over the Beijing-Tianjin-Hebei Region in China. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(3): 835-844.

[31] Remer, L.A., et al., 2013. MODIS 3 km aerosol product: algorithm and global perspective. Atmospheric Measurement Techniques, 6(7): 1829-1844.

5.1.2. Space-Time Machine Learning Models to Analyze COVID-19 Pandemic Lockdown Effects on Aerosol Optical Depth over Europe

Abstract

The recent COVID-19 pandemic affected various aspects of life. Several studies established the consequences of pandemic lockdown on air quality using satellite remote sensing. However, such studies have limitations, including low spatial resolution or incomplete spatial coverage. Therefore, in this paper we propose a machine learning-based scheme to solve the pre-mentioned limitations by training an optimized space-time extra trees model for each year of the study period. The results have shown that our trained models reach a prediction accuracy up to 95% when predicting the missing values in the MODIS MCD19A2 Aerosol Optical Depth (AOD) product. The outcome of the mentioned scheme was a geo-harmonized atmospheric dataset for aerosol optical depth at 550 nm with 1km spatial resolution and full coverage over Europe. As an application, we used the proposed machine learning based prediction approach in AOD levels analysis. We compared the mean AOD levels between the lockdown period from March to June in 2020 and the mean AOD values of the same period for the past 5 years. We found that AOD levels dropped over most European countries in 2020, while increased in several eastern and western countries. The Netherlands had the most significant average decrease in AOD levels (19%), while Spain had the highest average increase (10%). Moreover, we analyzed the relationship between relative percentage difference of AOD and four meteorological variables. We found a positive correlation between AOD and relative humidity and a negative correlation between AOD and wind speed. The value of the proposed prediction scheme is further emphasized by taking into consideration that the reconstructed dataset can be used for future air quality studies concerning Europe.

Keywords

Aerosol Optical Depth; CAMS; COVID-19; Machine Learning; MODIS

---

1. Introduction

The Severe Acute Respiratory Syndrome-COronaVIrus Diseases 2019 (SARS-COVID-19) pandemic made humanity reconsider how to adapt daily activities. By late June 2020, the EU average infection rate was around 160 per million inhabitants [1]. In general, most European countries started applying restrictions in March 2020. These restrictions included lockdown,

contain, various kinds of curfew, mandatory face masks, etc. By the 18th of March 2020, more than 250 million people in Europe were in lockdown [2].

Despite the unfortunate losses in human lives and economy, there could be a bright side to this pandemic when it comes to air quality. Some studies showed that air quality has improved under the applied restrictions. For example, only two weeks of lockdown have reduced urban air pollution in Spain, with essential differences among pollutants. The most considerable reduction was predestined for black carbon and Nitrogen Dioxide (NO_2) by 45%−51% [3].

According to data released in 2019−2020 by the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA), NO_2 was reduced up to 30% in some regions that were highly affected by COVID-19 lockdowns such as Wuhan in China, Italy, Spain, and USA [4]. Similar results were found in Poland when comparing air quality observations for the year of 2020 in five major cities with the same periods in the previous two years. In addition, AOD concentrations were reduced in April and May of 2020 by nearly 23% and 18% as compared to 2018–2019 [5].

During the lockdown in China, a significant drop in NO_2 (-37%), SO_2 (-64%) and AOD (-8%) for the year 2020 when compared with 11 years mean (2009–2019) [6]. Another study of the Eastern part of China, where AOD levels are usually high (AOD > 0.7), showed that the emission of pollutants in the first three months of 2020, has decreased when compared to the same period of the previous year [7]. Over India, the AOD level was greatly decreased (~45%) during the COVID-19 lockdown periods as compared to the mean AOD level in the previous 20 years [8]. Similarly, significant reductions in black carbon concentration (~8.4%) and AOD (10.8%) were observed in southern India during the first lockdown period (25th March to 14th April 2020) when compared to the pre-lockdown period (1st to 24th March 2020) over the selected measuring location [9].

In this study, we will focus on AOD, which is defined as a measure of the columnar atmospheric aerosol content. High AOD concentrations have a negative impact on all living things by affecting the respiratory system and reducing naked eye visibility. AOD is measured either from ground-based stations or retrieved by satellites measurements. AOD satellite-based products provide a vast spatial coverage when compared to the limited number of ground stations [10].

Due to the correlation between AOD and particulate matter (PM), AOD satellite products are commonly used to retrieve surface PM [11–13]. This justifies the increasing interest in AOD satellite products. Many sensors retrieve AOD at different spatial and temporal resolution [14], such as the Total Ozone Mapping Spectrometer (TOMS) [15], the Ozone Monitoring Instrument (OMI) [16], the Seaviewing Wide Field-of-view Sensor (SEAWIFS) [17], the Geostationary

Operational Environmental Satellite (GOES) [18], the Advanced Himawari Imager (AHI) [19], the Multiangle Imaging SpectroRadiometer (MISR) [20], and the widely used Moderate Resolution Imaging Spectroradiometer (MODIS) which we used in our study.

MODIS instrumentations have been carried on both the Terra and Aqua satellites in sun-synchronous polar orbits, since 1999 and 2002, respectively. They can record earth's surface reflectance and emittance with 2330km swath every one to two days [21]. MODIS measures 36 spectral bands between 0.4 and 14.4 μm wavelengths at many different spatial resolutions that provide a great opportunity to study the aerosol thickness and parameters characterizing aerosol size from space with good accuracy and on a worldwide scale.

MODIS provides various AOD products based on different aerosol retrieval algorithms, the most common algorithms are the Dark Target (DT) [22,23], the Deep Blue (DB) [24,25], and the Multi-Angle Implementation of Atmospheric Correction for MODIS (MAIAC) [26] which is the algorithm used to generate the MODIS MCD19A2 product with 1km spatial resolution.

However, AOD satellite-based products have a great number of gaps due to cloud cover and snow reflectance. An analysis of the spatial and temporal distribution of clouds retrieved by MODIS over 12 years of continuous observations from the Terra satellite and over 9 years from the Aqua satellite showed that clouds cover ~67% of the Earth's surface worldwide and ~55% over land [27]. To solve this issue, it has become common to use machine learning and deep learning algorithms in developing models that fill the gaps in satellite-based products either by removing the clouds [28], applying spatiotemporal interpolation [29], or merging different sources of data to predict gaps-free images [30]. Therefore, in this study we propose a machine learning-based scheme to fill the gaps in MODIS MAIAC AOD retrievals and to generate daily full coverage, high resolution AOD maps over Europe. Such maps will minimize time series analysis bias and uncertainty while investigating the influence of COVID-19 lockdown on AOD levels.

2. Material and Data

2.1. Study Area and Period

The study area is shown in Fig.1. It includes the "Continental EU" hence EEA (European Economic Area) and the United Kingdom, Switzerland, Serbia, Bosnia and Herzegovina, Montenegro, Kosovo, North Macedonia and Albania [31]. In this paper we refer to the area of study as "Europe" located inside this coordinates box 26° W, 72° N, 42° E, and 36° S. The total study area covers 13,391,504 of 1km grid cells; 5,450,009 of the total cell number are located over land. The study period covers the months of March–June from the years 2015–2020.

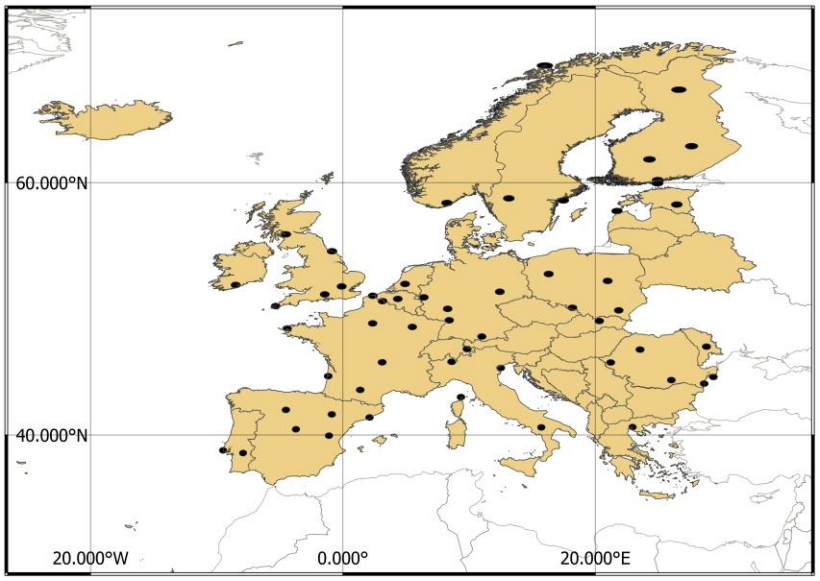


Figure 1. The study area with AERONET stations shown as black dots.

2.2. Data

In this section we summarize different data used throughout our study.

2.2.1. MODIS data

MCD19A2 daily product from MODIS collection 6 was released and made publicly available on 30 May 2018. It is generated from both the Aqua and Terra satellites and delivered in Hierarchical Data Format [26]. MCD19A2 has 1km spatial resolution and uses the MAIAC algorithm that utilizes time series (TMS) analysis, a set of image-based, and pixel processing to enhance the precision of cloud recognition, AOD, and other atmospheric rectification [32,33]. Daily MODIS MCD19A2 data were downloaded, and two science datasets (SDS) were extracted; AOD green band (at 550 nm) and AOD quality assurance layer (AOD_QA) which will be used to retrieve only pixels with the best quality. We created daily mosaics that cover the study area.

2.2.2. Copernicus Atmosphere Monitoring Service (CAMS) data

In this study, modelled AOD at 550 nm data with 80km spatial resolution that is produced by the European center for medium-range weather forecasts Atmospheric Composition Reanalysis 4 (EAC4) was used to fill the gaps in the MODIS MCD19A2 product. Reanalysis merges model data with worldwide observations into a compatible dataset generated by an atmospheric model that uses the laws of physics and chemistry. EAC4 estimates modelled AOD every 3 hours using the 4D-Var assimilation method [34].

2.2.3. Digital elevation model

The elevation of the grid cells was added as a land predictor in our study. The Japan Aerospace Exploration Agency (JAXA) provides a worldwide digital surface model for scientific research and other geospatial services. It provides a horizontal resolution (~30 m) by the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM), which was carried on the Advanced Land Observing Satellite "ALOS" [35]. Data can be accessed from (https://www.eorc.jaxa.jp/ALOS/).

2.2.4. Ground-based AOD data

NASA's Aerosols Robotic Network (AERONET) is considered one of the most reliable aerosol networks [36]. AERONET measures direct solar and sky radiance in various channels every 15 minutes at the local point to compute columnar AOD at intervals from 350 to 1020 nm with low expected uncertainties ranging between 0.01 to 0.02 under cloud-free conditions [37]. There are several categories of AERONET data: level 1.0 (unscreened), level 1.5 (cloud screened), and level 2.0 (cloud screened and quality assured).

In this study, AERONET level 2.0 quality assurance observations were used from 57 stations over Europe as shown in Fig.1. Since AERONET stations do not measure AOD at 550 nm, available measurements at the nearest two wavelengths to 550 nm (440 or 500 nm as λ_1 and 675 nm as λ_2) for each station were interpolated to 550 nm using the Ångström's turbidity equation represented in Equation (1) [21,38].

$$\tau_a(\lambda) = \beta\lambda^{-\alpha} \quad (1)$$

Where $\tau_a(\lambda)$ is the AOD at λ wavelength in micrometers, β is the Angstrom's turbidity coefficient, and α is the band index represented in Equation (2).

$$\alpha = -\frac{\ln(\tau_a(\lambda_1)/\tau_a(\lambda_2))}{\ln(\lambda_1/\lambda_2)} \quad (2)$$

AOD values at two different wavelengths λ_1, λ_2 are related by Equation (3).

$$\tau_a(\lambda_1) = \tau_a(\lambda_2) * \left(\frac{\lambda_1}{\lambda_2}\right)^{-\alpha} \quad (3)$$

2.2.5. European Centre for Medium-Range Weather Forecasts reanalysis (ECMWF)

ERA-5 is the fifth generation of ECMWF reanalysis for the global climate and weather. Hourly data between 10 a.m. and 2 p.m. of u and v wind components, total precipitation, 2m surface temperature for the months March–June of the years 2015–2020 with 0.1° spatial resolution were extracted from the ERA-5 land hourly data. Relative humidity data between 10 a.m. and 2 p.m. at 0.25° spatial resolution was extracted from the ERA-5 monthly averaged data.

All used data were reprojected to the European Terrestrial Reference System 1989 (EPSG:3035), using 1km grid cell with bilinear interpolation method for $CAMS_{AOD}$ and ECMWF data and the cubic convolution for the ALOS elevation model. All values of $MODIS_{AOD}$, $CAMS_{AOD}$ and elevations were assigned to the closest grid cell.

Table 1. Summary of data used in this study.

Product	Spatial resolution	Temporal resolution	Layer
MODIS MCD19A2	1km	Daily	AOD-055 Quality Assurance (QA)
CAMS	80km	3 h	Total aerosol optical depth at 550 nm
ALOS DSM	30m	-	Elevations
AERONET	-	~ 15 min	Level 2.0
ECMWF ERA-5	0.1°	Hourly	Wind U and V components Total precipitation 2m surface temperature
ECMWF ERA-5	0.25°	Monthly	Relative humidity

3. Methodology

In this study, we have created a Geo-Harmonized Atmospheric Dataset for Aerosol optical depth (GHADA) that covers the study area. Three stages were applied to generate GHADA: First, we merged the Terra and Aqua datasets of the MODIS MCD19A2 product by applying a simple average for all pixels that passed the quality assurance criteria ($QA_{CloudMask}$ = Clear and $QA_{AdjacencyMask}$ = Clear) of this product. Second, we created a machine learning model for every year of the study period to predict AOD values over the study area. MCD19A2 high-quality retrievals were used as the dependent variable, and since the Terra satellite is passing locally around 10:30 a.m. and the Aqua satellite passes around 1:30 p.m., we used the modelled AOD from CAMS at the closest three times per day to the satellites passing (9 a.m., 12 p.m., and 3 p.m.). In addition, the spatiotemporal information for the grid cells were used as independent variables. Finally, we filled MODIS MCD19A2 gaps with the predicted AOD by merging the outputs from stages one and two. We validated the daily maps of GHADA with ground-based observation and then we utilized this dataset to analyze how the COVID-19 lockdown has affected AOD levels over Europe during the period of March–June 2020, by comparing AOD levels for this period with the average AOD levels in the last 5 years (2015–2019) for the same months.

4. Space-time models

In this section, we propose a novel approach based on the Extremely Randomized Trees (ET) to predict the missing AOD values in the MODIS MCD19A2 product. first, we illustrate the principles of the ETs and discuss their suitability for the AOD prediction problem. Second, we describe in detail the proposed ET training and parameters setting for AOD prediction.

4.1. Extra trees algorithm

ET is a tree-based ensemble learning method used in our study to deal with the supervised regression and create prediction models for AOD. The idea behind ET is to strongly randomize the selection of both attribute and cut point while splitting a tree node. Unlike the widely used random forest algorithm that chooses the optimum split, ET chooses it randomly which further reduces bias and variance. When needed, the latter algorithm creates independent randomized trees of learning sample output values[38].

The number of attributes that are randomly selected at each node (K) and the minimum sample size for splitting a node (n_{min}) are the two main parameters in ET splitting process. This procedure is applied several times with the whole learning dataset to create an ensemble model that aggregates the predictions of the decision trees to obtain the final estimation, by majority vote in classification problems and arithmetic average in regression problems. In addition to accuracy, ET has high computational efficiency [39], which is required when dealing with big data problems.

4.2. Improved spatiotemporal information

To determine the spatial and temporal correlation between $MAIAC_{AOD}$ and $CAMS_{AOD}$, we included the following independent variables. For space, we used both the elevations of the grid cells and the great circle distance (D) between each grid cell and a reference point on a sphere identified by their latitudes and longitudes using the haversine approach (Equations (4)–(6)). For time, we used the day of the year (DOY) to calculate the radian time (Rt) for the grid cells on different days in a year to improve model handling of the seasonal cycle, Equation (7) [40].

$$\theta = f(\lambda_{i,t}, \varphi_{i,t}) = \text{haversin}(\varphi_1 - \varphi_2) + \cos(\varphi_1) * \cos(\varphi_2) * \text{haversin}(\lambda_1 - \lambda_2) \quad (4)$$

$$\text{haversin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2} \quad (5)$$

$$D_{i,t} = r * \text{archaversin}(\theta) = 2 * r * \arcsin(\sqrt{\theta}) \quad (6)$$

$$Rt_{i,t} = \cos\left(2\pi * \frac{DOYi,t}{T}\right) \quad (7)$$

Where θ is the central angle between two points in space, φ_1 and φ_2 denote the geographical latitudes in radians of two points in space, λ_1 and λ_2 denote the geographical longitudes in radians of two points in space, r denotes the Earth's radius in km, DOY represents the day of the year, T represents the total number of days in the year, for every grid cell (i) on day (t).

For each year between 2015–2020, the model was built using Equation (8).

$$AOD_{i,t} = f(\text{CAMS-9}_{i,t}, \text{CAMS-12}_{i,t}, \text{CAMS-15}_{i,t}, D_{i,t}, H_{i,t}, Rt_{i,t}) \quad (8)$$

Where for each grid cell (i) on day (t): $AOD_{i,t}$ is the target AOD value, CAMS-x represents the AOD value extracted from CAMS at hour x, $D_{i,t}$ represents the great circle distance, $H_{i,t}$ represents the elevation, $Rt_{i,t}$ represents the temporal information identified by the radian time.

5. Results

In this section we present the results of the space-time ET models when predicting the MAIAC AOD values. Then we utilize these models to generate AOD maps over the study area. Validation process is also stated below. Finally, these maps were used to analyze the effects of COVID-19 lockdowns on AOD levels as discussed in section 5.4.

5.1. Models

Due to the great number of $MODIS_{AOD}$ - $CAMS_{AOD}$ pairs over land of the study area (on average 380 million pairs per year), representative subsets consisting of ~10% of the whole population (all $MODIS_{AOD}$ - $CAMS_{AOD}$ pairs per year) were chosen using the Kolmogorov–Smirnov test to be used as learning dataset for a space-time model for each year. Then for each learning dataset, we used the k-fold cross validation (where k=5) to train and validate each model, in this method, the learning dataset is divided into 5 folds, which means 80% of the pairs in the learning dataset are used as training set for the model and the remaining 20% are used for validation, this procedure is repeated 5 times to test the model on each fold. Based on learning curve results, we found that increasing the learning dataset size to 15% will only increase the accuracy of the models by less than 1% and the curve reaches plateau beyond this percentage. Therefore, to decrease the computational complexity, we used ~10% of the whole population as learning dataset. In other words, a learning dataset size of 10% is enough to reach satisfactory accuracy for each year of the study period. The optimized models (number of trees = 30, maximum depth of the tree = 50) were tested on the remaining ~90% (approximately 340 million pairs) of the population.

The results of the trained models for each year are summarized in Table.2. All models achieved high accuracies when predicting MAIAC AOD with a correlation of determination (R^2) ranging between 92.5% to 95% and root mean squared errors from 0.016 to 0.02. These high achieved

accuracies with the relatively small errors show the efficiency of our space-time models in predicting the missing AOD values and emphasize the appropriateness of exploitation modelled AOD with improved spatiotemporal information in improving satellite AOD data.

Table 2. Results of the space-time extremely randomized models used to predict the missing AOD in the MODIS MCD19A2 product for each year of the study period.

Year	R-squared (%)	RMSE	MAE
2015	95	0.017	0.011
2016	94.3	0.018	0.011
2017	93.8	0.018	0.011
2018	92.5	0.02	0.012
2019	92.9	0.019	0.012
2020	94.1	0.016	0.010

Feature importance was calculated based on the reduction in sum of squared errors whenever a variable is chosen to split. Mean importance scores were calculated for all selected input variables of the models (see Fig.2.). $CAMS_{AOD}$ at 12:00 p.m. is the most influential variable accounting for ~33% of $MODIS_{AOD}$ estimates. The other two modelled AOD at 9:00 a.m. and 15:00 contributed by 18% and 24%, respectively. The radian time and the great circle distance had almost the same influence (10–10.4%). Finally, the elevation had the lowest influence with ~5% on $MODIS_{AOD}$ estimates.

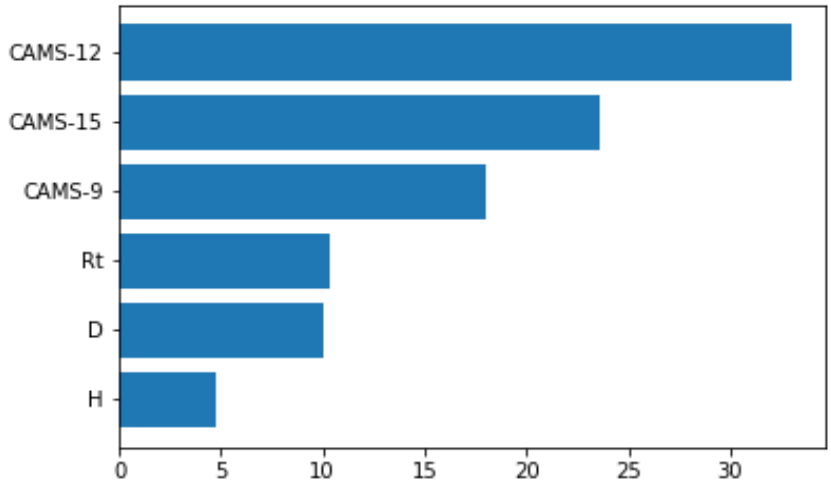
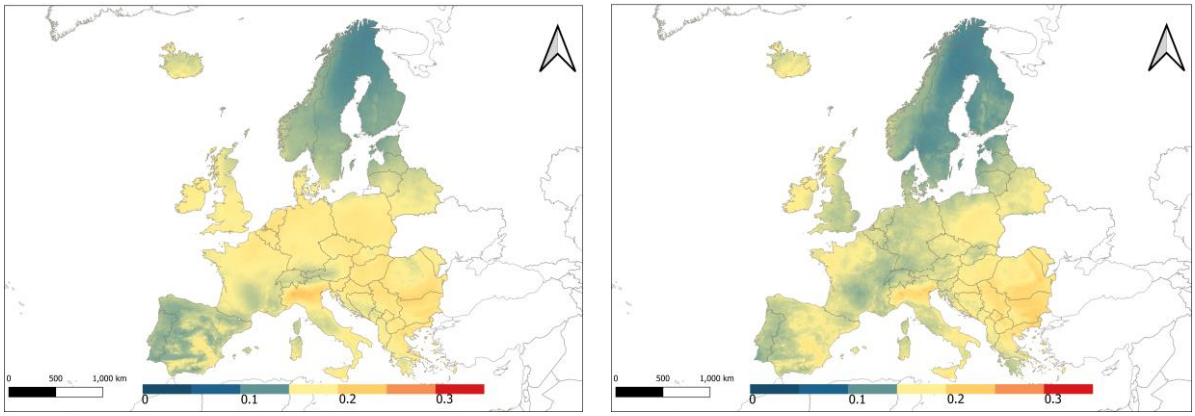


Figure 2. Mean importance scores (%) of independent variables to AOD estimates for the space-time extremely randomized models.

5.2. AOD Maps

We used the optimized space-time models to predict the missing values in the daily MCD19A2 data of the study period. Then we used these predictions to fill the gaps in this product. The outputs of the previous processes were a daily AOD maps with 1km spatial resolution and full coverage over Europe for the period of March–June in the years 2015–2020. To analyze the COVID-19 lockdown effects on AOD levels, we calculated the average AOD levels for the months March-June of the years 2015–2019 and compared these levels with the same period of the year 2020 (see Fig.3.). Moreover, we generated daily AOD maps for the period of January 2018–June 2020 to validate GHADA through all seasons and not solely during the chosen lockdown months.



(**a**) (**b**)

Figure 3. The average AOD values for the months March–June of (a) the years 2015–2019 and (b) of the year 2020 during the chosen lockdown period.

5.3. Validation with AERONET

With the assumption that the aerosol column is relatively uniform within a certain time-space boundary [41], the validation of satellite-based AOD products is usually performed between AOD retrievals within the spatiotemporal window and the corresponding AERONET observations [42]. An acceptable accuracy of AOD products can be achieved when 66% of retrievals fall within expected error envelopes (EE) [23,43]. We used for validation the average AERONET level 2.0 quality assurance observations between 10 a.m. and 2 p.m. from 57 stations across Europe during the period of January 2018–June 2020. We chose two spatial diameters 20km and 50km with AERONET stations in the center for validation and statistical analysis that extensively uses root-mean-square error (RMSE), mean absolute error (MAE), expected error (EE)

envelopes, and the fraction of AOD retrievals of the total number (N) falling within EE envelope (Equations (9)–(13)).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum(AOD_{GHADA} - AOD_{AERONET})^2} \quad (9)$$

$$\text{MAE} = \frac{1}{N}\sum |\, AOD_{GHADA} - AOD_{AERONET} \,| \quad (10)$$

$$\text{Bias} = \frac{1}{N}\sum(AOD_{GHADA} - AOD_{AERONET}) \quad (11)$$

$$\text{EE} = \pm\,(0.05 + 0.15 * \text{AOD}_{\text{AERONET}}) \quad (12)$$

$$\text{AOD}_{\text{AERONET}} - |\text{EE}| \leq \text{AOD}_{\text{GHADA}} \leq \text{AOD}_{\text{AERONET}} + |\text{EE}| \quad (13)$$

The statistical analysis between daily GHADA maps and AERONET observations has shown similar validation results for the two chosen spatial diameters with ~84% of the samples falling within the EE, good correlations R~76–77% and relatively small RMSE ~ 0.066–0.067, refer to Table.3.

Fig.4 represents the density scatter plots for the validation of AOD at 550 nm from GHADA with the AERONET stations at the two chosen spatial diameters.

Table 3. Validation results of GHADA with AERONET at two spatial diameters, where N is the total number of sample points.

D (Km)	N	R	MAE	RMSE	Bias	EE(%)
20	10916	0.762	0.043	0.067	-0.014	83.7
50	12212	0.767	0.043	0.066	-0.014	83.7

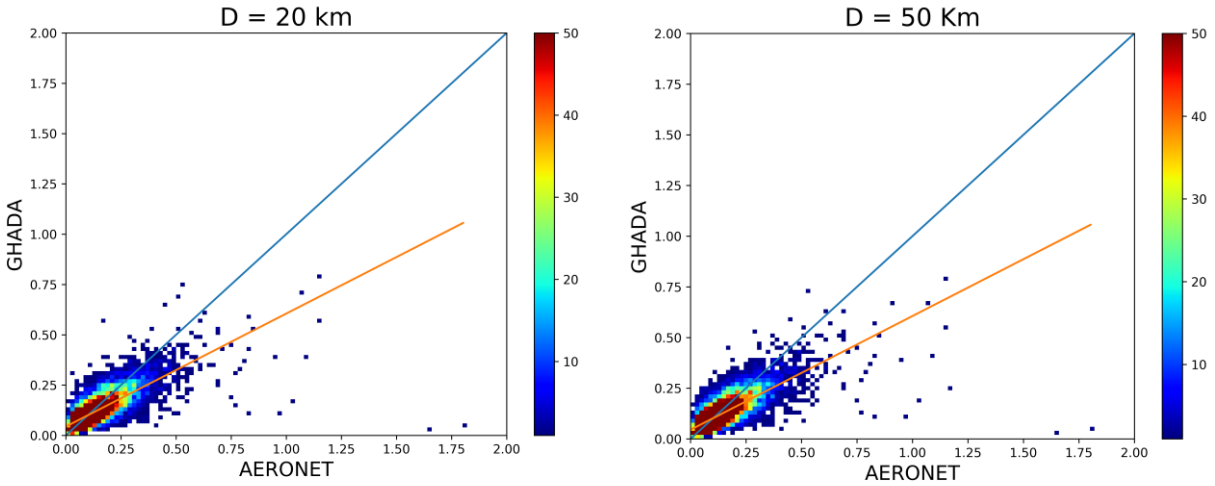


Figure 4. Density scatter plots of validation AOD at 550 nm from GHADA with 57 AERONET stations between 10 a.m. and 2 p.m. at two spatial diameters of 20km and 50km. The colored scale bar stands for the frequency of occurrence.

5.4. AOD Relative Percentage Difference

The variations in AOD levels were calculated for each grid cell using the Relative Percentage Difference (RPD) Equation (14).

$$\mathrm{RPD} = \frac{AOD_{2020} - AOD_{2015-2019}}{AOD_{2015-2019}} * 100 \qquad (14)$$

Where AOD_{2020} is the mean AOD value in the study period of 2020 and $AOD_{2015–2019}$ is the mean AOD value for the study period covering 2015–2019. The changes are presented in Fig.5.

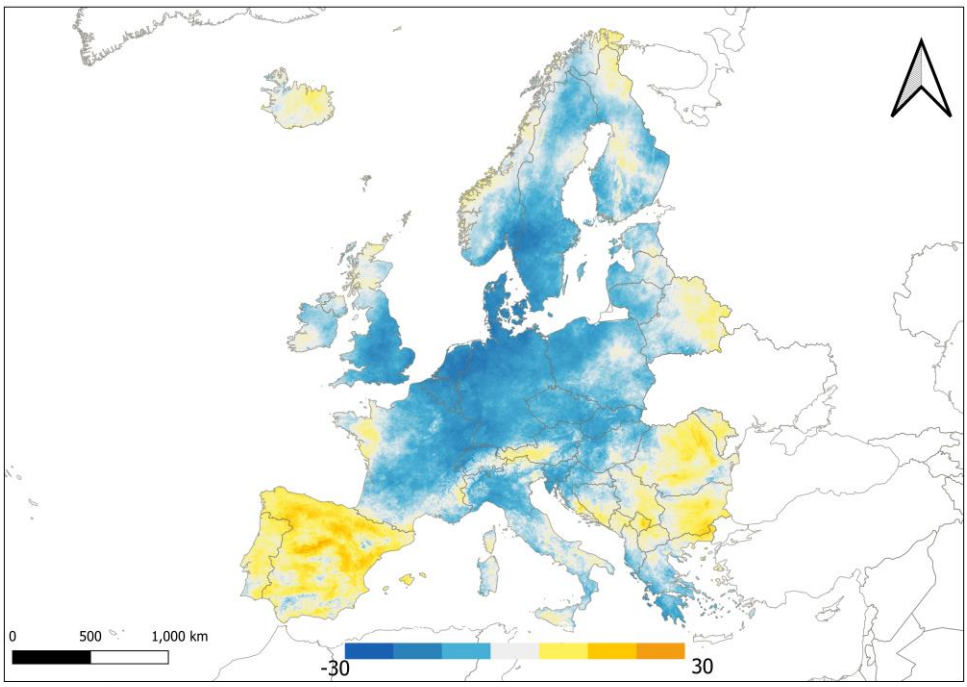


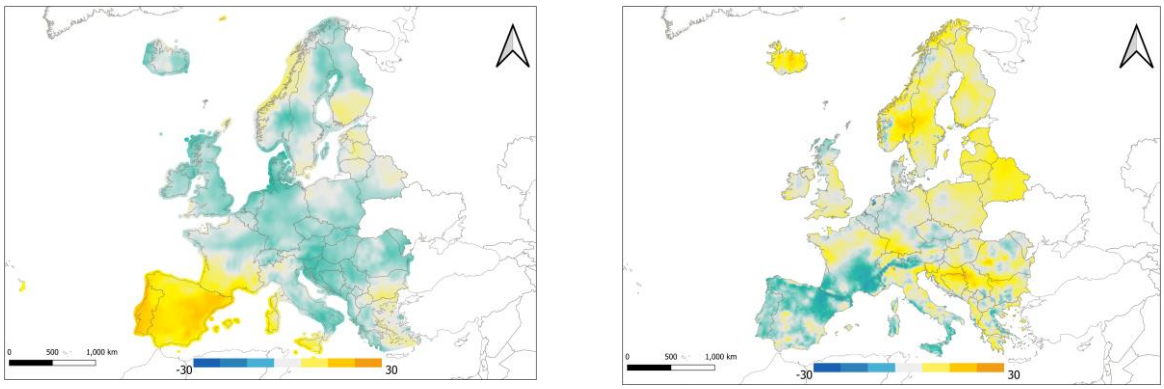
Figure 5. Relative percentage difference of AOD over Europe for the months March–June of the year 2020 and the same months of the previous 5 years.

6. Discussion

In this study, a machine learning-based scheme was used to overcome the limitations in time series analysis concerning AOD. A new dataset for AOD at 550 nm with full coverage over Europe and with 1km spatial resolution (GHADA) was built. We trained an extra trees model for each year (2015–2020) using the MODIS MCD19A2 as the target variable, and CAMS modelled AOD with improved spatiotemporal information as the independent variables. Results showed that the trained models had high accuracies ranging between 92.5–95% when estimating the missing $MAIAC_{AOD}$ retrievals. We compared the AOD_{550} from GHADA and surface observations at 57 AERONET sites over Europe, with two spatial diameters around these AERONET stations within the period of January 2018–June 2020. The overall comparison with ground-based measurements showed a good correlation, with bias as low as 0.014 and R ~0.76. Then we used GHADA to study the influence of COVID-19 pandemic lockdown on AOD levels over Europe in the months March–June by comparing it to AOD levels in the same months for the past five years (2015–2019). The most important advantage of our study when compared to similar work is that we used daily full coverage AOD maps with high spatial resolution when calculating the average AOD values before and after the lockdown. Such complete coverage reduces bias and uncertainty in such time series analysis. As shown above in Fig.5, we have found that AOD levels decreased

by 10–30% over most countries of the study area in 2020, mainly the countries located at the center of the analyzed area. While AOD levels increased over the countries that are located on the boundaries of the study area. In the west, AOD increased over Spain and Portugal, in the east, AOD increased over Romania, Bulgaria, Moldova, and Kosovo; in the north, the level slightly increased over Iceland. The decrease in AOD levels was the greatest in the Netherlands with an average decrease of 20%, while Spain had the highest average increase in AOD levels by 10%. It must be noted that the five AERONET stations in Spain included in this study did not reflect the average increase in AOD over the whole country, due to their limited spatial coverage.

As an attempt to justify the findings in areas of increased AOD, we investigated the relationship between the RPD in AOD for the months March–June of the year 2020 and the previous five years and the RPD for four meteorological variables (relative humidity, wind speed, surface temperature, and total precipitation) calculated for the times of MODIS satellites overpassing (10 a.m. to 2 p.m.). We found a close trend between relative humidity and AOD. Spain, Portugal, northern Norway, eastern Belarus, and southern Bulgaria had higher RPD in both AOD and relative humidity. Spain and Portugal had the highest increase of 10–23% in relative humidity. In agreement, areas of decreased humidity had lower RPD of AOD, however such correlation is to a lower extent than the effect of increased humidity. An exception to this finding is Romania where RPD in humidity was decreased however AOD was increased. Regarding wind speed, RPD has decreased by ~18% in Spain and Portugal where AOD had significant increase. Also, the northern part of Italy and the western part of Austria had a clear inverse trend between AOD and wind speed. Tha average relative humidity over Spain was 65% during the lockdown period of the year 2020. High relative humidity combined with low average wind speed of less than ~3 m/s play an important role in increasing AOD. Our findings are consistent with [44], where they associated higher humidity and lower wind speed with higher AOD. We found no direct relationship between RPD of neither surface temperature nor total precipitation and RPD of AOD, all of which strengthen the argument that lowering AOD is a consequence of the lockdown. Although we proved that AOD levels increased over Spain, other pollutants such as NO_2 were decreased which is attributed to the difference in the source of these pollutants as discussed elsewhere [44]. Fig.6. shows the RPD of relative humidity and RPD of wind speed between the lockdown months of the year 2020 and the same period of the previous 5 years.



(a) (b)

Figure 6. Relative percentage difference of (a) relative humidity and (b) wind speed over Europe between 10 a.m. and 2 p.m. for the months March–June of the year 2020 and the same months of the previous 5 years.

Nevertheless, it must be noted that the average AOD levels over Europe are relatively low (AOD < 0.3) compared to other more polluted regions, where more prominent differences in AOD levels can be observed, for example as published in [8] where AOD levels over India were investigated. In addition, the extent of restrictions imposed and the adherence to them may contribute to the significance of the change in AOD levels.

7. Conclusion

The advancement of machine learning algorithms provides solutions for AOD satellite-based data drawbacks such as low spatial resolution and gaps caused by persistent clouds, cloud contamination, and high surface reflectance, and opens new horizons for studies that shall influence decision making. A machine learning-based scheme was used to enhance time series analysis of AOD over the study period. Space-time extremely randomized trees models were built to fill the gaps in the MCD19A2 product of the moderate imaging spectroradiometer (MODIS). The output was a geo-harmonized atmospheric dataset for aerosol optical depth (GHADA) with complete coverage of 1km spatial resolution over Europe. Up to our knowledge, GHADA is the first dataset with this coverage and resolution for Europe, and we are the first to analyze how the COVID-19 affected AOD levels over Europe with gaps-free AOD maps at high spatial resolution.

We compared AOD levels during the chosen lockdown period to the mean AOD values during the same period in the previous 5 years. We found a general decrease trend in the countries located at the center of the study area with the Netherlands scoring the highest average decrease. In contrast, AOD levels increased in the eastern and western European countries as it is distinctly visible in Kosovo and Spain, respectively. We found a correlation between high humidity and

low wind speed with AOD increase, which justifies such increase in countries like Spain and Portugal. We excluded surface temperature and total precipitation as contributing factors to the detected changes in AOD levels, which in return makes COVID-19 lockdown the major cause for the decrease in AOD levels.

Once GHADA is made publicly accessible, it can be used to investigate air quality over Europe with 1km spatial resolution and improve time series analysis, overcoming the gaps encountered during such studies. The lockdown that happened due to the pandemic generally lowered AOD levels; however, such lockdown is not the ultimate solution to control AOD levels. Cleaner sources of energy and road transport are needed to maintain lower levels of AOD and good air quality. Based on our obtained results, we recommend utilizing machine learning to solve time series analysis limitations, to conduct various applications concerning air quality.

References

1. Boffey, D. US Visitors Set to Remain Banned from Entering EU. Guardian **2020**. Available online: https://www.theguardian.com/world/2020/jun/29/us-visitors-set-to-remain-banned-from-entering-eu (accessed on 1 May 2021).

2. Henley, J.; Oltermann, P. More Than 250 m People Now in Lockdown in EU as Germany and Belgium Adopt Measures. Guardian **2020**. Available online: https://www.theguardian.com/world/2020/mar/18/coronavirus-lockdown-eu-belgium-germanyadopt-measures (accessed on 1 May 2021).

3. Tobias, A.; Carnerero, C.; Reche, C.; Massague, J.; Via, M.; Minguillon, M.C.; Alastuey, A.; Querol, X. Changes in air quality during the lockdown in Barcelona (Spain) one month into the SARS-CoV-2 epidemic. Sci. *Total Environ*. **2020**, 726, 138540.

4. Muhammad, S.; Long, X.; Salman, M. COVID-19 pandemic and environmental pollution: A blessing in disguise? Sci. *Total Environ*. **2020**, 728, 138820.

5. Filonchyk, M.; Hurynovich, V.; Yan, H. Impact of COVID-19 lockdown on air quality in the Poland, Eastern Europe. *Environ. Res*. **2021**, 198, 110454.

6. Soni, M.; Verma, S.; Jethava, H.; Payra, S.; Lamsal, L.; Gupta, P.; Singh, J. Impact of COVID-19 on the Air Quality over China and India Using Long-term (2009–2020) Multi-satellite Data. *Aerosol Air Qual. Res*. **2020**, 21, e200295.

7. Filonchyk, M.; Hurynovich, V.; Yan, H.; Gusev, A.; Shpilevskaya, N. Impact Assessment of COVID-19 on Variations of SO2, NO2, CO and AOD over East China. Aerosol Air Qual. Res. **2020**, 20, 1530–1540.

8. Ranjan, A.K.; Patra, A.K.; Gorai, A.K. Effect of lockdown due to SARS COVID-19 on aerosol optical depth (AOD) over urban and mining regions in India. Sci. *Total Environ*. **2020**, 745, e141024.

9. Kalluri, R.O.R.; Gugamsetty, B.; Tandule, C.R.; Kotalo, R.G.; Thotli, L.R.; Rajuru, R.R.; Palle, S.N.R. Impact of aerosols on surface ozone during COVID-19 pandemic in southern India: A multi-instrumental approach from ground and satellite observations, and model simulations. *J. Atmos. Sol. Terr. Phys*. **2021**, 212, 105491.

10. Liu, Y.; Sarnat, J.A.; Coull, B.A.; Koutrakis, P.; Jacob, D.J. Validation of Multiangle Imaging Spectroradiometer (MISR) aerosol optical thickness measurements using Aerosol Robotic Network (AERONET) observations over the contiguous United States. *J. Geophys. Res*. **2004**, 109.

11. Shen, H.; Li, T.; Yuan, Q.; Zhang, H. Estimating Regional Ground-Level PM2.5 Directly from Satellite Top-of-Atmosphere Reflectance Using Deep Belief Networks. *J. Geophys. Res. Atmos*. **2018**, 123, 13875–13886.

12. Yang, Q.; Yuan, Q.; Yue, L.; Li, T.; Shen, H.; Zhang, L. The relationships between PM2.5 and aerosol optical depth (AOD) in mainland China: About and behind the spatio-temporal variations. *Environ. Pollut*. **2019**, 248, 526–535.

13. Zhang, H.; Hoff, R.M.; Engel-Cox, J.A. The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and PM2.5 over the United States: A geographical comparison by U.S. Environmental Protection Agency regions. *J. Air Waste Manag. Assoc*. **2009**, 59, 1358–1369.

14. Ajtai, N.; Mereuta, A.; Stefanie, H.; Radovici, A.; Botezan, C.; Zawadzka-Manko, O.; Stachlewska, I.S.; Stebel, K.; Zehner, C. SEVIRI Aerosol Optical Depth Validation Using AERONET and Intercomparison with MODIS in Central and Eastern Europe. *Remote Sens*. **2021**, 13, 844.

15. Torres, O.; Bhartia, P.K.; Sinyuk, A.;Welton, E.J.; Holben, B.N. Total Ozone Mapping Spectrometer measurements of aerosolabsorption from space: Comparison to SAFARI 2000 ground-based observations. J. Geophys. Res. **2005**, 110.

16. Torres, O.; Tanskanen, A.; Veihelmann, B.; Ahn, C.; Braak, R.; Bhartia, P.K.; Veefkind, P.; Levelt, P. Aerosols and surface UV products from Ozone Monitoring Instrument observations: An overview. *J. Geophys. Res*. **2007**, 112.

17. Sayer, A.M.; Hsu, N.C.; Bettenhausen, C.; Ahmad, Z.; Holben, B.N.; Smirnov, A.; Thomas, G.E.; Zhang, H. SeaWiFS Ocean Aerosol Retrieval (SOAR): Algorithm, validation, and comparison with other data sets. *J. Geophys. Res*. **2012**, 117.

18. Knapp, K.R.; Frouin, R.; Kondragunta, S.; Prados, A. Toward aerosol optical depth retrievals over land from GOES visible radiances: Determining surface reflectance. *Int. J. Remote Sens*. **2005**, 26.

19. Lim, H.; Choi, M.; Kim, J.; Kasai, Y.; Chan, P.W. AHI/Himawari-8 Yonsei Aerosol Retrieval (YAER): Algorithm, Validation and Merged Products. *Remote Sens*. **2018**, 10, e699.

20. Kahn, R.A.; Gaitley, B.J.; Garay, M.J.; Diner, D.J.; Eck, T.F.; Smirnov, A.; Holben, B.N. Multiangle Imaging SpectroRadiometer global aerosol product assessment by comparison with the Aerosol Robotic Network. *J. Geophys. Res*. **2010**, 115.

21. Sun, L.; Wei, J.; Bilal, M.; Tian, X.; Jia, C.; Guo, Y.; Mi, X. Aerosol Optical Depth Retrieval over Bright Areas Using Landsat 8 OLI Images. *Remote Sens*. **2016**, 8, 23.

22. Levy, R.C.; Mattoo, S.; Munchak, L.A.; Remer, L.A.; Sayer, A.M.; Patadia, F.; Hsu, N.C. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech*. **2013**, 6, 2989–3034.

23. Remer, L.A.; Kaufman, Y.J.; Tanré, D.; Mattoo, S.; Chu, D.A.; Martins, J.V.; Li, R.-R.; Ichoku, C.; Levy, R.C.; Kleidman, R.G.; et al. The MODIS Aerosol Algorithm, Products, and Validation. *J. Atmos. Sci*. **2005**, 62, 947–973.

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

25. Hsu, N.C.; Tsay, S.; King, M.D.; Herman, J.R. Aerosol Properties Over Bright-Reflecting Source Regions. IEEE Trans. Geosci. *Remote Sens*. **2004**, 42, 557–569.

26. Lyapustin, A.;Wang, Y.; Laszlo, I.; Kahn, R.; Korkin, S.; Remer, L.A.; Levy, R.; Reid, J.S. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *J. Geophys. Res*. **2011**, 116.

27. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. IEEE Trans. *Geosci. Remote Sens*. **2013**, 51, 3826–3852.

28. Meraner, A.; Ebel, P.; Zhu, X.X.; Schmitt, M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. *Photogramm. Remote Sens*. **2020**, 166, 333–346.

29. Yang, J.; Hu, M. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Sci. Total Environ*. **2018**, 633, 677–683.

30. 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 PM2.5 Concentrations across Great Britain. *Remote Sens*. **2020**, 12, 3803.

31. Open Data Science Europe. Geo-Harmonizer Project Implementation Plan 2020–2022; Open Data Science Europe: Wageningen, The Netherlands, 2020.

32. Lyapustin, A.;Wang, Y.; Korkin, S.; Huang,W. MODIS Collection 6 MAIAC algorithm. Atmos. Meas. Tech. **2018**, 11, 5741–5765.

33. Lyapustin, A.;Wang, Y.; Laszlo, I.; Korkin, S. Improved cloud and snow screening in MAIAC aerosol retrievals using spectral and spatial analysis. *Atmos. Meas*. Tech. **2012**, 5, 843–850.

34. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys*. **2019**, 19, 3515–3556.

35. 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**, 2, 71.

36. Holben, B.N.; Eck, T.F.; Slutsker, I.; Tanré, D.; Buis, J.P.; Setzer, A.; Vermote, E.; Reagan, J.A.; Kaufman, Y.J.; Nakajima, T.; et al. AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization. *Remote Sens. Environ*. **1998**, 66, 1–16.

37. Holben, B.N.; Tanre, D.; Smirnov, A.; Eck, T.F.; Slutsker, I.; Abuhassan, N.; Newcomb,W.W.; Schafer, J.S.; Chatenet, B.; Lavenu, F.; et al. An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET. *J. Geophys. Res*. **2001**, 106, 12067–12097.

38. Ångström, A. The parameters of atmospheric turbidity. *Tellus* **1964**, 16, 64–75.

39. Geurts, P.; Ernst, D.;Wehenkel, L. Extremely randomized trees. *Mach. Learn*. **2006**, 63, 3–42.

40. 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 PM2.5 estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys*. **2020**, 20, 3273–3289.

41. Anderson, T.L.; Charlson, R.J.;Winker, D.M.; Ogren, J.A.; Holmén, K. Mesoscale Variations of Tropospheric Aerosols. *J. Atmos. Sci*. **2003**, 60, 119–136.

42. Martins, V.S.; Lyapustin, A.; de Carvalho, L.A.S.; Barbosa, C.C.F.; Novo, E.M.L.M. Validation of high-resolution MAIAC aerosol product over South America. *J. Geophys. Res. Atmos*. **2017**, 122.

43. Levy, R.C.; Remer, L.A.; Kleidman, R.G.; Mattoo, S.; Ichoku, C.; Kahn, R.; Eck, T.F. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. *Atmos. Chem. Phys*. **2010**, 10, 10399–10420.

44. Acharya, P.; Barik, G.; Gayen, B.K.; Bar, S.; Maiti, A.; Sarkar, A.; Ghosh, S.; De, S.K.; Sreekesh, S. Revisiting the levels of Aerosol Optical Depth in south-southeast Asia, Europe and USA amid the COVID-19 pandemic using satellite observations. *Environ. Res*. **2021**, 193, 110514.

5.1.3. Machine Learning Based Approach Using Open Data to Estimate $PM_{2.5}$ over Europe

Abstract

Air pollution is currently considered one of the most serious problems facing humans. Fine particulate matter with a diameter smaller than 2.5 micrometres ($PM_{2.5}$) is a very harmful air pollutant that is linked with many diseases. In this study, we created a machine learning based scheme to estimate $PM_{2.5}$ using various open data such as satellite remote sensing, meteorological data, and land variables to increase the limited spatial coverage provided by ground-monitors. A space-time extremely randomised trees model was used to estimate $PM_{2.5}$ concentrations over Europe, this model achieved good results with an out-of-sample cross-validated R^2 of 0.69, RMSE of 5 μg/m3, and MAE of 3.3 $\mu g/m^3$. The outcome of this study is a daily full coverage $PM_{2.5}$ dataset with 1 km spatial resolution for the three-year period of 2018–2020. We found that air quality improved throughout the study period over all countries in Europe. In addition, we compared $PM_{2.5}$ levels during the COVID-19 lockdown during the months March–June with the average of the previous 4 months and the following 4 months. We found that this lockdown had a positive effect on air quality in most parts of the study area except for the United Kingdom, Ireland, north of France, and south of Italy. This is the first study that depends only on open data and covers the whole of Europe with high spatial and temporal resolutions. The reconstructed dataset will be published under free and open license and can be used in future air quality studies.

Keywords

$PM_{2.5}$, AOD, Machine learning, Europe, Open data

---

1. Introduction

Air quality monitoring is one of the most important fields when it comes to the individual's health due to the high risks related to its low quality. Fine particulate matter is an air pollutant that consists of liquid and solid molecules such as acid condensates, sulphates, and nitrates that have negative effects on human health [1]. The harmful effects of these particles vary depending on the concentrations, time exposure, and the particulate diameter. Risks are higher when the diameter gets smaller; $PM_{2.5}$ can penetrate deep into the lungs and may reach the blood circulation causing dangerous diseases such as cardiovascular problems, diabetes, prenatal disorder, and even mortality [2–5]. The effects are more notable in urban areas, where higher

population density can be found, and more exposure will occur [6]. The form of the urban area plays an important role in the concentration of $PM_{2.5}$ [7].

The U.S. Environmental Protection Agency (EPA) has set an annual average standard of 12 $\mu g/m^3$ and a daily (24 h) of 35 $\mu g/m^3$ for $PM_{2.5}$ and when the amounts of these pollutants in the ambient air exceed these limits that could cause serious health issues [8]. The revised Directive 2008/50/EC of the European Parliament (EP) and of the Council on ambient air quality and cleaner air for Europe set limit values of annual $PM_{2.5}$ to 25 $\mu g/m^3$ since 1 January 2015 and not to exceed 20 $\mu g/m^3$ since 1 January 2020. $PM_{2.5}$ ground-based monitors are used to measure $PM_{2.5}$ with high accuracy. These stations are considered the backbone in almost all analyses related to these particles. However, the high cost of establishing these monitors limits the overall spatial coverage and the researchers who are focusing on air quality were seeking new methodologies to increase the spatial coverage, so they have a better understanding on larger geographical scales. Numerous techniques were used to increase $PM_{2.5}$ spatial coverage, in other words, to estimate the pollutant concentrations in the areas where no monitors exist. Examples of that are interpolation techniques that count only on the ground stations [9,10]. The accuracy of these interpolations is highly related to the spatial distribution of the stations; although they can have good estimations in the areas that are surrounded by the network stations, they will probably fail to have good estimations where there is a lack of the stations [9]. Land use regression (LUR) models were also used to analyse pollution, particularly in densely populated areas [11,12].

Satellite remote sensing provides wide spatial coverage compared to the spatial coverage obtained from ground monitors. Aerosol optical depth (AOD) is an air quality indicator that can be observed from satellite remote sensing, and it is defined as the measure of the columnar atmospheric aerosol content. Numerous studies have found a positive correlation between satellite-based AOD and surface particulate matter [13,14]. Researchers have utilised satellite AOD to estimate $PM_{2.5}$ by developing different types of models such as physical models that were built based on the physical relationship between AOD and surface $PM_{2.5}$ [15]. Statistical methods which train the relationship between AOD and $PM_{2.5}$ using different statistical models [16,17] are suitable for the regions with a sufficient number of ground stations since they require a large amount of training data [18]. The generalised additive model (GAM) empowers the AOD–PM relationship by adding meteorological and land use information [19]. In the last few decades, artificial intelligence models have been applied to estimate $PM_{2.5}$ and were found to give a better description of the complex non-linear relationship between $PM_{2.5}$, AOD, and other independent variables than the previously mentioned methods [18] based on the usage of machine learning algorithms [20–22] or deep neural networks [23,24]. These algorithms utilise satellite

observations, various modelled meteorological variables, population, land use, land cover, etc., to estimate $PM_{2.5}$. The importance of the inputs differs from one area to another, but generally, they can enhance $PM_{2.5}$ estimations since counting solely on AOD to estimate near-surface particulate matter values is not sufficient [25]. AOD without other variables was not enough to provide good $PM_{2.5}$ estimations over Europe [26]. In Great Britain, AOD was not among the 15 most important variables when predicting $PM_{2.5}$ levels [20]. Satellite AOD are more correlated with surface PM when the aerosols are well mixed within the planetary boundary layer height (PBLH) [9]. A global study found that 69% of the total AOD are within the PBLH [27], other studies have shown that temperature plays an important role in capturing AOD and understanding its vertical distribution that improves PM analysis [28]. Moreover, a higher humidity atmosphere is likely to have higher AOD without affecting the levels of $PM_{2.5}$ [9]. Other meteorological variables that affect $PM_{2.5}$ are the precipitation that showed a negative correlation in some areas [29] and a positive correlation in other parts of the world [30], and wind speed (WS) that also has different effects from one area to another [30,31].

In this study, we report the modelling of spatiotemporal heterogeneity of $PM_{2.5}$ using machine learning to generate daily estimations of $PM_{2.5}$ over the European Union member states, together with the United Kingdom, Iceland, Liechtenstein, Norway, Switzerland, Albania, Bosnia and Herzegovina, Kosovo, Montenegro, North Macedonia, and Serbia [32].

We will refer to the area of study as "Europe" located inside the coordinates box 26˚ W, 72˚ N, 42˚ E, and 36˚ S. The total study area covers 13,391,504 of 1 km grid cells; 5,450,009 of the total cell number are located over land. The study period covers the years 2018–2020 with full coverage of 1 km spatial resolution using various open data. In the following sections, we will introduce the study area and period and present the preliminary data that were tested while building the predicting model.

2. Primary Data

In this section, we will introduce the primary data we investigated while building the model. Not all these data were utilised while building the model. The chosen data can be found in Section 3.3.

2.1. $PM_{2.5}$ measurements

$PM_{2.5}$ observations were collected from 848 stations across Europe represented in Figure 1. Data was downloaded from OpenAQ which is a non-profit organisation that collects air quality data from different governmental and research institutions and provides it to the users [33].

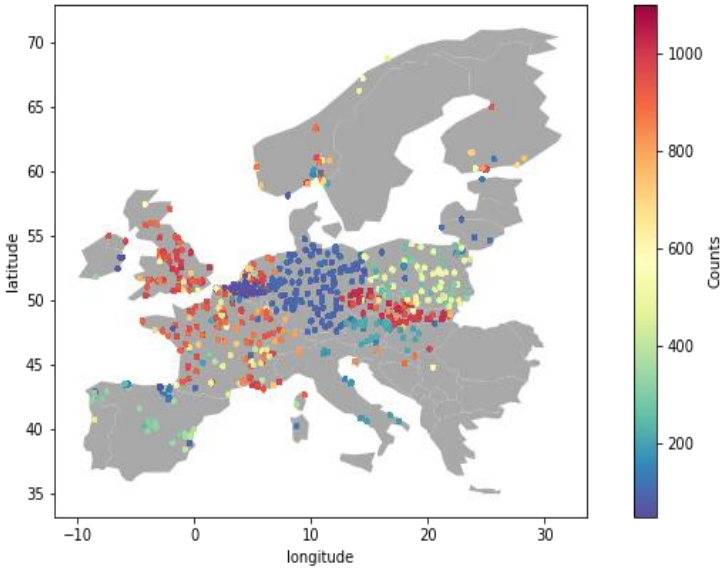


Figure 1. The location of $PM_{2.5}$ ground stations with the number of valid measurements used in this study.

For each station, data between 10 a.m. and 2 p.m. local time were averaged where there are at least 2 available observations to be consistent with MODIS satellites overpassing.

We identified a skewed distribution for $PM_{2.5}$ as shown in Figure 2, we calculated the 25th percentile (Q1), the 75th percentile (Q2) of the dataset, and the inter-quartile range (IQR = Q3 − Q1). All $PM_{2.5}$ values that are higher than 2 × (Q3 + 3 × IQR which is refer as outer fence [34]) were removed, which counted less than 1% of the total data. The number of valid $PM_{2.5}$ observations was 123,248 in 2018, 143,048 in 2019, and 158,964 in 2020 totalling 425,260 observations throughout the study period.

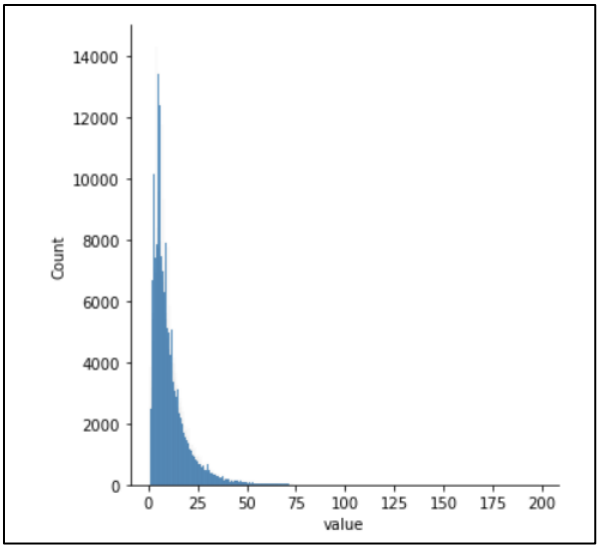


Figure 2. The distribution of the measured $PM_{2.5}$ used in this study

2.2. AOD data

AOD data were downloaded from GHADA, which is a Geo-Harmonized Atmospheric Dataset for Aerosol optical depth at 550 nm [35]. It contains daily estimations of AOD_{550} over Europe with 1 km spatial resolution. GHADA was built based on the MODIS MCD19A2 product [36] and modelled AOD data from Copernicus Atmosphere Monitoring Service (CAMS) [37] that were used to overcome the high percentage of gaps found in the MCD19A2 product. This dataset showed good results when validated with NASA's Aerosols Robotic Network (AERONET).

2.3. Meteorological data

Meteorological data of the following variables wind component u, wind component v, PBLH, total column water vapour, total perception, evaporation, surface pressure, and temperature at 2 m (T2m) were collected from ERA5-Land which is a reanalysis dataset offering a consistent view of the development of land parameters over several decades with a spatial resolution of ~9 km. ERA5-Land was produced by replaying the land component of the European Centre for Medium-Range Weather Forecasts ERA5 climate reanalysis [38]. Relative humidity was collected from ERA5 with 0.25 × 0.25 horizontal resolution.

2.4 Digital elevation model

The Japan Aerospace Exploration Agency (JAXA) provides a worldwide digital surface model with a horizontal resolution of ~30 m by the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM), which was carried on the Advanced Land Observing Satellite "ALOS" [39]. Data were accessed on the 8 March 2021 from https: //www.eorc.jaxa.jp/ALOS/.

2.5. Normalised difference vegetation index

MODIS Terra satellite provides a monthly normalised difference vegetation index (NDVI) product called MOD13A3 [40]. It has 1 km spatial resolution, and it quantifies vegetation presence with values ranging between −1 and 1. NDVI is commonly expressed as shown in Equation (1):

$$NDVI = \frac{NIR-Red}{NIR+Red} \quad (1)$$

Where NIR and Red are spectral reflectance values in the near-infrared and red wavelengths.

2.6. Land cover

Land cover data were extracted from the 2018 CORINE Land Cover (CLC) inventory that was built based on ortho-rectified satellite images having a spatial resolution ranging from 5 m to 60 m and were aggregated into 100 m. We grouped the original 44 CLC classes into seven level 1 classes defined as: agricultural areas, artificial areas, continues urban areas, discontinues urban areas, forests, industrial areas, and water surfaces. Then, we calculated the percentage of each class in every 1 × 1 km2 grid cell.

2.7. Population data

Population data was extracted from the Visible Infrared Imaging Radiometer Suite (VIIRS) night-time lights (NTL) data by averaging the monthly data of the year 2019.

3. Methodology

3.1 *Data pre-processing*

All data were reprojected to the European Terrestrial Reference System 1989 (EPSG:3035) that uses metres as measuring units. This system is used for statistical mapping and other purposes which requires a true area representation, using a 1 km grid cell with bilinear interpolation method for ECMWF data and the cubic convolution for the ALOS elevation model. In addition, we calculated WS based on the two wind *u* and *v* components.

A spatio-temporal dataset was created by extracting the information from all input data at the locations of $PM_{2.5}$ stations. The Julian day, month, and year were added as the temporal information; longitude and latitude were added as the spatial information. The generated dataset was used to train and test the model.

3.2. *Model Development*

We first analysed the linear relationship between the primary independent variables and $PM_{2.5}$ values. PBLH was negatively correlated to $PM_{2.5}$ with Pearson correlation of r = −0.24. Most of the

meteorological variables were also negatively related to $PM_{2.5}$ with r = −0.2 for WS, r = −0.15 for T2m, r = −0.13 for RH, and r = −0.1 for TP. AOD and evaporation had the highest positive correlation with $PM_{2.5}$ r = 0.14. Based on this initial data exploratory analysis, we excluded some primary inputs that had high correlation with other inputs such as skin temperature, which was correlated to T2m with r = 0.93. We tested linear models to estimate $PM_{2.5}$. These models suffered from underfitting issues and failed to describe the relationship between the independent variables and $PM_{2.5}$. Therefore, we used a more complex algorithm called Extremely Randomised Trees (ET).

ET is a very similar decision tree-based ensemble method to the widely used Random Forest (RF). Both algorithms are composed of large number of trees, where the final decision is obtained from the prediction of every tree by majority vote in classification problems and arithmetic average in regression problems. Both algorithms have the same growing tree procedure and selecting the partition of each node. Additionally, both algorithms randomly choose a subset of input features.

ET, on the other hand, strongly randomises the selection of both attribute and cut point while splitting a tree node using the whole learning sample to grow the trees which adds randomisation, making it a more robust algorithm against overfitting. From computational point of view, the complexity of the tree growing procedure is on the order of N log N with respect to learning sample size [41]. The main parameters in the ET splitting process are the number of attributes that are randomly selected at each node and the minimum sample size for splitting a node. For further information on how the ET algorithm operates refer to Table 1 in [41]. In addition to accuracy, the ET algorithm has higher computational efficiency than the RF algorithm since it chooses the splits randomly and does not look for the optimum split as the latter one [41]. The number of estimators (number of trees in the forest), the maximum depth of the trees, the number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node were the main parameters while tuning our model.

Table 1. The chosen independent variables used to build the ET model.

Name of the variable	Unit	Minimum	Maximum	Mean	STD
$PM_{2.5}$	$\mu g/m^3$	2	80	11.81	9.26
Aerosol optical depth	-	0.01	3.12	0.13	0.08
PBLH	m	73.90	3420.17	933.39	463.59
WS	m/sec	0.23	18.12	3.88	2.13
T2m	K	249.86	314.15	287.03	8.17
Relative Humidity	%	0.04	110.82	68.53	22.93

Total precipitation	mm	0	8	0.1	0.3
Total Column Water Vapour	Kg/m^2	0.95	50.61	16.76	7.88
NDVI	-	-0.3	0.73	0.25	0.12
Evaporation	mm	-0.744	0.065	-0.164	0.109
Elevation	m	-3.88	914.26	151.66	156.01

To reduce model complexity due to the large number of independent variables we excluded the input variables based on the feature importance in the ET algorithm. Besides the spatio-temporal information, we used $PM_{2.5}$ with the independent variables that are shown in Table 1 to develop our model.

3.3. Model Validation

3.3.1. Sample based cross validation

Cross validation (CV) is a common method to analyse the model performance and detect potential overfitting problems where the model achieves high accuracy on the training set and performs badly on new data or the test set. We applied a 10-fold CV where all samples in the training dataset were randomly divided into 10 equal subsets. Then, in each round, 9 subsets were used to fit the model, and the remaining subset was used for testing the model performance [42]. This approach is used widely in PM studies [20,21,43–45].

3.3.2. Spatial and temporal 10-fold cross validation

In this validation, we divided the samples based on two factors. For the spatial 10-fold cross validation we splatted the data based on the location of the stations, the stations were divided randomly into 10 folds. In each fold, the model was trained on the samples from 90% of the stations and the samples from the remaining 10% for testing. For the temporal 10-fold cross validation, we divided the samples into 10 folds based on the Julian day and applied the cross validation in a similar way to the previously mentioned one.

4. Results

The results of sample-based, spatial, and temporal 10-fold cross validation are shown in Table 2. The density scatter plot for the sample-based cross validation is shown in Figure 3.

Table 2. R^2, RMSE and MAE of the sample-based 10-CV, spatial 10-CV, and the temporal 10-CV.

10-CV	**R^2**	**RMSE**	**MAE**
Sampl-based	0.69	5.0	3.3
Spatial	0.69	4.9	3.2
Temporal	0.53	6.1	4.1

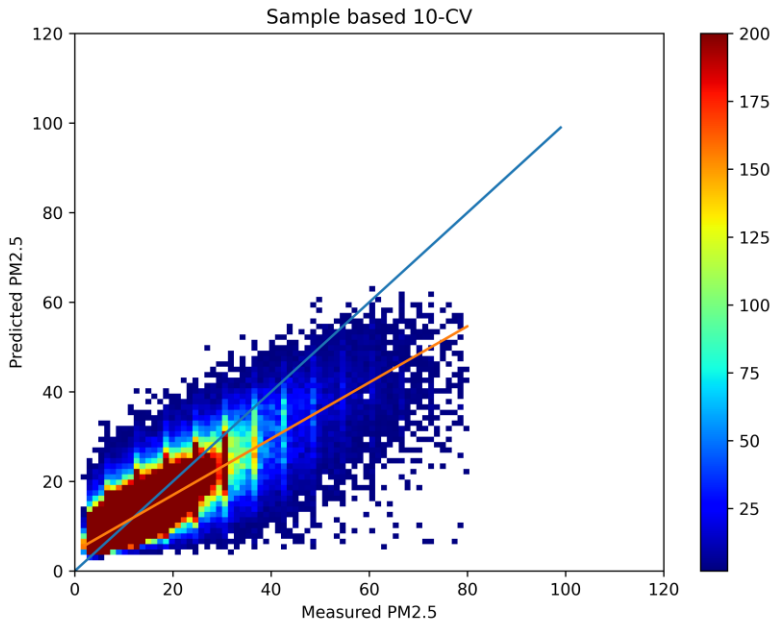


Figure 3. Density scatter plot of the sample based 10-CV results of the model.

It must be noted that $PM_{2.5}$ levels in general are low in Europe when compared to more polluted areas and this is reflected by the low RMSE we obtained in our study when compared to some studies outside Europe with higher R^2 values [44,45]. Our model proved its efficiency in predicting $PM_{2.5}$ when our results (out-of-sample R^2 = 0.69, RMSE = 5 μg/m3) were comparable with results obtained from a recent study over a smaller geographic area in Europe (Great Britain; out-of-sample R^2 = 0.77, RMSE = 4 μg/m3) [20]. It is also noted that the model underestimates high $PM_{2.5}$ values (>40 μg/m3) since such values are not abundant over our study area.

To justify the difference in the model performance spatially and temporally, we applied site-based cross validation where we used samples from one station as the test set, and the samples from all remaining stations were used to train the model. We applied this method to analyse the model performance spatially, since the standard 10-CV may not be able to detect potential spatial overfitting [18].

The results are shown in Figure 4. The model performs well in most of the locations in Central Europe with an average R^2~0.7. A total of 63% of all stations in Europe have $R^2 > 0.6$. The accuracy of the model is lower in the northern and southern parts of Europe. However, the RMSE and MAE are relatively small even in the northern and southern parts.

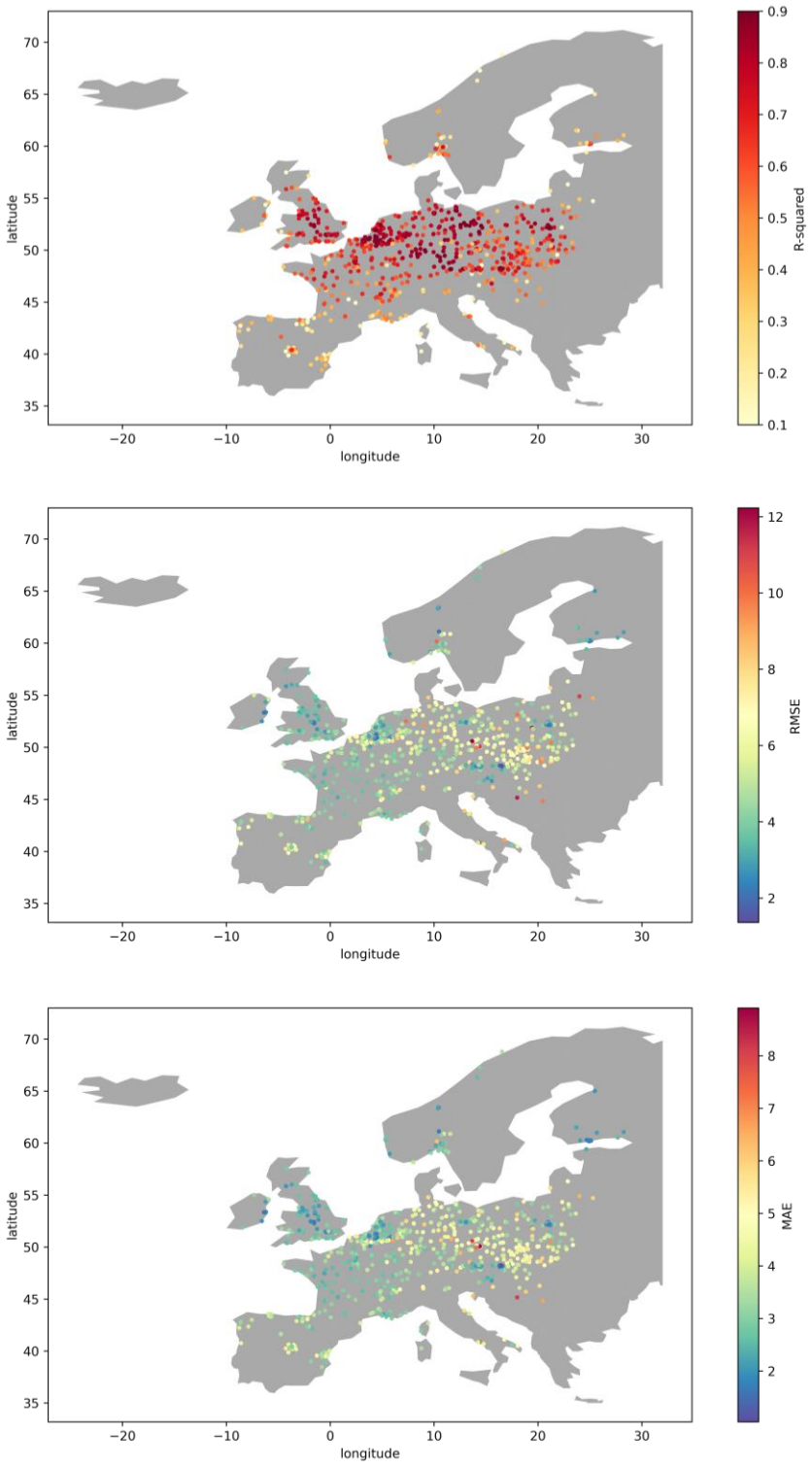


Figure 4. Spatial distribution of the site-based cross validation of coefficient of determination, the root mean square error, and the mean absolute error.

5. Creating $PM_{2.5}$ maps

Daily $PM_{2.5}$ maps during MODIS satellite overpassing were created for the period 2018–2020 over Europe. Figure 5 shows the average $PM_{2.5}$ for the year 2018, 2019, and 2020.

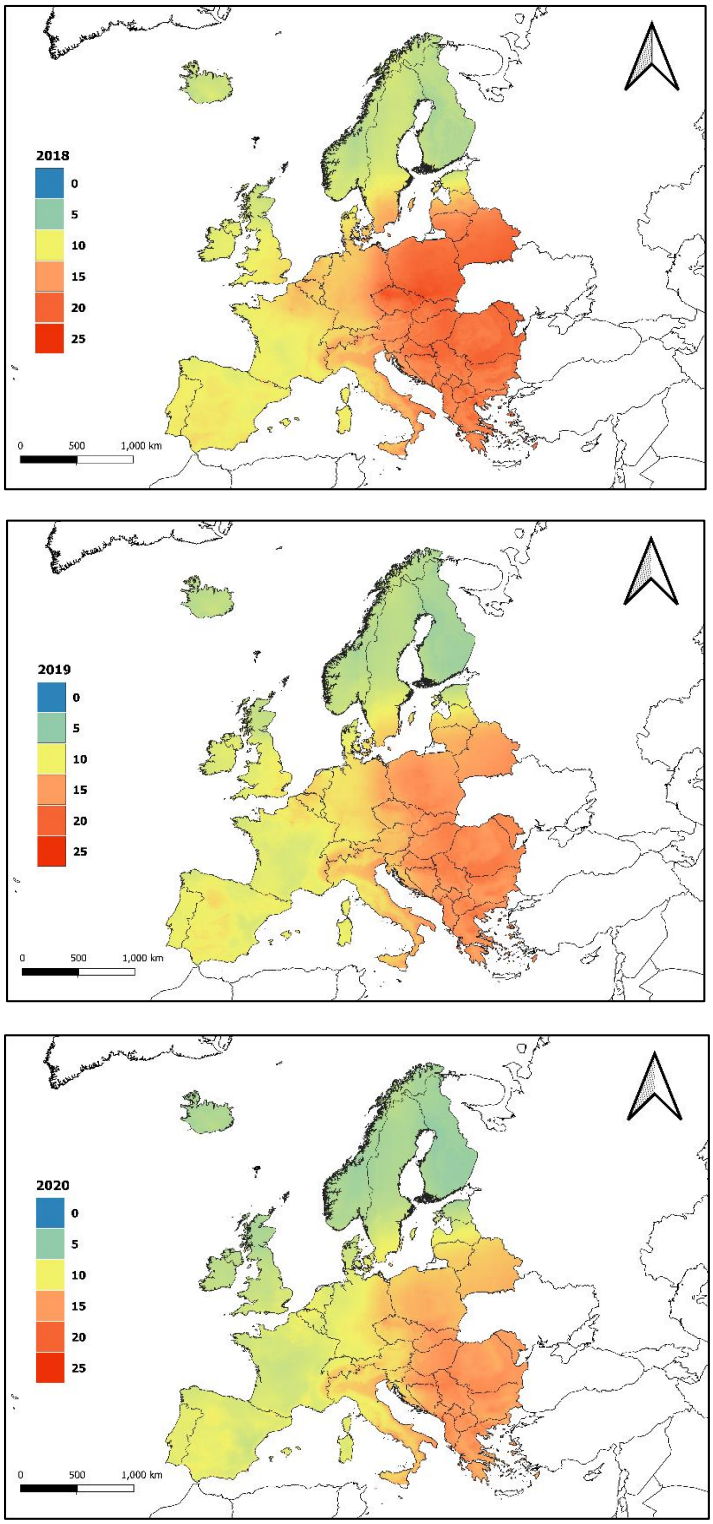


Figure 5. The average $PM_{2.5}$ for the years 2018, 2019, and 2020 over Europe.

A significant decline in $PM_{2.5}$ levels has occurred over Europe throughout the study period. Poland had the highest $PM_{2.5}$ average level in the year 2018 with an average level ~ 19.5 $\mu g/m^3$, in 2019 Romania had the highest average ~ 16.5 $\mu g/m^3$ whereas Serbia had the highest average in 2020 with an average ~ 15.8 $\mu g/m^3$. Finland had the lowest $PM_{2.5}$ average level in all three years with 7.1 in 2018, 6.3 in 2019, and 5.8 in 2020. Comparing the results of the average $PM_{2.5}$ levels for the years 2018, 2019, and 2020 were highly compatible with the reports of the European Environment Agency (EEA). According to EEA the highest $PM_{2.5}$ concentrations were found in central and eastern Europe and northern Italy. For the central and western parts, the main reason of high $PM_{2.5}$ is the usage of solid fuels with older vehicle compared to other parts of Europe [46], besides using the solid fuels for heating as was found in Poland [47]. For the northern part of Italy, the high levels of $PM_{2.5}$ are due to the combination of a high density of anthropogenic emissions and meteorological conditions [46, 48]. Furthermore, Milan, the largest city in the north of Italy previously reported levels of $PM_{2.5}$ exceeding the safety limit set by the EU [49].

As an application, we used the proposed machine learning based prediction approach in $PM_{2.5}$ levels analysis to study the effect of the COVID-19 lockdown (March to June of the year 2020) on air quality over Europe. As an attempt to verify the influence of the lockdown on air quality, we compared the average $PM_{2.5}$ of the previous 4 months (November to December in 2019 and January to February 2020) and the following 4 months (July to October 2020) to the 4 months of the lockdown by calculating the relative percentage difference (RPD). By doing so we masked the general improvement trend in air quality over Europe. RPD calculated using Equation 2.

$$\text{RPD} = \frac{\text{PM2.5 (lockdown)} - \text{PM2.5 avg(before lockdown,after lockdown)}}{\text{PM2.5 avg(before lockdown,after lockdown)}} * 100 \quad (2)$$

We found a significant improvement in air quality over Europe except for UK, Ireland, north of France, and south of Italy as shown in Figure 6. Our results are in agreement with another study over Poland (Eastern Europe), where the air quality represented by $PM_{2.5}$ has significantly improved in the months of March to April in 2020 when the authors compared to the same months from the previous two years [50]. Interestingly, the unusual increase in $PM_{2.5}$ levels in the UK was consistent with what was reported in [51] as the authors justified such increase by unusual meteorological conditions. The latter conditions may also justify the increase in $PM_{2.5}$ over northern France. In Italy, where people were spending most of their time at home, the increased house heating during the lockdown period limited the decrease in $PM_{2.5}$ levels besides the effects of the agriculture sector that kept performing during the lockdown [52].

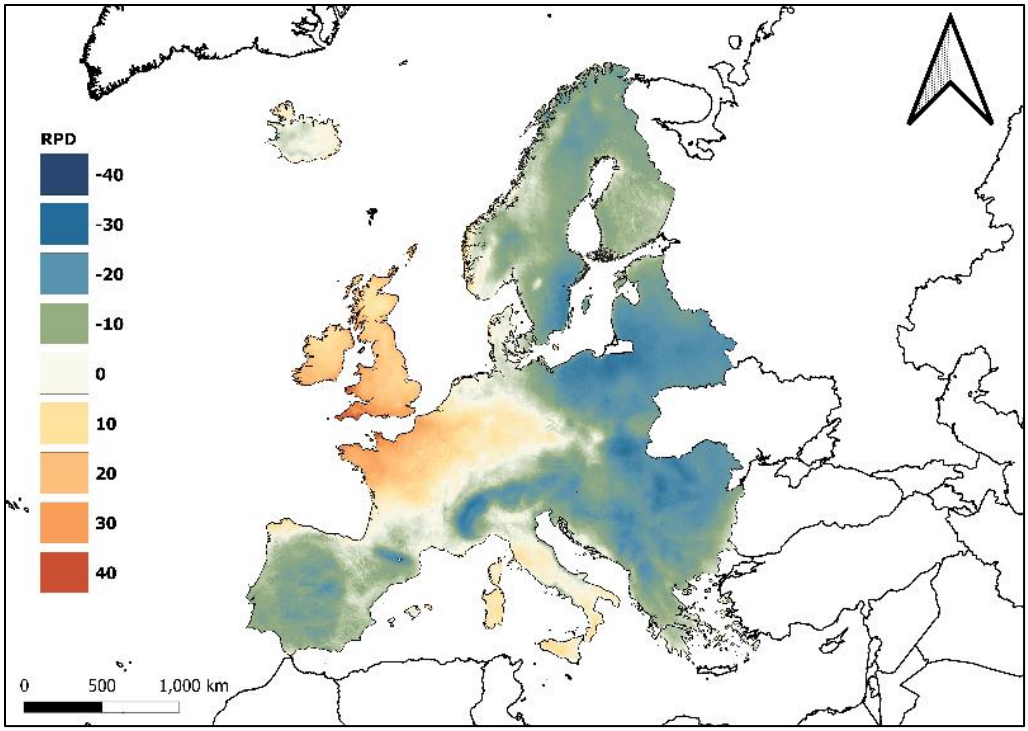


Figure 6. Relative percentage difference of $PM_{2.5}$ for the lock down period of the year 2020 with the average of the previous 4 months and the following 4 months.

In this study, we proposed the first machine learning-based scheme to estimate $PM_{2.5}$ levels over Europe with high spatial resolution of 1 km. We trained an extra trees model using observed $PM_{2.5}$ from 848 stations as the target variable. AOD, different meteorological variables, land variables and NDVI as the independent variables.

The sample based 10-fold CV showed that our model underestimates high $PM_{2.5}$ values (> 40 $\mu g/m^3$) which may limit the model ability to detect hazard situations. This underestimation occurred since high $PM_{2.5}$ values were not common over our study area as shown in Figure 2. The spatial cross validation showed that the model estimates $PM_{2.5}$ with a higher R^2 in the areas with high ground stations density the compared to the areas with a lower density. The occurred spatial overfitting is expected to happen due to spatially unbalanced data.

In the Central Europe (Czech Republic, Poland, Slovakia, and surrounding areas) the model performed with a higher R^2 compared to the Northern and Southern parts of Europe. However, the RMSE in the Central Europe was comparably higher than the ones in the prementioned parts. This is due to the fact that the average $PM_{2.5}$ values in the Central Europe is higher and have more variations than the Northern and Southern parts. The highest RMSE in the Central Europe can be found in three stations in the Czech Republic. These stations are located near mining areas were higher $PM_{2.5}$ values compared to other stations that are mostly located in

urban areas. This issue can be potentially solved by including a detailed land cover data with an appropriate classification for each station which is usually difficult to achieve on a large scale such as in our study.

Having unbalanced spatial-temporal data made the modelling more complex than other studies which focused on smaller areas with well-balanced data and with similar instruments in measuring $PM_{2.5}$ values. However, by tuning the parameters in the model we were able to achieve acceptable results for most parts of our study area. The effect of the chosen independent variables in estimating $PM_{2.5}$ differs across the study area. We analysed the spatial potential relationships of the independent variables in estimating $PM_{2.5}$ by calculating features importance in four parts of Europe. North-West (latitude > 50 and longitude < 10), North-East (latitude > 50 and longitude > 10), South-West (latitude < 50 and longitude < 10) and South-East (latitude < 50 and longitude > 10). AOD and PBLH had the most feature importance in all parts of Europe with an average of 10.4% and 14.1% respectively. WS and temperature had more effect in estimating $PM_{2.5}$ in the south of Europe compared to the north. Rh had more importance in estimating $PM_{2.5}$ in the western part of Europe compared to the Eastern part.

Table 3. Shows the effects of AOD and the most important meteorological variables on $PM_{2.5}$ estimates. We tried to train multiple models based on the area. However, this approach did not improve the overall performance over the whole study area.

Table 3. The effects (%) of AOD and the most important meteorological variables on $PM_{2.5}$ estimations in the four chosen parts of our study area.

Independent variable	North-West	North-East	South-West	South-East
AOD	13.25	8.81	10.43	9.11
BLH	15.89	15.22	14.98	10.41
T2m	8.62	6.25	10.13	10.71
Rh	6.41	3.99	5.82	4.71
E	3.58	5.99	3.44	7.96
WS	5.18	4.25	7.32	5.82
TCWV	4.469	3.63	4.55	4.07

6. Conclusion

In this study we developed a spatio-temporal machine learning model to estimate daily $PM_{2.5}$ levels for the years 2018–2020 with 1 km spatial resolution over Europe using open data from multiple sources like remote sensing satellite-based products, meteorological reanalysis datasets, and other land variables.

The developed model was used to estimate $PM_{2.5}$ values over 5,450,009 land cells (1 km^2) for a 3-year period (1096 days) totalling more than 5.973 billion estimations with a good sample-based CV coefficient of 0.69, RMSE of 5 $\mu g/m^3$, and MAE of 3.3 $\mu g/m^3$.

We calculated the yearly average of $PM_{2.5}$ levels and we found that $PM_{2.5}$ values have dropped in almost all parts of Europe during the study period.

The full coverage dataset of $PM_{2.5}$ that we produced can be used to investigate air quality over Europe with higher spatial resolution compared to the available products which may provide better understanding in time series analysis in this field.

References

1. Li, L.; Losser, T.; Yorke, C.; Piltner, R. Fast Inverse Distance Weighting-Based Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate Matter PM2.5 in the Contiguous U.S. Using Parallel Programming and k-d Tree. *Int. J. Environ. Res. Public Health* **2014**, 11, 9101–9141.
2. Crippa, M.; Janssens-Maenhout, G.; Guizzardi, D.; Van Dingenen, R.; Dentener, F. Contribution and uncertainty of sectorial and regional emissions to regional and global $PM_{2.5}$ health impacts. *Atmos. Chem. Phys*. **2019**, 19, 5165–5186.
3. Pascal, M.; Falq, G.; Wagner, V.; Chatignoux, E.; Corso, M.; Blanchard, M.; Host, S.; Pascal, L.; Larrieu, S. Short-term impacts of particulate matter (PM10, PM10–2.5, PM2.5) on mortality in nine French cities. *Atmos. Environ*. **2014**, 95, 175–184.
4. Liu, C.; Chen, R.; Sera, F.; Vicedo-Cabrera, A.M.; Guo, Y.; Tong, S.; Coelho, M.S.Z.S.; Saldiva, P.H.N.; Lavigne, E.; Matus, P.; et al. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *N. Engl. J. Med*. **2019**, 381, 705–715.
5. Martins, N.R.; da Graça, G.C. Impact of PM2.5 in indoor urban environments: A review. *Sustain. Cities Soc*. **2018**, 42, 259–275.
6. Baklanov, A.; Molina, L.T.; Gauss, M. Megacities, air quality and climate. *Atmos. Environ*. **2016**, 126, 235–249.
7. Mao, X.; Wang, L.; Pan, X.; Zhang, M.; Wu, X.; Zhang, W. A study on the dynamic spatial spillover effect of urban form on PM2.5 concentration at county scale in China. *Atmos. Res*. **2022**, 269, 106046.
8. Environmental Protection Agency 40 CFR Part 50 Review of the National Ambient Air Quality Standards for Particulate Matter. Available online: https://cfpub.epa.gov/ncea/ (accessed on 19 December 2021).

9. Lee, H.J. Advancing Exposure Assessment of PM2.5 Using Satellite Remote Sensing: A Review. *Asian J. Atmos. Environ*. **2020**, 14, 319–334.
10. Deng, L. Estimation of PM2.5 spatial distribution based on kriging interpolation. In Proceedings of the First International Conference on Information Sciences, Machinery, Materials and Energy, Chongqing, China, 11–13 April **2015**; Volume 126.
11. Vienneau, D.; De Hoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ*. **2010**, 44, 688–696.
12. Briggs, D.J. The use of GIS to evaluate traffic-related pollution. *Occup. Environ. Med*. **2006**, 64, 1–2.
13. You, W.; Zang, Z.; Pan, X.; Zhang, L.; Chen, D. Estimating PM2.5 in Xi'an, China using aerosol optical depth: A comparison between the MODIS and MISR retrieval models. *Sci. Total Environ*. **2015**, 505, 1156–1165.
14. Yao, F.; Si, M.; Li, W.; Wu, J. A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. *Sci. Total Environ*. **2017**, 618, 819–828.
15. Zhang, Y.; Li, Z. Remote sensing of atmospheric fine particulate matter (PM2.5) mass concentration near the ground from satellite observation. *Remote Sens. Environ*. **2015**, 160, 252–262.
16. Kanabkaew, T. Prediction of Hourly Particulate Matter Concentrations in Chiangmai, Thailand Using MODIS Aerosol Optical Depth and Ground-Based Meteorological Data. *EnvironmentAsia*. **2013**, 6, 65–70.
17. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res. Earth Surf*. **2009**, 114.
18. Ma, Z.; Dey, S.; Christopher, S.; Liu, R.; Bi, J.; Balyan, P.; Liu, Y. A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data. *Remote Sens. Environ*. **2021**, 269, 112827.
19. Liu, Y.; Paciorek, C.J.; Koutrakis, P. Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environ. Health Perspect.* **2009**, 117, 886–892.
20. 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 PM2.5 Concentrations across Great Britain. *Remote Sens*. **2020**, 12, 3803.

21. 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 PM2.5 estimates across China using enhanced space–time extremely randomized trees. *Atmos. Chem. Phys*. **2020**, 20, 3273–3289.
22. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.S.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ*. **2018**, 636, 52–60.
23. Xiao, F.; Yang, M.; Fan, H.; Fan, G.; Al-Qaness, M.A.A. An improved deep learning model for predicting daily $PM_{2.5}$ concentration. *Sci. Rep*. **2020**, 10, 20988.
24. Li, L.; Girguis, M.; Lurmann, F.; Pavlovic, N.; McClure, C.; Franklin, M.; Wu, J.; Oman, L.D.; Breton, C.; Gilliland, F.; et al. Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. *Environ. Int*. **2020**, 145, 106143.
25. Van Donkelaar, A.; Martin, R.V.; Brauer, M.; Kahn, R.; Levy, R.; Verduzco, C.; Villeneuve, P.J. Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application. *Environ. Health Perspect*. **2010**, 118, 847–855.
26. Koelemeijer, R.; Homan, C.; Matthijsen, J. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmos. Environ*. **2006**, 40, 5304–5315.
27. Bourgeois, Q.; Ekman, A.M.L.; Renard, J.-B.; Krejci, R.; Devasthale, A.; Bender, F.A.-M.; Riipinen, I.; Berthet, G.; Tackett, J.L. How much of the global aerosol optical depth is found in the boundary layer and free troposphere? Atmos. *Chem. Phys*. **2018**, 18, 7709–7720.
28. Liu, B.; Ma, X.; Ma, Y.; Li, H.; Jin, S.; Fan, R.; Gong, W. The relationship between atmospheric boundary layer and temperature inversion layer and their aerosol capture capabilities. *Atmos. Res*. **2022**, 271, 106121.
29. Li, X.; Feng, Y.J.; Liang, H.Y. The Impact of Meteorological Factors on PM2.5 Variations in Hong Kong. IOP Conf. *Series Earth Environ. Sci*. **2017**, 78, 012003.
30. Wang, J.; Ogawa, S. Effects of Meteorological Conditions on PM2.5 Concentrations in Nagasaki, Japan. Int. J. *Environ. Res. Public Health* **2015**, 12, 9089–9101.
31. Wang, S.; Gao, J.; Guo, L.; Nie, X.; Xiao, X. Meteorological Influences on Spatiotemporal Variation of PM2.5 Concentrations in Atmospheric Pollution Transmission Channel Cities of the Beijing–Tianjin–Hebei Region, China. *Int. J. Environ. Res. Public Health* **2022**, 19, 1607.
32. Open Data Science Europe. Geo-Harmonizer Project Implementation Plan 2020–2022; Open Data Science Europe: Wageningen, The Netherlands, **2020**.
33. OpenAQ. Available online: https://openaq.org/ (accessed on 8 May 2022).

34. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Publishing Company: Boston, MA, USA, **1977**.
35. Ibrahim, S.; Landa, M.; Pešek, O.; Pavelka, K.; Halounova, L. Space-Time Machine Learning Models to Analyze COVID-19 Pandemic Lockdown Effects on Aerosol Optical Depth over Europe. *Remote Sens*. **2021**, 13, 3027.
36. Lyapustin, A.; Wang, Y.; Laszlo, I.; Kahn, R.; Korkin, S.; Remer, L.; Levy, R.; Reid, J.S. Multiangle implementation of atmospheric correction (MAIAC): Part 2. Aerosol algorithm. *J. Geophys. Res*. **2011**, 116.
37. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.-M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys*. **2019**, 19, 3515–3556.
38. Muñoz-Sabater, J.; Dutra, E.; Agustí-Panareda, A.; Albergel, C.; Arduini, G.; Balsamo, G.; Boussetta, S.; Choulga, M.; Harrigan, S.; Hersbach, H.; et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **2021**, 13, 4349–4383.
39. 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, 71–76.
40. Didan, K. MOD13A3 MODIS/Terra Vegetation Indices Monthly L3 Global 1 km SIN Grid V006 [Dataset]. NASA EOSDIS Land Processes DAAC. 2015. Available online: https://doi.org/10.5067/modis/mod13a3.006 (accessed on 14 March 2021).
41. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn*. **2006**, 63, 3–42.
42. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Trans. Pattern Anal. Mach. Intell*. **2009**, 32, 569–575.
43. 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, 477–489.
44. He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground PM2.5 in China via space-time regression modeling. *Remote Sens. Environ*. **2018**, 206, 72–83.
45. Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; Cribb, M. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sens. Environ*. **2019**, 231, 111221.
46. European Environment Agency. Available online: https://www.eea.europa.eu (accessed on 19 December 2021).

47. Baborska-Narozny, M.; Szulgowska-Zgrzywa, M.; Mokrzecka, M.; Chmielewska, A.; Fidorow-Kaprawy, N.; Stefanowicz, E.; Piechurski, K.; Laska, M. Climate justice: Air quality and transitions from solid fuel heating. *Build. Cities* **2020**, 1, 120–140.
48. Perrone, M.G.; Zhou, J.; Malandrino, M.; Sangiorgi, G.; Rizzi, C.; Ferrero, L.; Dommen, J.; Bolzacchini, E. PM chemical composition and oxidative potential of the soluble fraction of particles at two sites in the urban area of Milan, Northern Italy. *Atmos. Environ*. **2016**, 128, 104–113.
49. Perrone, M.; Larsen, B.; Ferrero, L.; Sangiorgi, G.; De Gennaro, G.; Udisti, R.; Zangrando, R.; Gambaro, A.; Bolzacchini, E. Sources of high PM2.5 concentrations in Milan, Northern Italy: Molecular marker data and CMB modelling. *Sci. Total Environ*. **2012**, 414, 343–355.
50. Filonchyk, M.; Hurynovich, V.; Yan, H. Impact of Covid-19 lockdown on air quality in the Poland, Eastern Europe. *Environ. Res*. **2020**, 198, 110454.
51. Jenkins, N.; Parfitt, H.; Nicholls, M.; Beckett, P.; Wyche, K.; Smallbone, K.; Gregg, D.; Smith, M. *Estimation of Changes in Air Pollution Emissions, Concentrations and Exposure during the COVID-19 Outbreak in the UK; Report for The Air Quality Expert Group*, on Behalf of Defra: Analysis of Air Quality Changes Experienced in Sussex and Surrey since the COVID-19 Outbreak; UK Air, Department for Food and Rural Affairs: London, UK, **2020**; 57p.
52. Pala, D.; Casella, V.; Larizza, C.; Malovini, A.; Bellazzi, R. Impact of COVID-19 lockdown on PM concentrations in an Italian Northern City: A year-by-year assessment. *PLoS ONE* **2022**, 17, e0263265.

5.1.4. $PM_{2.5}$ Estimation in the Czech Republic using Extremely Randomized Trees: A Comprehensive Data Analysis

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 $PM_{2.5}$ over whole Europe using unbalanced spatial-temporal data, the focus of this study was on estimating $PM_{2.5}$ specifically over the Czech Republic using more balanced dataset to train and evaluate the model. Moreover, the spatial autocorrelation between $PM_{2.5}$ 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 R^2 of the 10-CV for the new model was 16% higher than the previous one. Where R^2 reached 0.85 with RMSE=5.42 μg/m^3, MAE=3.41 μg/m^3, and bias=-0.03 μg/m^3. The developed spatiotemporal model was employed to generate comprehensive daily maps covering the entire study area throughout the 2018–2020 years. The temporal analysis showed that the levels of $PM_{2.5}$ 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.

Keywords: Air quality; $PM_{2.5}$; Artificial intelligence; Spatial autocorrelation; Czech Republic

---

1. Introduction

Atmospheric Particulate Matter (PM) with a diameter smaller than or equal to 2.5 microns ($PM_{2.5}$) 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 $PM_{2.5}$ 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 $PM_{2.5}$ to 20 μg/m^3 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 $PM_{2.5}$ 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 $PM_{2.5}$ fine particles came from local heating sources that used brown coal with a high SO_2 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 northeastern 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 $\mu g/m^3$ [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 PM_{10} 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 $PM_{2.5}$ concentrations. These studies have focused on developing various types of models to increase the limited spatial coverage that is provided by $PM_{2.5}$ 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 $PM_{2.5}$ is a type of aerosol, there is generally a positive correlation that made AOD a crucial factor in predicting $PM_{2.5}$ 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 $PM_{2.5}$ is estimated [14,17,18]. Moreover, a few studies considered the Spatial Autocorrelation (SA) of $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ concentrations increased the coefficient of determination (R^2) by ~9% [22].

In this study, we aimed to estimate the concentrations of $PM_{2.5}$ over the CZ during the years 2018, 2019, and 2020. CZ is a landlocked country that covers an area of 78870 square kilometers located in central Europe bordering Germany, Poland, Slovakia, and Austria.

2. Materials and methods

2.1. Dependent variable and primary independent variables

Daily $PM_{2.5}$ 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.

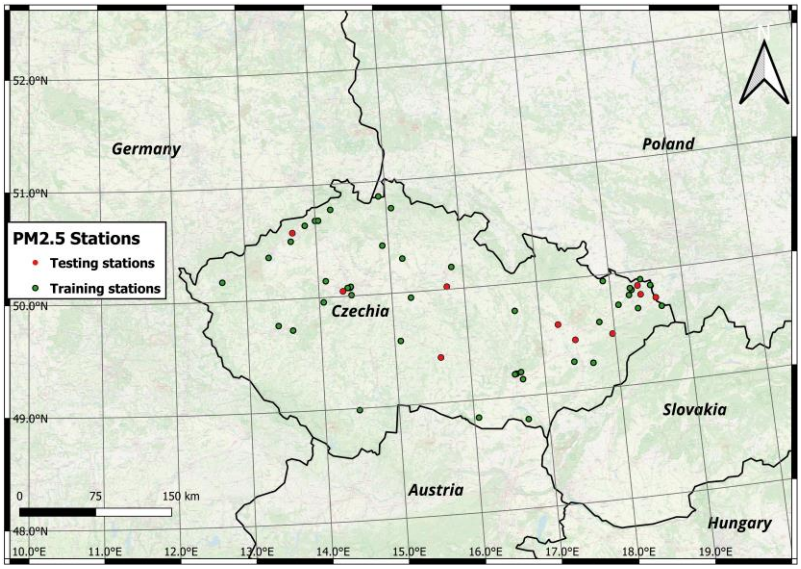


Fig 1. Study area with ground stations. The red dots represent the stations that were used to test the model and the green dots represent the stations that were used to train the model.

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 1km^2 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 $PM_{2.5}$ showed that PBLH and T2m were the most negatively correlated variables to $PM_{2.5}$ with Pearson correlation of -0.25 and -0.22 respectively. NDVI, TCWV, WS, RH, H, and TP also had negative correlations with $PM_{2.5}$. Whereas, E, AOD, NL, and RL had positive correlations with $PM_{2.5}$. The following table shows the primary data that were used in our study. All primary data was reprojected to the European Terrestrial Reference System 1989 (EPSG:3035) with a grid of 1 km^2 that covers the study area using bilinear interpolation for meteorological data and the cubic convolution for the elevation model.

Table 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

2.2. Model development

A machine learning algorithm was used with feature engineering techniques that were applied to train the $PM_{2.5}$ 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].

2.3. Feature engineering and model training

The temporal inputs were represented by the radian day and the year. The radian day will help the model to 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.

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

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; z_j is the value of the observation at all other stations where $i \neq j$; σ^2 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), Fig. 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. Fig. 2 shows feature importance of the primary inputs in the training set.

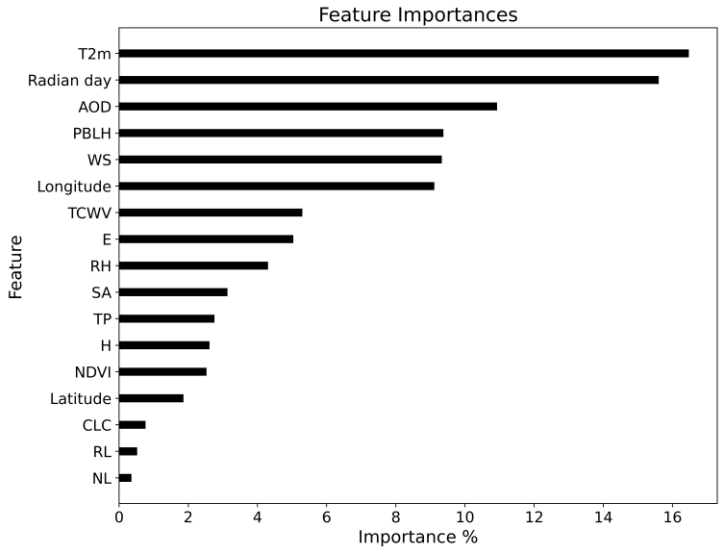


Fig 2. Feature importance calculated based on the training data.

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.

2.4. Model validation

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

2.4.1. 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/m³, and MAE=3.37 μg/m³.

2.4.2. 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/m³, MAE=3.41 μg/m³ and, bias=-0.03 μg/m³. Fig. 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/m³.

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.

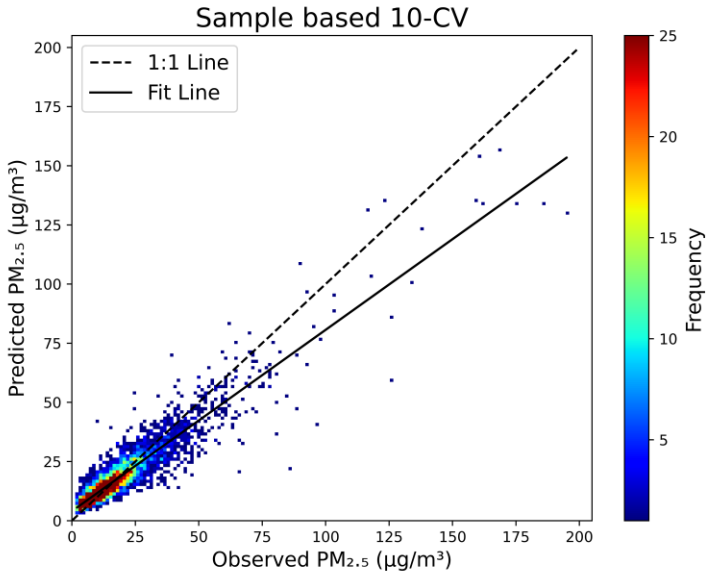


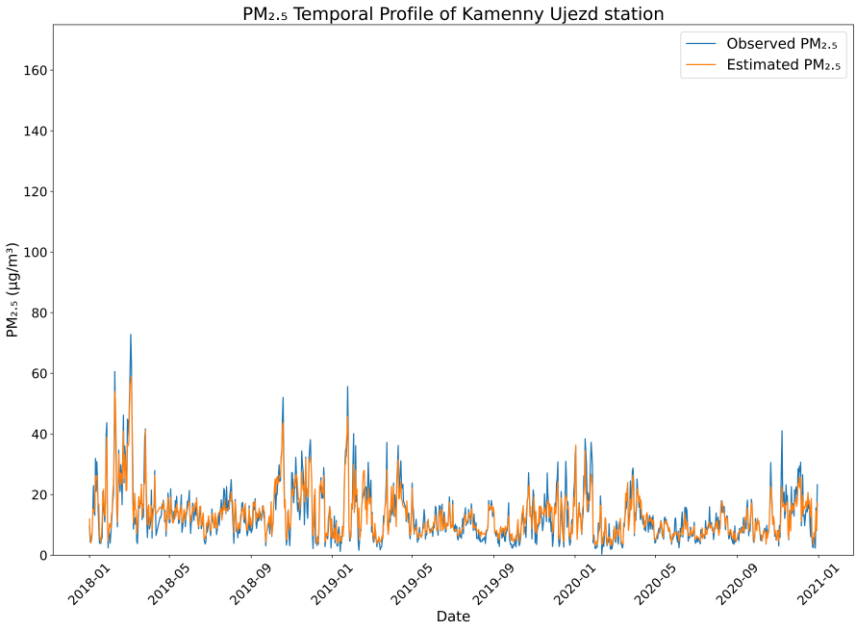
Fig 3. Density scatter plot for the 10-CV applied on all data.

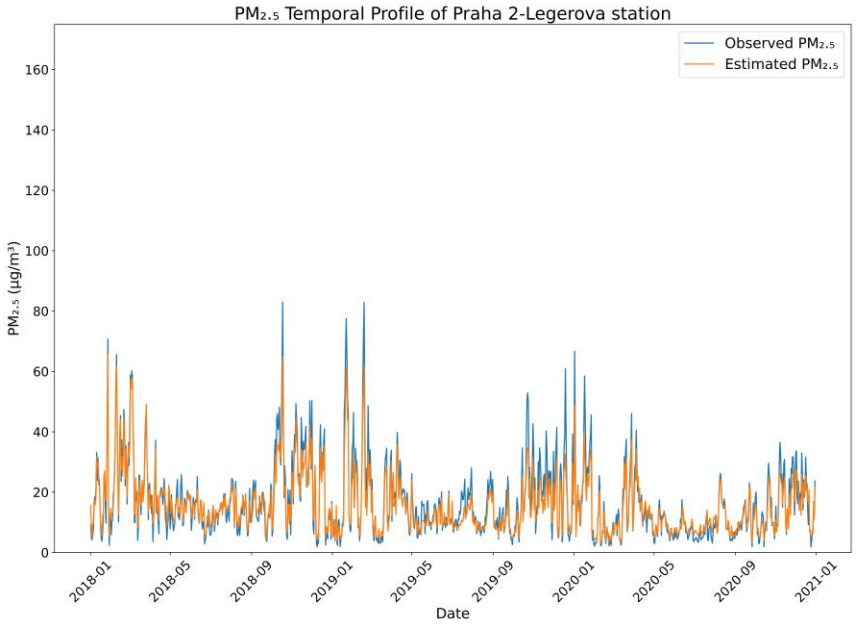
3. Results

3.1. 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. Fig. 4 represents the temporal profile for three stations with high, normal, and low $PM_{2.5}$ levels.





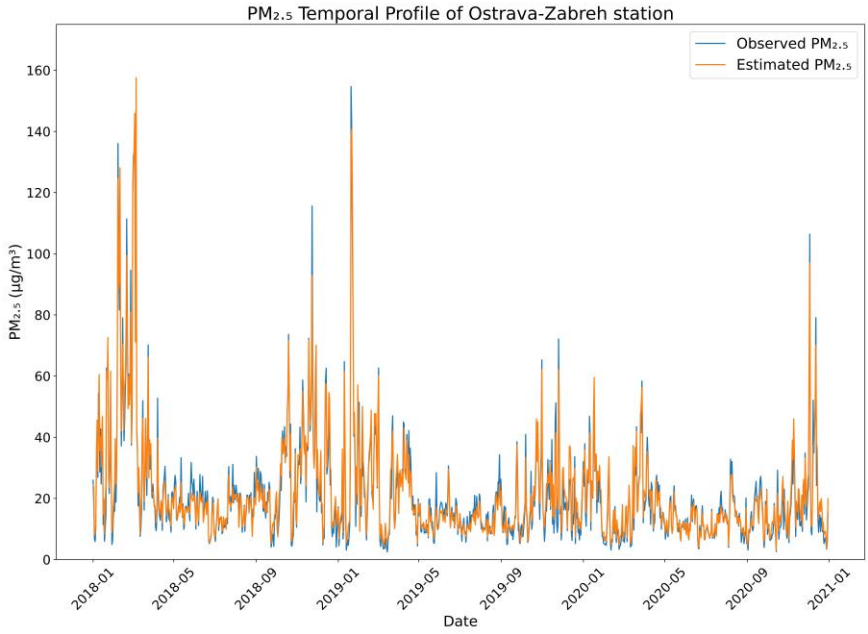


Fig 4. $PM_{2.5}$ temporal profile over three stations: Kamenny Ujezd station, Praha 2-Legerova station, and Ostrava-Zabreh station.

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.

3.2. Temporal and regional analysis

We calculated the average $PM_{2.5}$ levels for each year during the study period. In Fig. 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 $\mu g/m^3$ in 2018, 18 $\mu g/m^3$ in 2019 and 15.8 $\mu g/m^3$ in 2020. Karlovy Vary Region had the lowest $PM_{2.5}$ values with 16.4 $\mu g/m^3$ in 2018, 11.1 $\mu g/m^3$ in 2019, and 10.2 $\mu g/m^3$ in 2020. Besides, the Moravia-Silesian Region, $PM_{2.5}$ values exceeded 20 $\mu g/m^3$ in Zlín and Olomouc Regions with average values of 22.7 $\mu g/m^3$ and 22.2 $\mu g/m^3$ respectively during 2018. Good $PM_{2.5}$ levels <= 12 $\mu g/m^3$ were found in six regions in 2020, these regions are Plzeň, Karlovy Vary, Southern Bohemia, Vysočina, Central Bohemia, and Liberec.

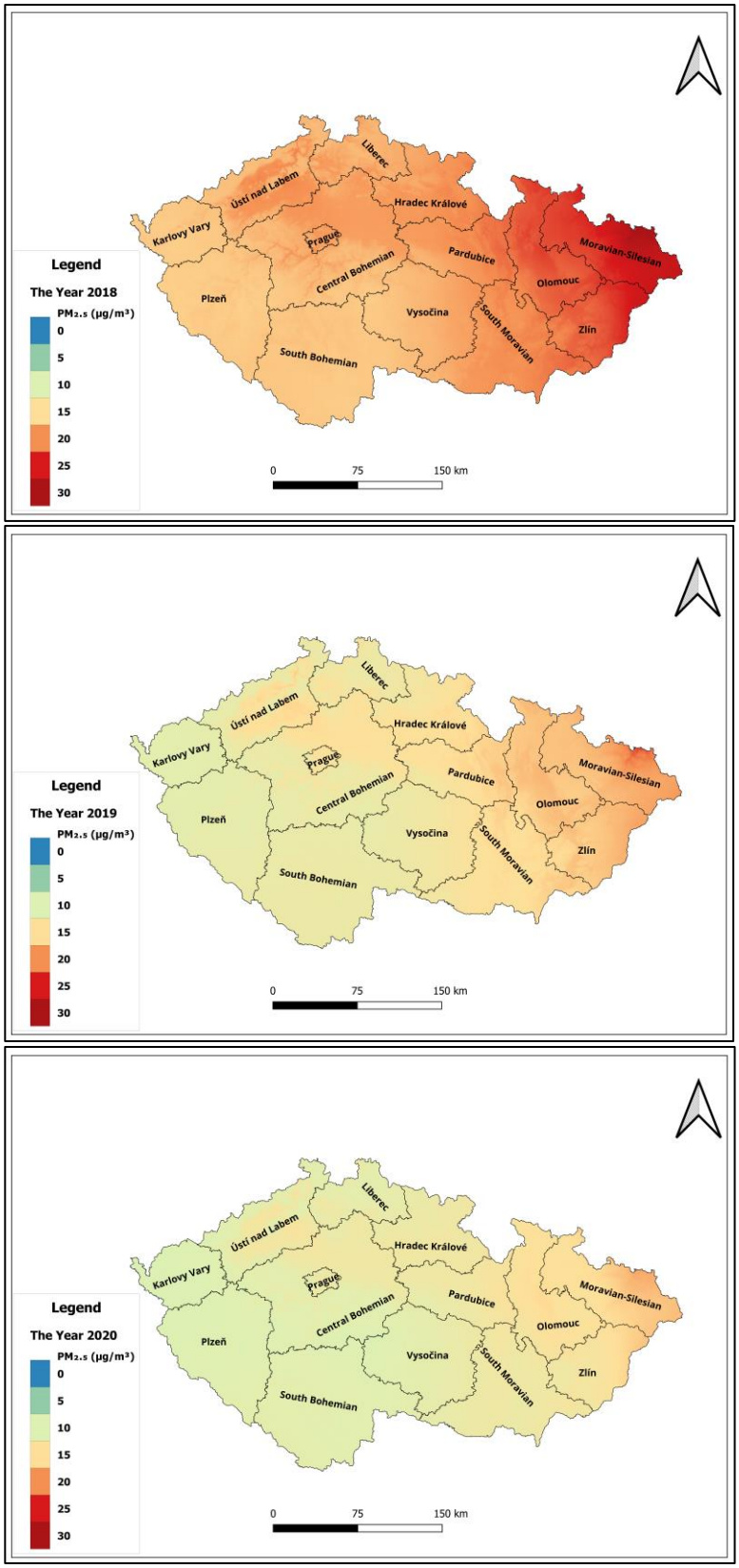


Fig 5. The average $PM_{2.5}$ levels over the Czech Republic in the years 2018, 2019, and 2020.

3.3. **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. Fig. 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.

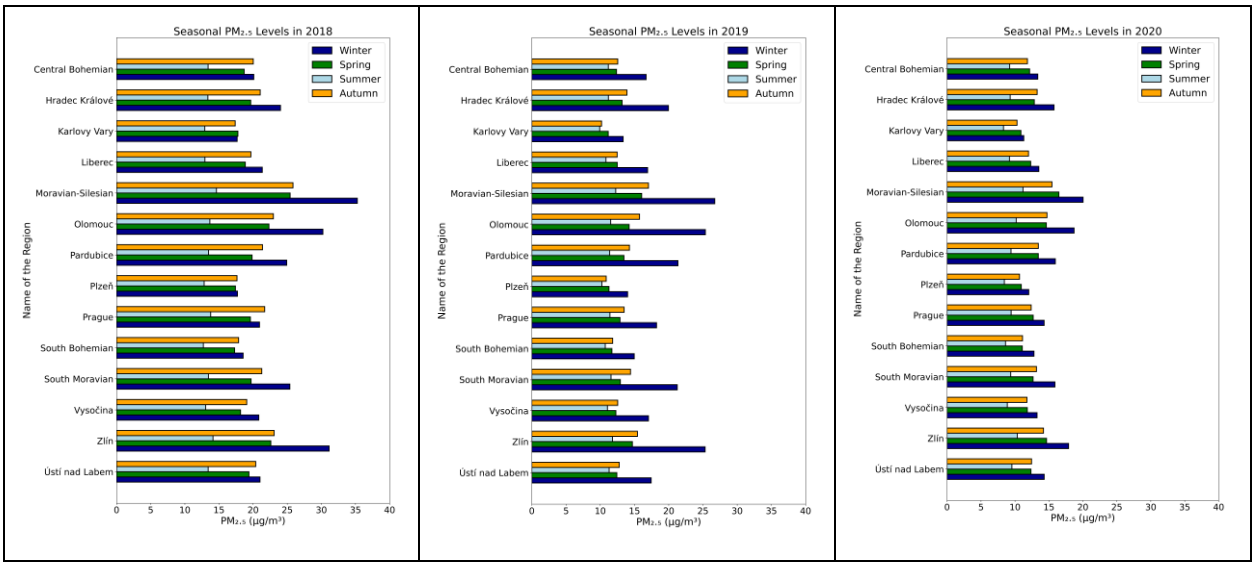


Fig 6. $PM_{2.5}$ seasonal analysis over the Czech Republic in 2018, 2019, and 2020.

4. Discussion

In this study, we used a tree-based machine learning algorithm called the Extra Trees to estimate $PM_{2.5}$ 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 R^2 obtained from the 10-fold cross-validation of the model developed specifically for CZ was 0.85, whereas the corresponding R^2 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 $PM_{2.5}$ values in new locations, achieving a high R^2 of 0.86 and a low RMSE of 5.61 $\mu g/m^3$.

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 $PM_{2.5}$ 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 $\mu g/m^3$. 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. $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ levels in the year 2020 due to decreased industrial activities and reduced transportation emissions [31]. The concentrations of $PM_{2.5}$ in 2020 were less than 20 $\mu g/m^3$ in all regions except the Moravian-Silesian Region during winter months. The yearly average $PM_{2.5}$ 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 $PM_{2.5}$ data for conducting air quality studies throughout Europe. Even though the western part of the country had low concentrations of $PM_{2.5}$, 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.

5. Conclusion

In this study, we estimated daily $PM_{2.5}$ 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 $PM_{2.5}$ predictive model. The model achieved high accuracy in estimating missing $PM_{2.5}$ values with R^2 of 0.85, RMSE of 5.42 $\mu g/m^3$, MAE of 3.41 $\mu g/m^3$, and bias of -0.03 $\mu g/m^3$. Leveraging machine learning techniques and incorporating auxiliary data in model construction can enhance our comprehension of both the temporal and spatial fluctuations in $PM_{2.5}$ 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 $\mu g/m^3$ and 35 $\mu g/m^3$ respectively. In contrast to 2018, $PM_{2.5}$ levels dropped over the whole Czech Republic during the next two years reaching acceptable levels that are less than 20 $\mu g/m^3$ 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 $PM_{2.5}$ concentrations throughout the study period.

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

5.2. Accessing published datasets

Three datasets were published based on our research. These datasets were made publicly available so other researchers may utilize them in their studies.

5.2.1. Geo-Harmonized Atmospheric Dataset for AOD over Europe

This dataset contains daily estimations of AOD_{550} over Europe with 1 km spatial resolution. It was built using machine learning where the MODIS MCD19A2 product with 1km spatial resolution was used as the target variable and modelled AOD data from Copernicus Atmosphere Monitoring Service (CAMS) with 80km spatial resolution and auxiliary data were used as independent variables.

Three stages were applied to generate this dataset: first, we applied a simple average for all available pixels that passed the quality assurance criteria ($QA_{CloudMask}$ = Clear, and $QA_{AdjacencyMask}$ = Clear) of the MCD19A2 product. Second, we created an Extra-trees machine learning model for every year between 2018-2020 to predict the missing AOD values in the MCD19A2 using modelled AOD with enhanced spatiotemporal information. Finally, we filled MODIS MCD19A2 gaps with the predicted AOD by merging the outputs from stages one and two. Daily AOD data for each year between 2018 – 2020 during the MODIS satellite overpassing (10:30 a.m. – 1:30 p.m.) are included in this dataset.

The AOD dataset was validated using AERONET observations and showed a good result with ~84% of the samples falling within the expected error envelopes defined for AOD products.

Data can be accessed: 10.5281/zenodo.5675427.

5.2.2. Geo-Harmonized $PM_{2.5}$ Dataset over Europe

A space-time extremely randomized trees model was used to estimate daily (between 10 a.m. and 2 p.m.) $PM_{2.5}$ concentrations with 1 km spatial resolution for a three-year period 2018–2020 over Europe. Satellite remote sensing, meteorological data, and land variables were used as the independent variables, $PM_{2.5}$ ground-observations were used as the dependent variable while building the model.

Data can be accessed: 10.5281/zenodo.6798975.

6. Discussion

The accuracy of artificial intelligence techniques in estimating air quality is contingent upon a multitude of influencing factors. This research examined AOD and $PM_{2.5}$ over the whole of Europe. The models that were developed to estimate AOD reached high R^2 ranging between 93-95%. These models were utilized to generate daily estimations to fill the gaps in the MCD19A2 MODIS product resulting a full coverage dataset for AOD with 1 km spatial resolution over Europe. This dataset was validated with AERONET observations and showed good results which made it a reliable source for AOD that can be utilized in other air quality studies. In the second part of this research, we used the full coverage AOD dataset, various auxiliary data, and unbalanced spatial-temporal $PM_{2.5}$ dataset to construct a machine learning model aimed at predicting $PM_{2.5}$ concentrations across Europe. The intricacies associated with handling the imbalanced dataset introduced complexities into the modeling procedure. Nevertheless, the model demonstrated commendable performance, particularly given the expansive geographical scope of the study. The results from this part of the study have raised questions that need more investigation.

- Can the utilization of openly available unbalanced $PM_{2.5}$ data be considered a dependable approach for obtaining results?
- If a model is developed specifically for a smaller region within Europe, can the results be deemed compatible when compared to a model designed for the entire continent?
- To enhance validation, is there room for further investigation into alternative methods, particularly those involving entirely unseen data?
- How significant is the influence of spatial autocorrelation (SA) on the outcomes of the study?
- Was the deployment of the model executed effectively and in a manner conducive to reliable results?

In addressing these inquiries, the author conducted a comparable $PM_{2.5}$ analysis within a more confined region, utilizing a well-balanced $PM_{2.5}$ dataset. Additionally, considerations were given to spatial autocorrelation and diverse validation methods. Furthermore, temporal profiles were generated to assess the deployment of the models.

A machine learning model was developed specifically for the Czech Republic. Daily $PM_{2.5}$ measurements for 2018, 2019, and 2020 were collected from the Czech Hydrometeorological Institute (CHMI). Moreover, the spatial autocorrelation between the ground-based station was taken into consideration while building the model. Different validation methods were used to evaluate the performance of the model, the test set was generated based on totally unseen location

for the model. The model reached $R^2 = 0.86$ while predicting values in new locations. Finally, the model deployment used in this research was examined and showed nearly perfect overlap while creating the temporal profiles between the estimated and observed $PM_{2.5}$ values. We compared the results obtained for the Cech Republic from both models and they were highly compatible.

7. Results

In this section, the results of this research with the goals mentioned in section 3 are presented.

- Acquire a deeper understanding of MODIS algorithms:

In the first publication, the performance of three MODIS algorithms over the Czech Republic was analyzed. During this work the author got familiar with the data model and structure of MODIS and how to validate the data with AERONET observations. The products that were tested had a low spatial resolution of 10 km and they suffered from a great number of gaps, these findings prompted the author to explore alternative options.

Another MODIS product called MCD19A2 that was obtained separately from both the Terra and Aqua satellites was used in the second publication. MCD19A2 has a high spatial resolution of 1 km when compared to the previous products which were tested.

- Increase the spatial coverage of MODIS AOD products:

To increase the spatial coverage of AOD obtained from the MCD19A2 product, data from both the Terra and Aqua satellites were merged and an image processing workflow was applied to extract only the retrievals with high quality assurance utilizing the QA sub dataset that is included in the MCD19A2 product. Merging the data of both satellites increased the spatial coverage. However, many gaps still existed. To overcome this issue, various machine learning algorithms were explored starting with the simple linear algorithm till more complex ones. A machine learning approach was developed based on an ensemble method called the Extra Trees to estimate AOD values and fill the gaps in the MCD19A2 MODIS product. Modelled AOD and auxiliary data were used as inputs to the AOD predictive model. Different statistical methods were applied to deal with the large datasets of AOD over Europe.

- Establish the first AOD dataset with full coverage of 1 km spatial resolution over Europe :

A Geo-Harmonized Atmospheric Dataset for AOD was generated with a full coverage of 1 km spatial resolution over Europe. This dataset was validated with AERONET observations, and good results were achieved, making this dataset the first verified dataset with full coverage of 1 km spatial resolution over Europe.

- Analyze the effects of the COVID-19 lockdowns on AOD levels over Europe:

After generating the daily AOD maps over Europe, the effects of the COVID-19 lockdowns on AOD levels were analyzed taking into consideration the relationship between AOD and other meteorological variables.

- Establish the first $PM_{2.5}$ dataset with full coverage of 1 km spatial resolution over Europe:

In the third publication, the full coverage AOD dataset and various inputs with open $PM_{2.5}$ data collected by ground-based monitors were utilized to estimate $PM_{2.5}$ over Europe. The main issue was dealing with unbalanced spatial-temporal data of $PM_{2.5}$ over the study area. Many statistical methods were applied to overcome the limitations this data suffered. The outcome of the third publications was the first full coverage $PM_{2.5}$ dataset with 1 km spatial resolution over the whole of Europe.

- Analyze the effects of the COVID-19 lockdowns on $PM_{2.5}$ levels over Europe:

In the third publication, the effects of the COVID-19 lockdowns on $PM_{2.5}$ concentrations over the study area were analyzed.

- Include the spatial autocorrelation of the ground based $PM_{2.5}$ observations while developing the machine learning model:

In the fourth paper, the spatial autocorrelation of the ground based $PM_{2.5}$ observations was calculated based on the Local Moran Index and included while developing the machine learning model to estimate $PM_{2.5}$ over the Czech Republic.

- Compare the results of $PM_{2.5}$ obtained for whole Europe with the results from the model developed for the Czech Republic:

The final part in this research was to compare the results obtained for the whole Europe using open unbalanced $PM_{2.5}$ data with a smaller landlocked area like the Czech Republic that has more spatial-temporal balanced data. Higher overall accuracy of the model developed for the Czech Republic was achieved. However, the results obtained for the Czech Republic based on the last model were compatible with the results obtained from the model designed for the whole Europe.

8. Conclusion

In this research, the author aimed to generate comprehensive and high-resolution air quality datasets to overcome the limited spatial coverage provided by ground-based monitors. We tested various machine learning methods, starting with basic ones like linear models and progressively advanced to more intricate techniques including ensemble methods. We found that by harnessing satellite remote sensing data, meteorological variables, auxiliary information in conjunction with machine learning techniques has allowed us to successfully transcend the limitations inherent within each individual dataset. We generated the first full coverage datasets with fine spatial

resolution of 1 km for both AOD and $PM_{2.5}$ over the entire Europe. This has enabled a more detailed understanding of air pollution dynamics obviating the necessity for additional ground stations and thereby yielding cost reduction. By employing a combination of GIS techniques and advanced statistical analyses, the study has provided valuable insights into air pollution patterns and trends across the whole of Europe. We were able to identify pollution hotspots, this information is critical for policymaking to reduce the harmful impacts of poor air quality on public health and the environment. The analysis of the generated datasets showed seasonal and temporal variations in air pollution concentrations. These insights can help decision makers understand the underlying factors that affect air quality in order to implement appropriate strategies for air pollution management. The validation techniques we followed ensured the reliability of the machine learning model which we utilized to generate the datasets. This will increase confidence and strengthen the potential for future use in environmental studies. Our work conclusively established the adequacy of open data and open-source software in creating these air quality datasets. We also actively supported the open data policy, promoting transparency and collaboration.

9. Future work

The methodology we applied while developing the machine learning models can be applied in similar air quality studies to estimate other pollutants such as O_3, NO_2, SO_2, CO. Generating a full coverage AOD datasets with the availability of ground monitors are the main factors to apply our methodology. Different auxiliary data can be tested and as we found in this research, including dynamic variables that vary temporally throughout the study period will enrich the machine learning models and improve their accuracy. Delving deeper into spatial autocorrelation when studying $PM_{2.5}$ at large scales, exploring additional factors, and employing advanced spatial analysis techniques could yield richer insights into the distribution dynamics.

10. References

[1] Pope, A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate. JAMA. 2002 Mar 6;287(9):1132-41.

[2] Gakidou, E.; Afshin, A.; Abajobir, A.A.; Abate, K.H.; Abbafati, C.; Abbas, K.M.; Abd-Allah, F.; Abdulle, A.M.; Abera, S.F.; Aboyans, V.; et al, 2017. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or

clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet, 390, 1345–1422.

[3] S. Segura, V. Estell_es*, M.P. Utrillas, J.A. Martínez-Lozano, 2017. Long term analysis of the columnar and surface aerosol relationship at an urban European coastal site. Atmospheric Environment, 167, 309e322.

[4] Zhang, L., M. Zhang, and Y.B. Yao., 2019. Multi-Time Scale Analysis of Regional Aerosol Optical Depth Changes in National-Level Urban Agglomerations in China Using Modis Collection 6.1 Datasets from 2001 to 2017. Remote Sensing, 11(2).

[5] Kaufman, Y.J., et al., 1997. Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect. Journal of Geophysical Research-Atmospheres, 102(D14): 16815-16830.

[6] Cheng, A.Y.S., M.H. Chan, and X. Yang., 2006. Study of aerosol optical thickness in Hong Kong, validation, results, and dependence on meteorological parameters. Atmospheric Environment, 40(24): 4469-4477.

[7] Brown, J.S., Gordon, T., Price, O., Asgharian, B., 2013. Thoracic and respirable particle definition for human health risk assessment. Part. Fibre Toxicol. 10e12. http:// dx.doi.org/10.1186/1743-8977-10-12.

[8] IPCC (2007), Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by S. Solomon et al., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[9] Poschl, U., 2005. Atmospheric aerosols: Composition, transformation, climate and health effects. Angewandte Chemie-International Edition, 44(46): 7520-7540.

[10] Remer, L.A., et al., 2005. The MODIS aerosol algorithm, products, and validation. Journal of the Atmospheric Sciences, 62(4): 947-973.

[11] Sun, L., et al., 2016. A Universal Dynamic Threshold Cloud Detection Algorithm (UDTCDA) supported by a prior surface reflectance database. Journal of Geophysical Research-Atmospheres, 121(12): 7172-7196.

[12] Tanre, D., M. Herman, and Y.J. Kaufman., 1996. Information on aerosol size distribution contained in solar reflected spectral radiances. Journal of Geophysical Research-Atmospheres, 101(D14): 19043-19060.

[13] Tanre, D., et al., 1997. Remote sensing of aerosol properties over oceans using the MODIS/EOS spectral radiances. Journal of Geophysical Research-Atmospheres, 102(D14): 16971-16988.

[14] Kaufman, Y.J., D. Tanre, and O. Boucher., 2002. A satellite view of aerosols in the climate system. Nature, 419(6903): 215-223.

[15] Li, Z., et al., 2009. Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: a review and perspective. Annales Geophysicae, 27(7): 2755-2770.

[16] Bilal, M., et al., 2013. A Simplified high resolution MODIS Aerosol Retrieval Algorithm (SARA) for use over mixed surfaces. Remote Sensing of Environment, 136: 135-145.

[17] Levy, R.C., et al., 2010. Global evaluation of the Collection 5 MODIS dark-target aerosol products over land. Atmospheric Chemistry and Physics, 10(21): 10399-10420.

[18] Chu, D.A., et al., 2002. Validation of MODIS aerosol optical depth retrieval over land. Geophysical Research Letters, 29(12).

[19] Remer, L.A., et al., 2002. Validation of MODIS aerosol retrieval over ocean. Geophysical Research Letters, 29(12).

[20] Holben, B.N., et al., 1998. AERONET - A federated instrument network and data archive for aerosol characterization. Remote Sensing of Environment, 66(1): 1-16.

[21] King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. IEEE Trans. Geosci. Remote Sens. 2013, 51, 3826–3852.

[22] Sahoo, T., Patnaik, S., 2008. Cloud removal from satellite images using auto associative neural network and stationarywevlet transform. International Conference on Emerging Trends in Engineering and Technology.

[23] Meraner, A.; Ebel, P.; Zhu, X.X.; Schmitt, M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. Photogramm. Remote Sens. 2020, 166, 333–346.

[24] Yang, J.; Hu, M. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. Sci. Total Environ. 2018, 633, 677–683.

[25] 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 PM2.5 Concentrations across Great Britain. Remote Sens. 2020, 12, 3803.

[26] Crippa, M., Janssens-Maenhout, G., Guizzardi, D., van Dingenen, R., & Dentener, F., **2019**. Contribution and uncertainty of sectorial and regional emissions to regional and global PM2.5 health impacts. *Atmospheric Chemistry and Physics*, 19(7), 5165–5186. doi:10.5194/acp-19-5165-2019.

[27] Pascal, M., et al., **2014**. Short-term impacts of particulate matter (PM10, PM10-2.5, PM2.5) on mortality in nine French cities. *Atmospheric Environment*, 95. doi:10.1016/j.atmosenv.2014.06.030.

[28] Liu, C., et al., **2019**. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *New England Journal of Medicine*, 381(8). doi:10.1056/nejmoa1817364.

[29] Martins, N. R., & Carrilho da Graça, G., **2018**. Impact of PM2.5 in indoor urban environments: A review. *Sustainable Cities and Society*, 42. doi:10.1016/j.scs.2018.07.011.

[30] Lee, H. J., **2020**. Advancing Exposure Assessment of PM2.5 Using Satellite Remote Sensing: A Review. *Asian Journal of Atmospheric Environment*, 14(4). doi:10.5572/ajae.2020.14.4.319.

[31] Deng, L., **2015**. Estimation of PM2.5 Spatial Distribution Based on Kriging Interpolation. In Proceedings of the First International Conference on Information Sciences, Machinery, *Materials, and Energy*, Vol. 126. doi:10.2991/icismme-15.2015.370.

[32] You, W., Zang, Z., Pan, X., Zhang, L., & Chen, D., **2015**. Estimating PM2.5 in Xi'an, China using aerosol optical depth: A comparison between the MODIS and MISR retrieval models. *Science of the Total Environment*, 505. doi:10.1016/j.scitotenv.2014.11.024.

[33] Yao, F., Si, M., Li, W., & Wu, J., **2018**. A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. *Science of the Total Environment*, 618. doi:10.1016/j.scitotenv.2017.08.209.

[34] Ma, Z., et al., **2022**. A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data. *Remote Sensing of Environment*, 269. doi:10.1016/j.rse.2021.112827.

[35] J. Wei et al., "Improved 1 km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees," Atmospheric Chemistry and Physics, vol. 20, no. 6, 2020, doi: 10.5194/acp-20-3273-2020.

[36] Chen, G., et al., **2018**. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological, and land use information. *Science of the Total Environment*, 636. doi:10.1016/j.scitotenv.2018.04.251.

[37] Xiao, F., Yang, M., Fan, H., Fan, G., & Al-qaness, M. A. A., **2020**. An improved deep learning model for predicting daily PM2.5 concentration. Scientific Reports, 10(1). doi:10.1038/s41598-020-77757-w.

[38] Li, L., et al., **2020**. Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. *Environment International*, 145. doi:10.1016/j.envint.2020.106143.

[39] Van Donkelaar, A., et al., **2010**. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. Environmental Health Perspectives, 118(6). doi:10.1289/ehp.0901623.

[40] Koelemeijer, R. B. A., Homan, C. D., & Matthijsen, J., **2006**. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment*, 40(27). doi:10.1016/j.atmosenv.2006.04.044.

11. List of publications

Publications related to the dissertation

- Ibrahim, S., & Halounová, L. Statistical Study of MODIS Algorithms in Estimating Aerosol Optical Depth Over the Czech Republic. *Stavební Obzor - Civil Engineering Journal*. **2019**, 28(4).

  https://doi.org/10.14311/CEJ.2019.04.0043.

- 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 Sensing. **2021**, 13(15).

  https://doi.org/10.3390/rs13153027.

- Ibrahim, S., Landa, M., Pešek, O., Brodský, L., & Halounová, L. Machine Learning-Based Approach Using Open Data to Estimate $PM_{2.5}$ over Europe. *Remote Sensing.* **2022**, Vol. 14, Page 3392, 14(14), 3392.
  https://doi.org/10.3390/RS14143392.

- Ibrahim, S., Landa, M., Matoušková, E., Brodský, L., & Halounová, L. $PM_{2.5}$ Estimation in the Czech Republic using Extremely Randomized Trees: A Comprehensive Data Analysis. *Stavební Obzor - Civil Engineering Journal,*. **2023**. (Accepted).

Publications are not related to the dissertation thesis:

- Matoušková E, Pavelka K, Ibrahim S. Creating a Material Spectral Library for Plaster and Mortar Material Determination. *Materials*. **2021**; 14(22):7030.
  https://doi.org/10.3390/ma14227030.